# Bayesian Inference in Phylogeny

Trang Do
Morgan McCarty

# Bayesian Statistics – In a Nutshell

a statistical estimation method based on **Bayes' Theorem**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- P(A|B): **posterior probability**; the probability of A, given the evidence B
- P(B|A): **likelihood function**; the probability of the evidence B given that A is true
- P(A): **prior probability**; the belief in A before taking into account B
- P(B): **the probability of the evidence**; in simple terms, a probabilistic relation between collected data and A

https://en.wikipedia.org/wiki/Bayesian_inference_in_phylogeny

# Bayesian Inference in Phylogeny

- P(A): the prior probability of a given tree occuring
- P(A|B): the posterior probability of a tree being correct, given the prior, the data, and the correctness of the likelihood model
- P(B|A): the likelihood of our data given the prior
- P(B): *as we will see this becomes significantly less important*

But how do we estimate the probability of a tree being correct?

# Markov Chain Monte Carlo (MCMC)

- Estimating the posterior distribution is very difficult because of P(B)
- Commonly we estimate it proportionally without P(B) $P(A \mid B) \propto P(B \mid A)P(A)$
- MCMC randomly walks through possible posteriors and estimates their fitness, eventually approximating the ideal posterior - P(B) is “unimportant”
- Estimating the posterior also requires calculating the likelihood which is a substep of the algorithm

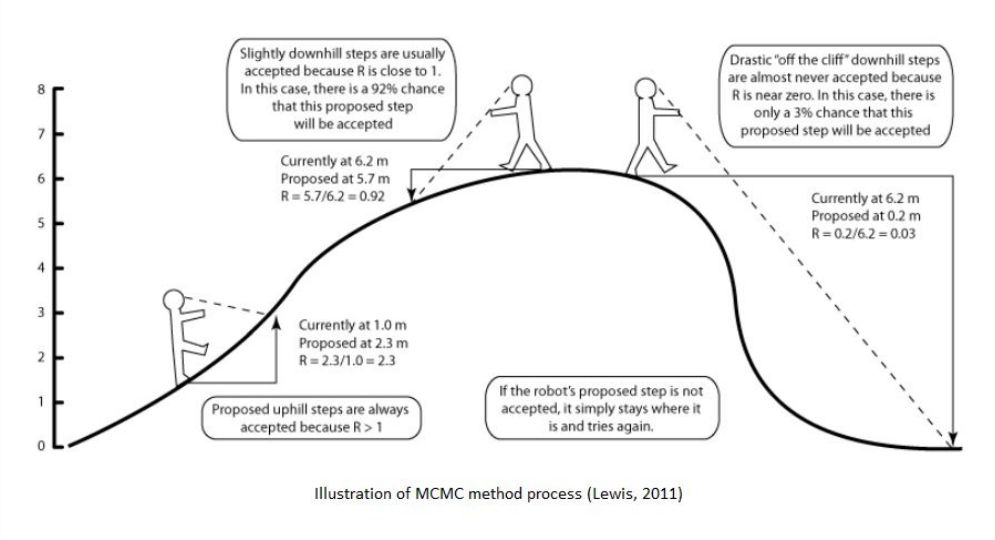


Illustration of MCMC method process (Lewis, 2011)

# Building a Tree using Bayesian Probability

## Data/Model Preparation

**What species are we interested in?**
Best fit model for the given data?
Any initial assumptions?

## Exploring Probabilities

**Run the MCMC process!**
Try out different tree constructions and see what sticks - best fit the given data :)

→ Preliminary tree!

