Phylogenetics

Morphological Substitution models

Laura Mulvey

laura.l.mulvey@fau.de

June 11 2024

Todays class

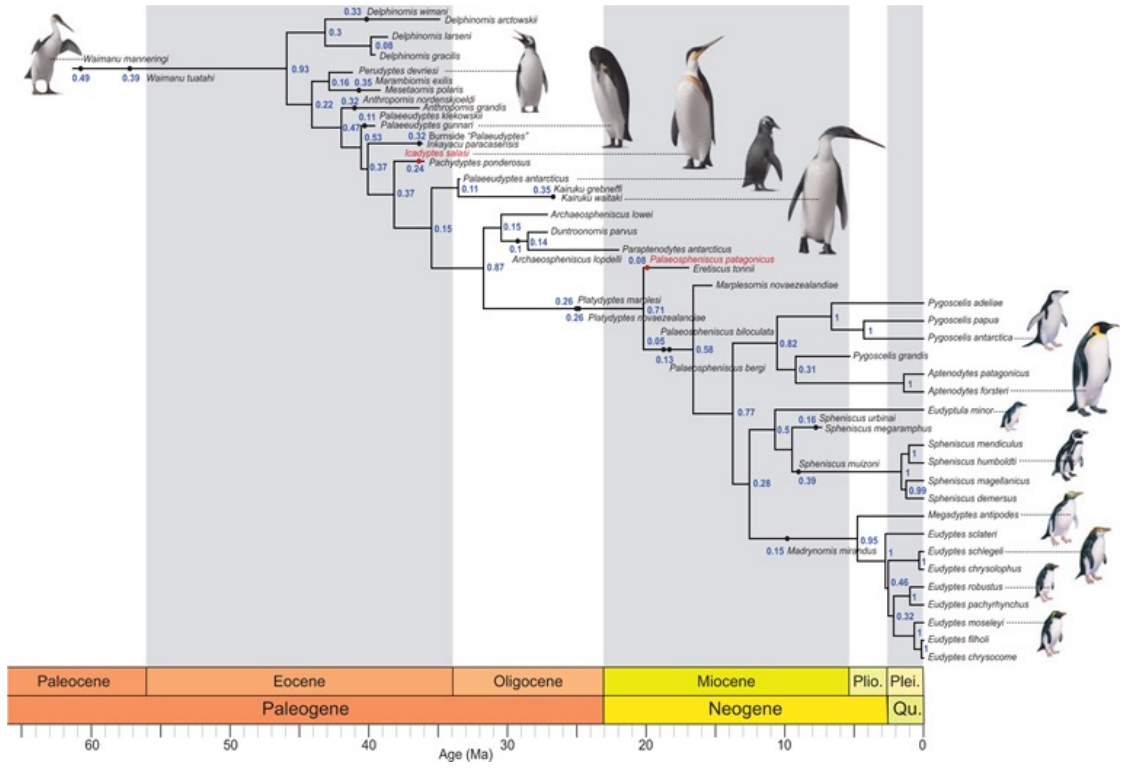
1. How can we use fossils in phylogenetic analysis?
2. Exercise 1
3. How do we choose a model to use?
4. Exercise 2

What is the benefit of incorporating fossil 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!



Gavryushkina et al 2016

Types of data

In a phylogenetic analysis you can use a combination of data types

- Molecular (mainly just for extant taxa)
- Morphological (for both extinct and extant taxa)
- A combination of molecular and morphological taxa (often referred to as total evidence)

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

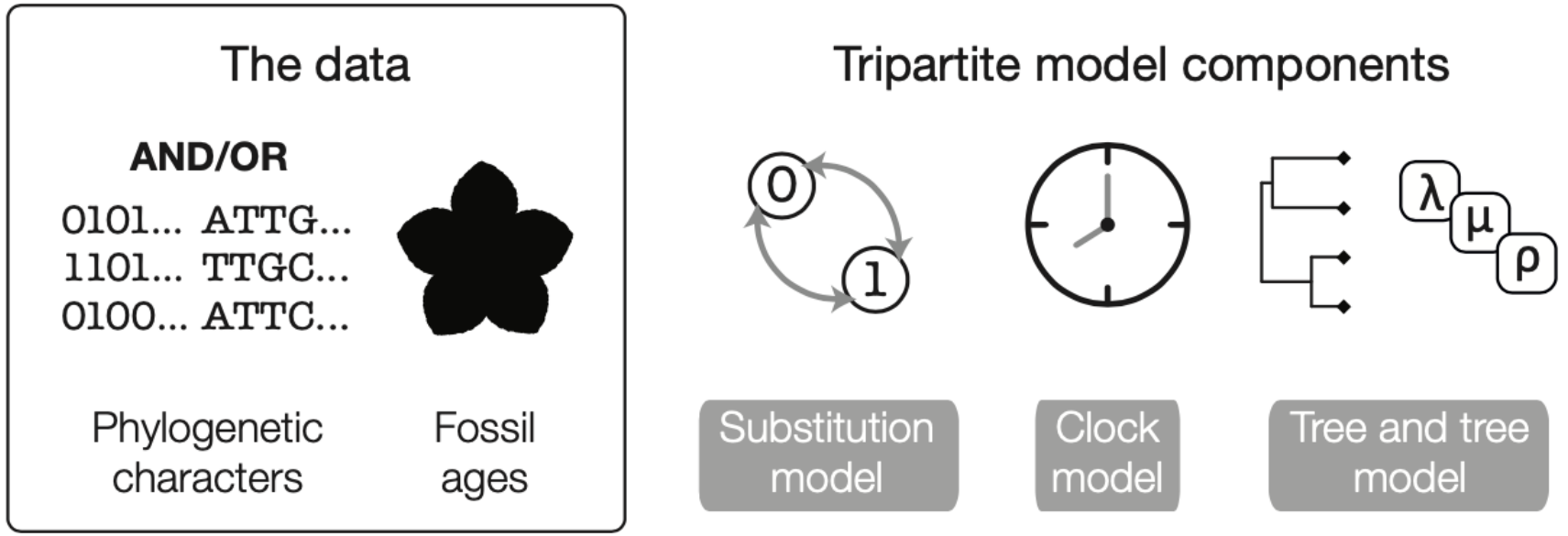
We can use morphological characters that are manually coded to describe species traits

