Lecture 2.2

Bayesian Phylogenetics I

The Bayesian framework

Bayesian phylogenetic analysis

- Bayesian phylogenetic analysis was developed in the mid 1990s
- Now one of the most widely used methods
- Bayes's theorem (1763)
- Reverend Thomas Bayes

Image probably not of Thomas Bayes

Contrast with frequentist statistics (likelihood)

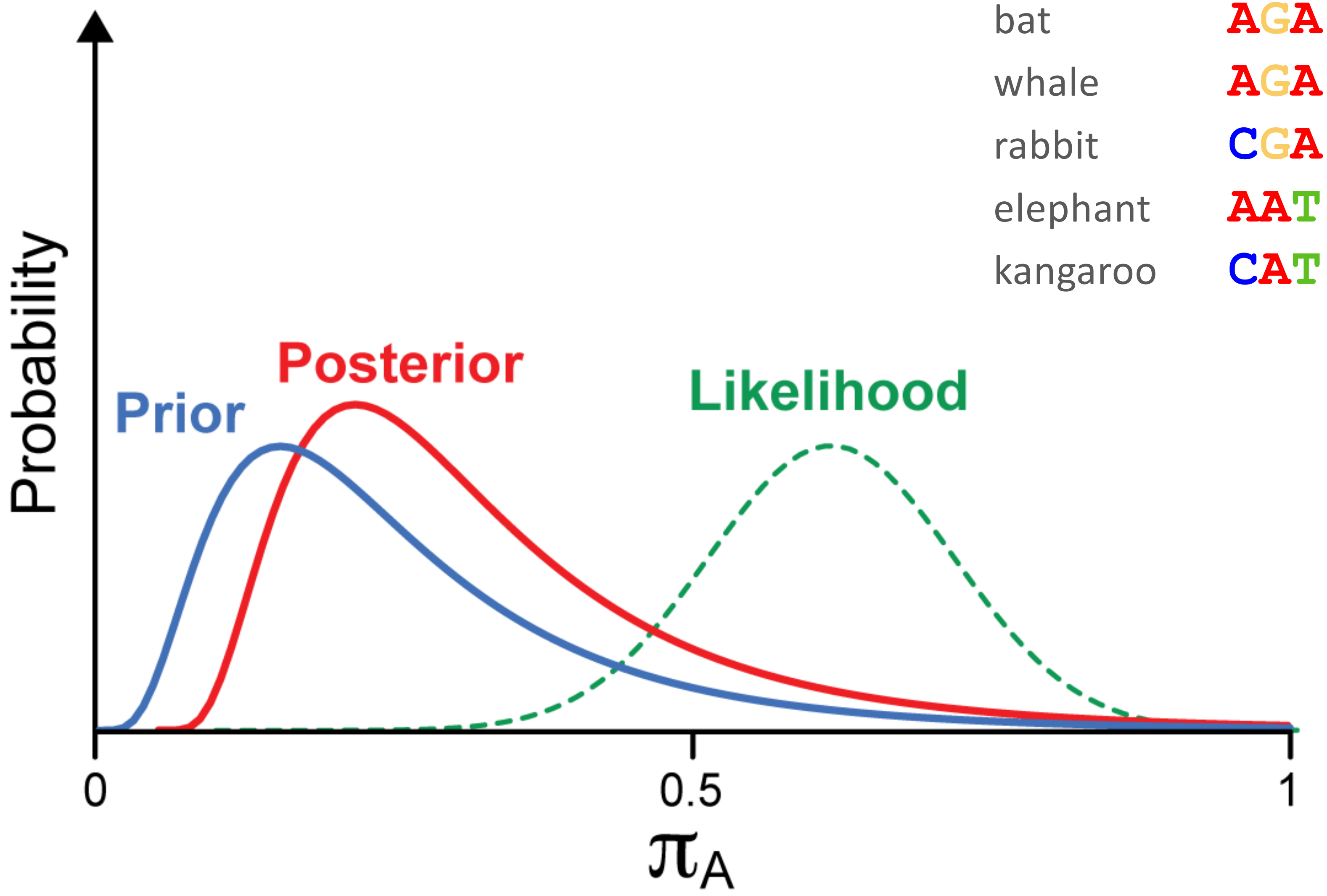
Bayesian phylogenetic analysis

$$\Pr(\theta \mid D) \propto \Pr(\theta) \; \Pr(D \mid \theta)$$

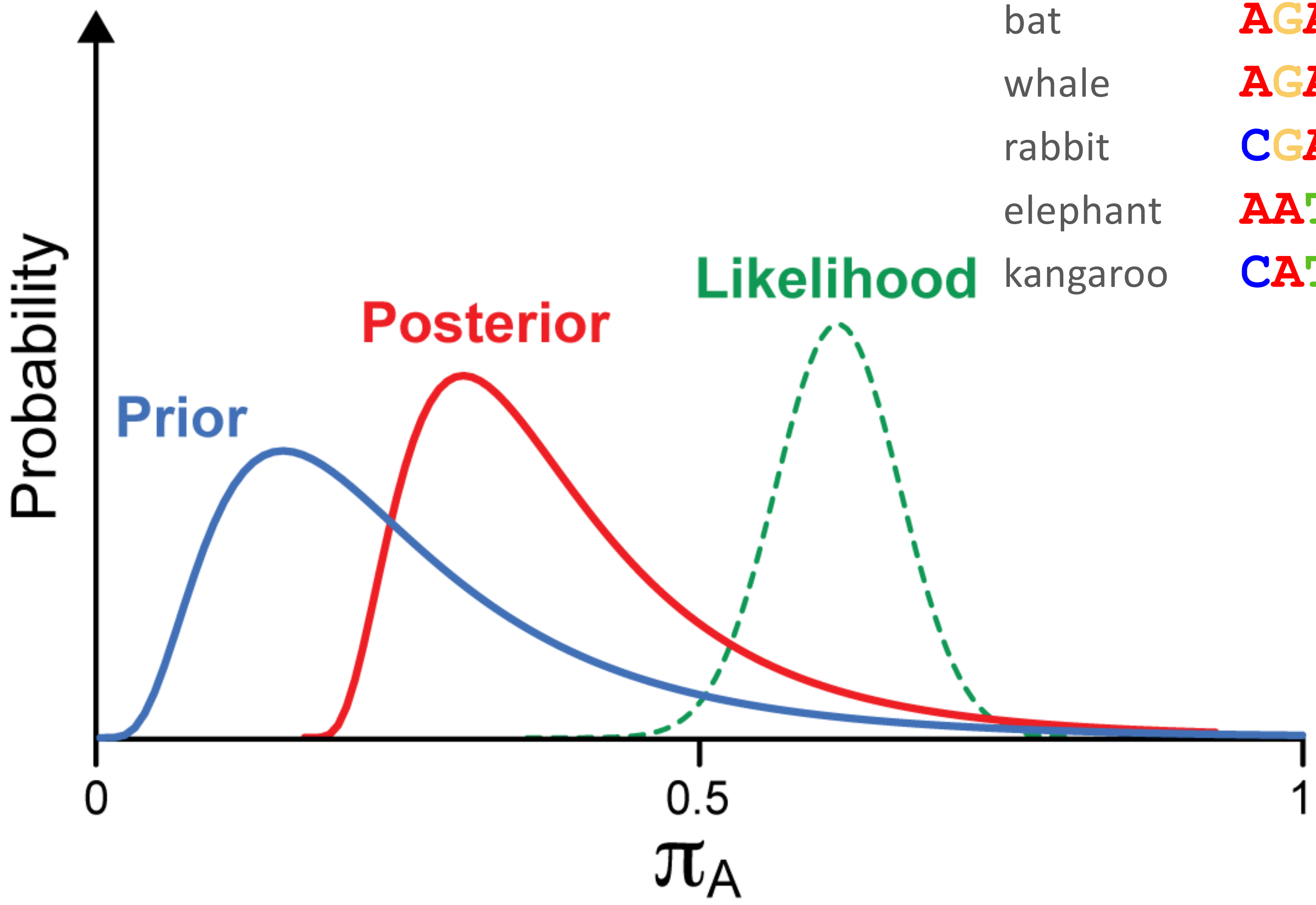
- Parameters have distributions
- Before the data are observed, each parameter has a prior probability distribution
    - Reflect our prior expectations (and uncertainty) about values of parameters (without knowledge of the data)
- Likelihood of the data is computed
- Prior probability distribution is combined (updated) with the likelihood to yield the posterior probability distribution