## Assess Tree Construction

**Is the tree reliable?**
Check for consistency between tree generation simulation
If unsure how species evolve
→ compare different models against each other

# Advantages (Why Bayesian?)

- **More than just a best guess tree!**
  - Captures uncertainty when it comes to building trees
  - Constructs *many* trees compared to one best guess tree
- **Considers the weight of prior knowledge**
  - Gain more confidence in sparse/noisy dataset, or when relationships are unclear
- **Possibilities with different models of evolutions**
- **Capable of handling complex/mixed data**

# Disadvantages

## Bootstrap values vs posterior probabilities

- Bootstrap values tends to be lower/less extreme in bipartitioning
- Overconfidence in posterior probabilities? (Literature is spare, what this means is hard to say)

## Controversy of using prior probabilities

- Might incorporate subjective data selection

## Model choice

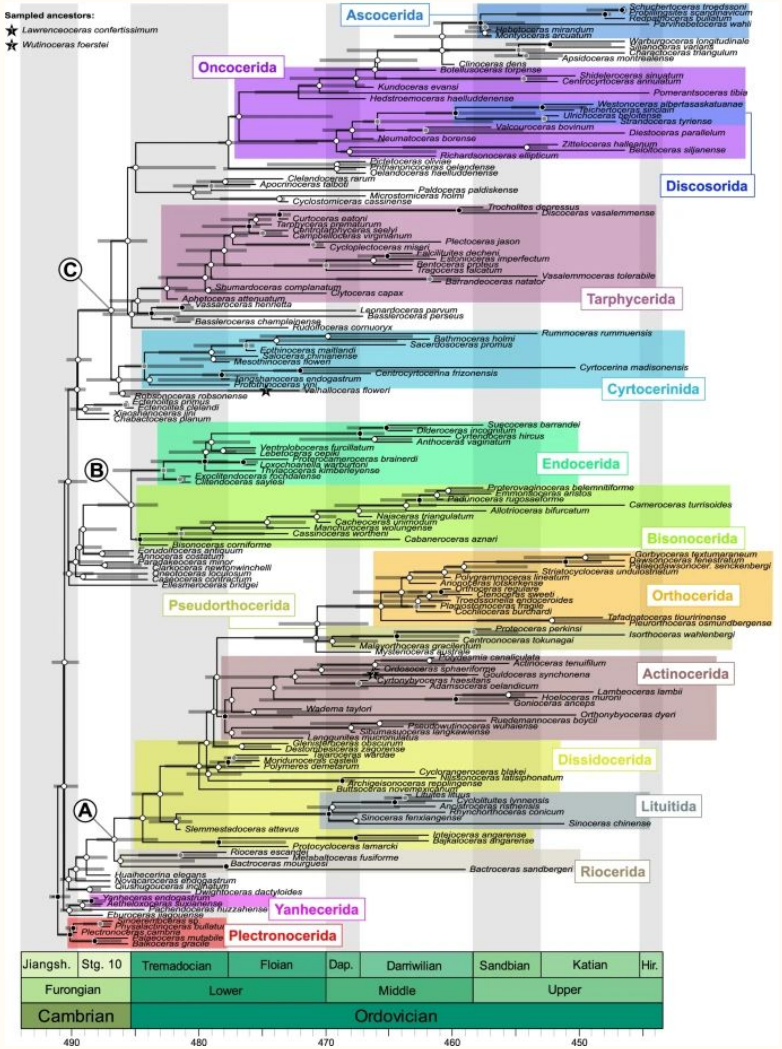
- Oversimplified models might result in higher posterior probabilities
- Not as intuitive to interpret compared to traditional tree construction

# Applications of Bayesian Inference

- Inference of phylogenies
- Inference and evaluation of uncertainty of phylogenies
- Inference of ancestral character state evolution
- Inference of ancestral areas
- Molecular dating analysis
- Model dynamics of species diversification and extinction
- Inference of phenotypic trait evolution

