

Lecture 2.1

Bayesian Phylogenetic Analysis

Simon Ho

The Bayesian framework

Phylogenetic methods

	Algorithm-based	Optimality criterion	Other
No explicit substitution model	Distance-based methods	Maximum parsimony	
$\begin{array}{cc} A & \longleftrightarrow & G \\ \updownarrow & \times & \updownarrow \\ C & \longleftrightarrow & T \end{array}$	Distance-based methods	Maximum likelihood	Bayesian inference

2

Bayesian phylogenetic analysis

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

MrBayes



BEAST 1

RevBayes

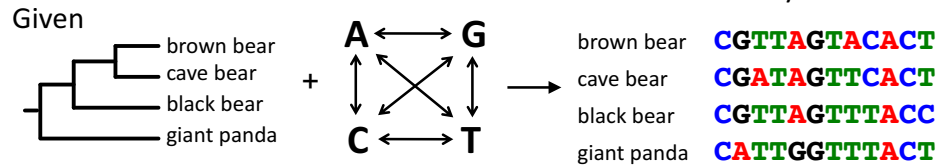


BEAST 2

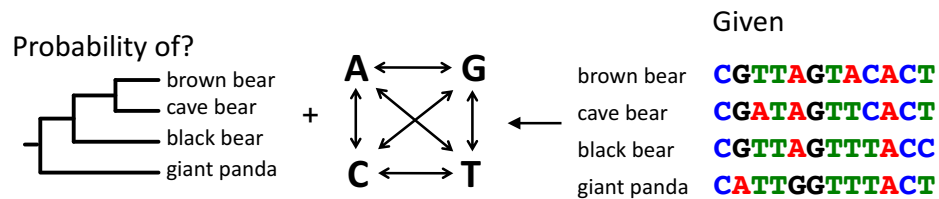
4

Bayesian phylogenetic analysis

Maximum likelihood



Bayesian inference



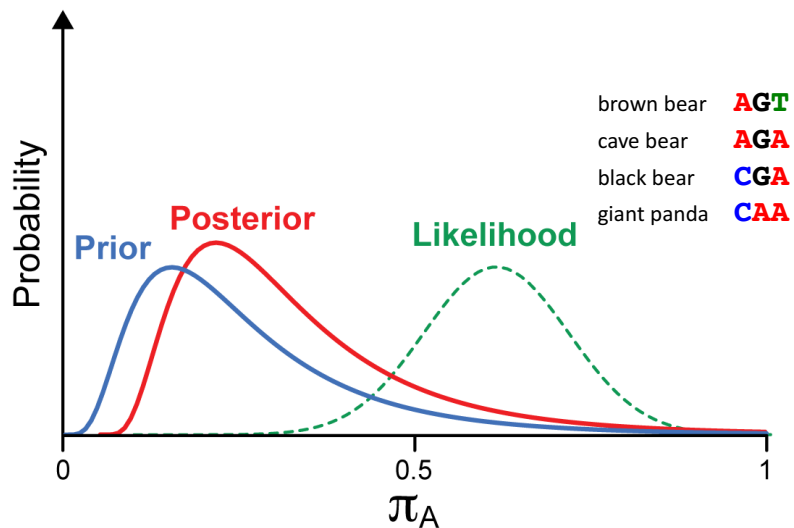
5