This is incredibly time consuming and meticulous type of data collection

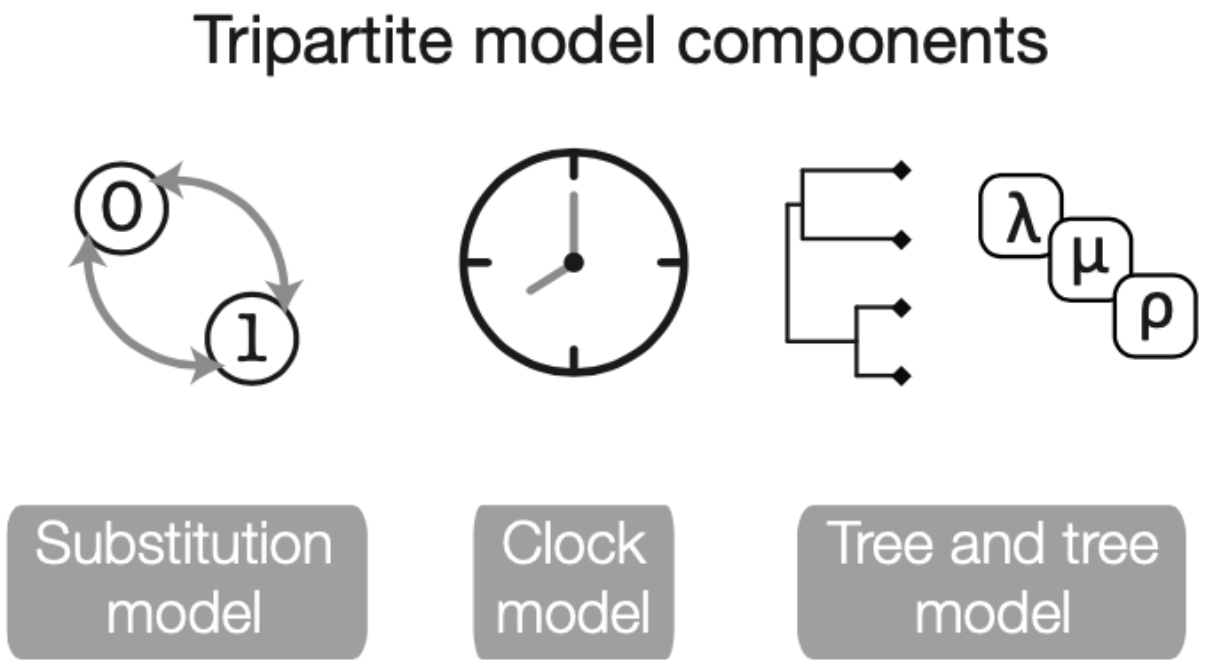
Using fossils in phylogenetic inference

What types of characters/traits would you use if you were making a morphological matrix

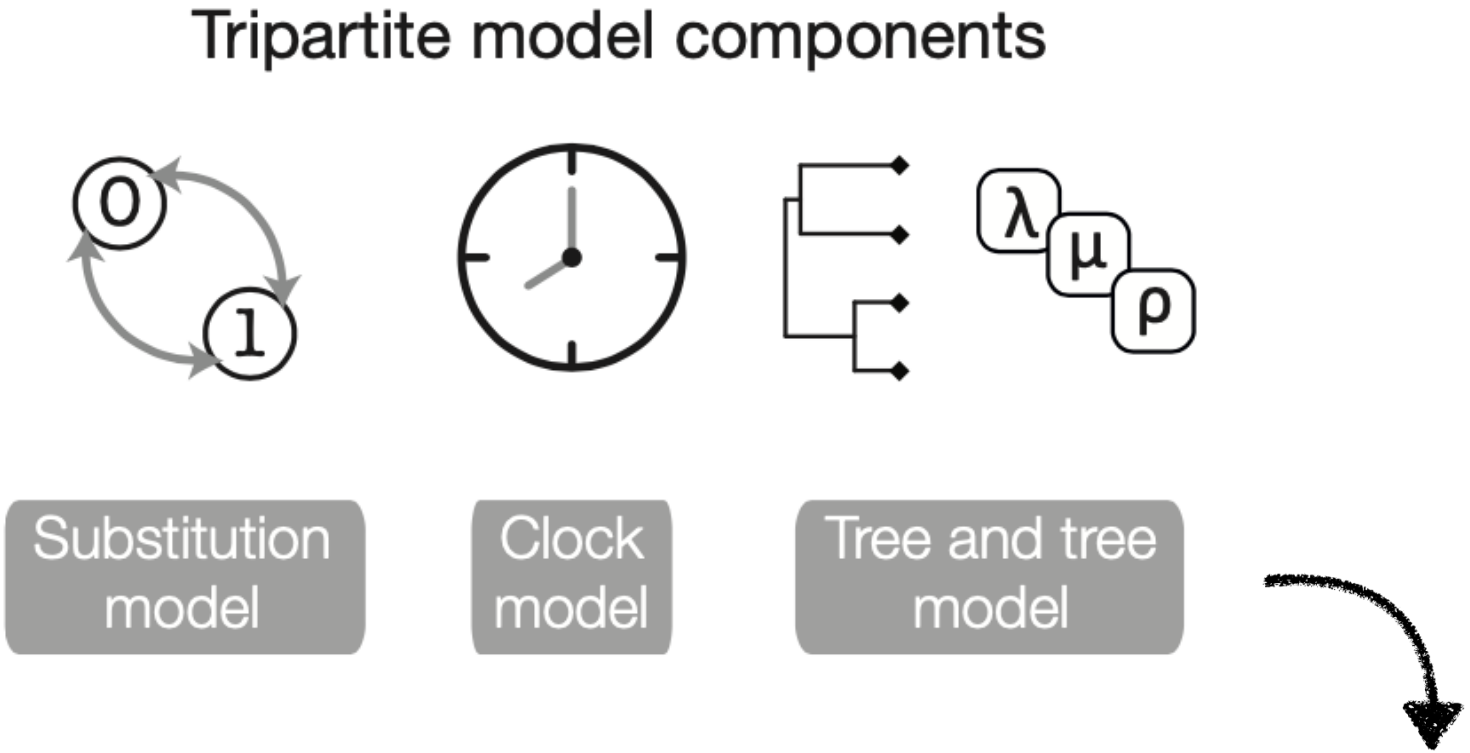
Bayesian Phylogenetic Analysis Components