Simple example

bat	AGA
whale	AGA
rabbit	CGA
elephant	AAT
kangaroo	CAT

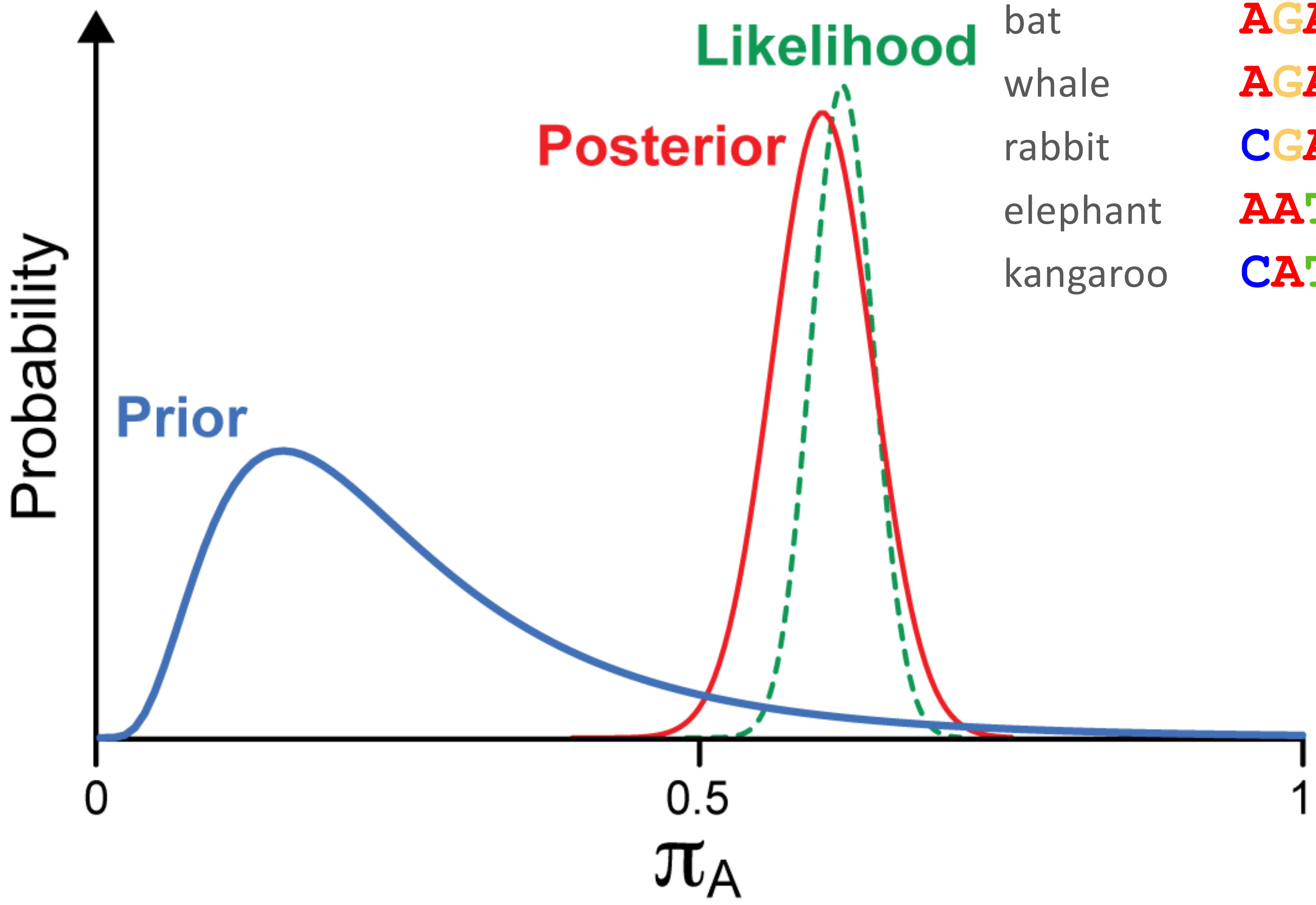


Simple example



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

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

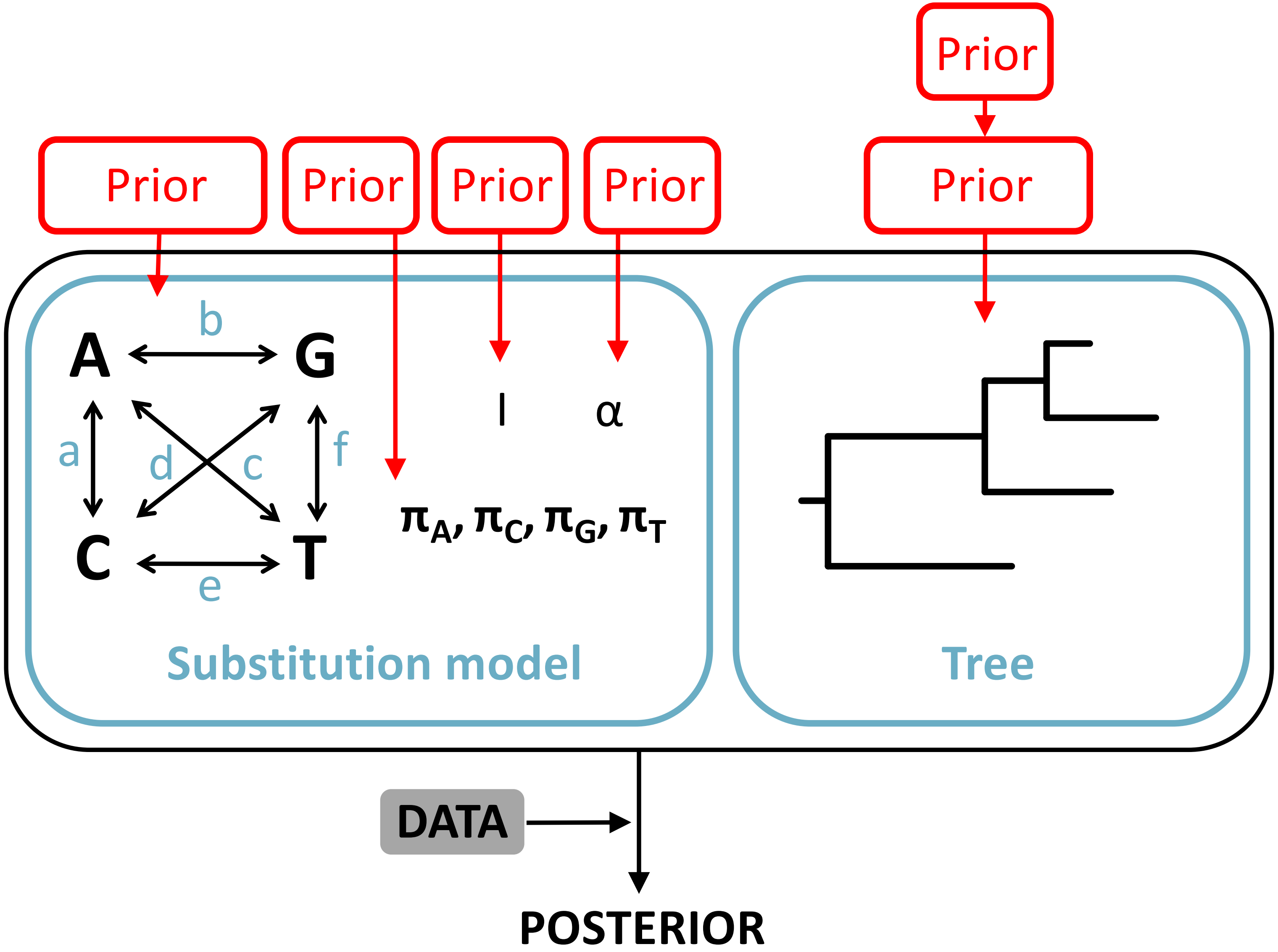
Posterior

This is what we want to estimate

Likelihood

Calculated from data

Bayesian hierarchical model

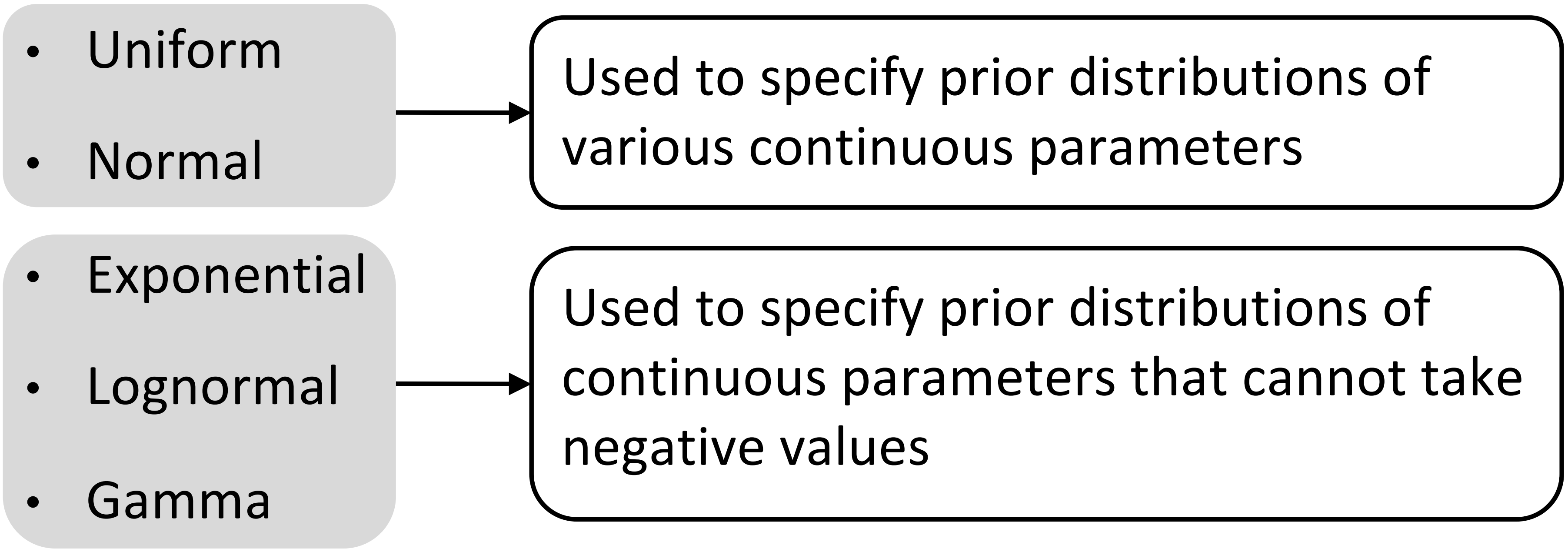


Priors

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

Continuous distributions



- Uniform
- Normal

Used to specify prior distributions of various continuous parameters

- Exponential
- Lognormal
- Gamma

Used to specify prior distributions of continuous parameters that cannot take negative values

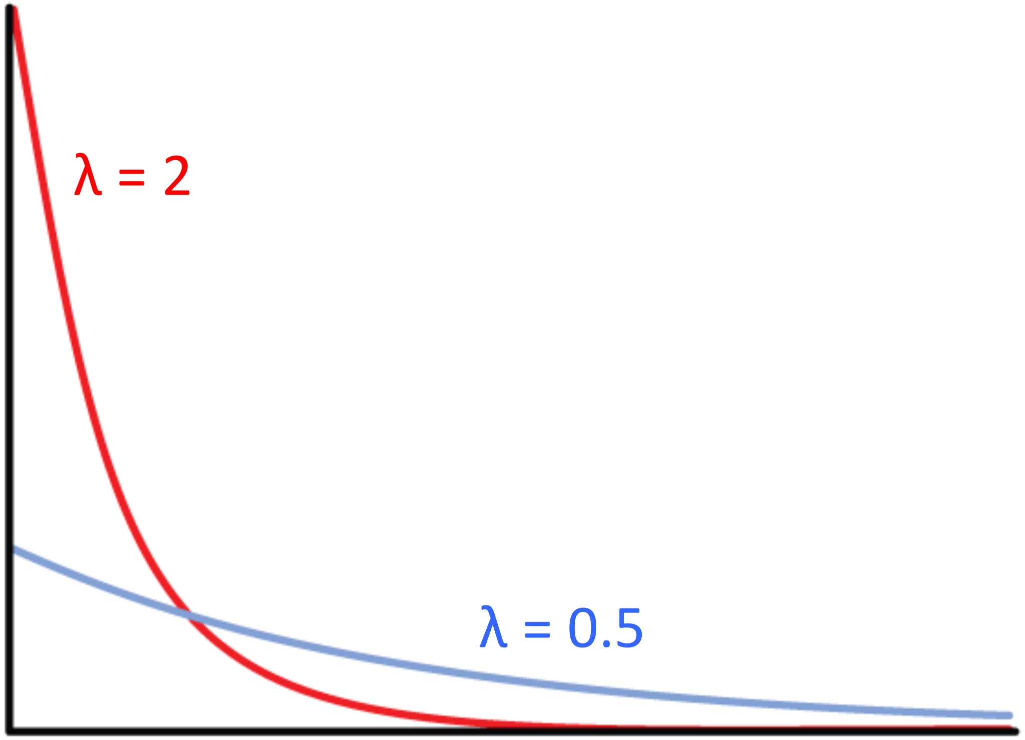
- Beta
- Dirichlet

Continuous distributions

- Uniform
- Normal
- **Exponential**
- Lognormal
- Gamma
- Beta
- Dirichlet

Parameters

- λ = rate of decay



Continuous distributions

- Uniform
- Normal
- Exponential
- Lognormal
- Gamma
- Beta
- Dirichlet

Used to specify the prior distributions of simplex parameters

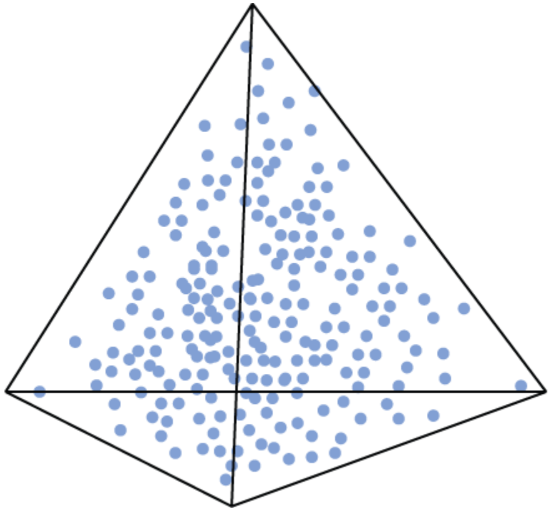
- Base frequencies
- Relative rates in GTR model