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**

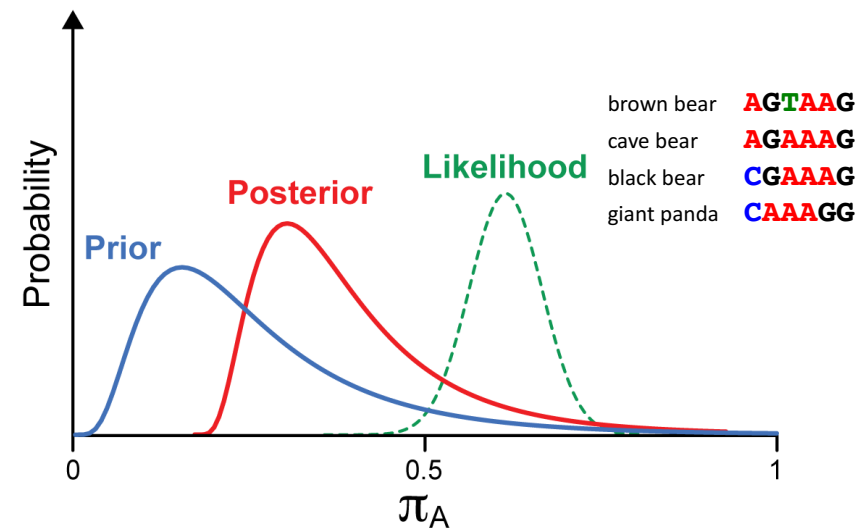
6

Simple example



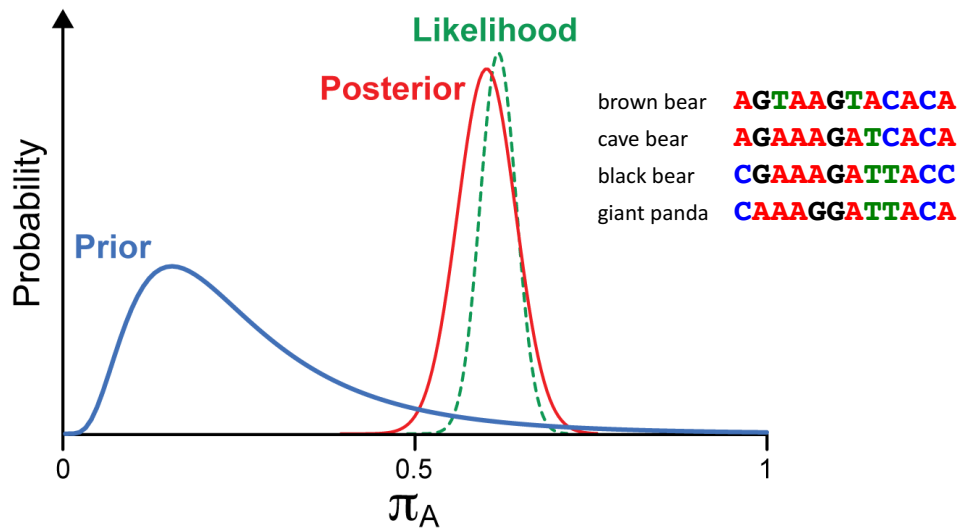
7

Simple example



8

Simple example



9

Bayesian inference

Prior
Specified by user,
independent of data

Likelihood
Calculated from data

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

Posterior
This is what we
want to estimate

normalising constant
marginal likelihood of the data
model likelihood

10

Bayesian inference

Prior prob of tree
Topology
Branch lengths

**Prior prob of substitution
model parameters**
Rate parameters
Base frequencies

**Prior prob of
branch rates**

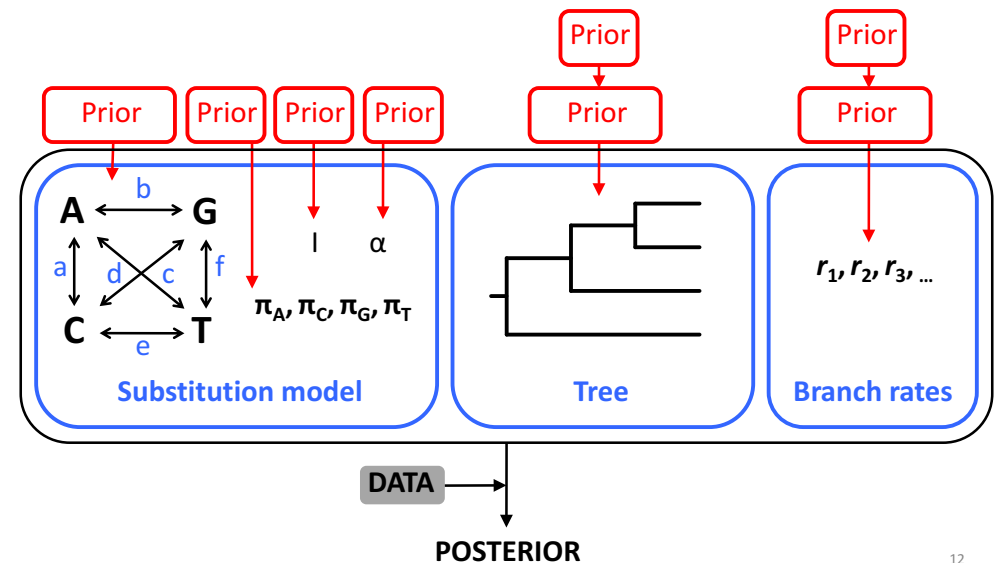
$$\Pr(\tau, M, r | D) = \frac{\Pr(\tau) \Pr(M) \Pr(r) \Pr(D | \tau, M, r)}{\Pr(D)}$$

