

Phylogenetics

Morphological Substitution models

Laura Mulvey

laura.mulvey@qmul.ac.uk

May 13 2025

Today's class

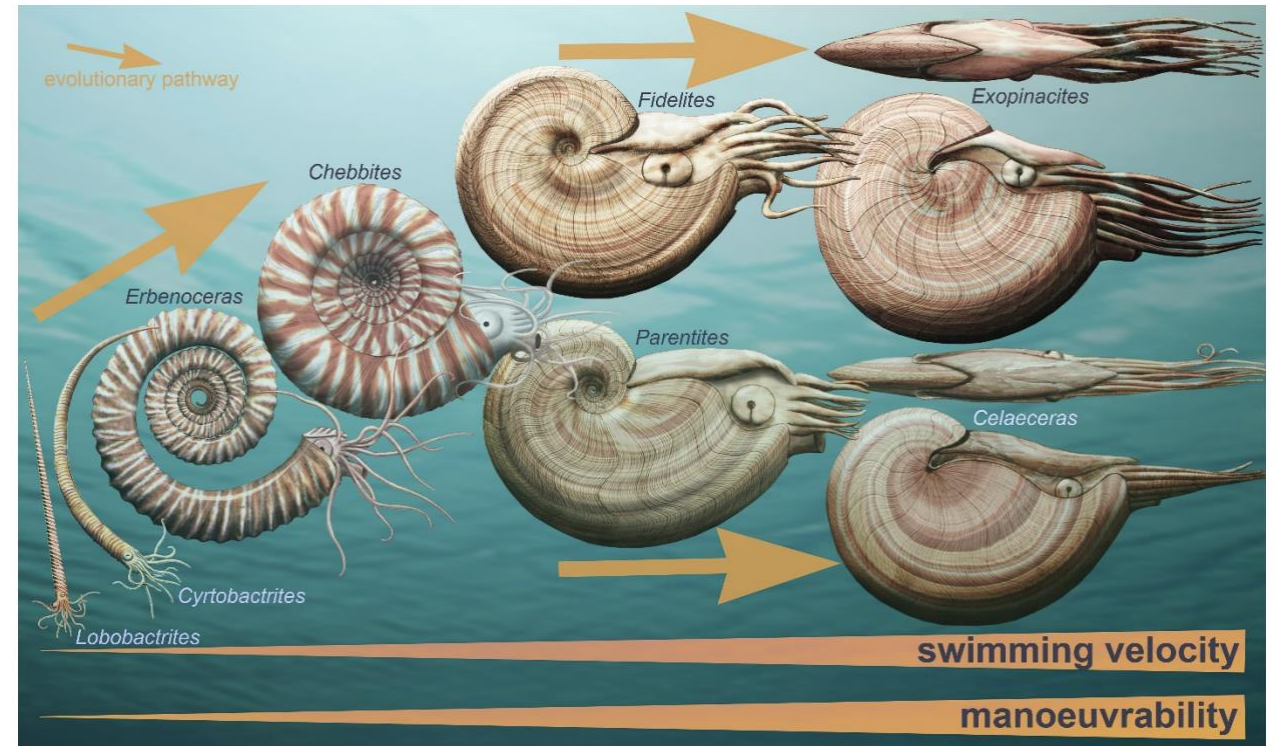


1. Morphological data & Substitution models
2. *Exercise 1: MorphoSim*
3. How do we choose a model to use?
4. *Exercise 2: PPS Tutorial*

Morphological data

Morphological data was the original type of information used in phylogenetic analysis

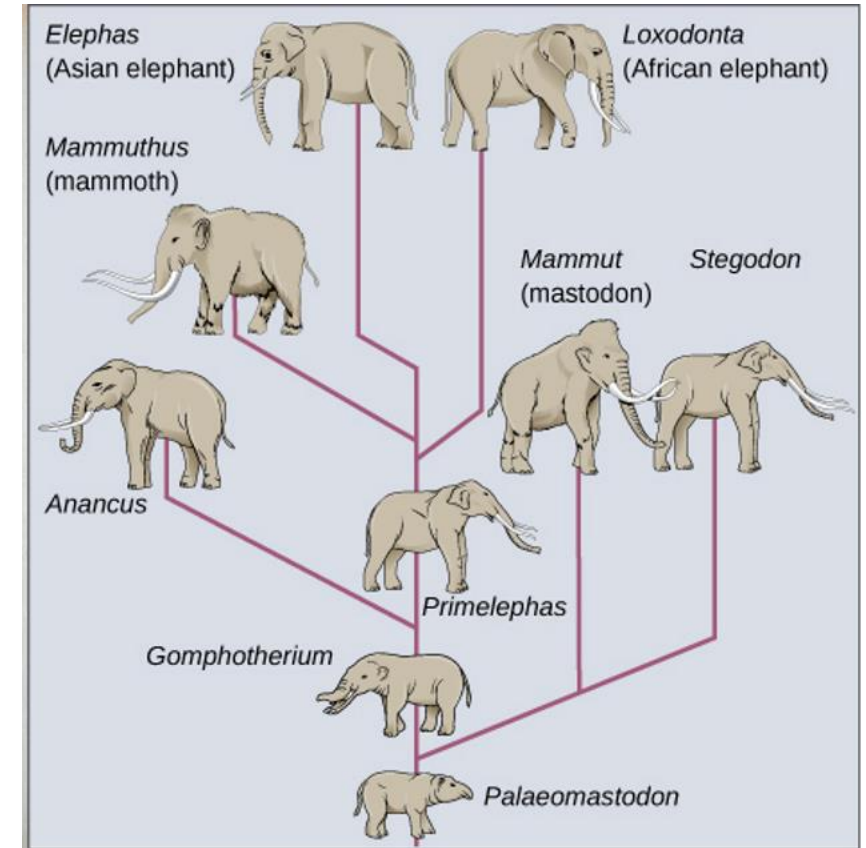
Fossils can be used to provide time calibrations, helps extant phylogeny, allows us to understand evolution through time



Morphological data

Discrete Characters: Morphological data often consist of discrete characters, such as the presence or absence of certain traits, or more complex multistate traits (e.g., number of limbs, type of leaf, presence of a particular bone structure)

Continuous Characters: Some morphological data can be continuous, such as measurements of body size, length of bones, or other quantitative traits



Morphological data

Discrete Characters: Morphological data often consist of discrete characters, such as the presence or absence of certain traits, or more complex multistate traits (e.g., number of limbs, type of leaf, presence of a particular bone structure)

Continuous Characters: Some morphological data can be continuous, such as measurements of body size, length of bones, or other quantitative traits

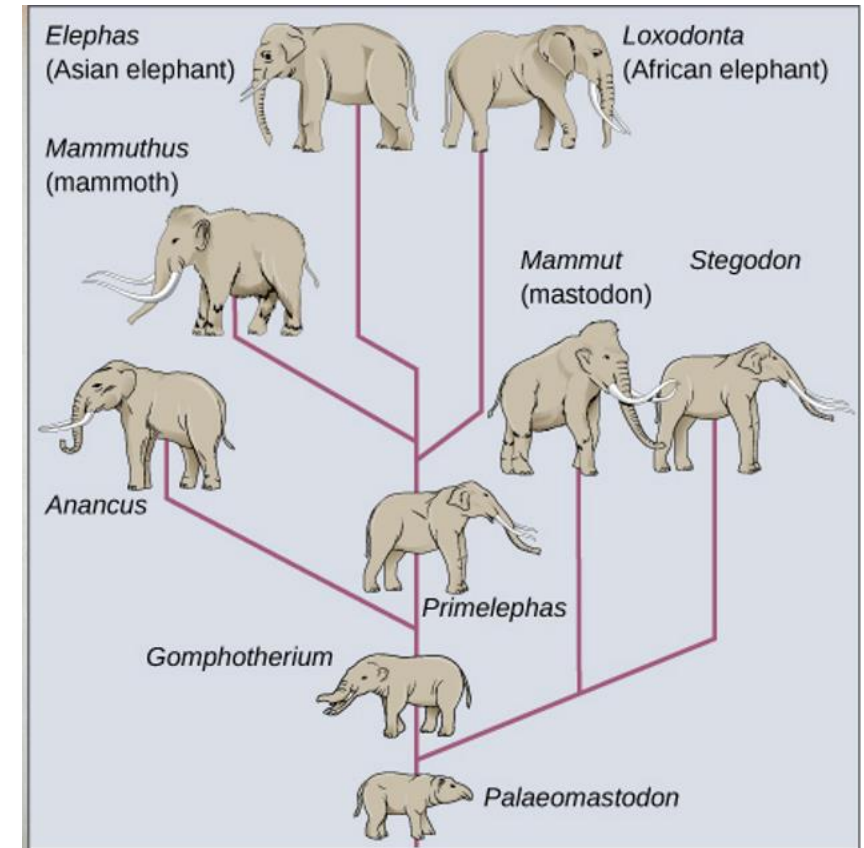
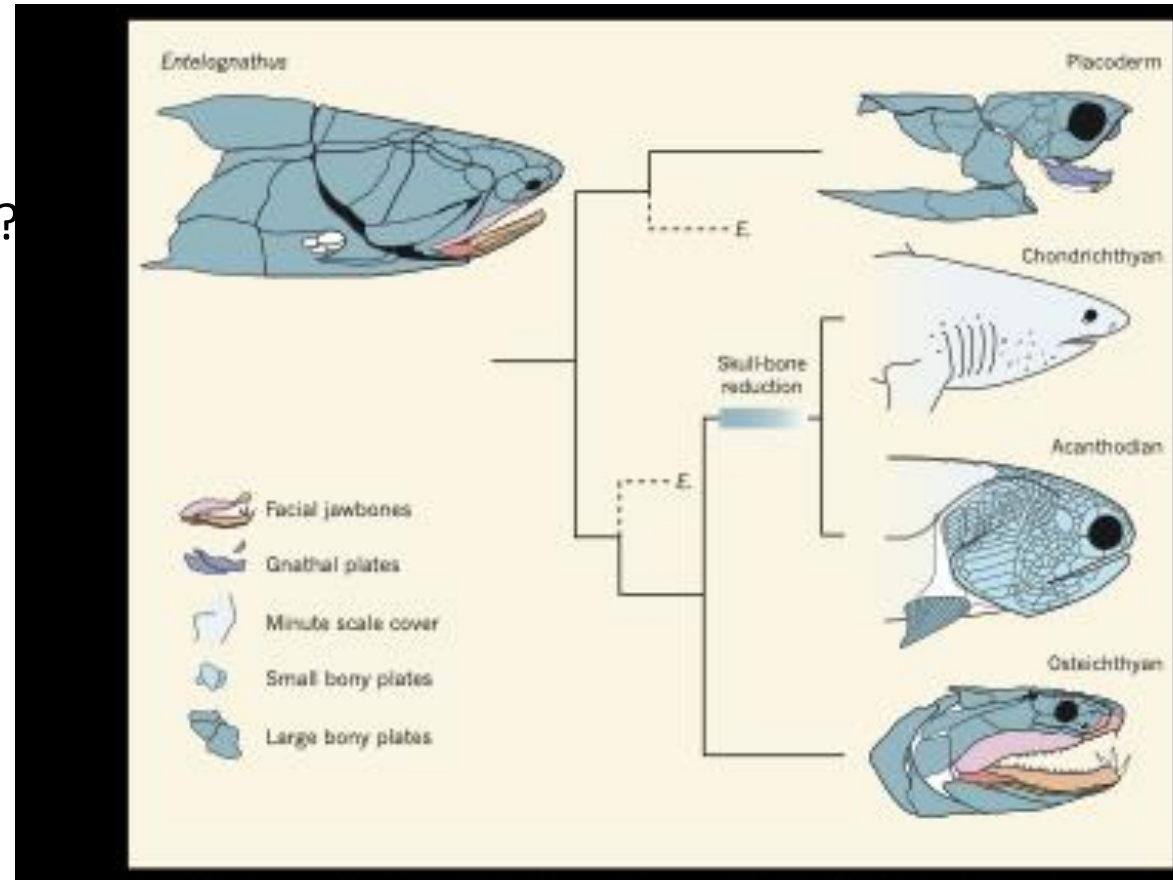


Image from
<https://www.zoologytalks.com/>

Morphological changes along a tree

We can observe changes in morphology between related taxa, how can we use this?