Continuous distributions

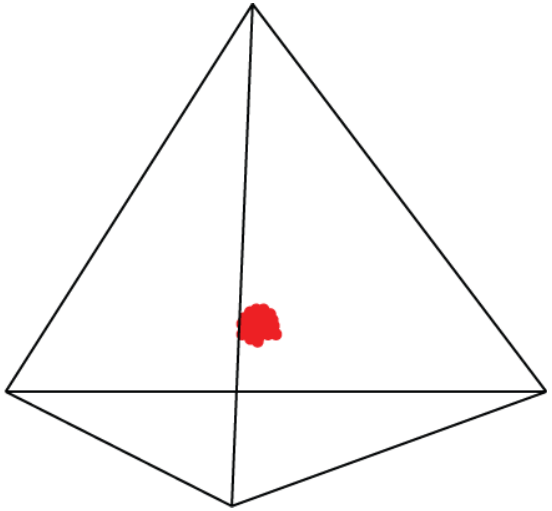
- Uniform
- Normal
- Exponential
- Lognormal
- Gamma
- Beta
- **Dirichlet**

Parameters

- α_1 , α_2 , … = shape parameters



$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$



$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 300$

Image modified from Paul Lewis's slides

Discrete distributions

- Bernoulli distribution
- Binomial
- Multinomial
- Poisson

Default priors

	BEAST2	***MrBayes***
Rate matrix parameters	Gamma(0.05,10)	Dirichlet(1,1,1,1,1,1)
Base frequencies	Uniform(0,1)	Dirichlet(1,1,1,1)
Shape parameter (α)	Exponential(1)	Exponential(2)
Proportion invariable	Uniform(0,1)	Uniform(0,1)

Can specify uninformative priors where appropriate

Tree Prior

Tree prior

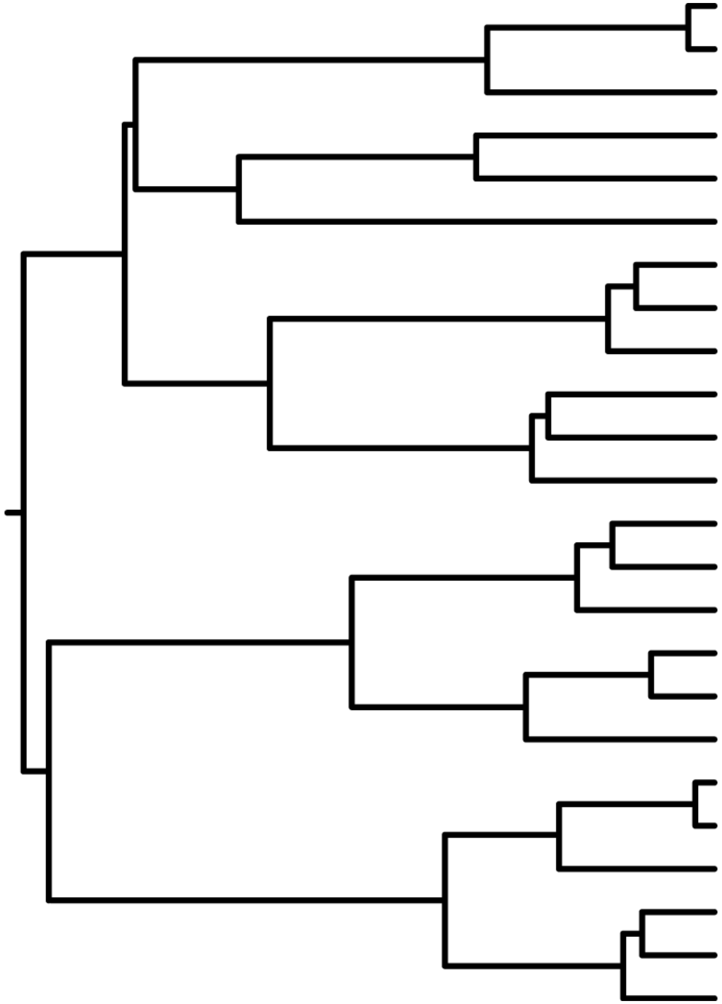
1. Use a **flat prior** (*MrBayes*)
    - All trees have equal probability
    - Also need a prior for branch lengths or node times

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

} **Priors on rooted trees**

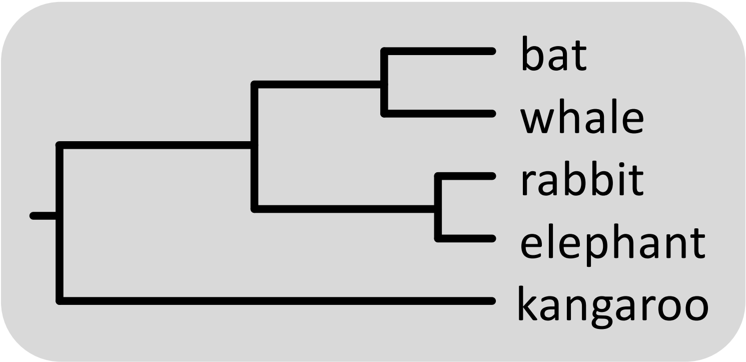
Speciation model

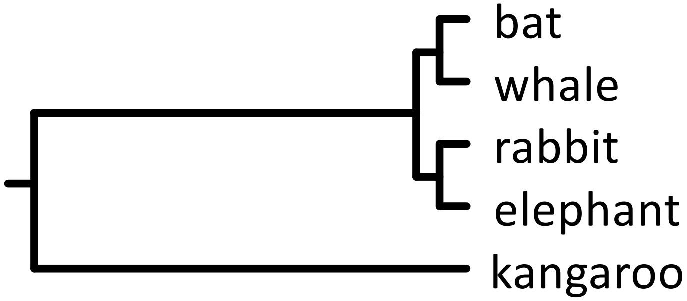
- Tree shape described by a stochastic branching process
- **Yule process**
  - The root lineage splits into two
  - Lineages split at a constant rate
  - Simulates speciation process
- **Birth-death process**
  - Allow lineages to go extinct

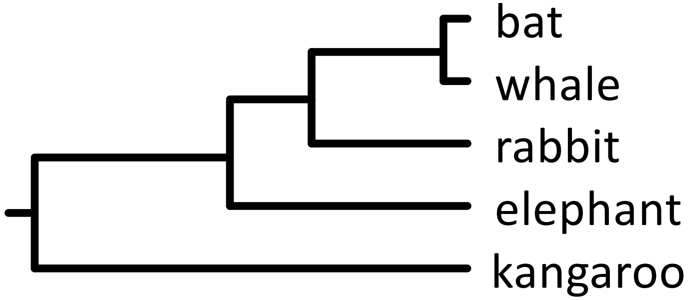


Speciation model

- Tree shape described by a stochastic branching process
- **Yule process**
  - The root lineage splits into two
  - Lineages split at a constant rate
  - Simulates speciation process
- **Birth-death process**
  - Allow lineages to go extinct

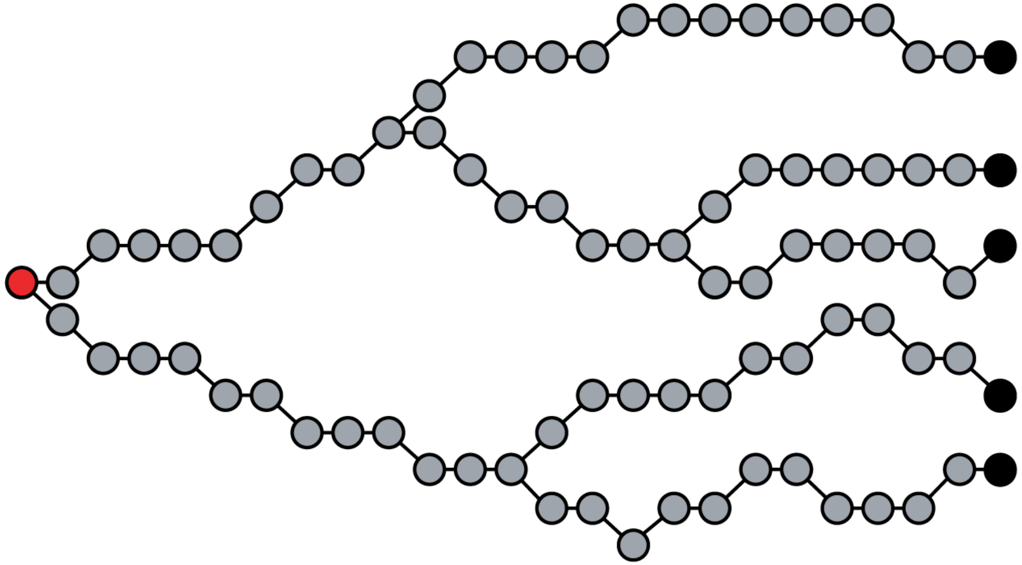




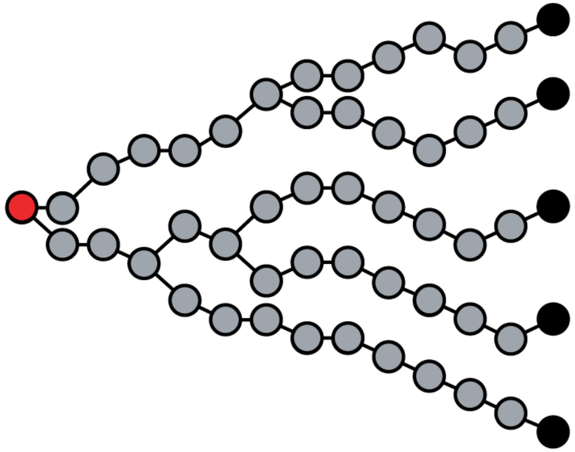


Coalescent model

Constant size



Exponential growth



Choosing a tree prior

- Test whether inferences are robust to the choice of tree prior

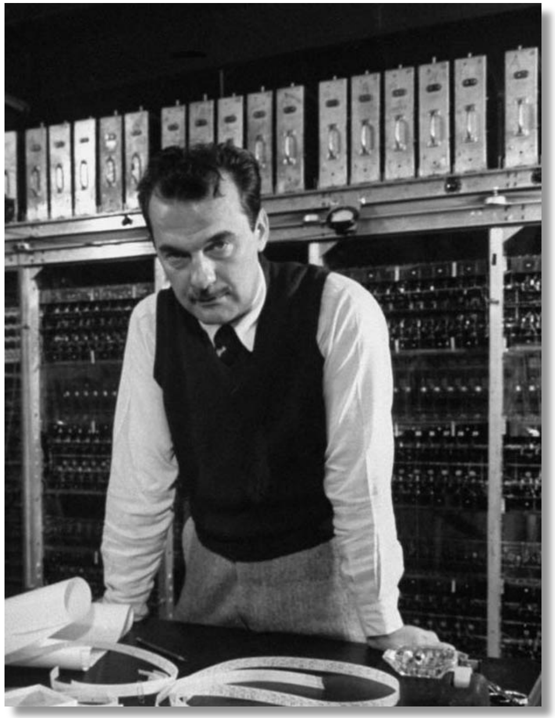
- Mixed data sets: multiple sequences from each species
  - Birth-death prior generally works well