Posterior
This is what we
want to estimate

Likelihood
Calculated from data

11

Bayesian hierarchical model



12

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

13

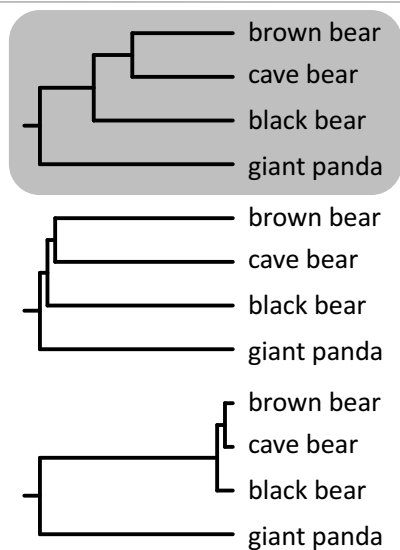
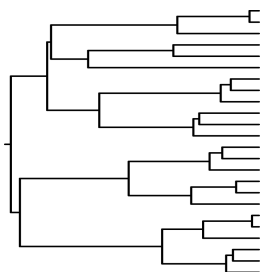
Priors

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

14

Tree prior: Among species

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

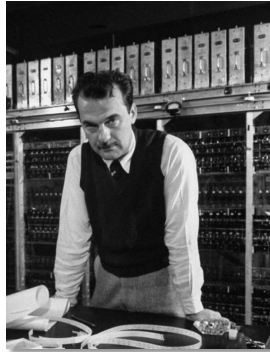


15

Markov Chain Monte Carlo Sampling

Estimating the posterior

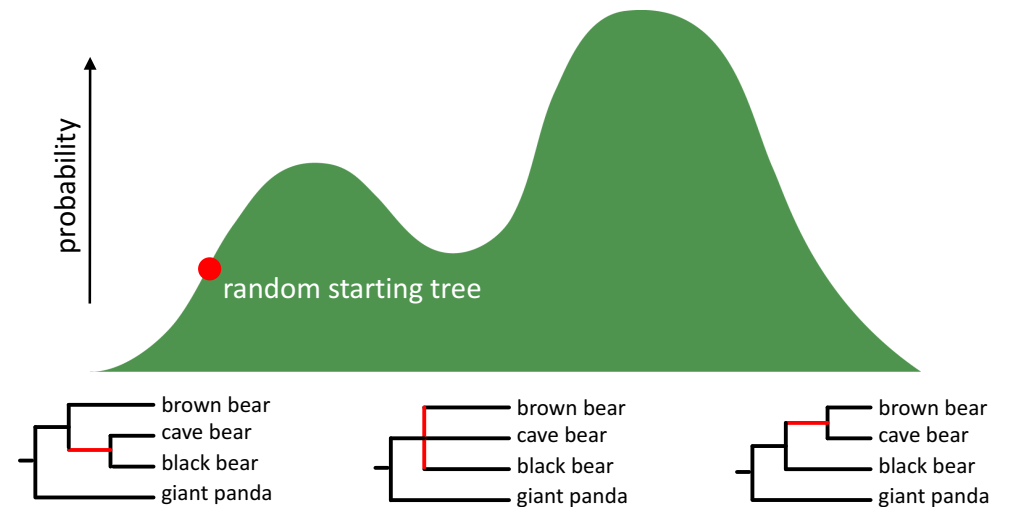
- Impossible to obtain the posterior directly
- Instead, the posterior can be estimated using **Markov chain Monte Carlo simulation**
- This is usually done using the **Metropolis-Hastings algorithm**