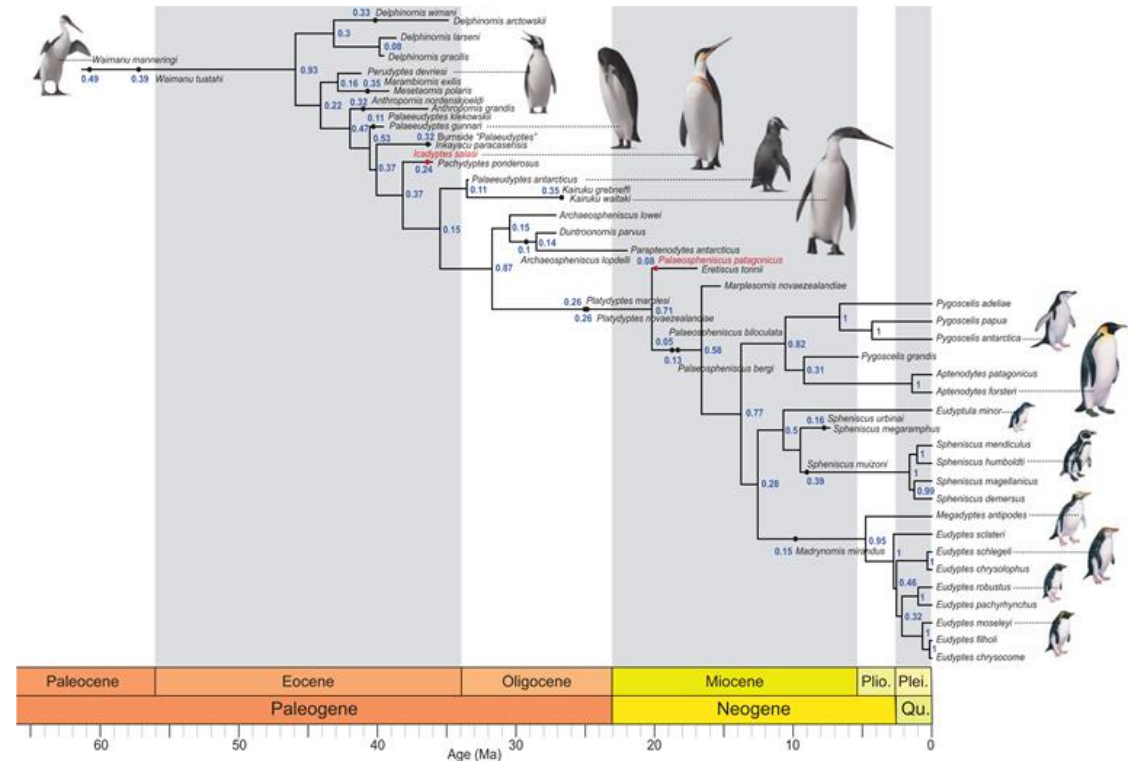


Combining morphological and molecular data?

Incorporating fossil information into an analysis allows us to use all the available information to understand their evolutionary history

Essential for dating a tree

It has been shown to improve our estimates even when we are mainly interested in the extant topology!



Using fossils in phylogenetic inference

How can we include information about fossils into an inference?

What is the character data the is available to us from the fossil record?



Using fossils in phylogenetic inference

We can use morphological characters that are manually coded to describe species traits

This is incredibly time consuming and meticulous type of data collection



100

[illegible]

Discrete Character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4	Used to describe more complex traits and can capture greater variation between taxa

Discrete Character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4	Used to describe more complex traits and can capture greater variation between taxa
Missing characters	?	Used when the specimen is either too decayed to determine whether it has a certain character trait or not, or we are missing the relevant part of the body

Discrete Character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4	Used to describe more complex traits and can capture greater variation between taxa
Missing characters	?	Used when the specimen is either too decayed to determine whether it has a certain character trait or not, or we are missing the relevant part of the body
Non-applicable	-	Used when the trait is not associated with a taxon. They represent a type of nested coding where the presence of the trait is defined in a different trait

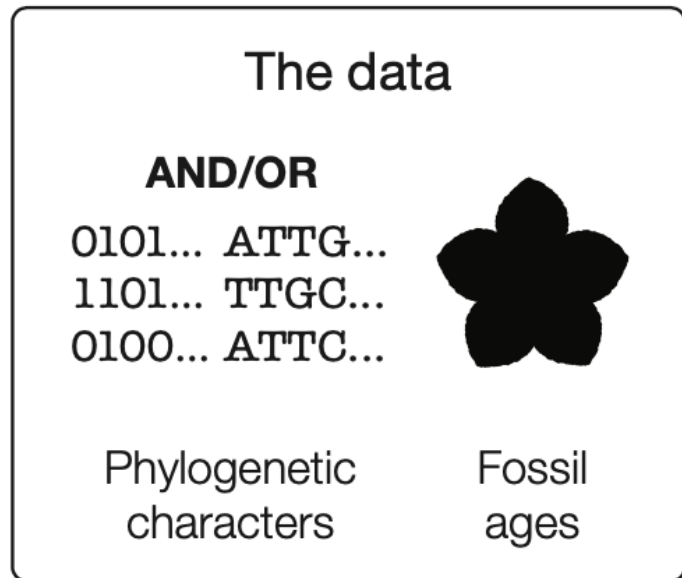
Discrete Character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4	Used to describe more complex traits and can capture greater variation between taxa
Missing characters	?	Used when the specimen is either too decayed to determine whether it has a certain character trait or not, or we are missing the relevant part of the body
Non-applicable	-	Used when the trait is not associated with a taxon. They represent a type of nested coding where the presence of the trait is defined in a different trait
Polymorphisms	0/1/2	Used when there are variations in a traits within species

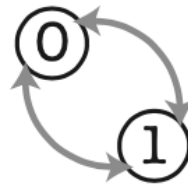
Discrete Character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4	Used to describe more complex traits and can capture greater variation between taxa
Missing characters	?	Used when the specimen is either too decayed to determine whether it has a certain character trait or not, or we are missing the relevant part of the body
Non-applicable	-	Used when the trait is not associated with a taxon. They represent a type of nested coding where the presence of the trait is defined in a different trait
Polymorphisms	0/1/2	Used when there are variations in a traits within species
Uncertain	0/1/2	Used when it is not clear which character trait is present in the taxon

