

# **Bayesian Phylogenetic Analysis in Practice**

**Simon Ho**

## **Part 1: Comparison of *MrBayes* and *BEAST***

## ***MrBayes vs BEAST***

- Popular software for Bayesian phylogenetic analysis
- Several key differences
  - Available models
  - Implementations

## ***MrBayes vs BEAST***



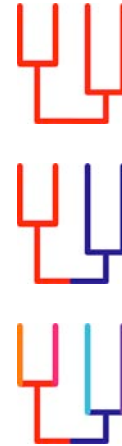
1. Strict clock or no clock
2. Unrooted trees (without clock)
3. No tree prior
4. No estimates of rates or dates (without clock)
5. Multiple MCMC chains



1. Strict clock or relaxed clock
2. Rooted trees
3. Tree prior used
4. Rates and dates
5. Single MCMC chain

# 1. Molecular clocks

- Strict or 'global' clock
  - Many programs / methods / algorithms
- Local clocks
  - Maximum-likelihood (*PAML*, *QDate*)
  - Mean path length (*Pathd8*)
- Relaxed clocks
  - Non-parametric rate smoothing (*r8s*)
  - Penalised likelihood (*r8s*)
  - Bayesian, fixed tree (*multidivtime*, *PhyBayes*)
  - Bayesian, tree co-estimated (*BEAST*)



## What is a 'relaxed clock'?

- Strict clock: rate identical in all branches
- Relaxed clock: rate allowed to vary among branches
  1. Autocorrelated relaxed clock: rates in adjacent branches are related
  2. Uncorrelated relaxed clock: rates identically and independently distributed among branches



## Autocorrelated relaxed clock

- Available in *multidivtime* and *PhyBayes*
- Treat substitution rate as a heritable trait, so that it can 'evolve' through the tree
- Rate is assumed to be tied to:
  - Life history traits (e.g., generation time, population size, body size)
  - DNA proofreading mechanisms
  - Cellular/biochemical environment

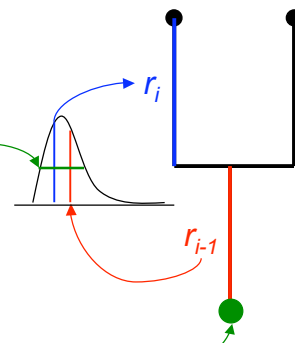
## Autocorrelated relaxed clock

- Model of autocorrelated rate change used to describe prior distribution of rates

- Lognormal  
 $\log(r_i) \sim N(\log(r_{i-1}), vt)$

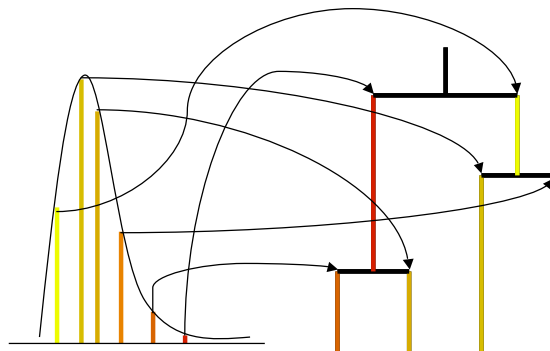
$v$  controls the s.d. of the distribution

Further assumption needs to be made about rate at the root



## Uncorrelated relaxed clock

- Available in *BEAST*



## Uncorrelated relaxed clock

- Models available
  - **Lognormal distribution**  
Most rates cluster around the mean
  - **Exponential distribution**  
Most rates are quite low
- Two statistics can be obtained:
  1. **Coefficient of variation of rates**  
Measures the rate variation among branches
  2. **Covariance of rates**  
Measures autocorrelation of rates between adjacent branches

## 2. Rooted vs unrooted

- *MrBayes* does not produce rooted trees (unless a strict clock is assumed)
  - Root with an outgroup (undesirable)
  - Midpoint-rooting (dodgy)
- *BEAST* produces rooted trees under strict- and relaxed-clock models
  - No need for an outgroup

## 3. Tree Prior

- *MrBayes* does not use tree priors
  - Instead uses prior on branch lengths
  - Branch lengths are identically and independently distributed
  - Default is exponential distribution (with a mean of 0.1)
- *BEAST* can implement several tree priors
  - Coalescent prior
  - Yule prior
- This will be discussed in more detail later in Part 3 of this talk

## 4. Rates and dates

- *MrBayes* cannot estimate rates and dates (unless a clock is assumed)
- *BEAST* will estimate rates and dates if calibrating information is included (more on this in Part 2 of this talk)