Bayesian Phylogenetic Analysis Components

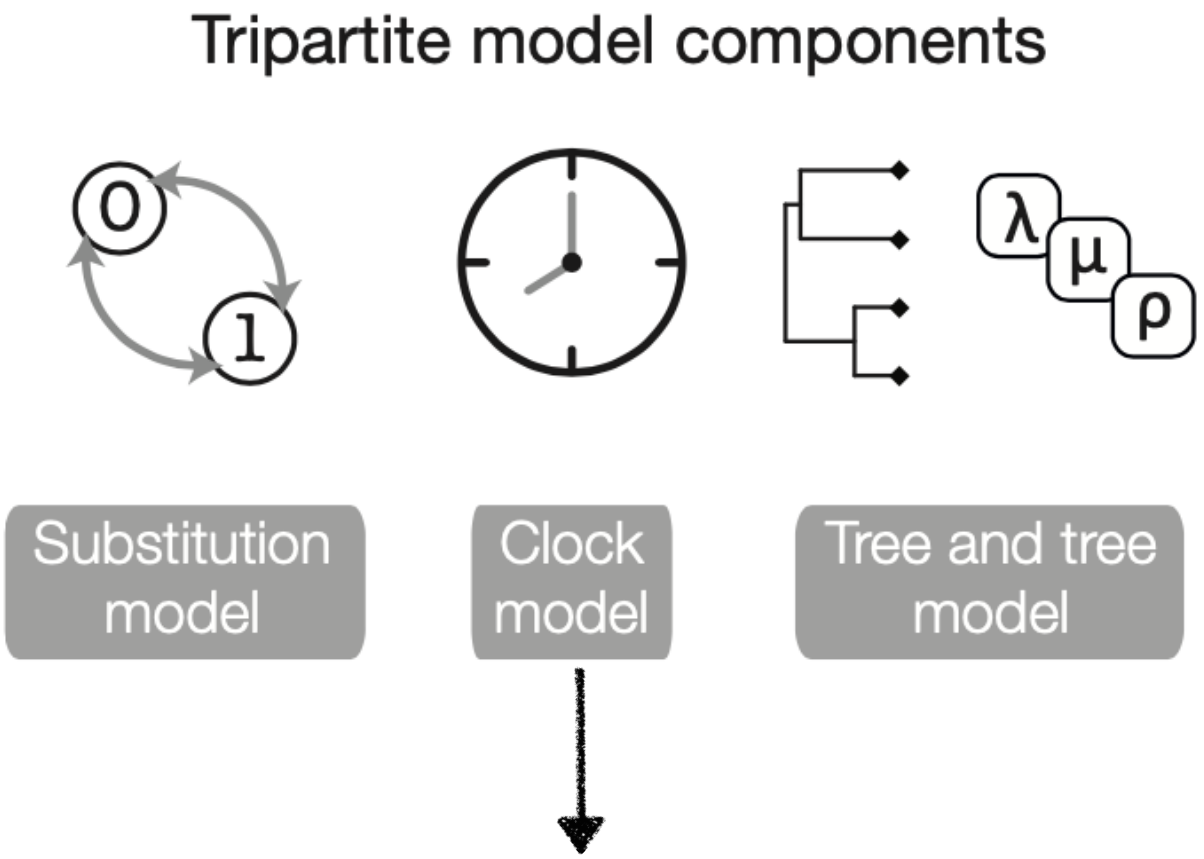


Bayesian Phylogenetic Analysis Components



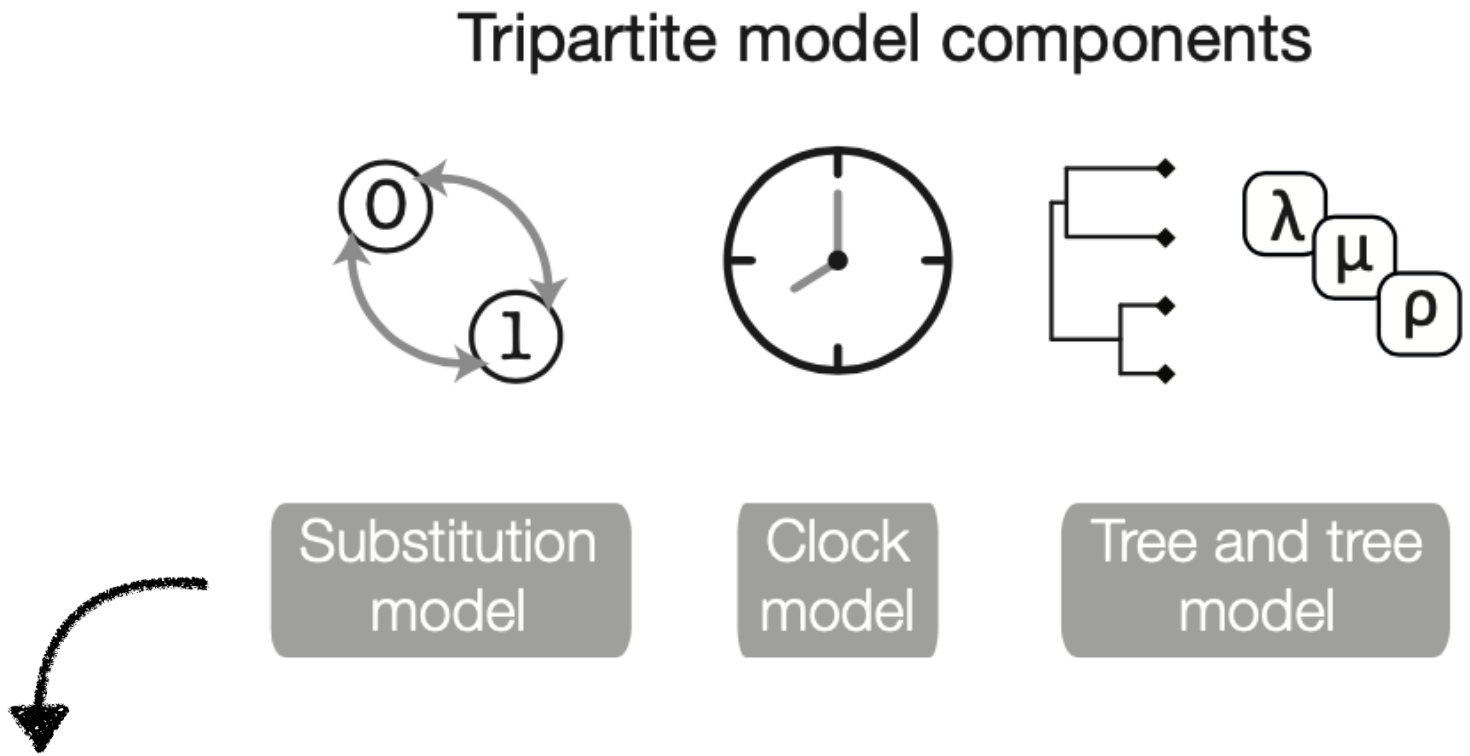