Bayesian Phylogenetic Analysis Components



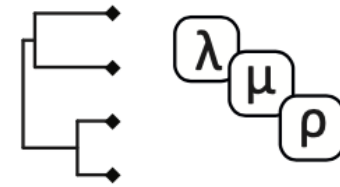
Tripartite model components



Substitution
model



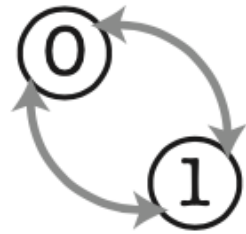
Clock
model



Tree and tree
model

Bayesian Phylogenetic Analysis Components

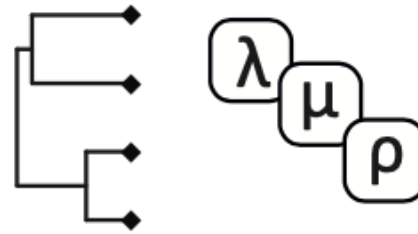
Tripartite model components



Substitution
model



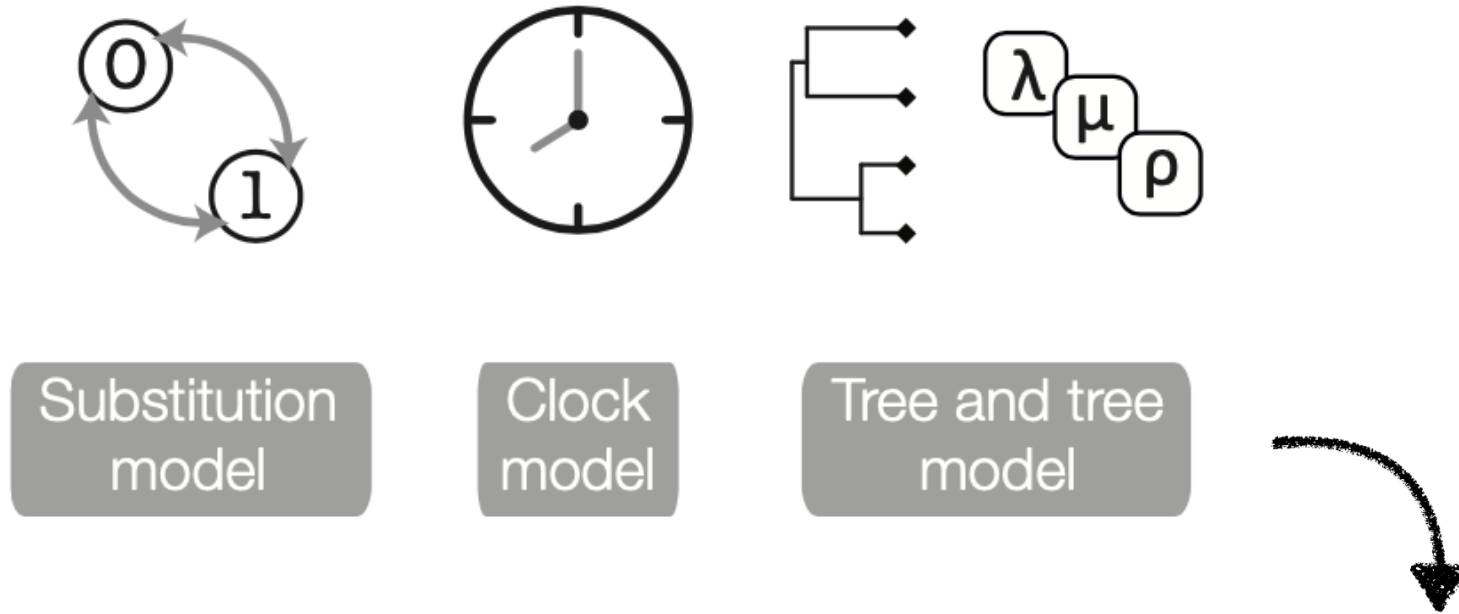
Clock
model



Tree and tree
model

Bayesian Phylogenetic Analysis Components

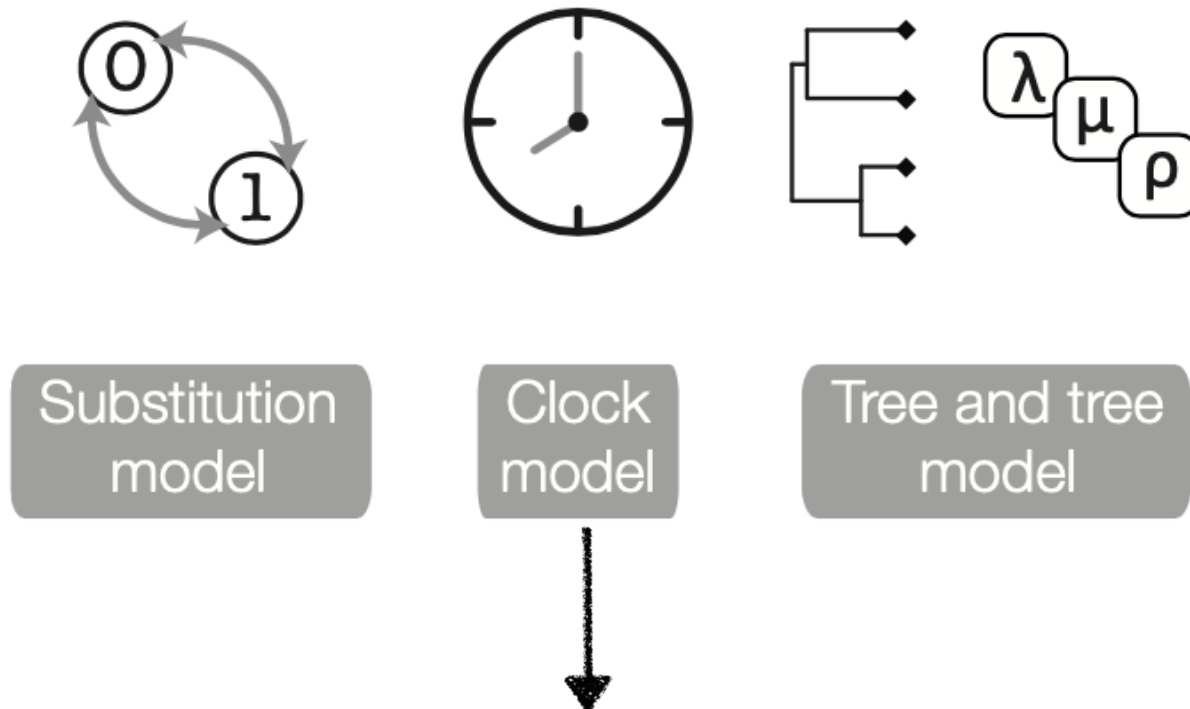
Tripartite model components



How have species originated, gone extinct and been sampled through

Bayesian Phylogenetic Analysis Components

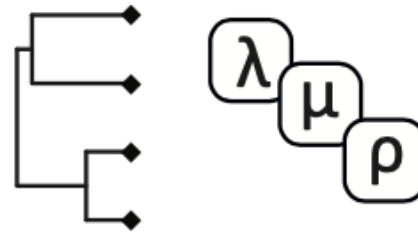
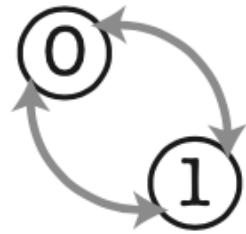
Tripartite model components



How have rates of evolution varied (or not) across the tree?

Bayesian Phylogenetic Analysis Components

Tripartite model components



Substitution
model

Clock
model

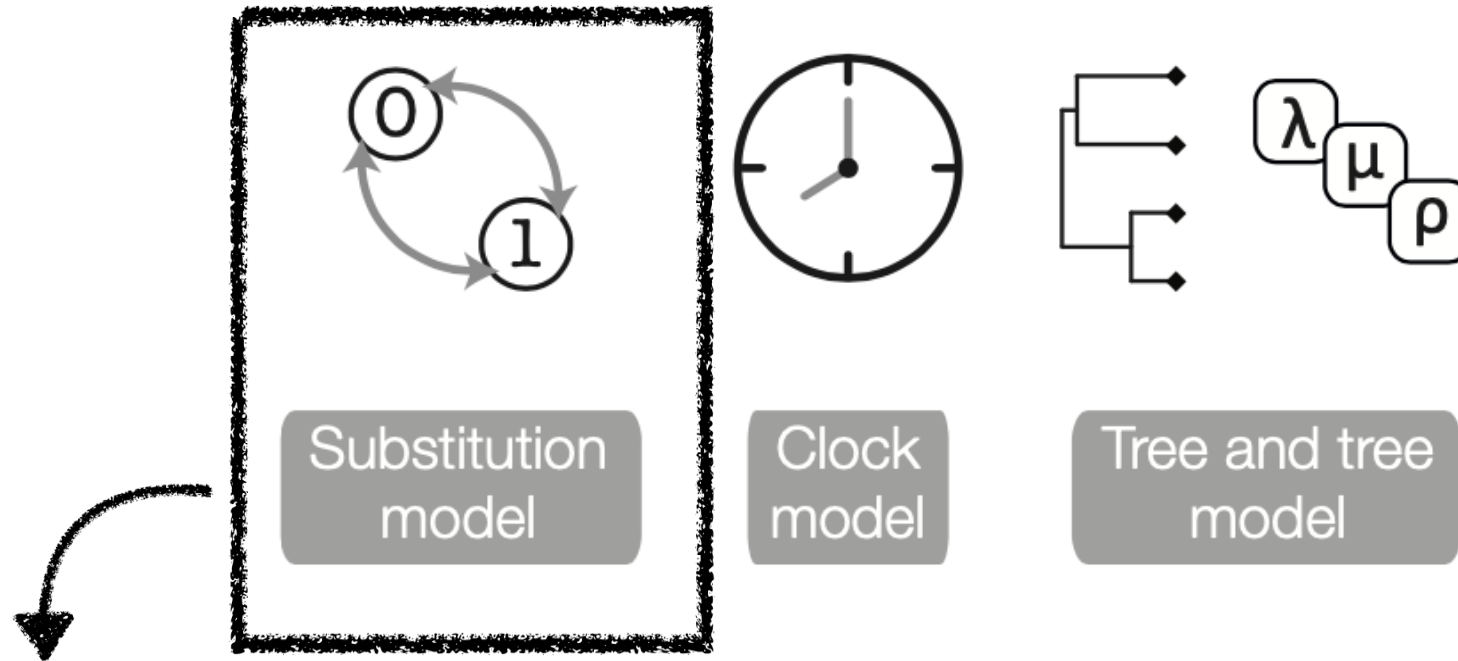
Tree and tree
model



How likely are we to observe a change between character states? e.g., $0 \rightarrow 1$

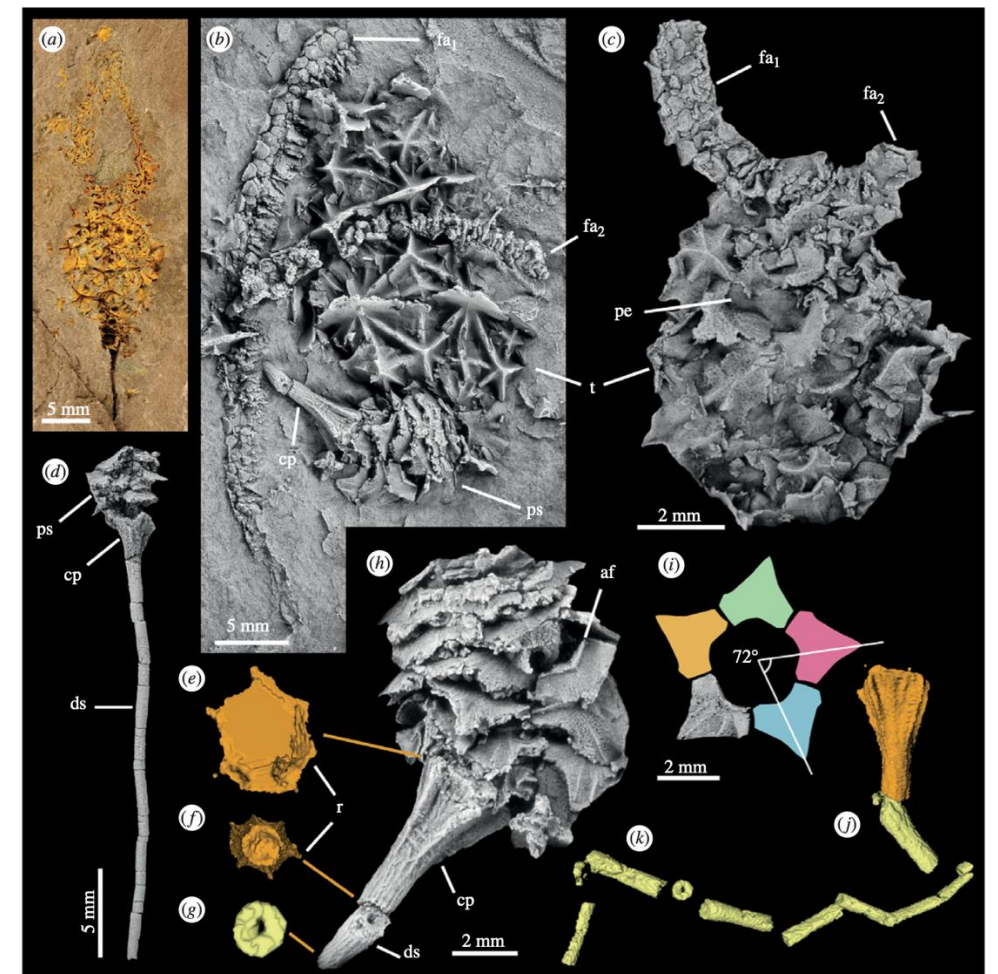
Bayesian Phylogenetic Analysis Components

Tripartite model components



How likely are we to observe a change between character states? e.g., $0 \rightarrow 1$