- Compare tree priors using Bayesian model selection

Posterior Distribution

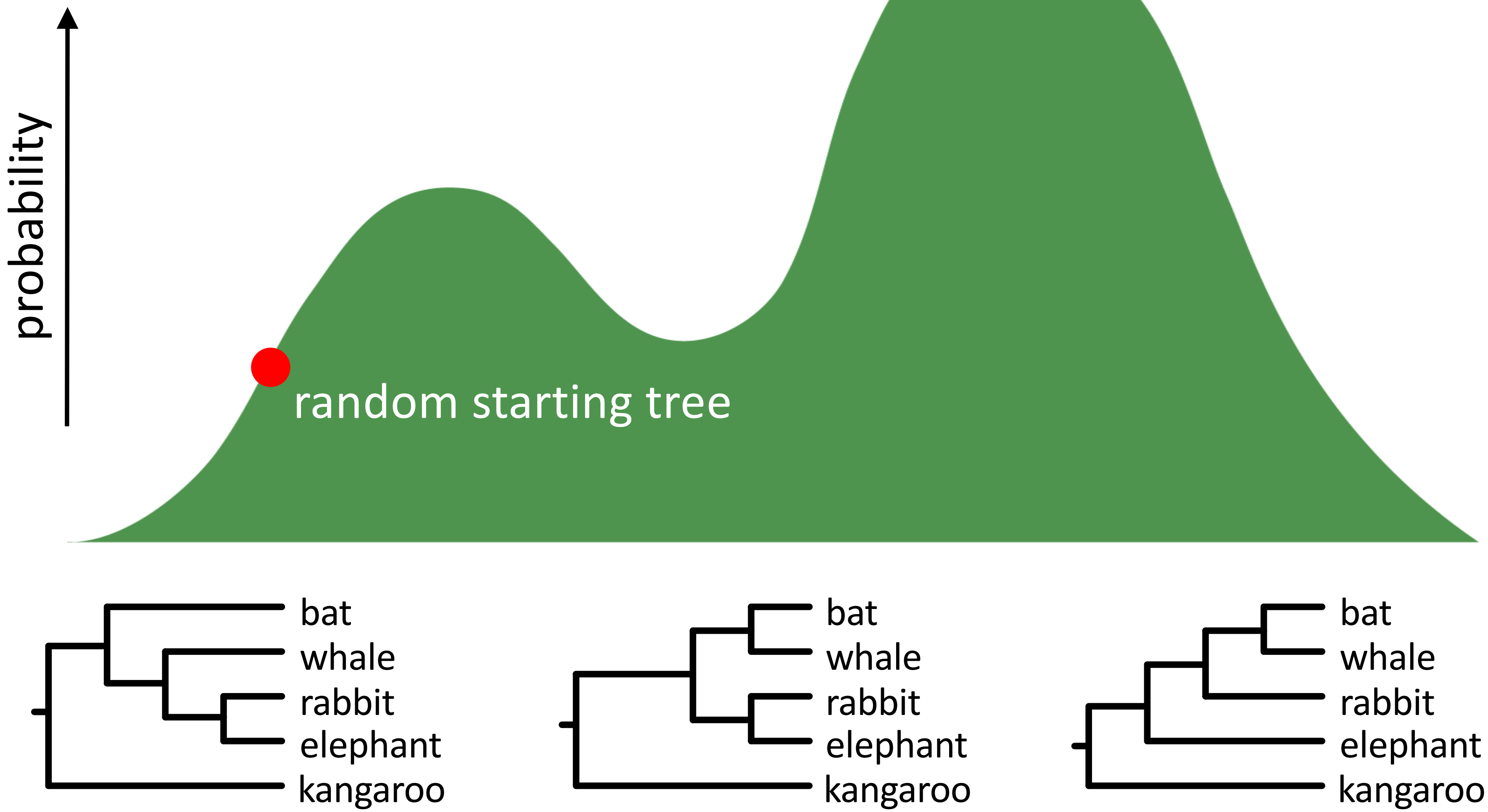
Estimating the posterior

- 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**

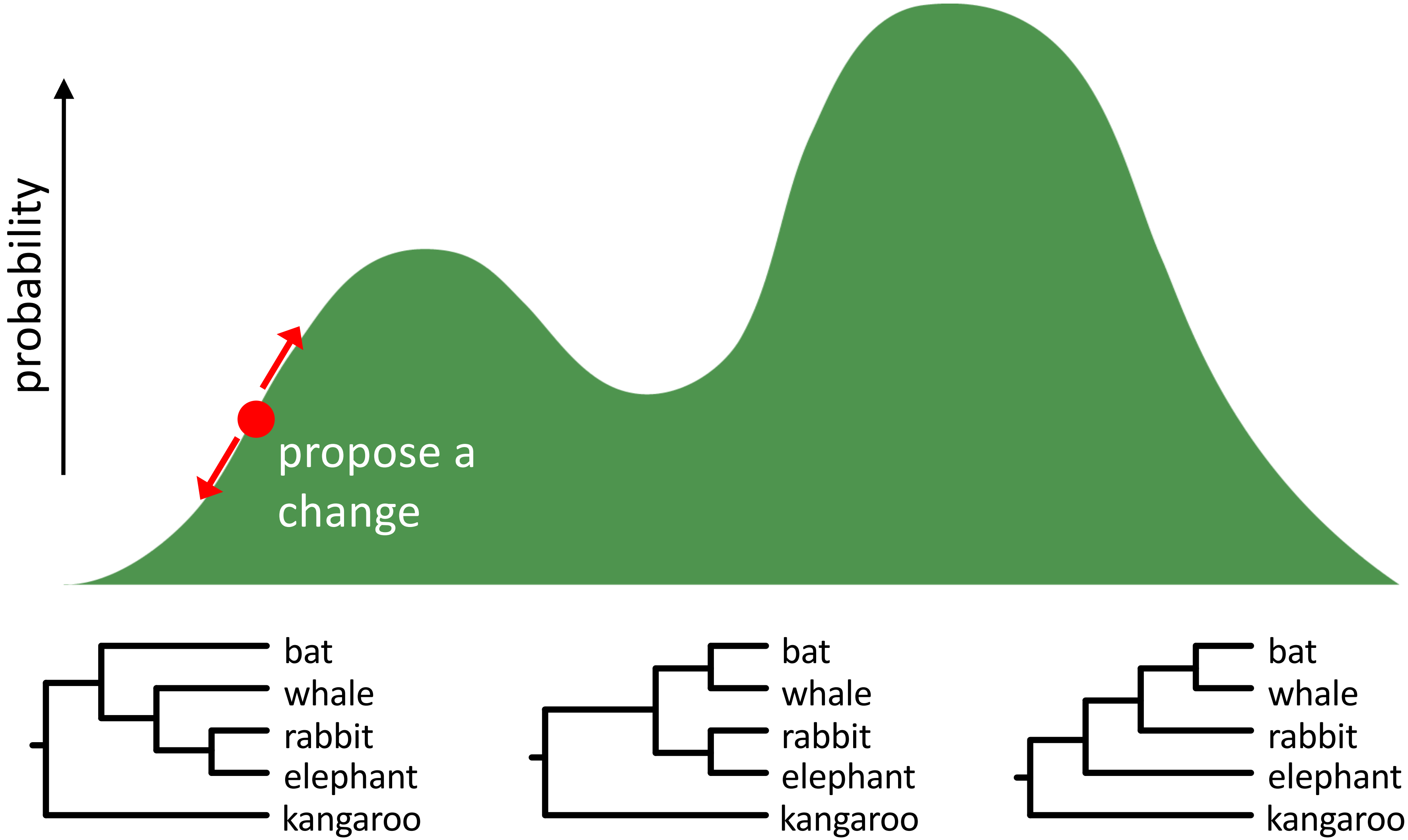


Nicholas Metropolis
Los Alamos, 1953

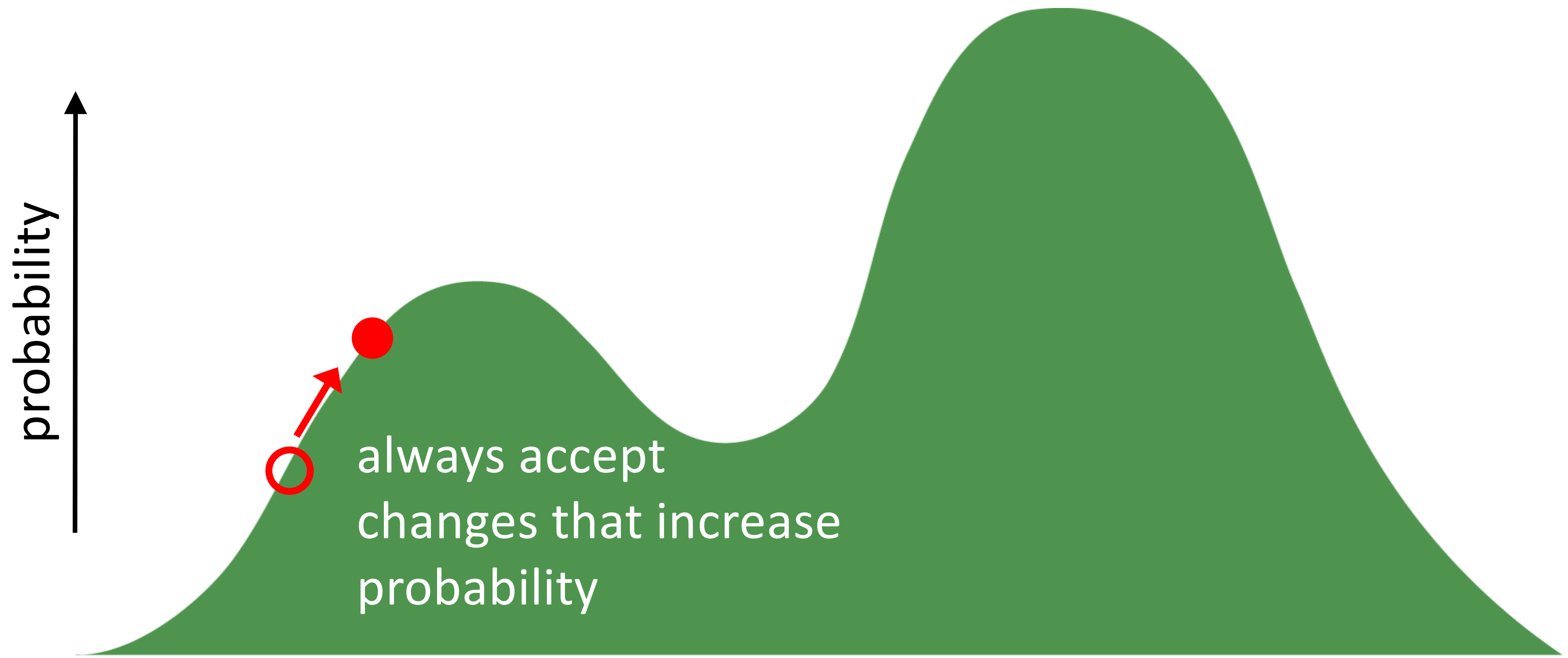
MCMC simulation



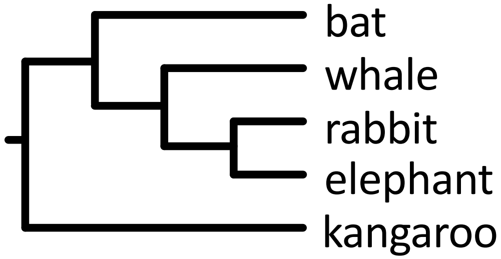