## 5. Metropolis-coupled MCMC

- *MrBayes* uses multiple MCMC chains
  - Analysis can be readily parallelised (one chain per computing processor)
  - Quicker convergence and better mixing
- *BEAST* only uses a single MCMC chain
  - Need several independent runs

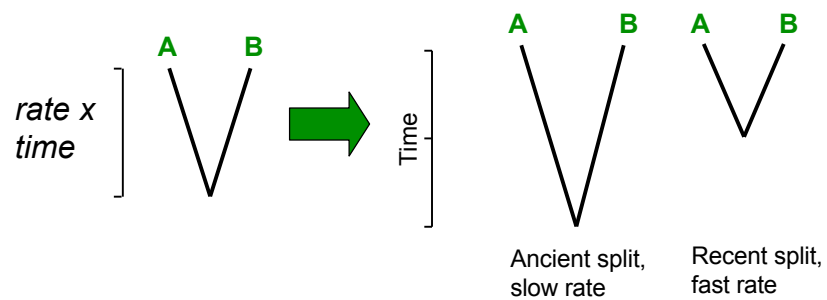
## Summary

- In practice, there are many differences between *MrBayes* and *BEAST*
- *BEAST* has most of the capabilities of *MrBayes*, plus a lot more
- Conclusion: Delete *MrBayes*

## Part 2: Calibrating Estimates of Rates and Divergence Times

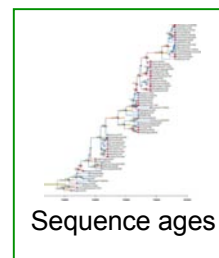
## Why do we need to calibrate?

- Phylogenetic methods usually estimate trees with branch lengths measured in substitutions per site
- Substitutions per site = rate x time



## Separating rate and time

- Information about rate
  - Substitution rate obtained from an independent study
- Information about time:



## Calibration: Fossil Record

- Fossil record provides estimates of divergence times



12% difference

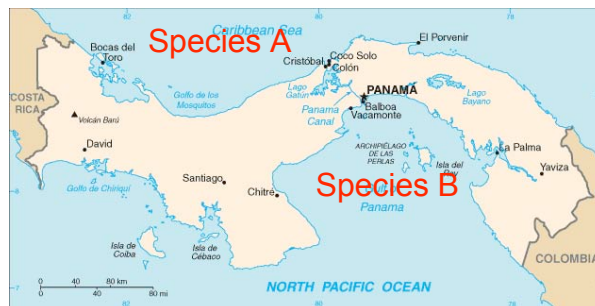
$$\text{Rate} = 12\% / 6 \text{ Myr}$$

$$= 2\% / \text{Myr}$$

6 Myr

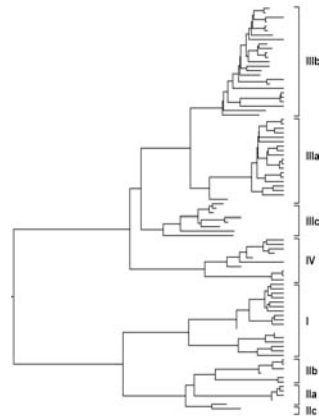
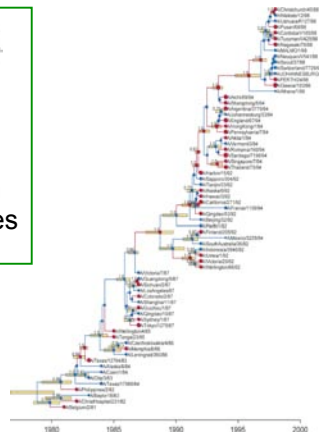
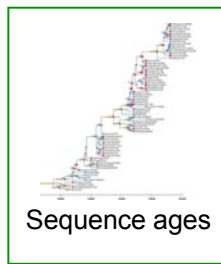
## Calibration: Biogeography

- Biogeographic events can be used to estimate time since divergence



## Calibration: Sequence ages

- Sequence ages provide sufficient age information



## Calibration types

1. Point calibrations (*all programs*)
2. Hard minimum/maximum bounds (*all programs*)
3. Soft minimum/maximum bounds (*PAML*)
4. Parametric prior distributions (*PAML, BEAST*)
  - Normal distribution
  - Lognormal distribution
  - Exponential distribution