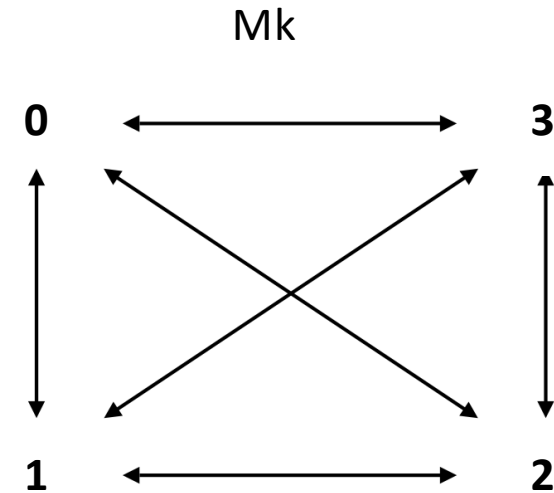
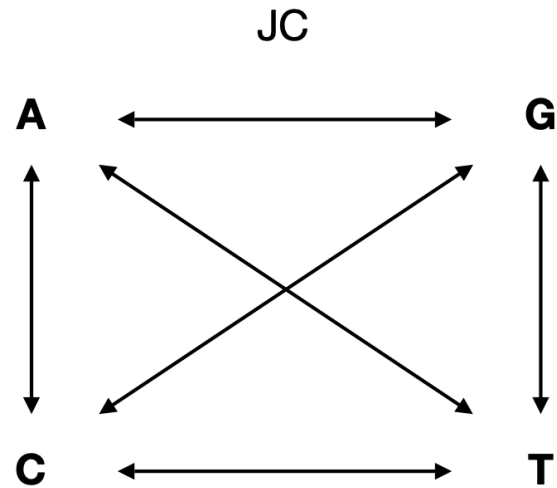
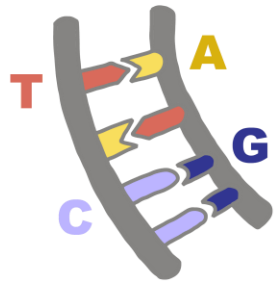
	Appendage branching pattern										Coring plate arrangement										Presence Absence									
001510010?00-100--00000000000																														
000500010?200100--00100100000																														
002500010?200100--0?100100000																														
00?5?0010?200100?-0??010110																														
0015000101201000430100011111																														
0015000101201010440111011111																														
??050? ??? ?201000440?11011111																														
01050?010-210000?501??010110																														
00020001002101003-1110010110																														
0002000100211001441121011111																														
000201111-210010?-??11011121																														
?103?0?11?1001104-0000010000																														
1005002110100010--0?00110?20																														
1005002000101010540?00110020																														



Dibrachicystis purujoensis

Cambrian stalked echinoderms show unexpected plasticity of arm construction
Zamora & Smith. 2012. Proc B

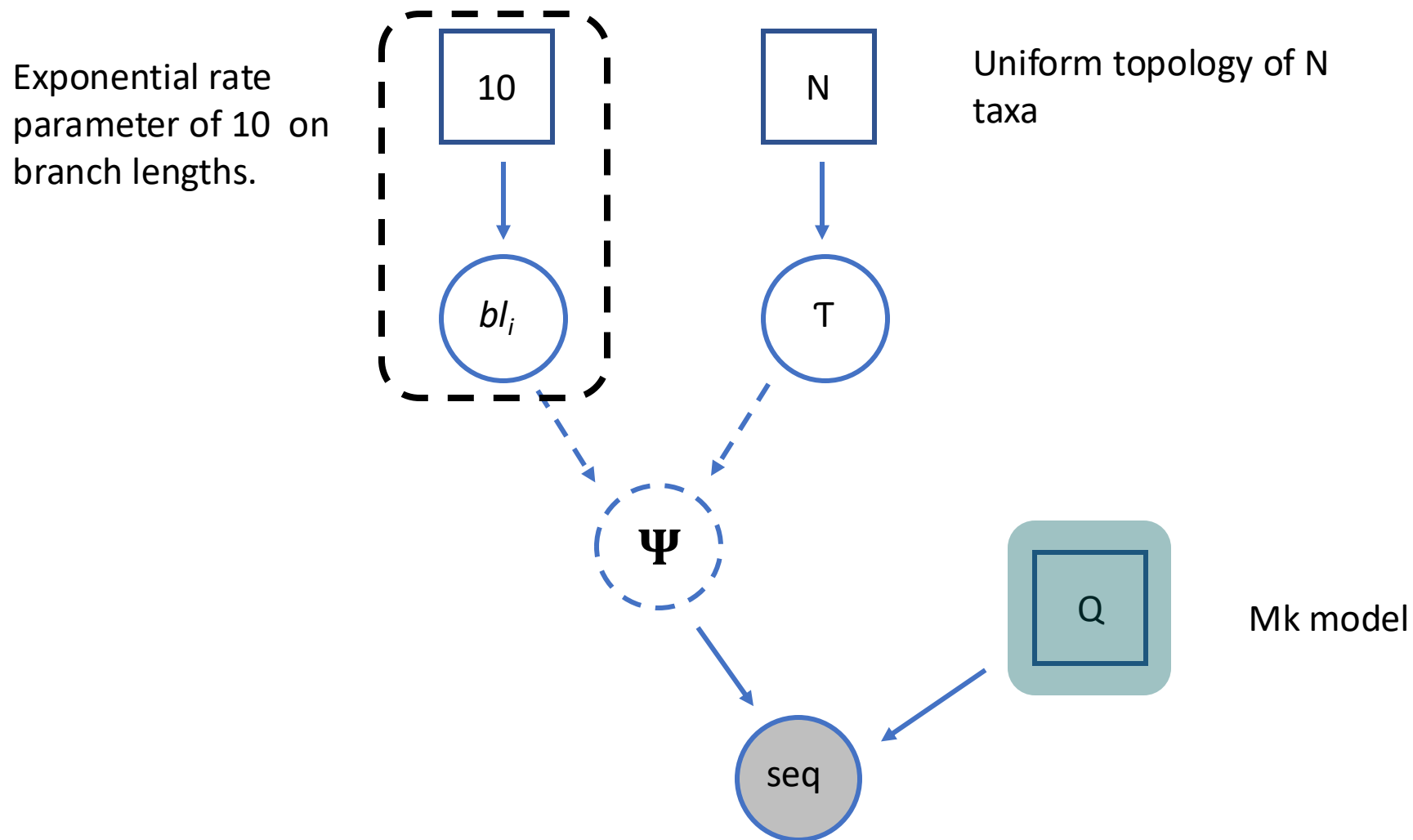
Substitution models for morphological data



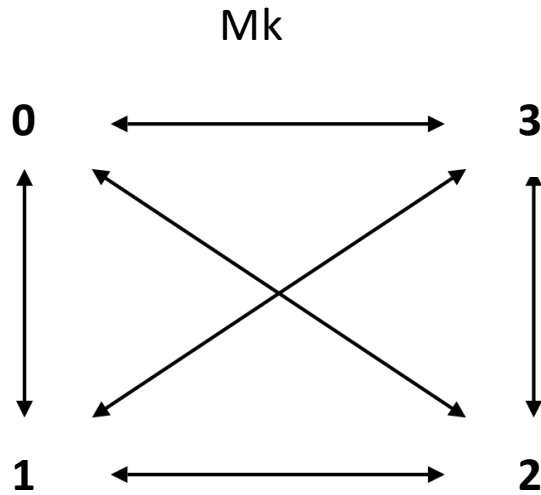
*4 state here as an example, can be any number from 2!

Line width represents the relative rate of change between different steps.

Mk Model



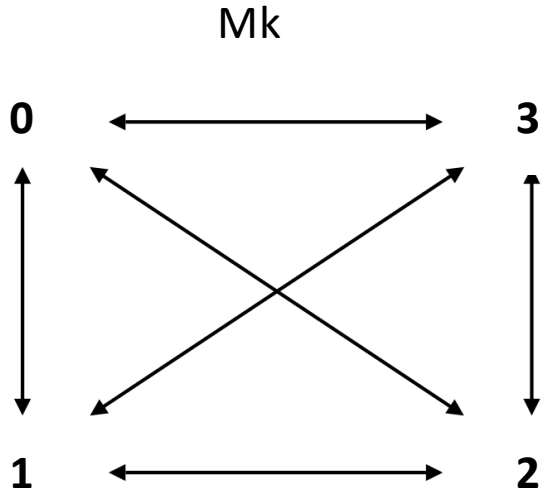
Substitution models for morphological data



$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} & \mu_{02} & \mu_{03} \\ \mu_{10} & -\mu_1 & \mu_{12} & \mu_{13} \\ \mu_{20} & \mu_{21} & -\mu_2 & \mu_{23} \\ \mu_{30} & \mu_{31} & \mu_{32} & -\mu_3 \end{pmatrix},$$

*4 state here as an example, can be any number from 2!

Substitution models for morphological data

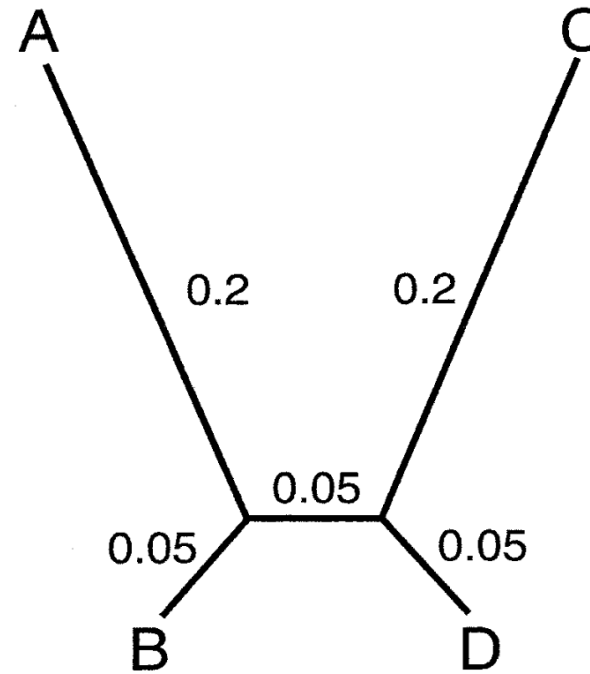


We can **add extensions** to the standard Mk model in a number of ways

*4 state here as an example, can be any number from 2!

Ascertainment Bias (V)

Conditions on the fact that all sites are variable



	True branch length	Mk (uncorrected)	Mkv (corrected)
Percent correct	—	74.0	99.8
Branch A	0.2	241,750 ($\pm 349,100$)	0.206 (± 0.060)
Branch B	0.05	0.43210 (± 0.13756)	0.050 (± 0.018)
Branch X	0.05	54.646 ($\pm 1,725.3$)	0.052 (± 0.023)
Branch C	0.2	143,950 ($\pm 228,910$)	0.206 (± 0.059)
Branch D	0.05	0.022 (± 0.054)	0.051 (± 0.019)

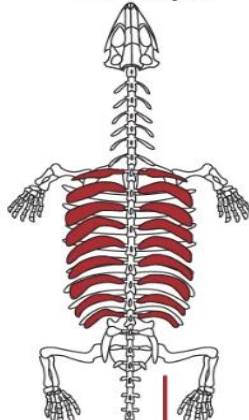
Across Site Rate Variation (+G)

Turtle shell evolution

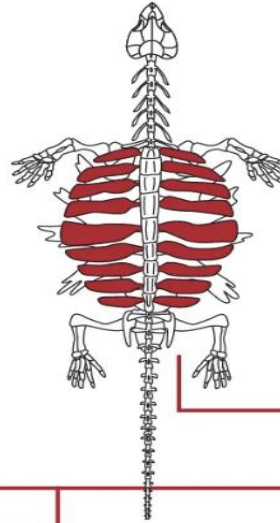
Eunotosaurus
~260 mya



Pappochelys
~240 mya



Odontochelys
~220 mya



Proganochelys
~210 mya



Across Site Rate Variation (+G)

	T1	T2
Taxa A	0	0
Taxa B	0	1
Taxa C	1	2

The transition rate
will impact branch
lengths

Slow rate of evolution



Fast rate of evolution

Relative to each other!