Nicholas Metropolis
Los Alamos, 1953

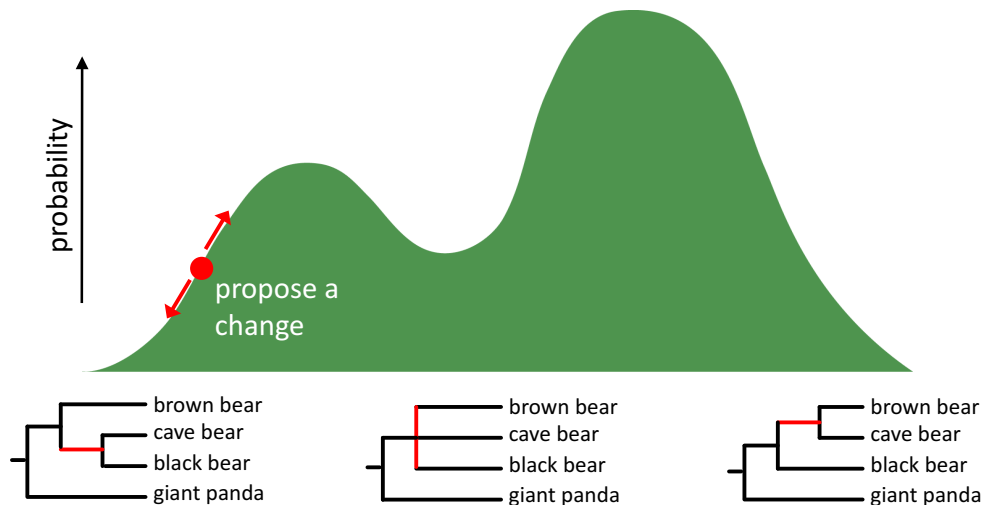
17

MCMC simulation



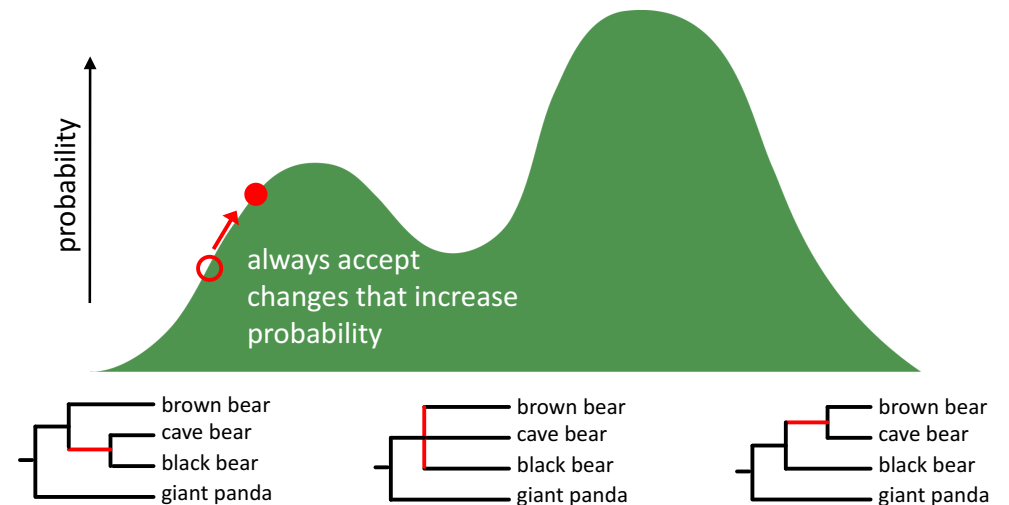
18

MCMC simulation