MCMC simulation

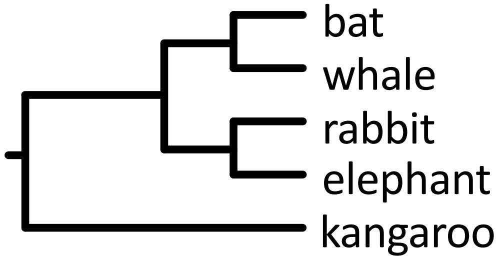


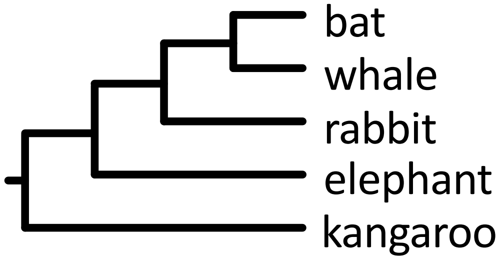
MCMC simulation



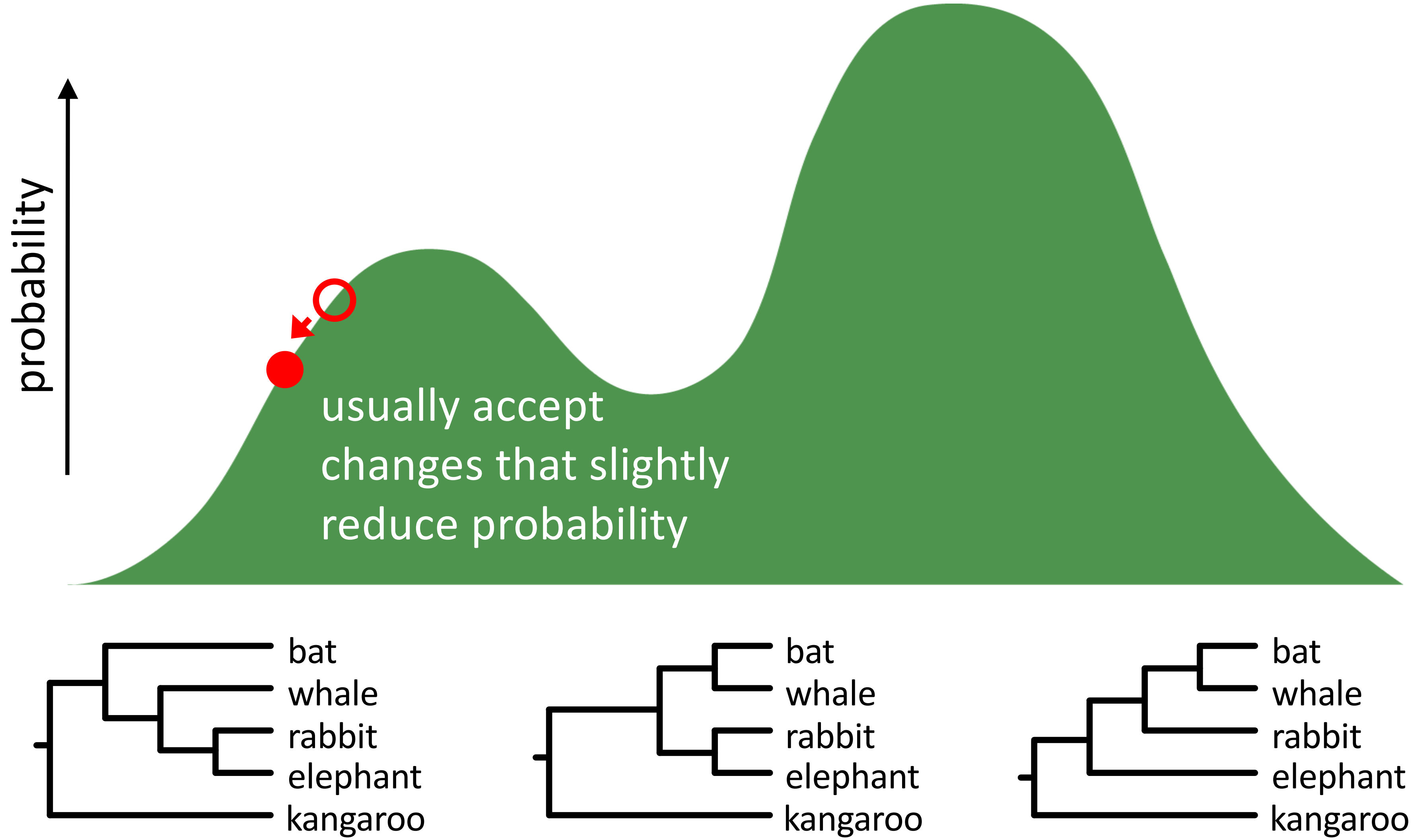
always accept
changes that increase
probability





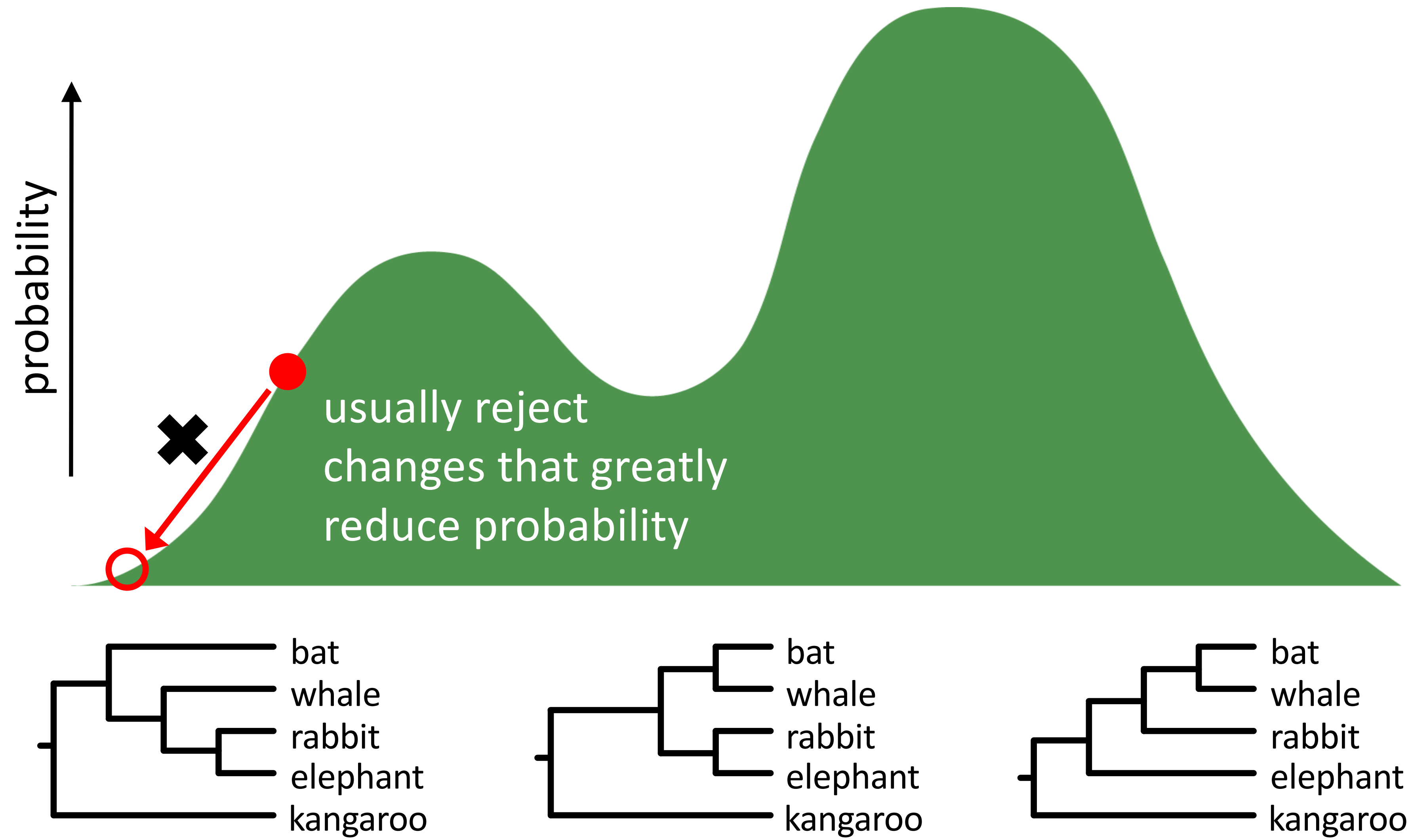


MCMC simulation



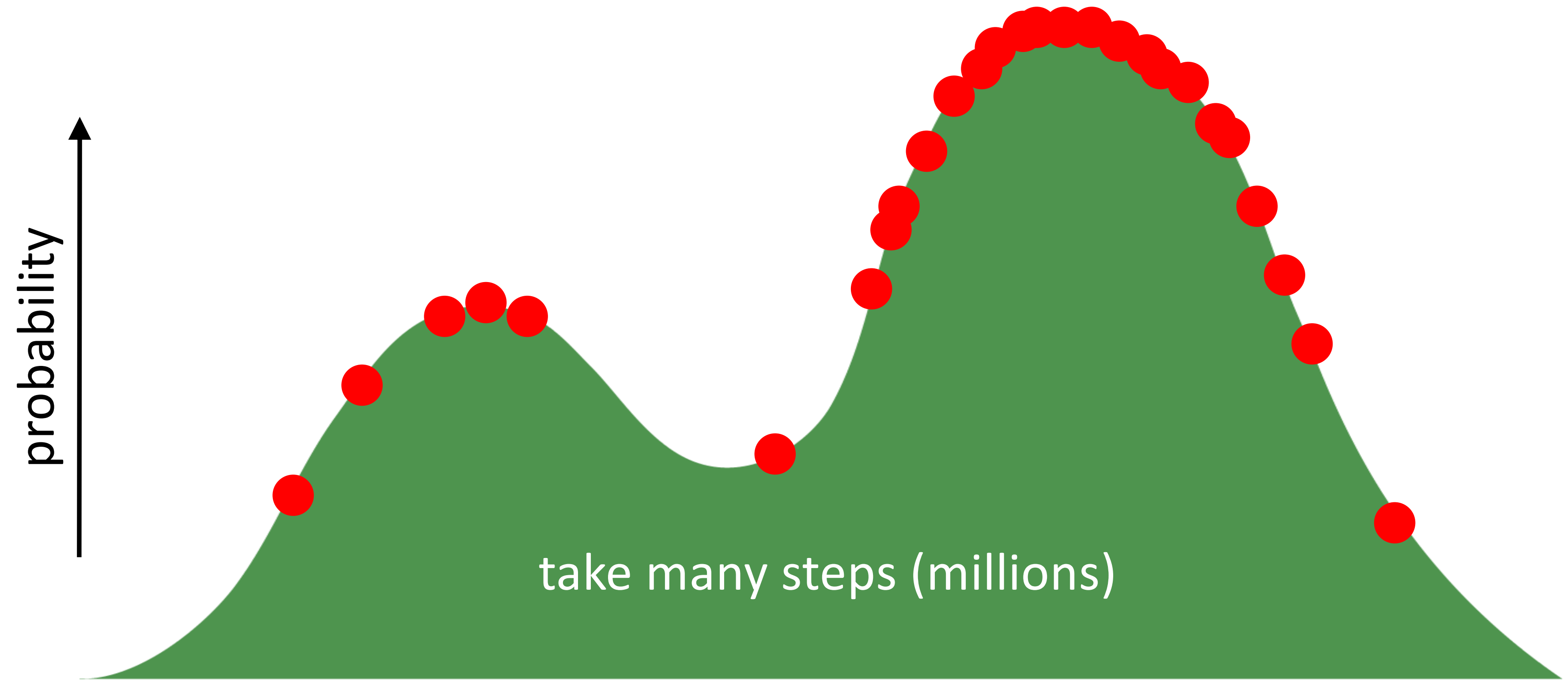
usually accept
changes that slightly
reduce probability