How have species originated, gone extinct and been sampled through

Bayesian Phylogenetic Analysis Components



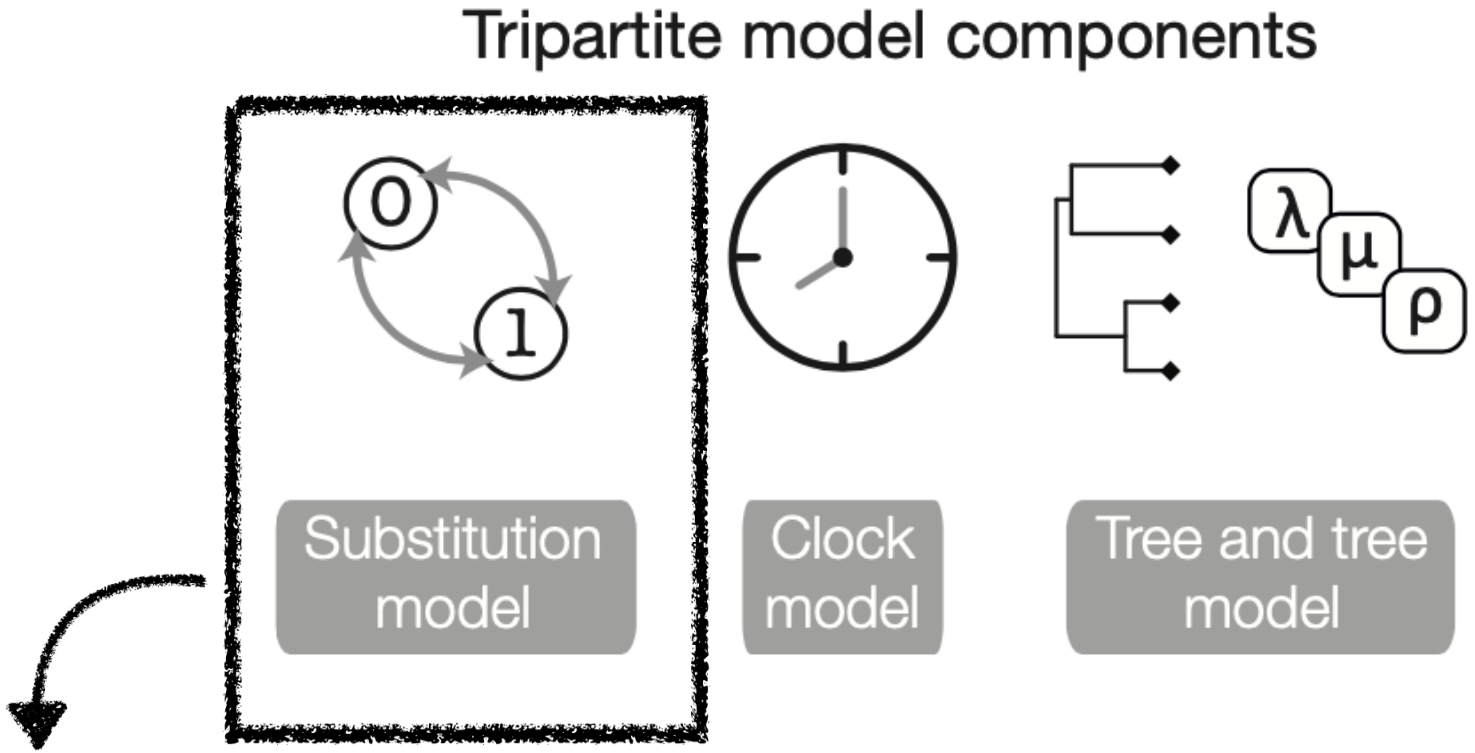
How have rates of evolution varied (or not) across the tree?