# 1. Point calibrations

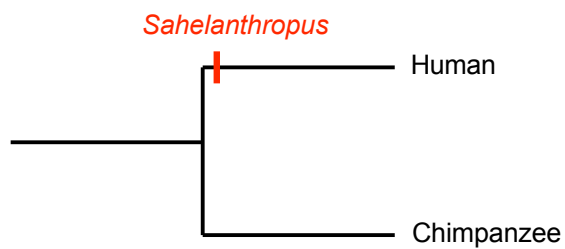
- Traditional usage: **point calibrations**
  - Birds-mammals 300 Ma ago
  - Human-chimp 6 Ma ago
- Often assume no error/uncertainty
- Artificial precision
- Erroneous calibrations will yield biased estimates

# Calibration errors

- Preservational bias
  - Hard parts
  - Environment, proximity to water bodies
  - Age
  - Sampling effort
- Taxonomic affinity
  - Fragmentary fossils
  - Extinct, stem lineages
- Stratigraphic and isotopic dating errors

## 2a. Minimum bounds

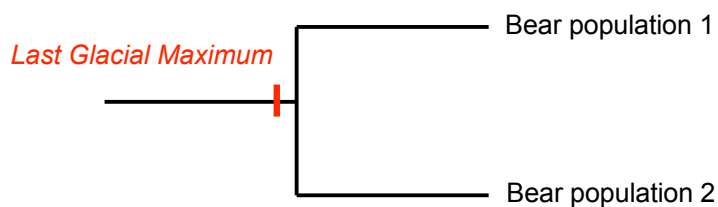
- Minimum Bounds



- Discards potentially useful information
- Inadequate information for divergence dating

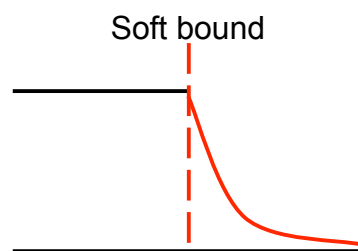
## 2b. Maximum bounds

- Necessary for most methods
- Compromise between:
  - Setting it too low: might exclude true date
  - Setting it too high: uninformative

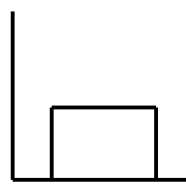


### 3. Soft Bounds

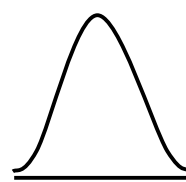
- Extension of hard bounds
- Assign non-zero probability to values outside bound
- Able to forgive calibration errors



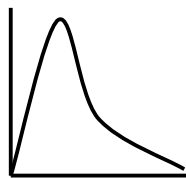
### 4. Probability Distributions



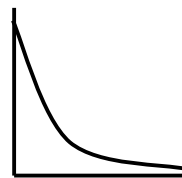
Uniform



Normal



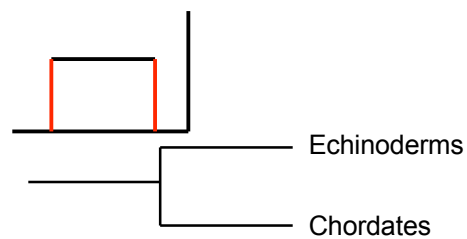
Lognormal



Exponential

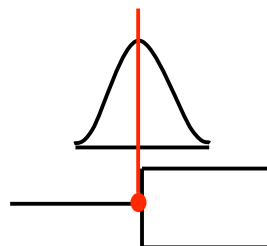
## 4a. Uniform Distribution

- Two parameters (minimum, maximum)
  - Origin of metazoan phyla
    - Minimum 545 Ma
    - Maximum 1,500 Ma



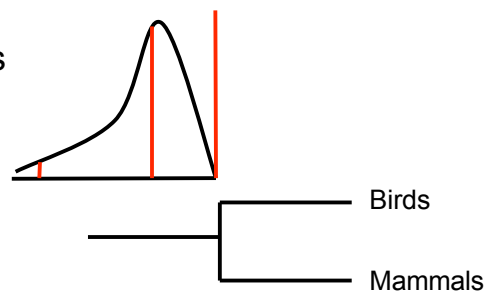
## 4b. Normal Distribution

- Two parameters (mean, s.d.)
  - Radiometric dating errors (approximately)
  - Secondary calibrations
- Some biogeographic calibrations
- Some fossil calibrations



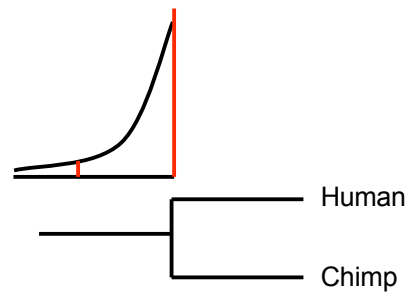
## 4c. Lognormal Distribution

- Three parameters (minimum, mean, s.d.)
  - Bird-mammal
    - Minimum 288 Ma
    - Mean 300 Ma
    - Maximum 310 Ma
- Most fossil calibrations