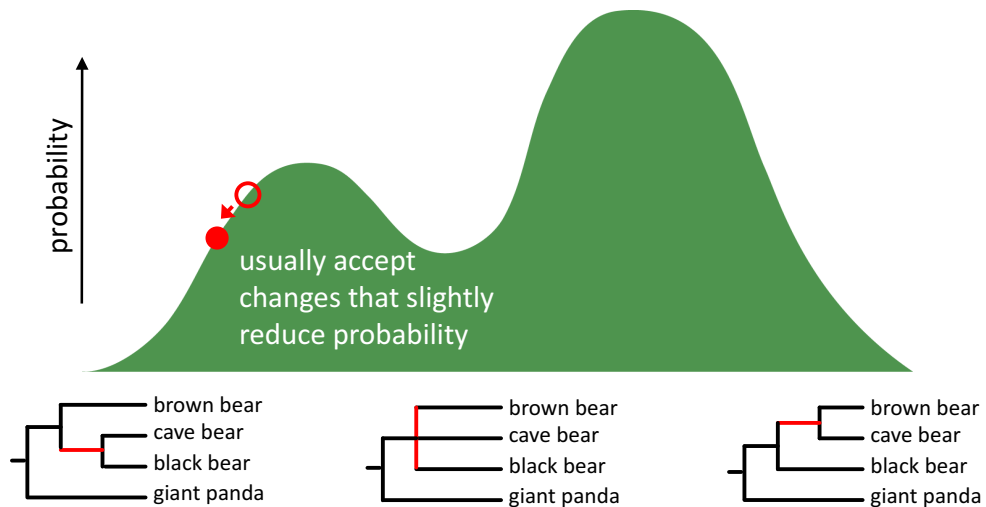
19

MCMC simulation



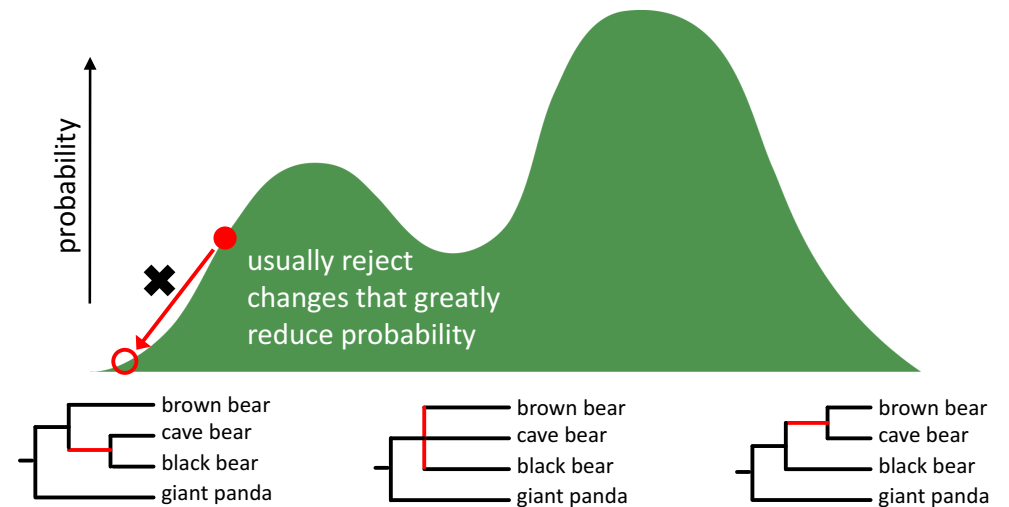
20

MCMC simulation



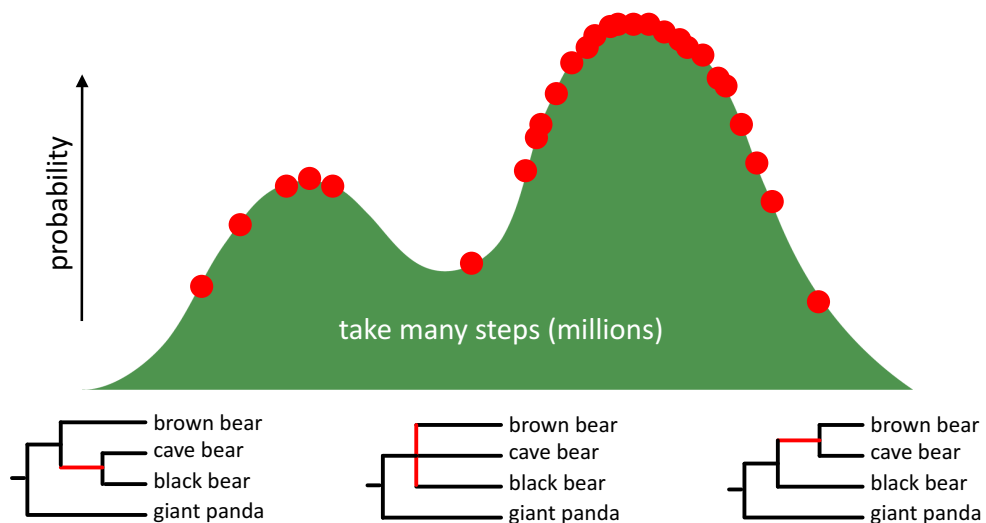
21

MCMC simulation