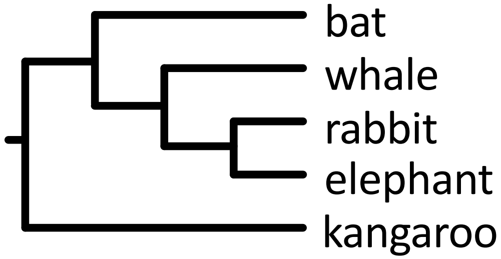
MCMC simulation

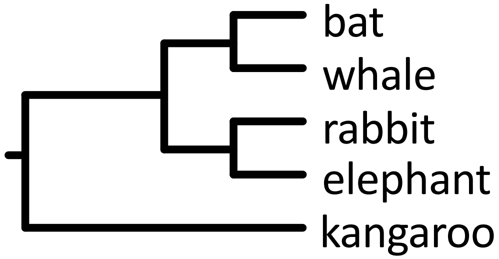


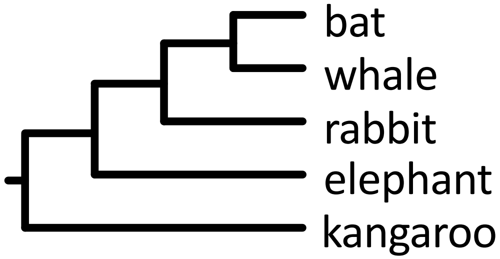
usually reject changes that greatly reduce probability