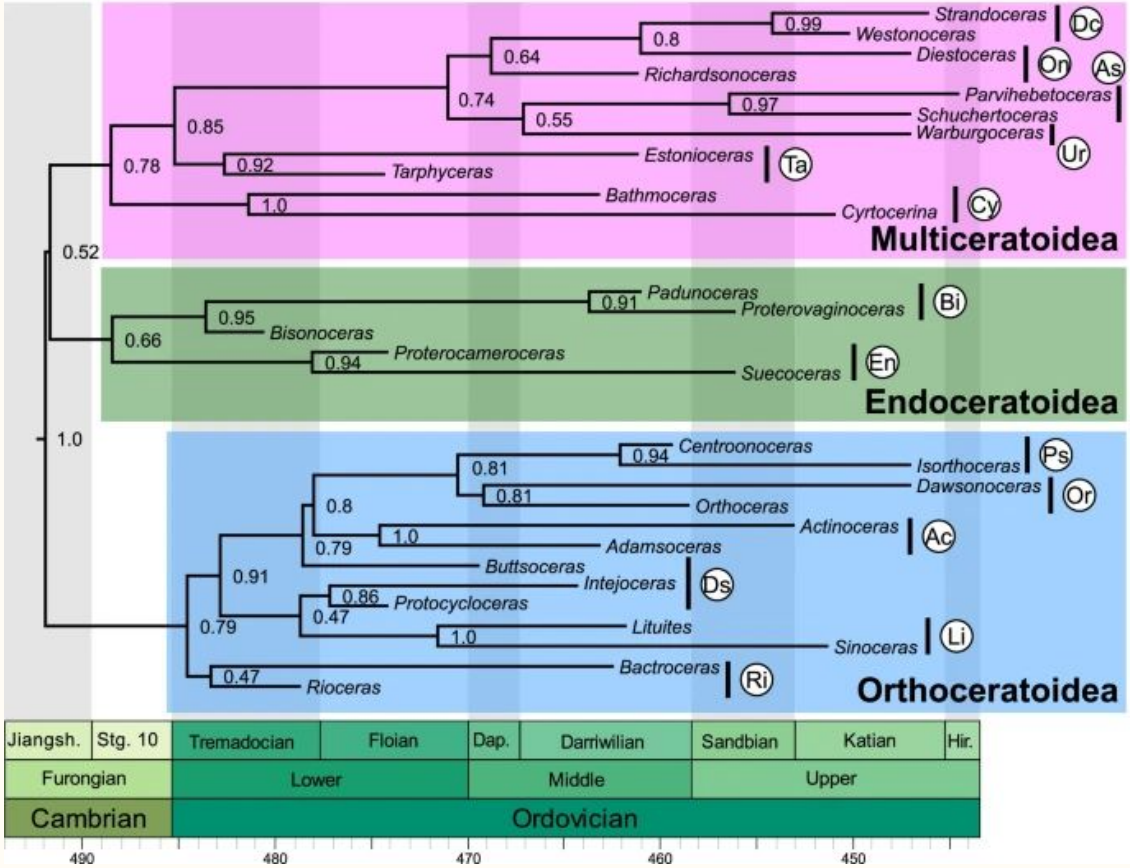
(Really anywhere where probabilistic measurements could be made)