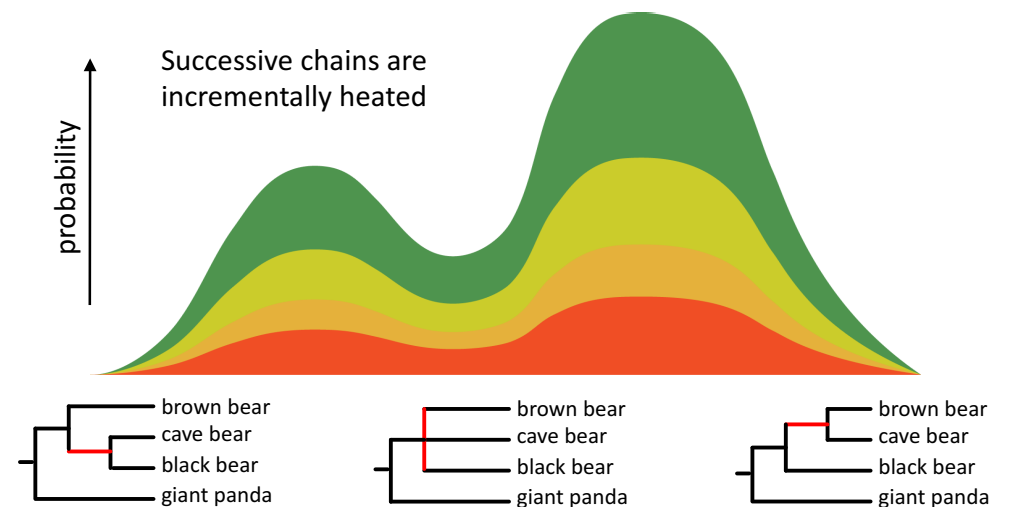
22

MCMC simulation



23

Metropolis-coupled MCMC



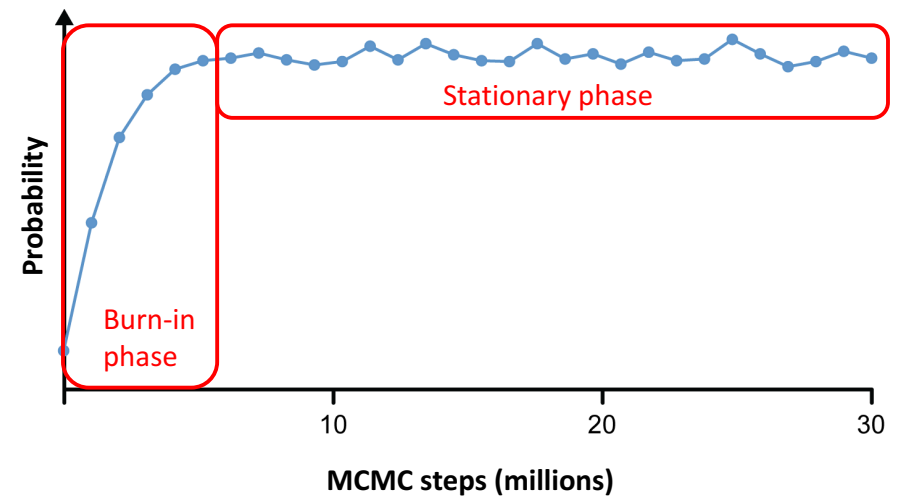
24

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*)