MCMC simulation

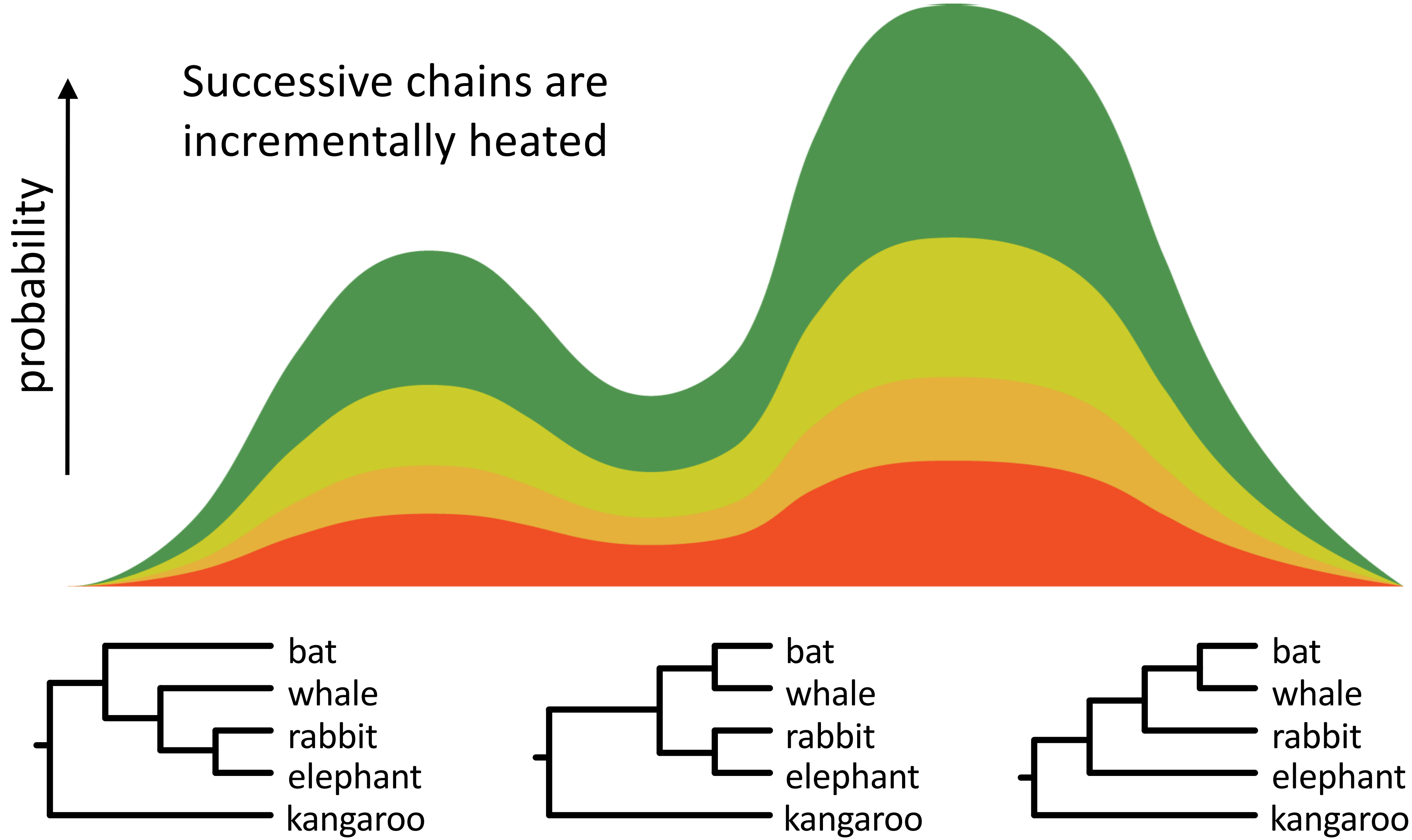








Metropolis-coupled MCMC

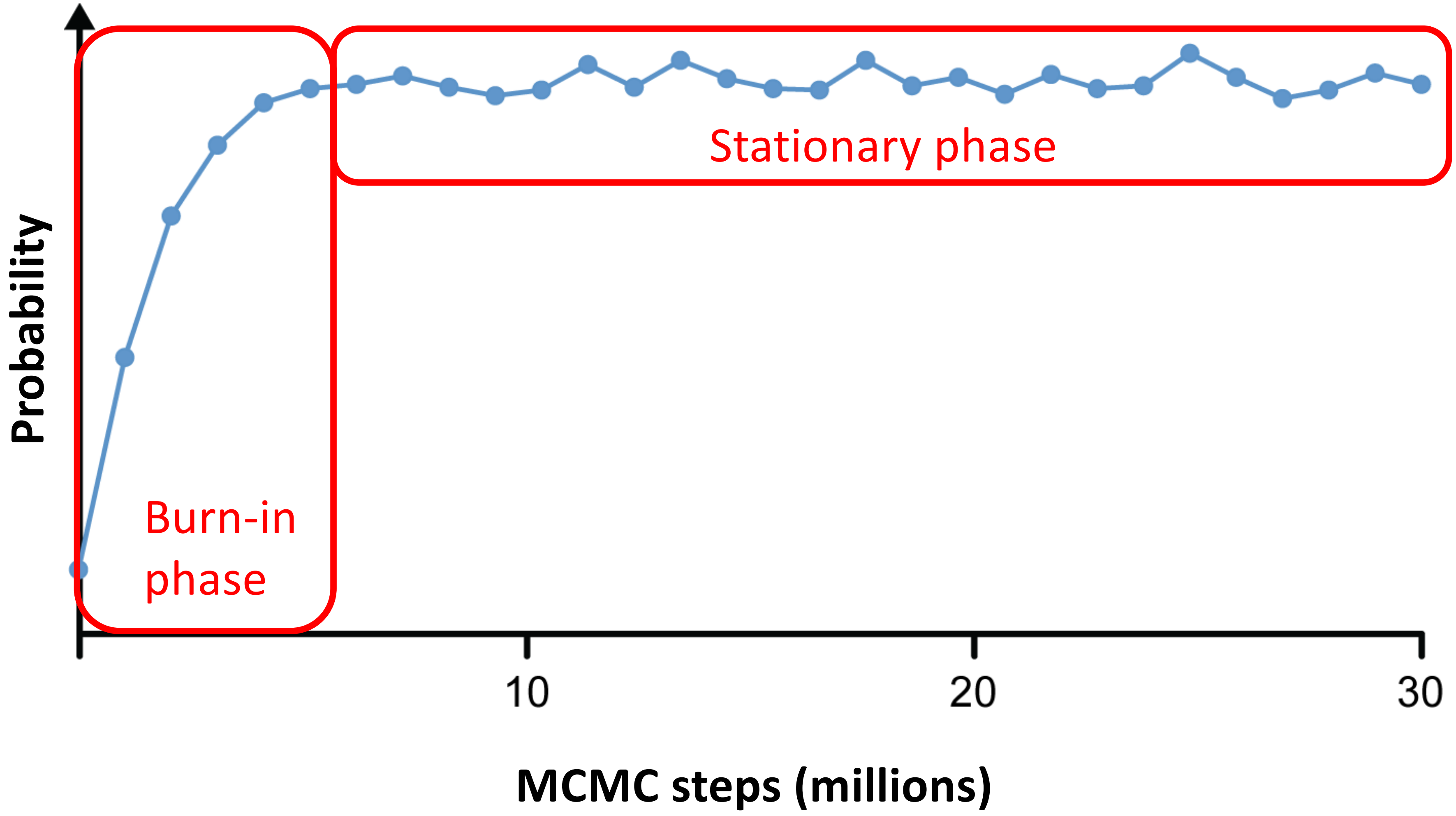


Successive chains are incrementally heated

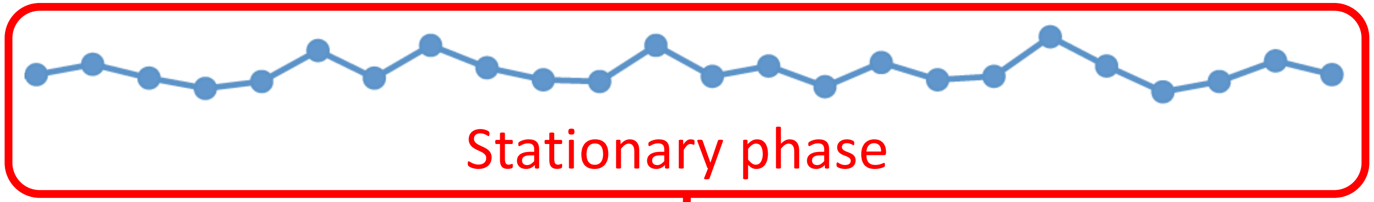
Samples from the MCMC

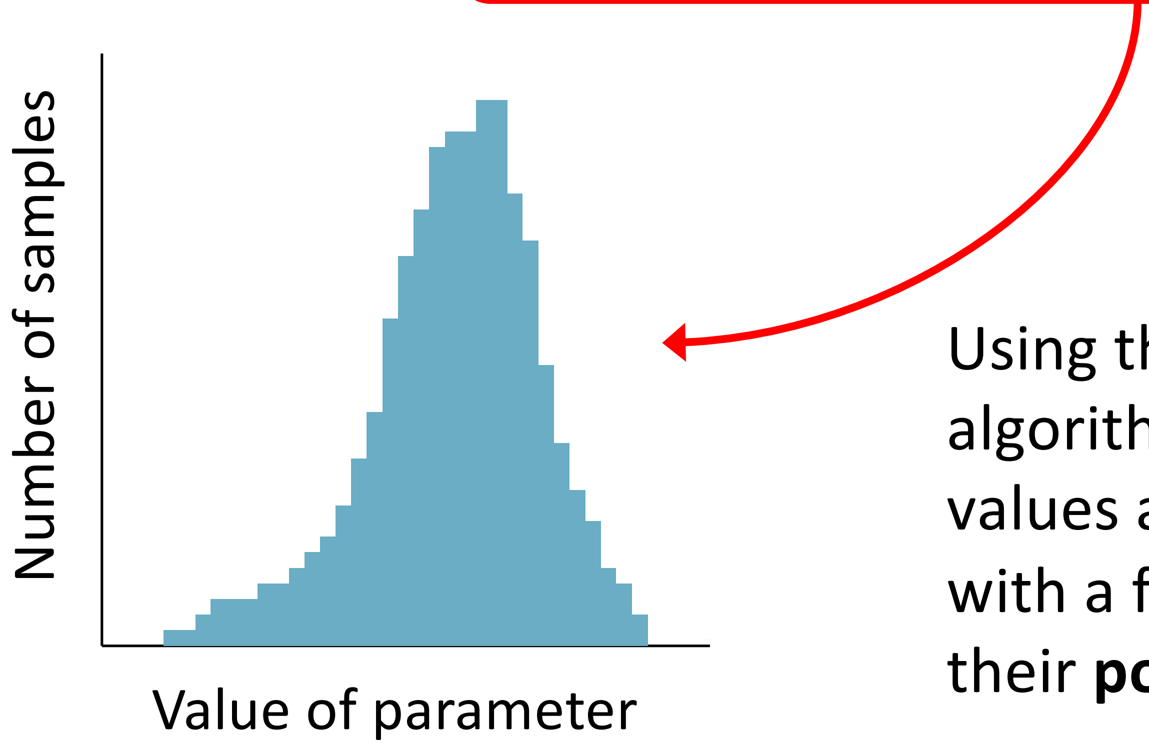
- Output from a Bayesian phylogenetic analysis:
  - A list of the **parameter values** visited by the Markov chain (.p file in *MrBayes*, .log file in *BEAST)*
  - A list of the **trees** visited by the Markov chain (.t file in *MrBayes*, .trees file in *BEAST*)

Samples from the MCMC



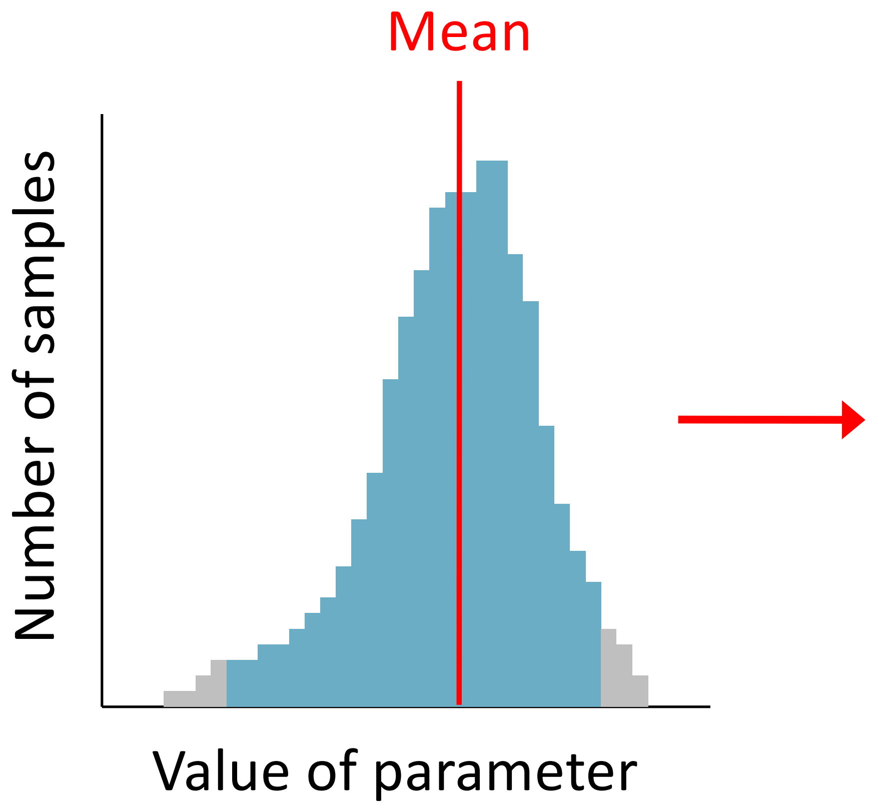
Samples from the MCMC





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



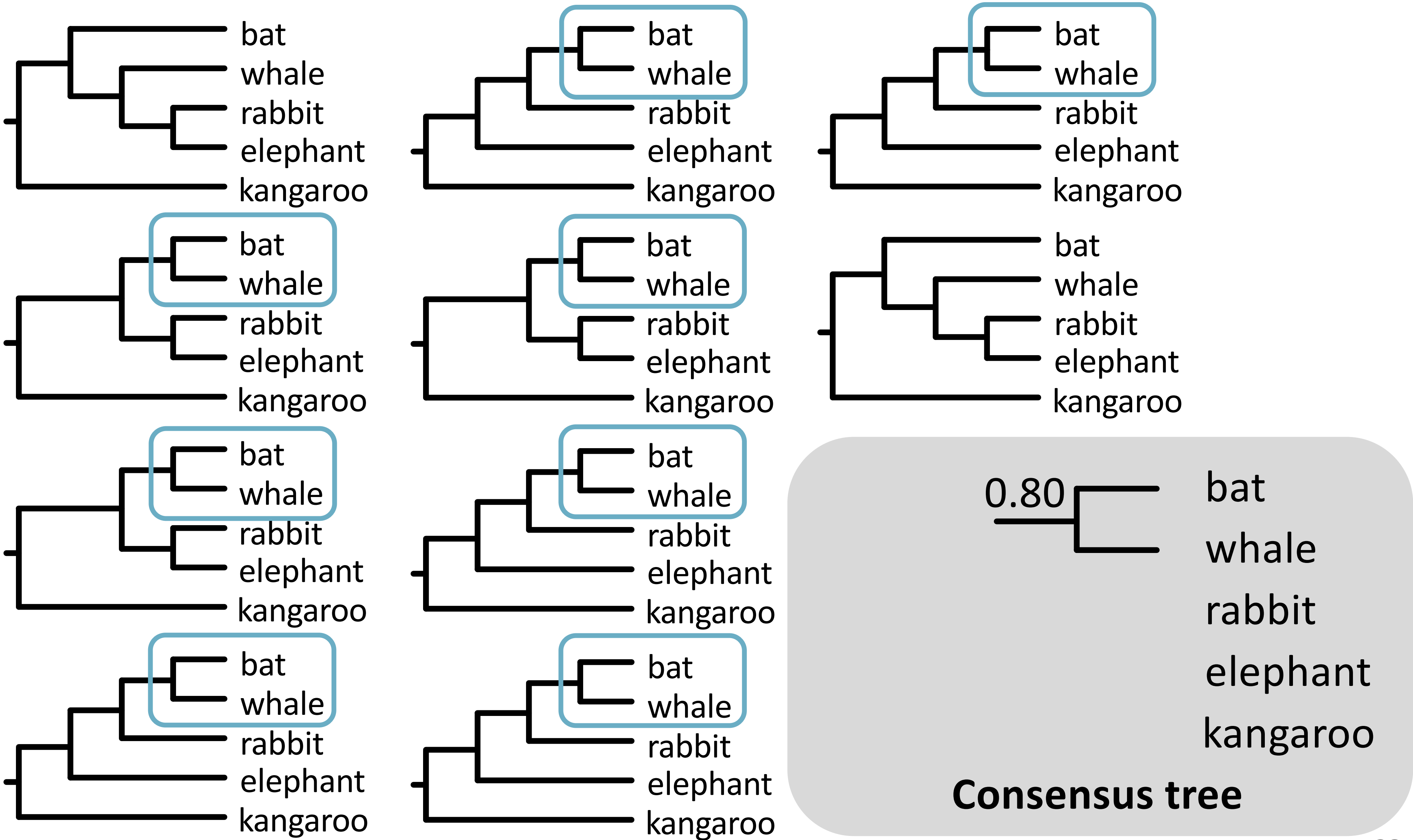
- Take the mean of the sampled values

  Mean posterior estimate

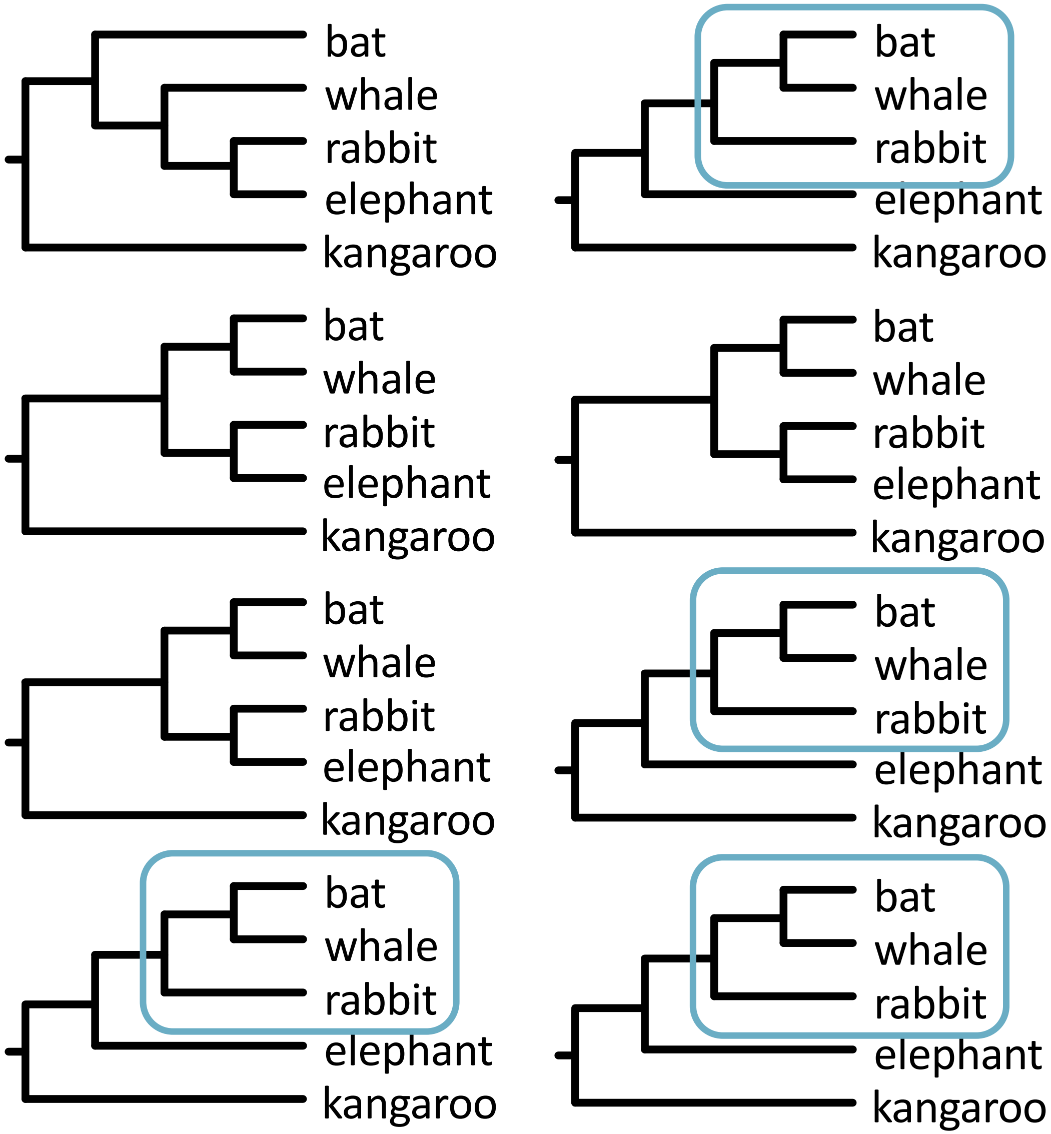
- Take the 'central' 95% of the sampled values