Across Site Rate Variation (+G)

What do we do?

	T1	T2
Taxa A	0	0
Taxa B	0	1
Taxa C	1	2

Allow these traits to evolve at different rates:

- Specify which traits evolve fast
- Use a gamma model to account for rate heterogeneity

Across Site Rate Variation (+G)

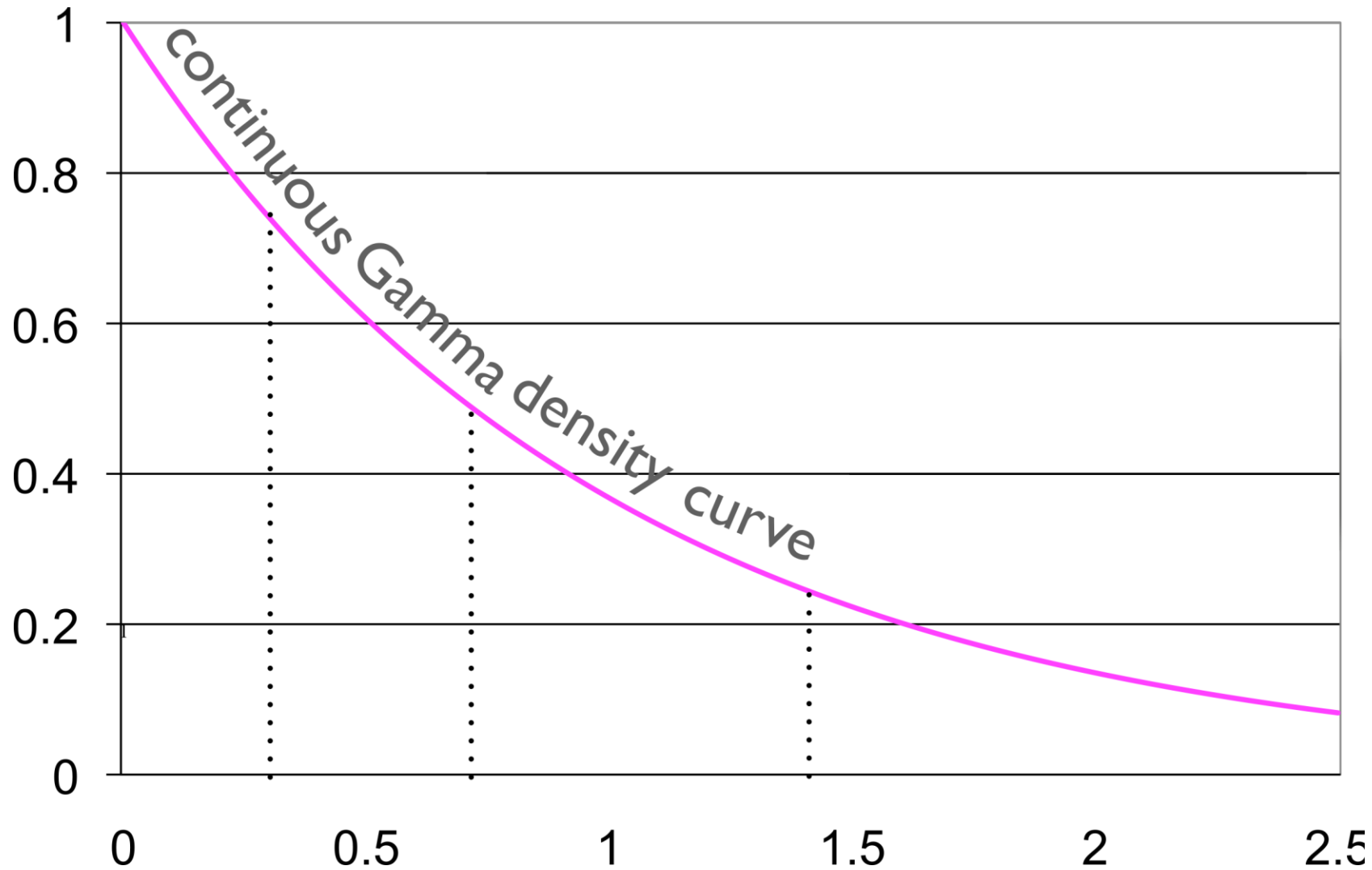
What do we do?

	T1	T2
Taxa A	0	0
Taxa B	0	1
Taxa C	1	2

Allow these traits to evolve at different rates:

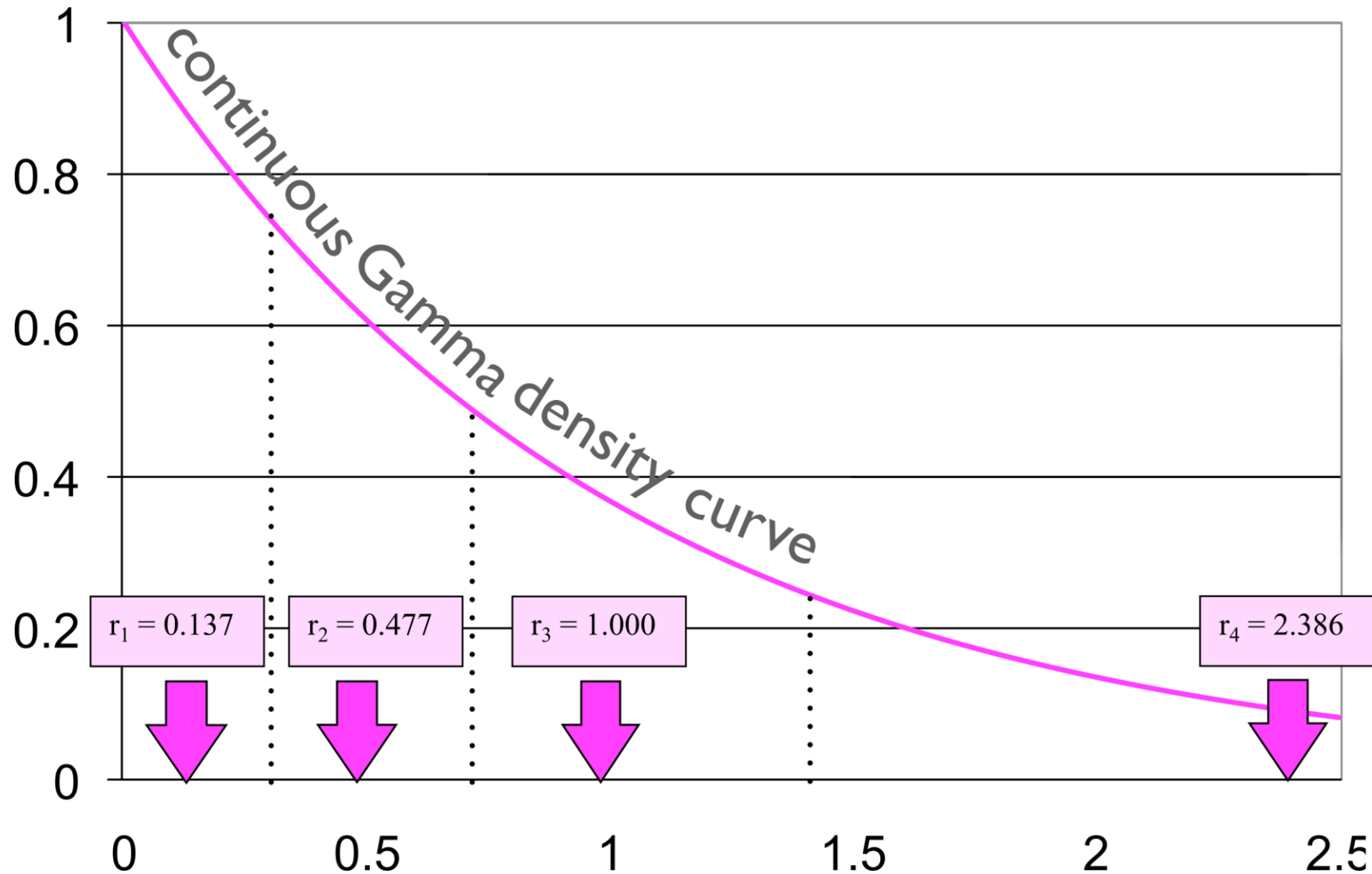
- Specify which traits evolve fast
- **Use a gamma model to account for rate heterogeneity**

Across Site Rate Variation (+G)



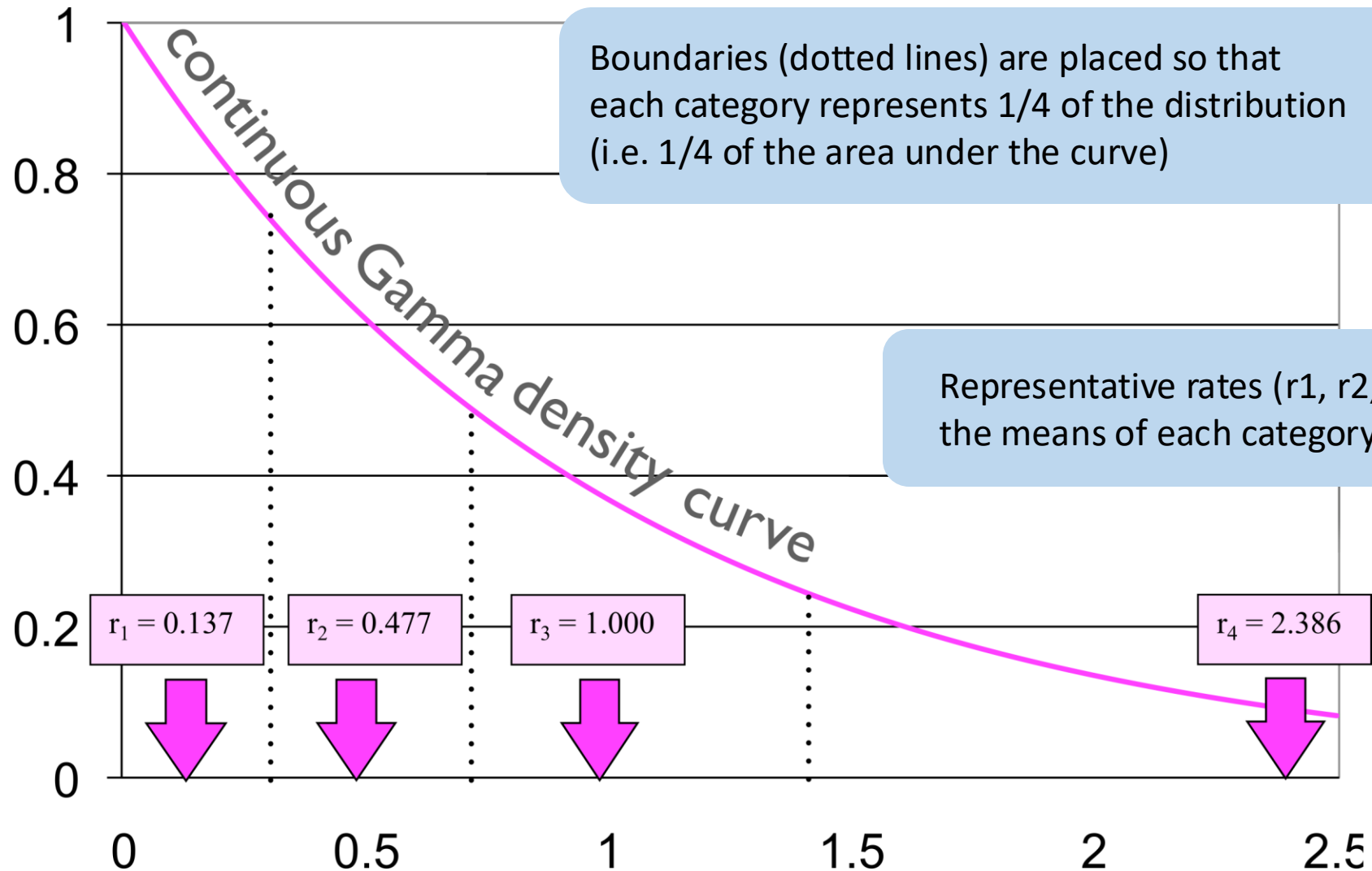
*Adapted from Paul
Lewis PhyloSeminar*

Across Site Rate Variation (+G)