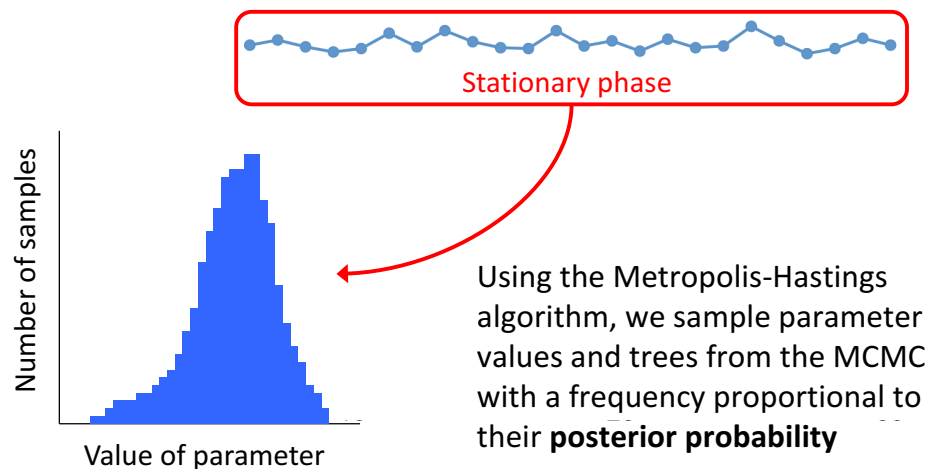
25

Samples from the MCMC



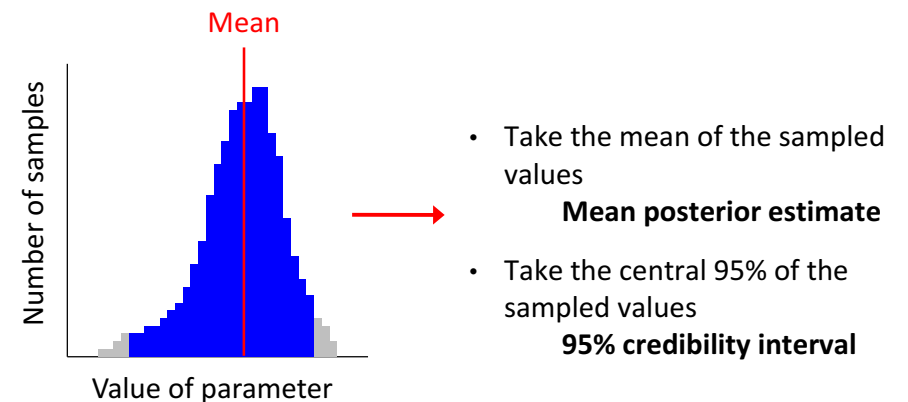
26

Samples from the MCMC



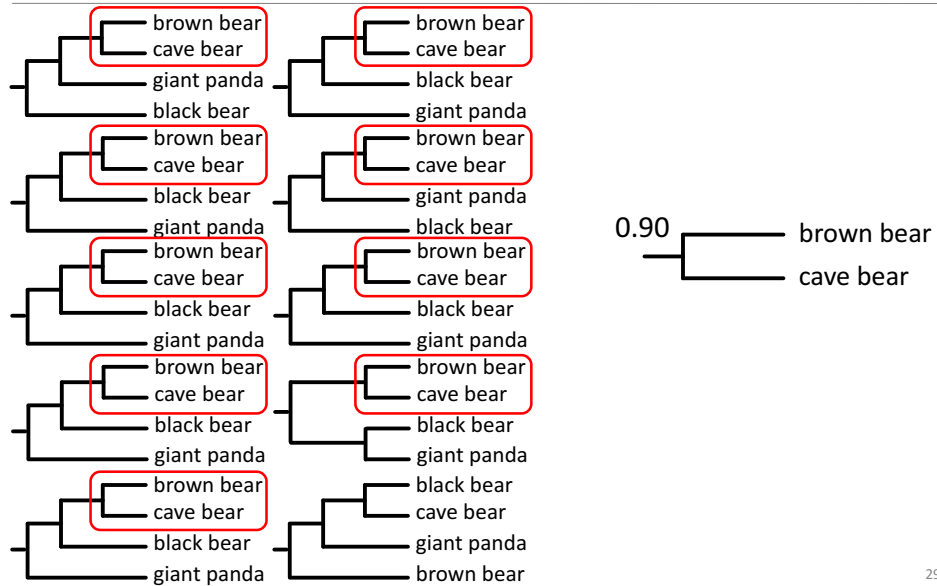
27

Samples from the MCMC



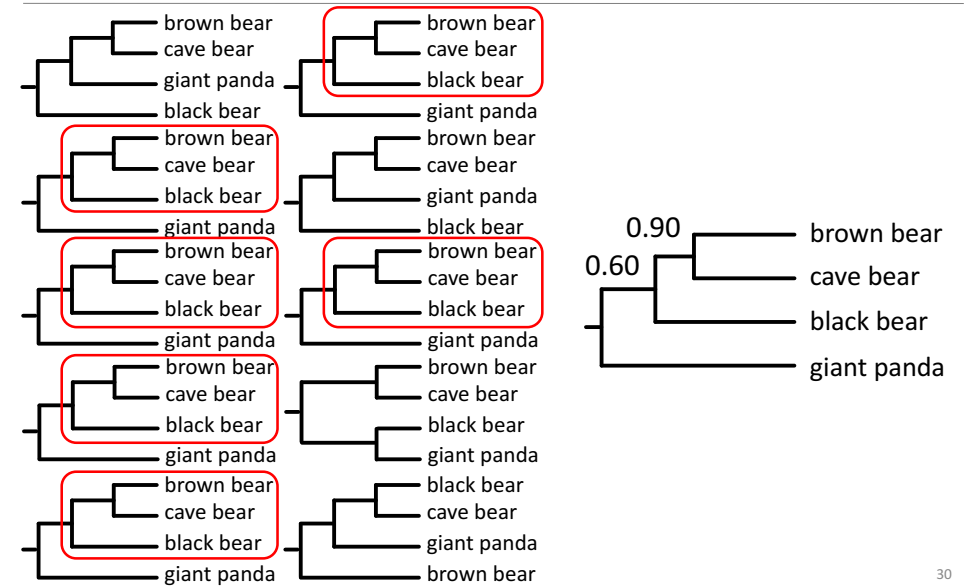
28

Samples from the MCMC



29

Samples from the MCMC



30

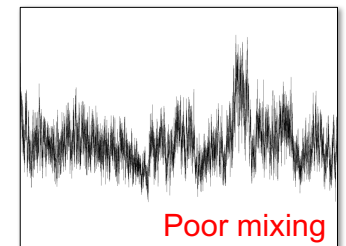
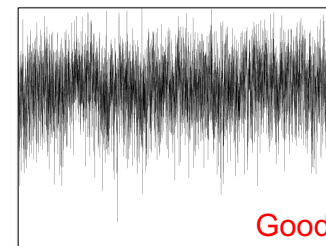
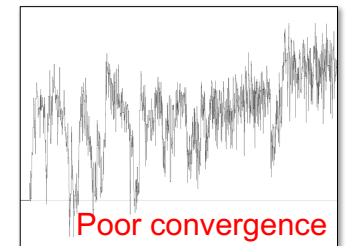
Samples from the MCMC

- **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

31

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?