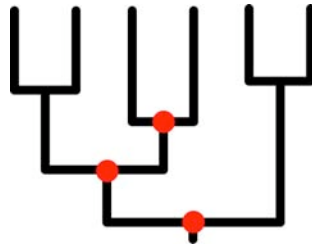
## 4d. Exponential Distribution

- Two parameters (minimum, mean)
- Some fossil calibrations



## Multiple Calibrations

- Using multiple calibrations is less risky
- Improves estimates using relaxed-clock methods
- Poor calibrations can be corrected in the analysis
  - Posteriors different from priors
- Calibrations interact with each other
- Calibrations can affect topological inference



## Summary

- In practice, how do I calibrate my dating analysis?
  - **Ideal:** Hard minimum, mean/mode, soft maximum
  - **Good:** Hard minimum, soft maximum
  - **Satisfactory:** Hard minimum

### Accounting for calibration uncertainty in phylogenetic analyses

Ho SYW and Phillips MJ

In review for *Systematic Biology*

## Part 3: Handling Intraspecific and Interspecific Data

### Intraspecific vs Interspecific

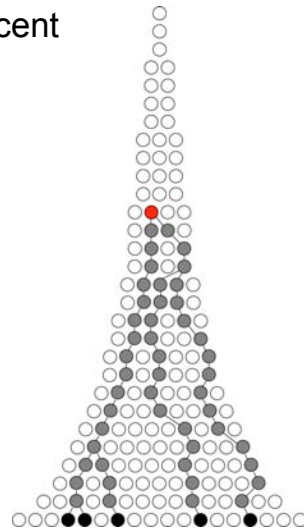
Intraspecific	Interspecific
<ul style="list-style-type: none"><li>• Sequences from individuals in a population</li><li>• Low sequence divergence</li><li>• Estimating genealogy</li><li>• Gene trees often incongruent</li><li>• Rate assumed to be constant throughout genealogy</li><li>• Tree shape described by coalescent process</li></ul>	<ul style="list-style-type: none"><li>• One sequence from each species</li><li>• Higher sequence divergence</li><li>• Estimating phylogeny</li><li>• Gene trees usually congruent</li><li>• Rate often assumed not to be constant throughout phylogeny</li><li>• Tree shape described by birth-death process</li></ul>

## Intraspecific data

- Individuals from a single population or single species
- Recent divergence times
- Different genes have different histories

## Intraspecific data

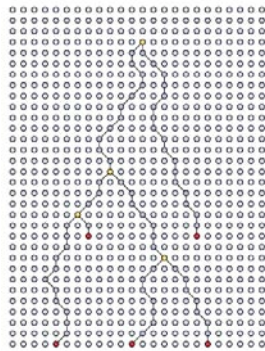
- Tree shape described by a coalescent process
- Trace relationships back in time from modern haplotypes