Adapted from Paul
Lewis PhyloSeminar

Across Site Rate Variation (+G)



Across Site Rate Variation (+G)

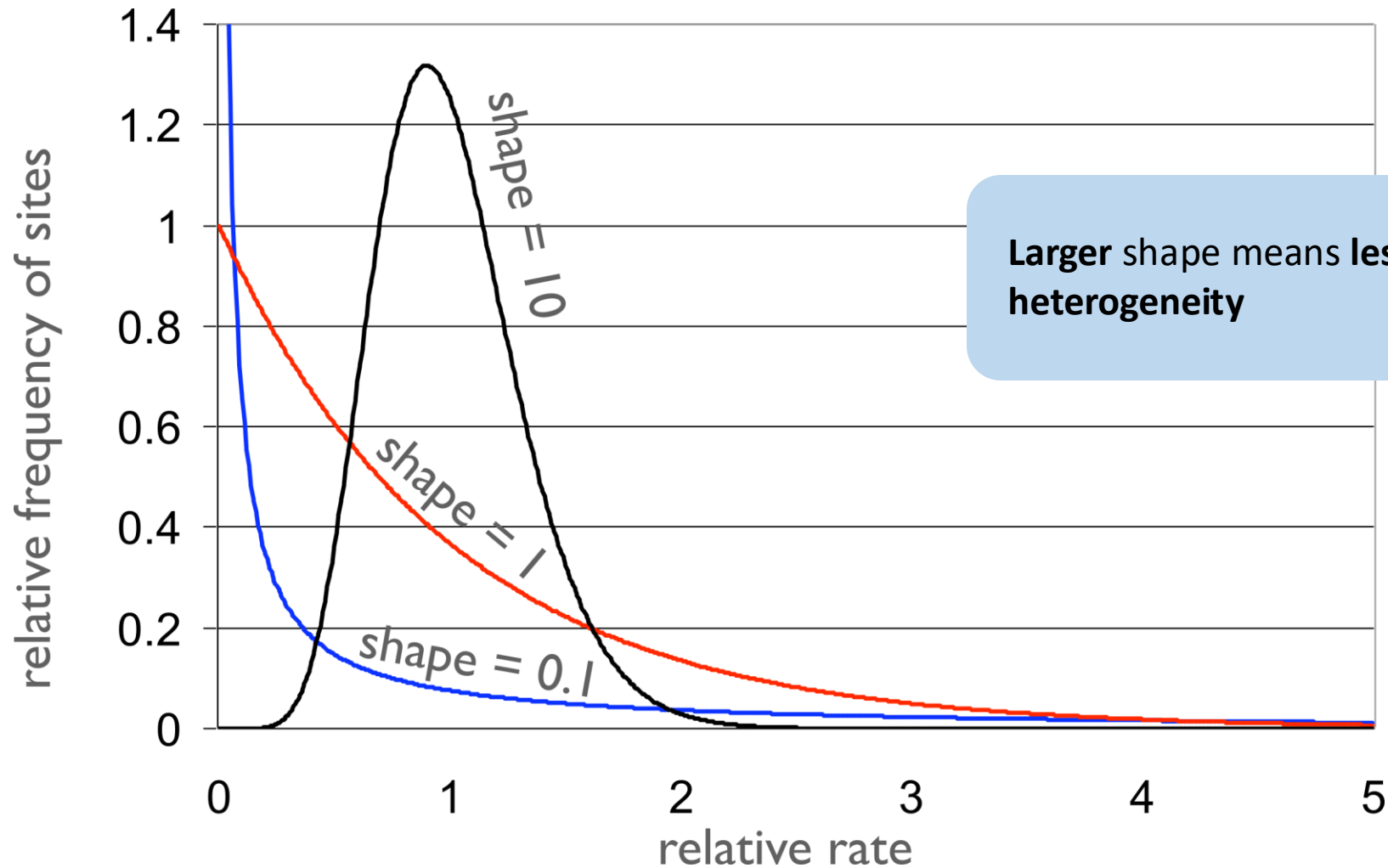
	T1	T2	
Taxa A	0	0	
Taxa B	0	1	
Taxa C	1	2	Faster R (R4)

Slower R (R1,2)

Allow each trait to evolve according to the rates drawn from the gamma distribution

One rate will fit the best and be the most influential for the likelihood calculation

Across Site Rate Variation (+G)



Partitioning the data

Grouping together parts of the alignment that have similar characteristics and or may have **evolved together** due to evolutionary pressures

The **defaults** in many phylogenetic software is to group by maximum observed state size

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} & \mu_{02} & \mu_{03} \\ \mu_{10} & -\mu_1 & \mu_{12} & \mu_{13} \\ \mu_{20} & \mu_{21} & -\mu_2 & \mu_{23} \\ \mu_{30} & \mu_{31} & \mu_{32} & -\mu_3 \end{pmatrix},$$

$$\begin{bmatrix} -\mu_0 & \mu_{01} \\ \mu_{10} & -\mu_1 \end{bmatrix} \quad \begin{bmatrix} -\mu_0 & \mu_{01} & \mu_{02} \\ \mu_{10} & -\mu_1 & \mu_{12} \\ \mu_{20} & \mu_{21} & -\mu_2 \end{bmatrix}$$



Partitioning the data

When should we partition our data?



Partitioning the data

When should we partition our data?

If we have presence (1) absence (0) traits partitioning will always be a logical approach: what would transitioning to state 2 in this scenario even mean?



Partitioning the data

When should we partition our data?

If we have presence (1) absence (0) traits partitioning will always be a logical approach: what would transitioning to state 2 in this scenario even mean?

We should be cautious for traits describing a trait – just because we do not observe a state 2 can we be absolutely certain there never was one?

Justifying partitioning schemes is very important as they have a major impact on inference results

Challenges with morphological data

Generalising assumptions across different traits is often not possible

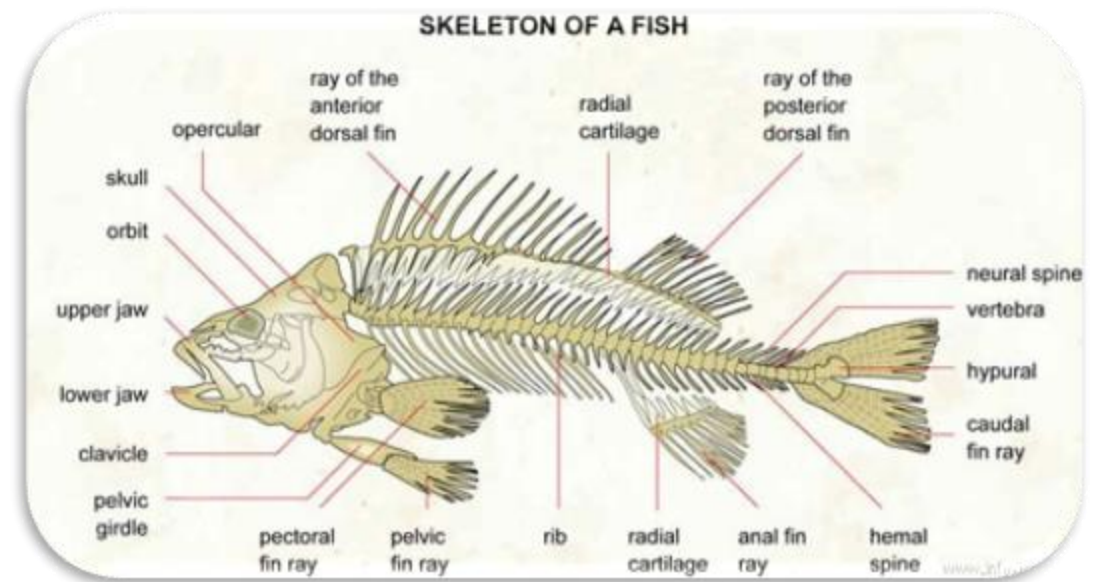
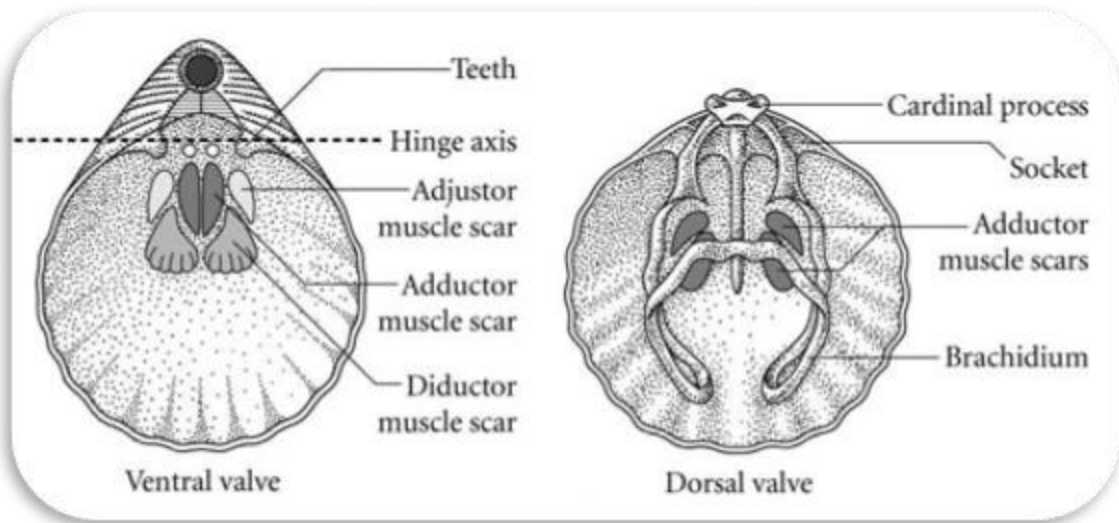
Modelling special characters in matrices

```
001510010?00-100--0000000000
000500010?200100--0010010000
002500010?200100--0?10010000
00?5?0010?200100?-0???010110
0015000101201000430100011111
0015000101201010440111011111
??050?????201000440?11011111
```

Challenges with morphological data

Morphological matrices are often quite small:

- Collection is very time consuming
- Number of characters available can be very small depending on the group





Exercise 1:

Simulate morphological data using the Shiny app for MorphoSim

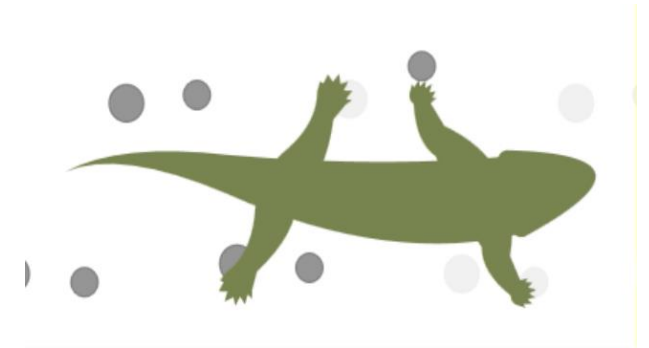
<https://github.com/fossilsim/MorphoSimShiny>

How do we choose a model and does it matter?