Bayesian Phylogenetic Analysis Components



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

Bayesian Phylogenetic Analysis Components

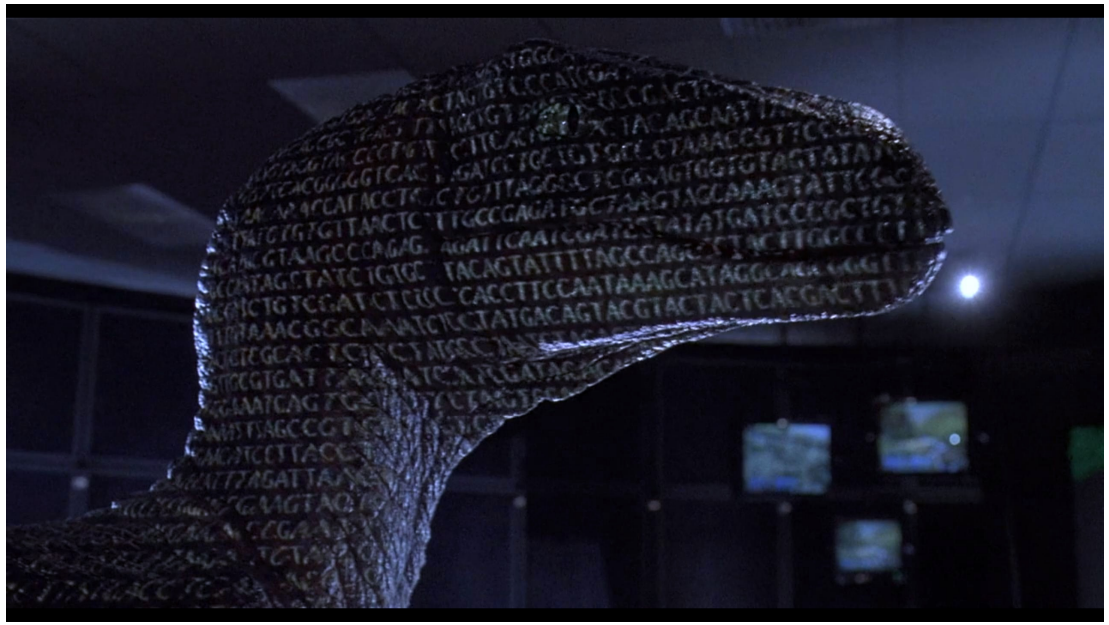


How likely are we to observe a change between character states? e.g., 0 → 1

Molecular Substitution models

JC substitution model

$$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$$



GTR substitution model

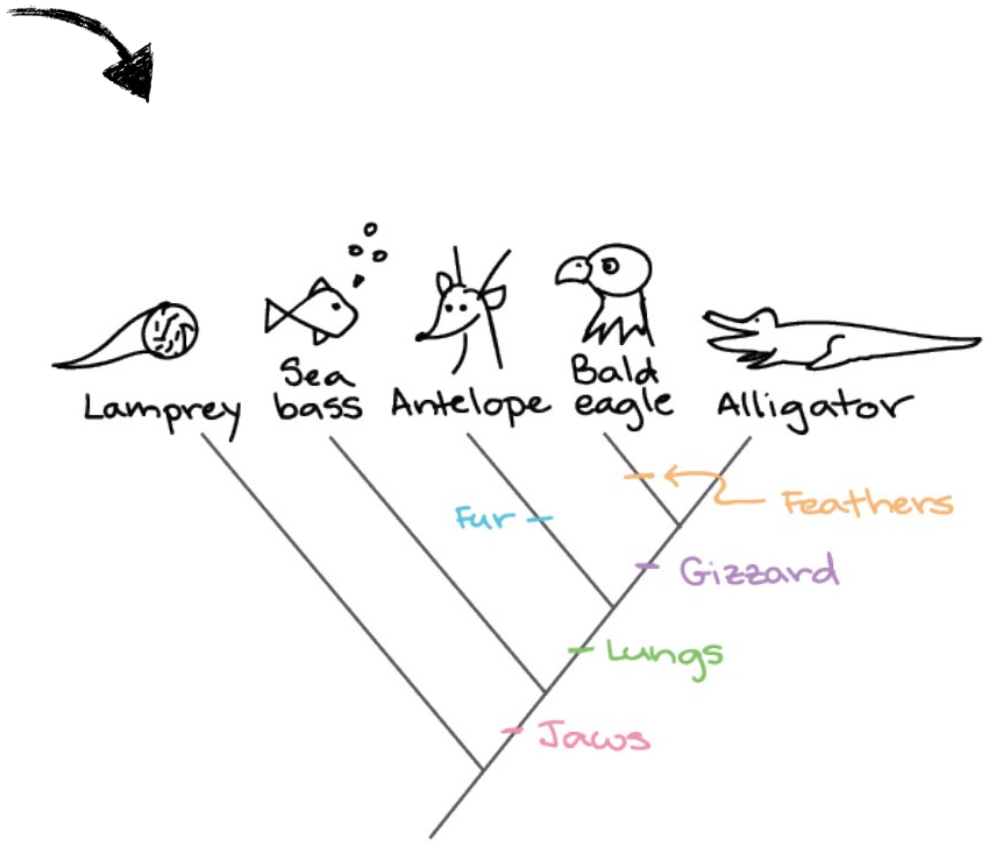
$$Q = \begin{pmatrix} * & \mu_{AG}\pi_G & \mu_{AC}\pi_C & \mu_{AT}\pi_T \\ \mu_{GA}\pi_A & * & \mu_{GC}\pi_C & \mu_{GT}\pi_T \\ \mu_{CA}\pi_A & \mu_{CG}\pi_G & * & \mu_{CT}\pi_T \\ \mu_{TA}\pi_A & \mu_{TG}\pi_G & \mu_{TC}\pi_C & * \end{pmatrix}$$