# Full maximum clade credibility tree of early cephalopod evolution

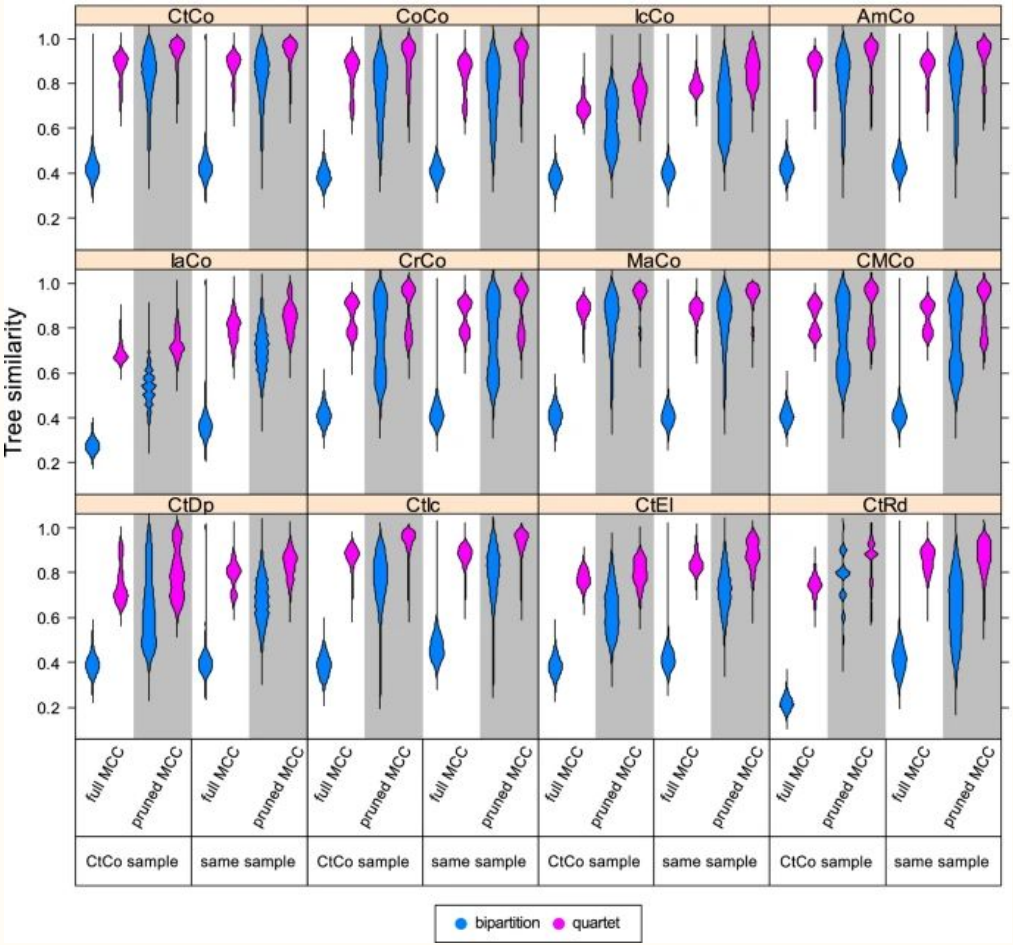



Pohle, A., Kröger, B., Warnock, R.C.M. et al. Early cephalopod evolution clarified through Bayesian phylogenetic inference. BMC Biol 20, 88 (2022). https://doi.org/10.1186/s12915-022-01284-5

# Pruned clade credibility tree of early cephalopod evolution




Pohle, A., Kröger, B., Warnock, R.C.M. et al. Early cephalopod evolution clarified through Bayesian phylogenetic inference. BMC Biol 20, 88 (2022). https://doi.org/10.1186/s12915-022-01284-5




# Distributions of tree comparisons.

Each set of posterior trees is compared to **four different single trees**: the full and the pruned MCC tree of the main analysis and the full and the pruned MCC tree resulting from the same posterior tree sample. Comparisons are made with bipartition (blue) and quartet (pink) similarity.

Pohle, A., Kröger, B., Warnock, R.C.M. et al. Early cephalopod evolution clarified through Bayesian phylogenetic inference. BMC Biol 20, 88 (2022). https://doi.org/10.1186/s12915-022-01284-5

# Understanding Bayesian Inference in Phylogeny

1. Which of the following is a key reason to use Bayesian inference for phylogeny?
    a. The dataset is exceptionally large
    b. The dataset appears very ambiguous and uncertain
    c. Other models have already been used to predict trees over this data
    d. There is an unlimited amount of time to analyze possible trees

# Understanding Bayesian Inference in Phylogeny

1. Which of the following is a key reason to use Bayesian inference for phylogeny?
   a. The dataset is exceptionally large
   b. The dataset appears very ambiguous and uncertain
   c. Other models have already been used to predict trees over this data
   d. There is an unlimited amount of time to analyze possible trees

# Understanding Bayesian Inference in Phylogeny

2. What do the values on the branches of the trees represent in Bayesian trees?
   a. Bootstrap values
   b. Prior Probabilities
   c. Posterior Probabilities
   d. Estimation coefficients

# Understanding Bayesian Inference in Phylogeny

2. What do the values on the branches of the trees represent in Bayesian trees?

a. Bootstrap values
b. Prior Probabilities
c. Posterior Probabilities
d. Estimation coefficients

# Questions?