Used 114 empirical **tetrapod** matrices and compared 7 different models

Looked at the impact on:

- branch lengths (**evolutionary distances**)
- Tree topology (**species relationships**)





Impact of model choice

Used 114 empirical **tetrapod** matrices and compared 7 different models

Looked at the impact on:

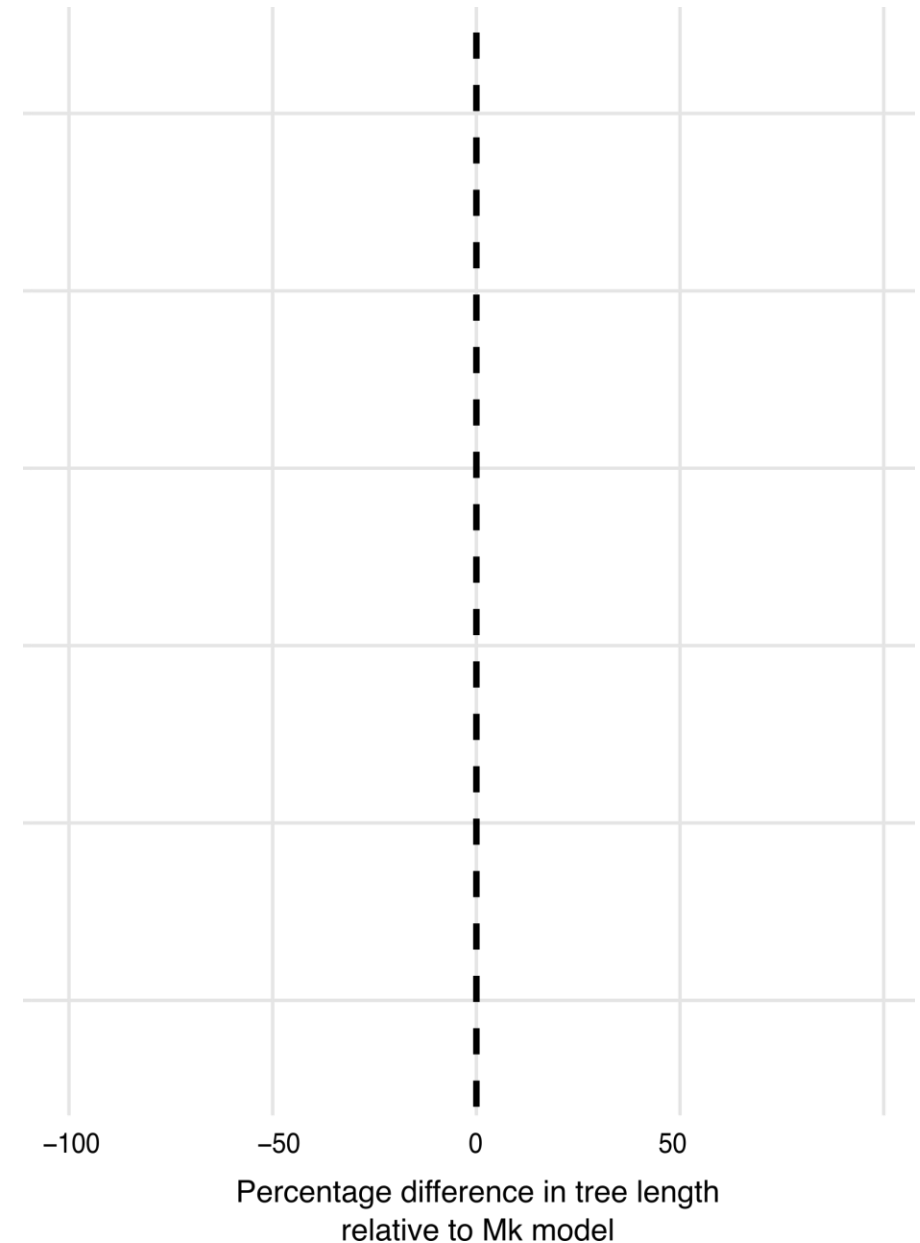
- branch lengths (**evolutionary distances**)
- Tree topology (**species relationships**)

Impact of model choice

Used 114 empirical **tetrapod** matrices and compared 7 different models

Looked at the impact on:

- branch lengths (**evolutionary distances**)
- Tree topology (**species relationships**)

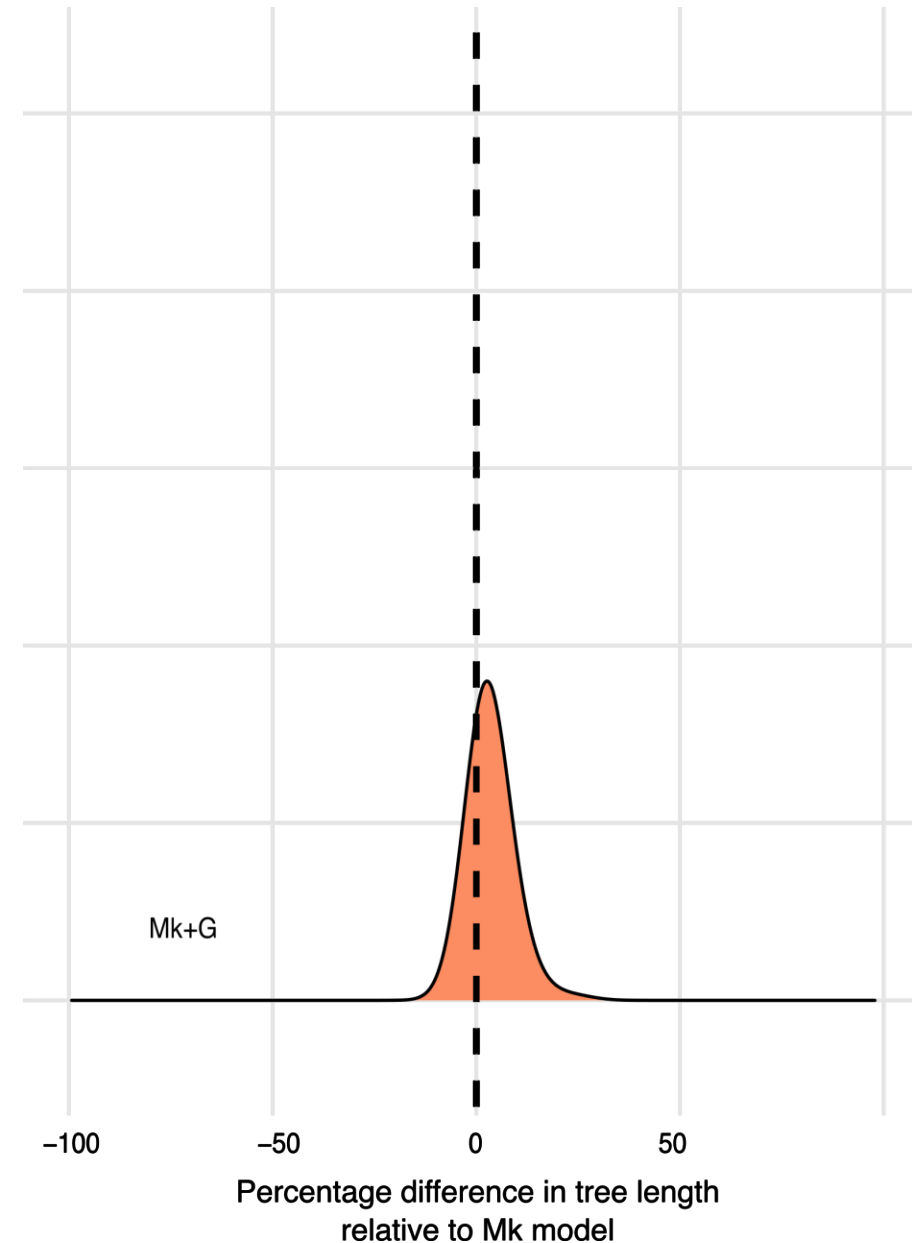


Impact of model choice

Used 114 empirical **tetrapod** matrices and compared 7 different models

Looked at the impact on:

- branch lengths (**evolutionary distances**)
- Tree topology (**species relationships**)