μ = substitution rate
Π = stationary frequency

Morphological data

	Lungs	Jaws	Feathers	Gizzards	Fur
taxa A	0	0	0	0	0
taxa B	1	1	0	0	1
taxa C	1	1	1	1	0
taxa D	1	1	0	1	0
taxa E	0	1	0	0	0

Note: the biological interpretation of a ”1” is not the same across the matrix here

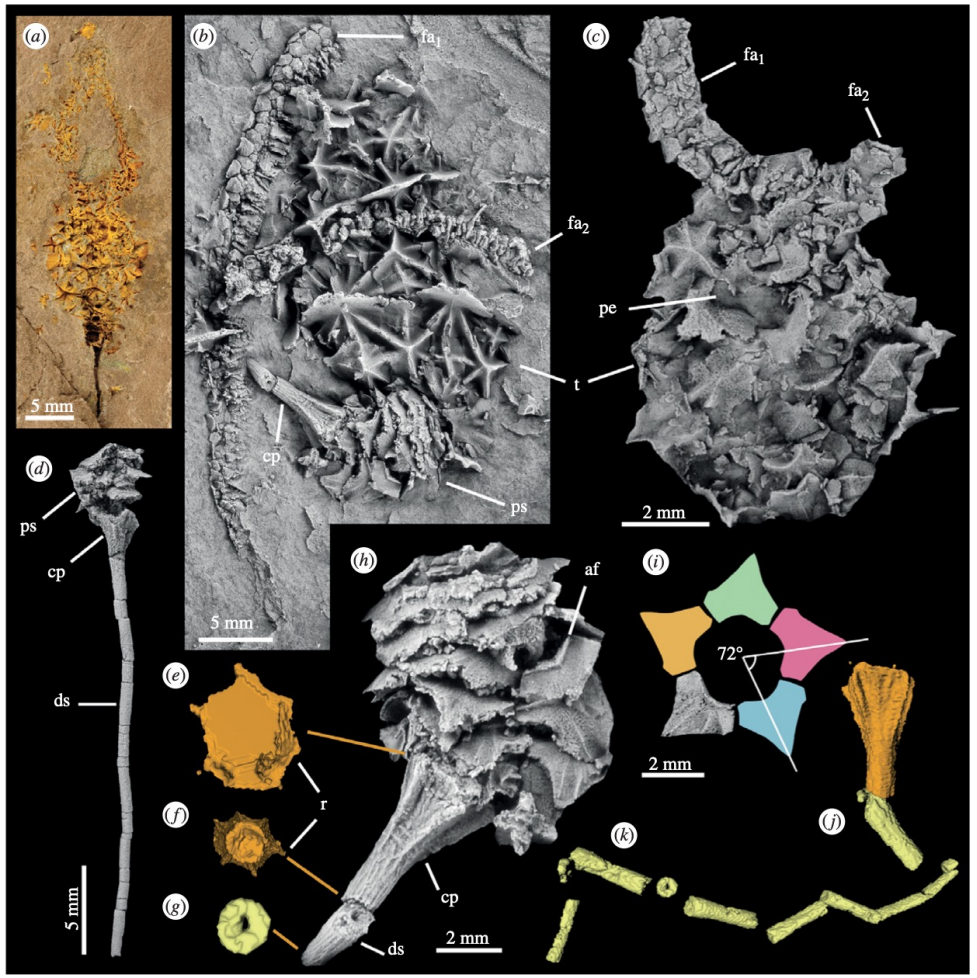


Appendage branching pattern

Coving plate arrangement

Presence Absence

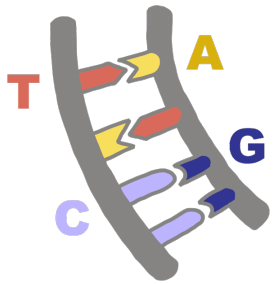
```
0015100l0?00-100--0000000000
0005000l0?200100--0010010000
0025000l0?200100--0?10010000
00?5?00l0?200100?-0???010110
0015000l0120100043010011111
0015000l0120101044011011111
??050????2010004402110111111
01050?010-210000?501??010110
0002000l002101003-1110010110
0002000l00211001441121011111
000201111-210010?-??11011121
?103?0?11?1001104-0000010000
1005002l0100010--0?00110?20
1005002000l010105402?00110020
```