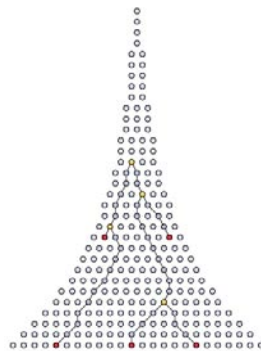
## Intraspecific data

- Different population histories will lead to different coalescent tree shapes

Constant size



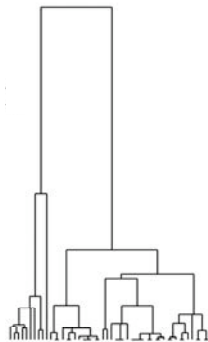
Exponential growth



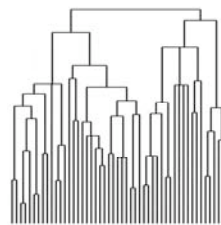
## Intraspecific data

- Different population histories will lead to different coalescent tree shapes

Constant size



Exponential growth

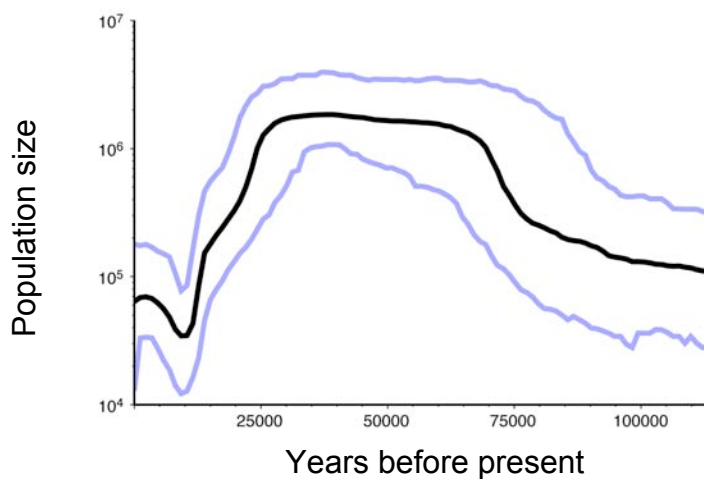


## Intraspecific data

- Coalescent model used to put a prior on the tree
- Different demographic models available:
  - Constant population
  - Exponential growth
  - Logistic growth
  - Boom-bust
- Parameters of these models are estimated
- Or can implement non-parametric “Bayesian skyline plot”

## Intraspecific data

- Bayesian skyline plot: Bison

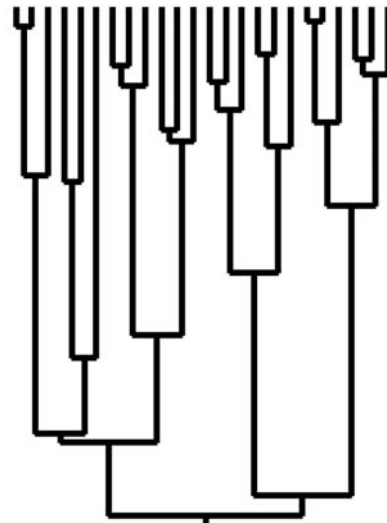


## Interspecific data

- One sequence from each species
- Deeper divergence times
- Gene trees should be congruent
- Tree shape described by birth-death process
  - 3 parameters: birth rate, death rate, sampling probability
  - 2 parameters: birth rate, death rate
  - 1 parameter: birth rate (Yule process)

## Interspecific data

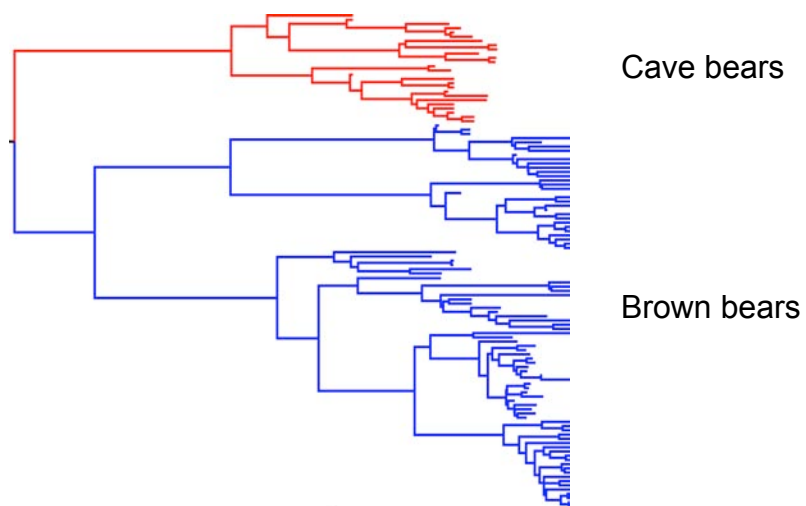
- Yule process:
  - One ancestral lineage splits into two
  - Lineages split at a constant rate
  - Simulates speciation process



## Combined data

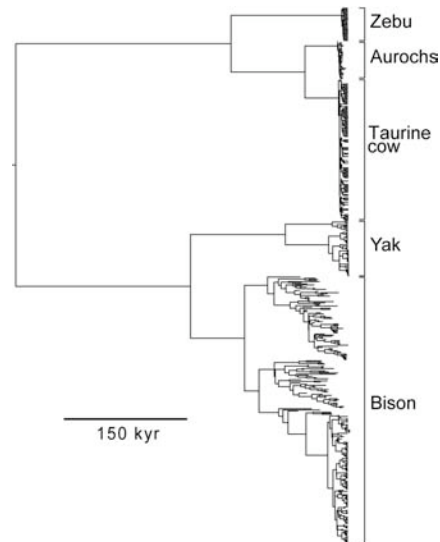
- Some data sets have multiple representatives of each species
- This is problematic because there is no appropriate tree prior
- A separate coalescent model can be used for each species in a single analysis

## Combined data



## Combined data

- Five bovine species
- One demographic model for each species



## Summary

- In practice, what's the difference between analysing intraspecific and interspecific data?
  - **Necessary**
    1. Different prior on the tree
    2. Multilocus for intraspecific, partitioned for interspecific
  - **Optional**
    1. Strict clock for intraspecific, relaxed clock for interspecific