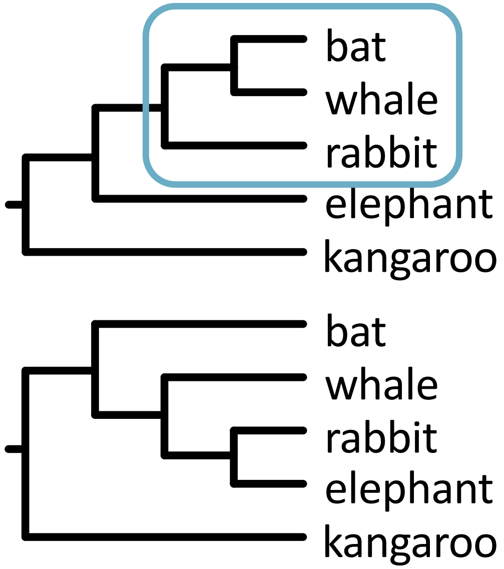
  95% credibility interval

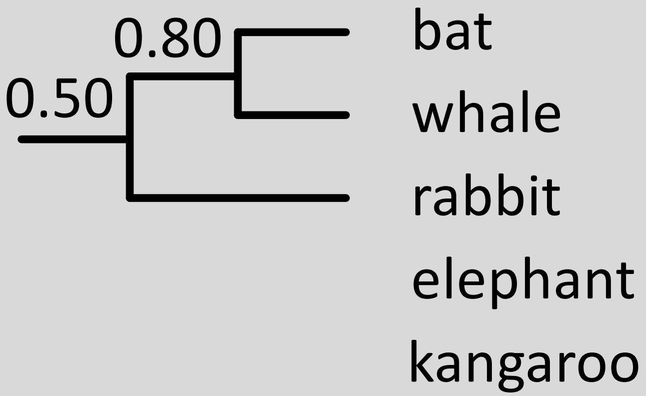
Samples from the MCMC



Samples from the MCMC

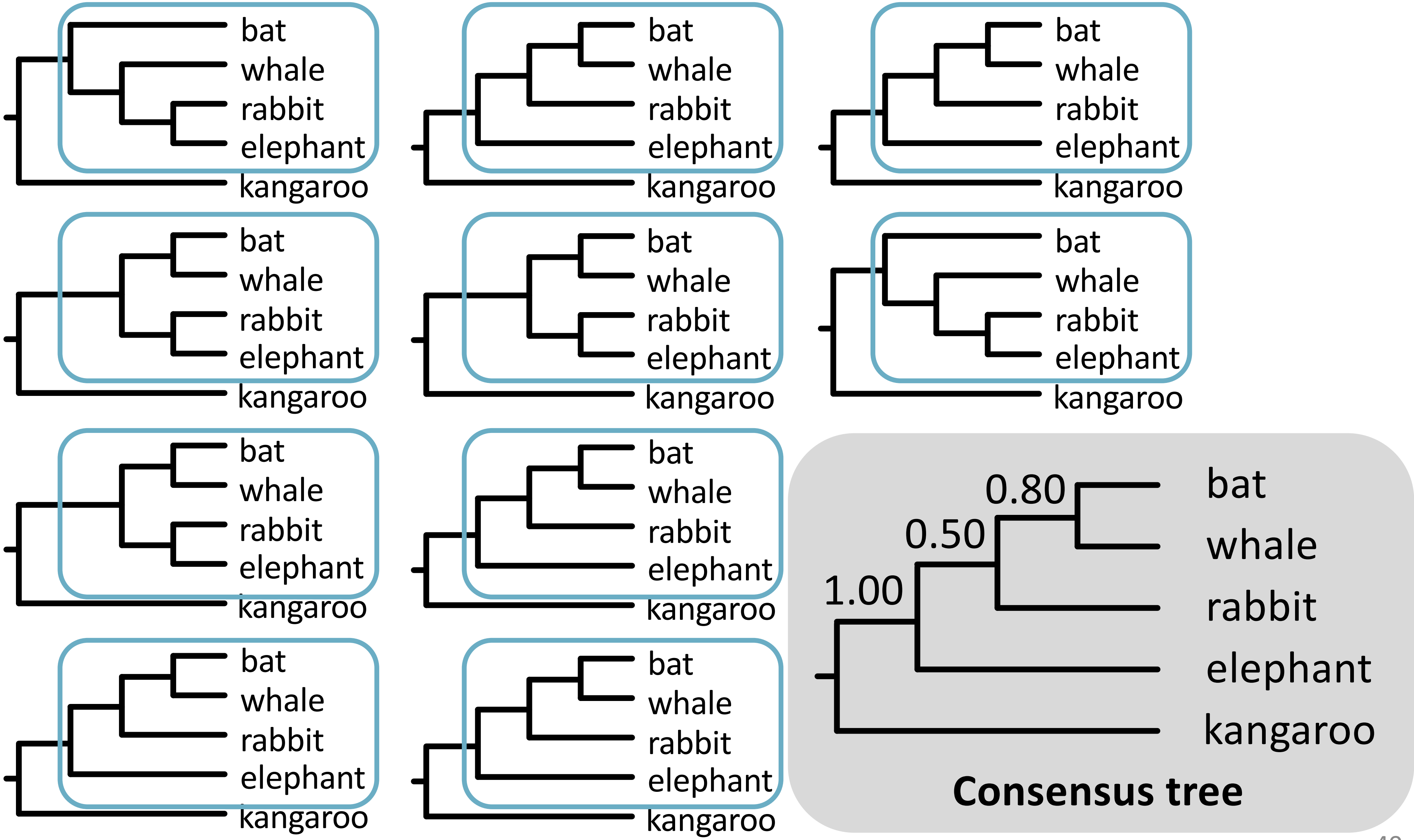






Consensus tree

Samples from the MCMC



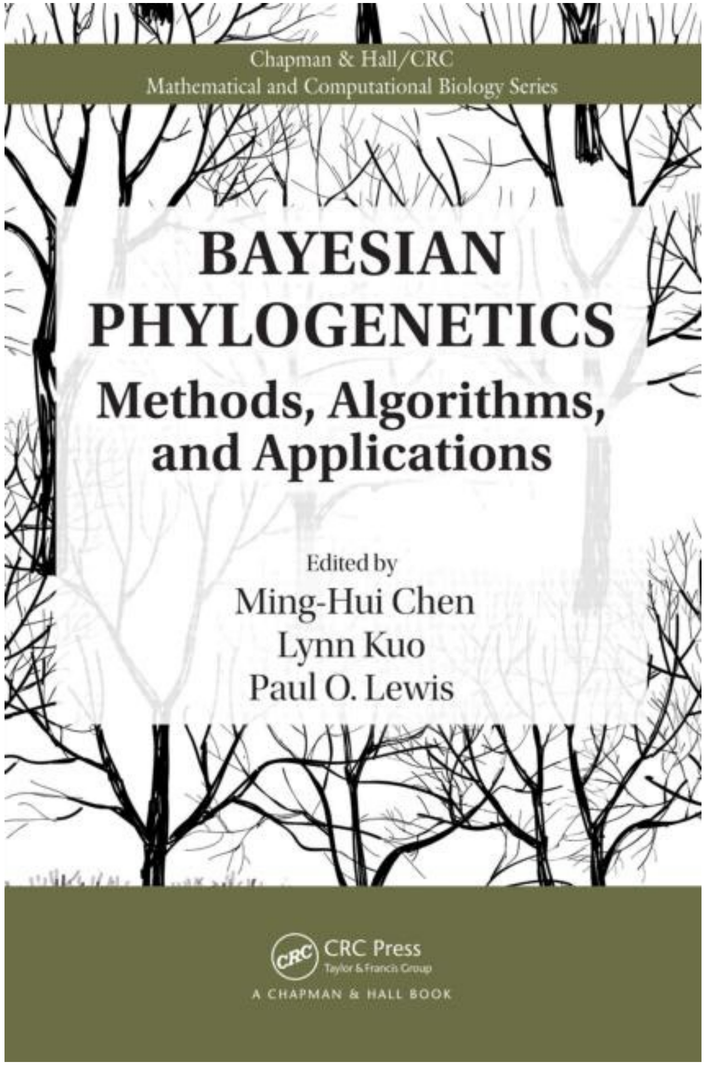
Samples from the MCMC

- **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

Useful references

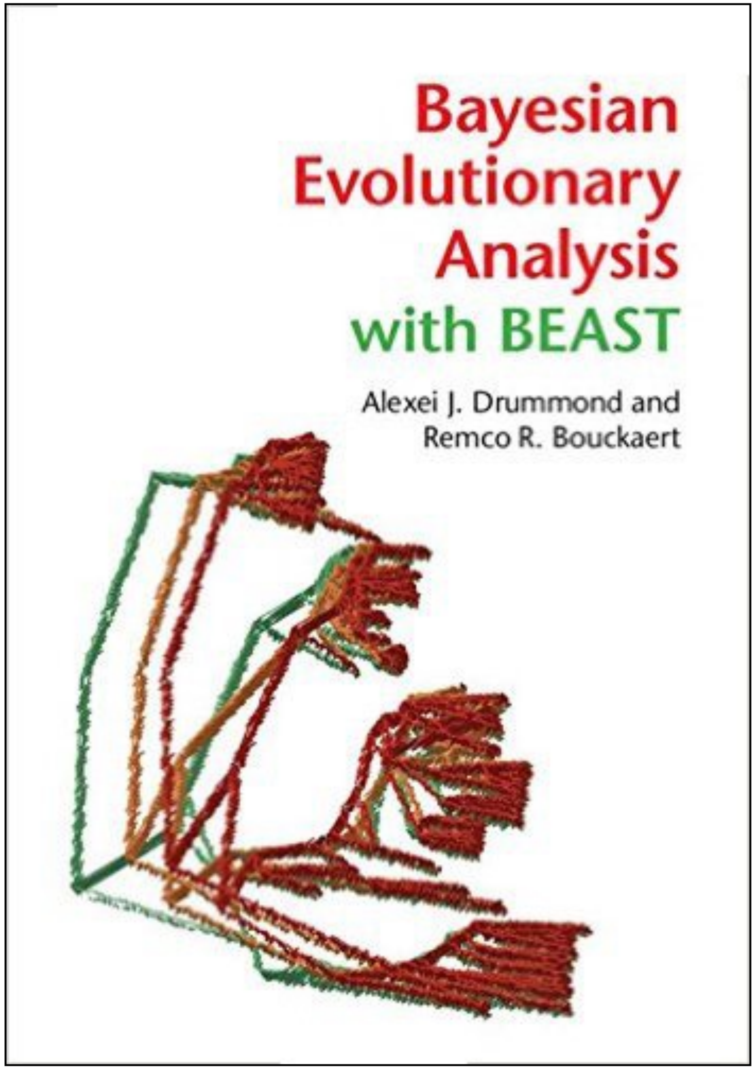


Chapman & Hall/CRC
Mathematical and Computational Biology Series

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