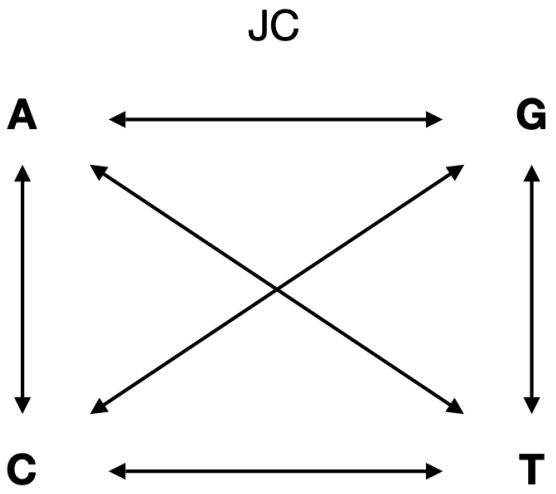


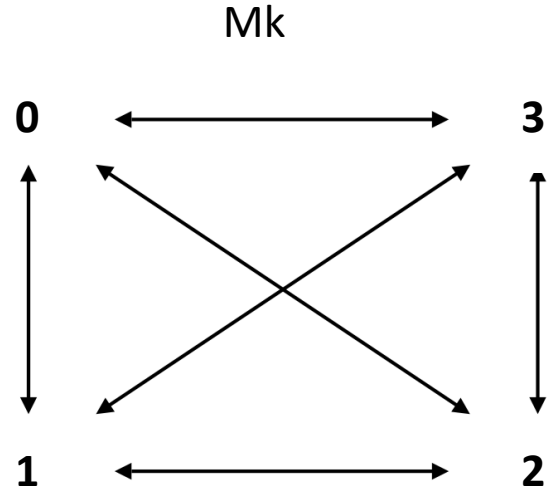
Dibrachicystis purujoensis

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

Substitution models for morphological data





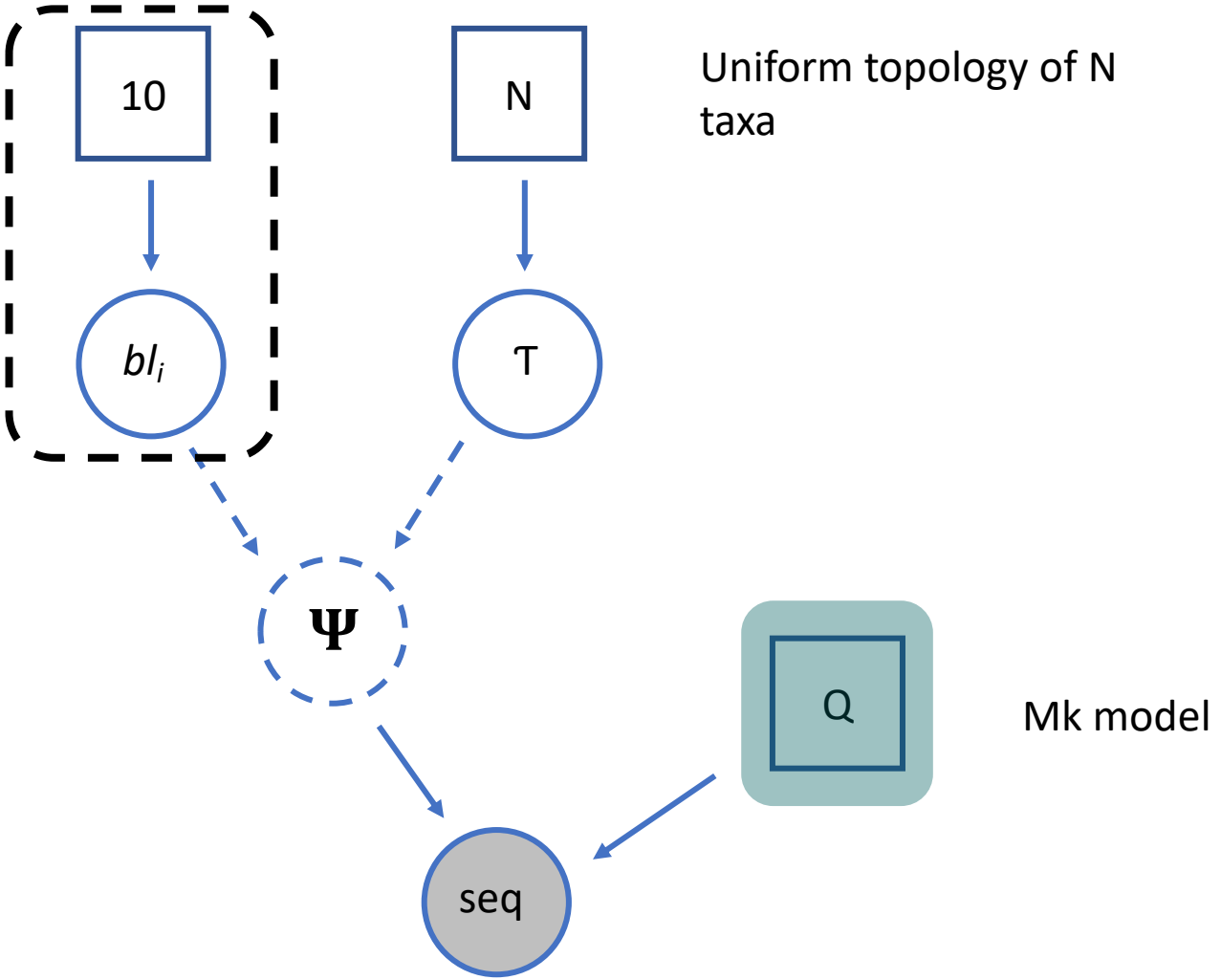


*4 state here as an example, can be any number from 2!

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

Mk Model

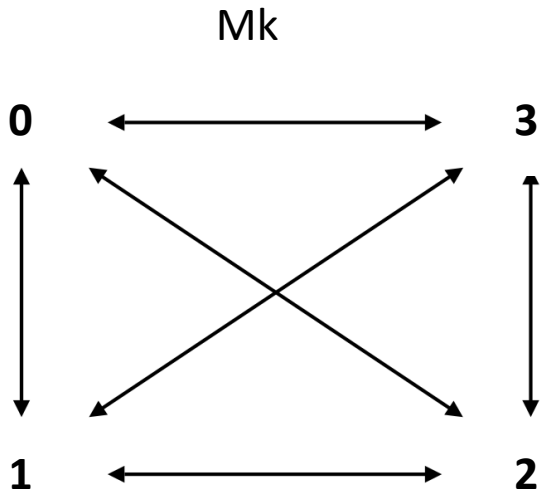
Exponential rate parameter of 10 on branch lengths.