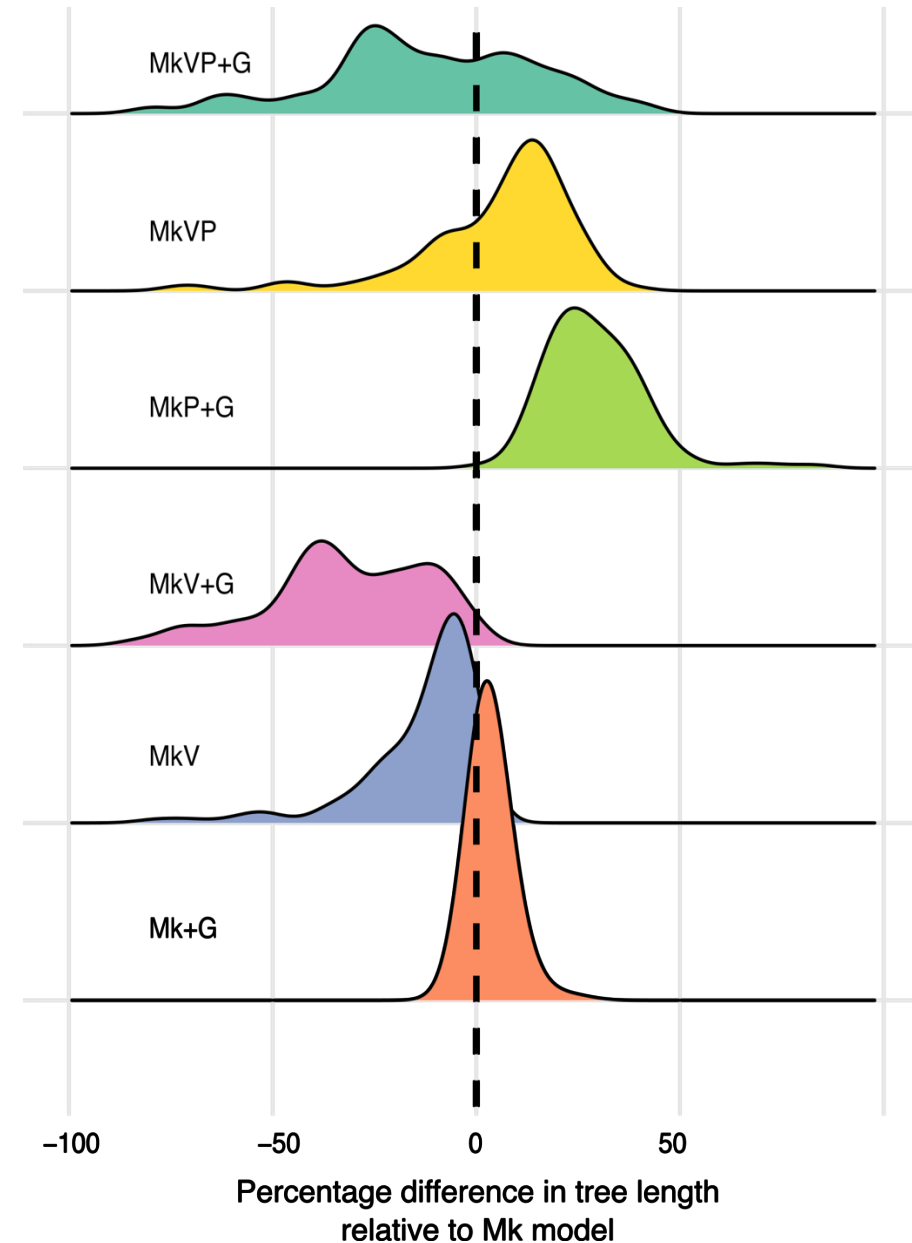
Impact of model choice

Used 114 empirical **tetrapod** matrices and compared 7 different models

Looked at the impact on:

- branch lengths (**evolutionary distances**)
- Tree topology (**species relationships**)

So how do we choose a model?





How to choose a model

Model selection using **bayes factors** is a common approach found in the literature

It relies on comparing the marginal likelihoods approximated from different models

How to choose a model

Model selection using **bayes factors** is a common approach found in the literature

It relies on comparing the marginal likelihoods approximated from different models

$$\begin{array}{c} \text{posterior} \\ \boxed{} \end{array} \quad P(\text{tree} \text{ with } 0,1 \text{ nodes} \mid \begin{array}{l} 0101\dots \\ 1101\dots \\ 0100\dots \end{array}) = \frac{\begin{array}{c} \text{likelihood} \quad \text{priors} \\ \boxed{} \quad \boxed{} \end{array} \quad P(\begin{array}{l} 0101\dots \\ 1101\dots \\ 0100\dots \end{array} \mid \text{tree} \text{ with } 0,1 \text{ nodes}) \quad P(\text{tree} \text{ with } 0,1 \text{ nodes})}{\begin{array}{c} \text{marginal probability} \\ \boxed{} \end{array} \quad P(\begin{array}{l} 0101\dots \\ 1101\dots \\ 0100\dots \end{array})}$$

How to choose a model

Model selection using **bayes factors** is a common approach found in the literature

It relies on comparing the marginal likelihoods approximated from different models

Table 6.16.1 The Scale for Interpreting Bayes Factors by Harold Jeffreys (1961)

Strength of evidence	$BF(M0, M1)$	$\log(BF(M0, M1))$	$\log_{10}(BF(M0, M1))$
Negative (supports M_1)	<1	<0	<0
Barely worth mentioning	1 to 3.2	0 to 1.16	0 to 0.5
Substantial	3.2 to 10	1.16 to 2.3	0.5 to 1
Strong	10 to 100	2.3 to 4.6	1 to 2
Decisive	>100	>4.6	>2

For a detailed description of Bayes factors see Kass and Raftery (1995)



Issues with model selection

It provides the **relative fit** of models

It makes a strong assumption a priori that one of the models fits your data

It cannot be used to determine between the fit of models that change the Q-matrix size, i.e., different partitioning schemes



Model adequacy

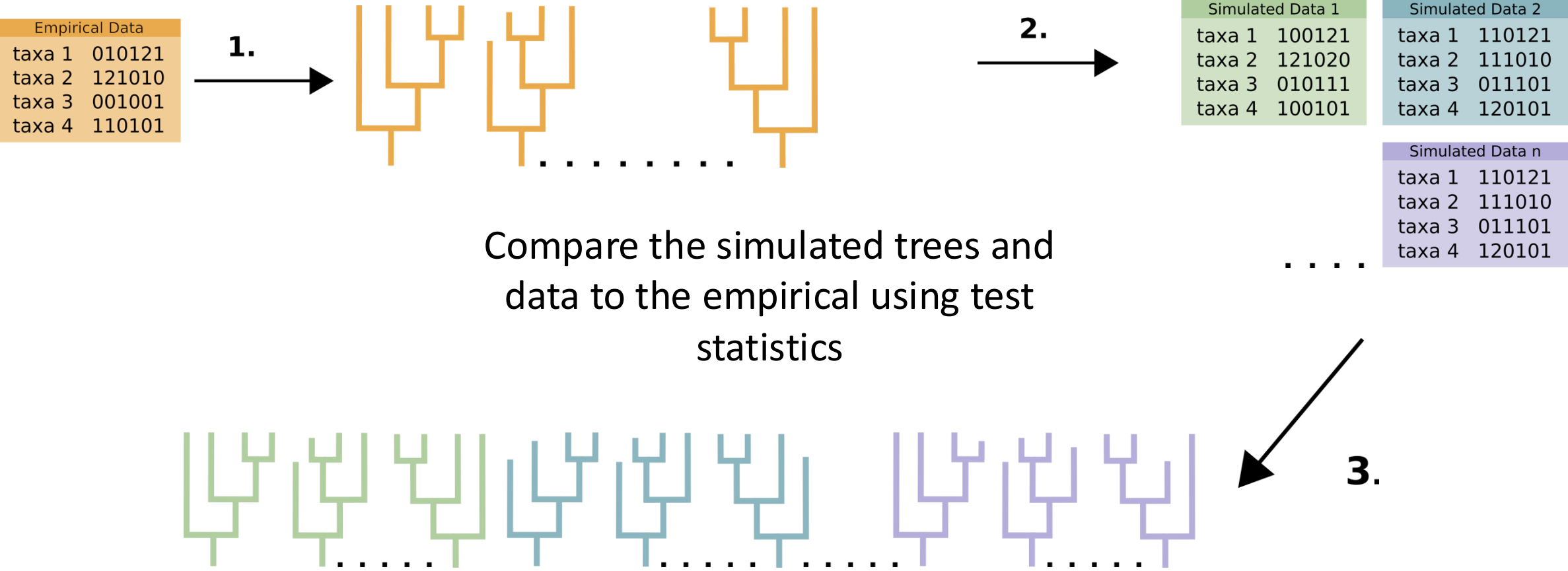
We know that none of our models are really true. Can we be sure that the chosen model captures the salient features of the evolutionary process and provides reliable inferences

Could the model and priors plausibly have given rise to the data

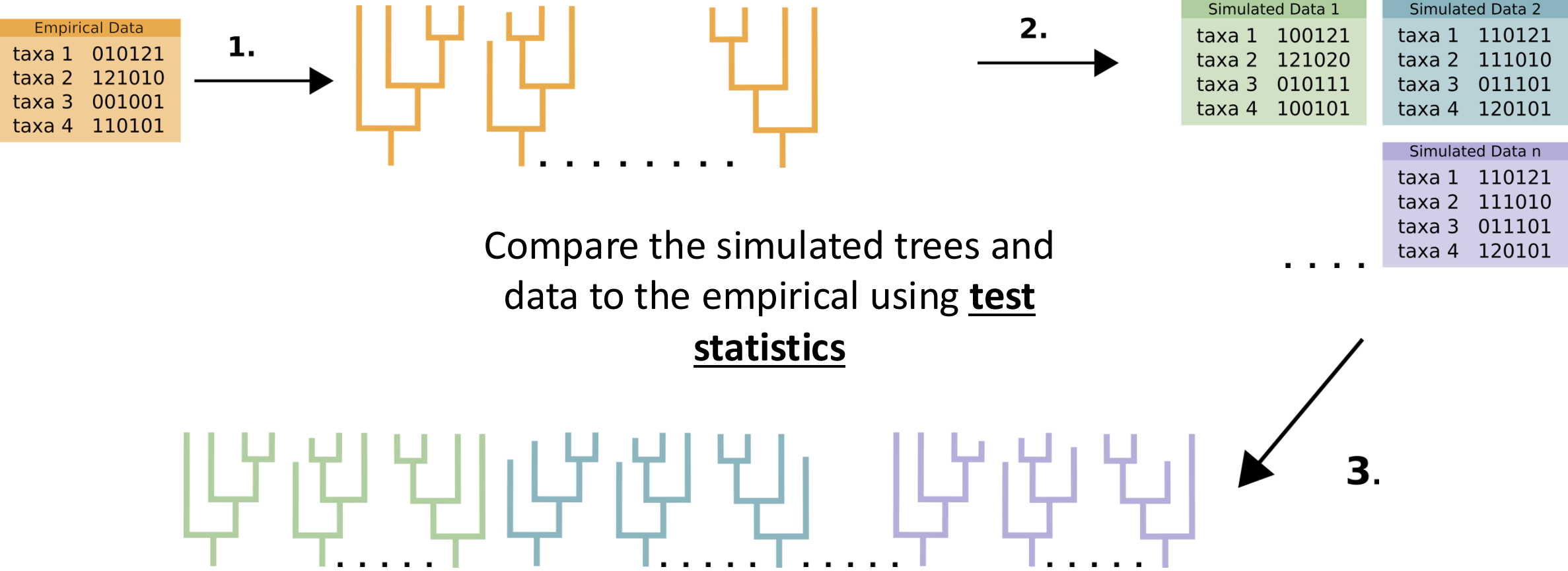
Allows us to ask whether **any** of our models are doing a good job describing the evolutionary processes that produced our data

Provides the **absolute fit** of our model to a data set

Posterior Predictive Simulations



Posterior Predictive Simulations





Test Statistics

A test statistic is a **numerical summary** of data.

A value that captures the characteristic of your data.

For PPS we use two test statistics: **Consistency Index** and **Retention Index**

These test statistics use both the data and the trees

Note: there may be more test statistics worth investigating but as of now these are the only two that are validated for use with morphological data, - see [Mulvey et al 2024](#) for more info



Exercise 2:

Carry out [posterior predictive simulations](#) for the data set in exercise one. Do either of the models fit our data?