32

Convergence

- Run at least 2 independent chains
- Posterior probabilities and likelihoods should be similar
- **Model parameters**
 - Check if estimates of model parameters are similar between runs

33

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

34

Advantages and Problems

Advantages

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

36

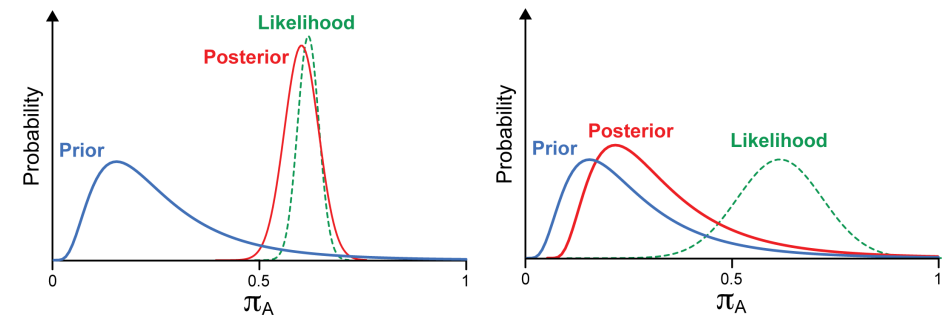
Nuisance parameters

- Integrate over 'nuisance' parameters
- Marginal distribution of a parameter of interest

37

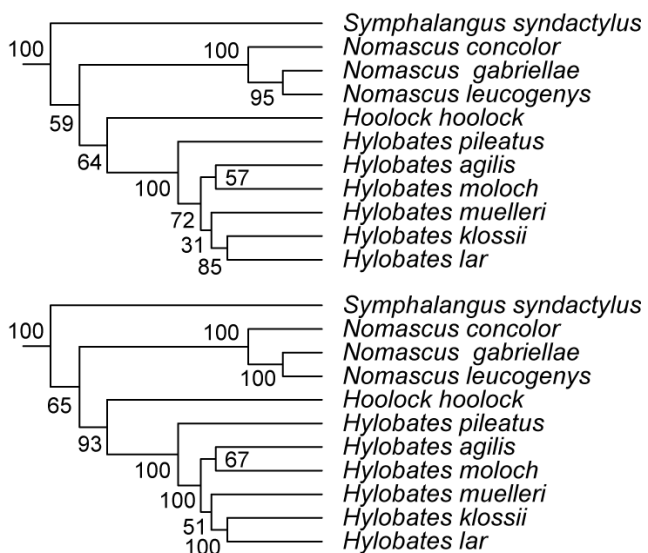
Influence of priors

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



38

Problems: Inflated support values?



Chatterjee et al. (2009) BMC Evol Biol

39

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

40



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*:
<http://beast2.org/beast-features/>

41

MrBayes



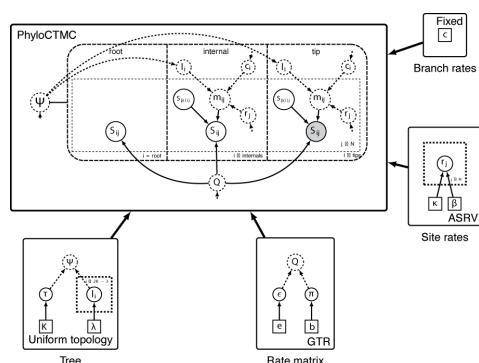
- Primarily designed for species-level data
- Simultaneous estimation of tree and node times
- Range of clock models
- Range of tree priors
- Multiple chains and MCMC diagnostics

42

RevBayes

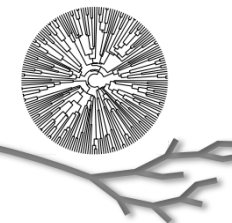
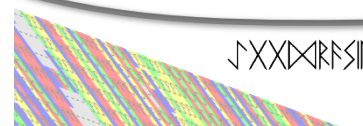


- Uses its own R-like language, Rev
- Interactive construction of graphical model
- Flexible and can be used for simulation and inference
- Ongoing development



43

EXABAYES



- 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*

44

Useful references