Uniform topology of N taxa

Mk model

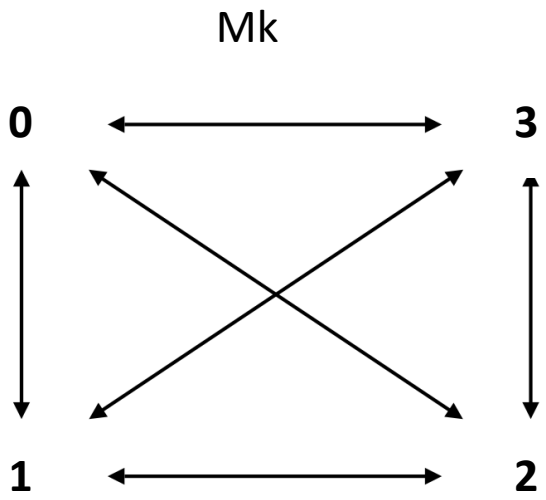
Substitution models for morphological data



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

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

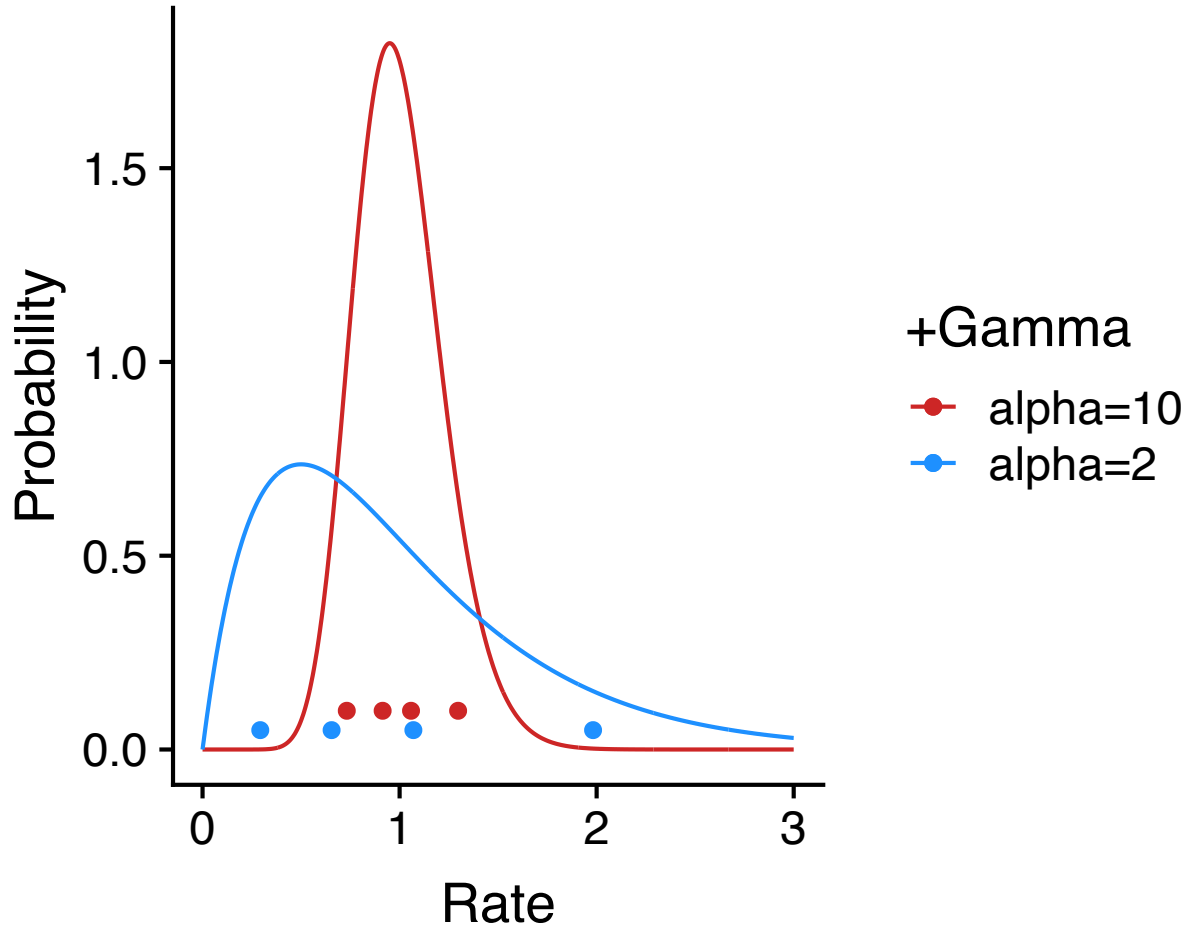
Substitution models for morphological data



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!

Across Site Rate Variation (+G)

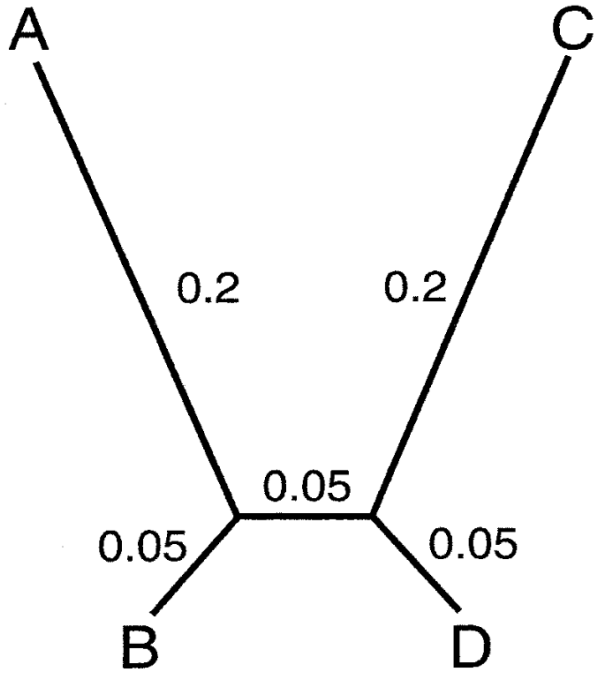


alpha = 10, the rates are similar
alpha = 2 the rates differ

This approach allows **faster evolving sites to evolve according to higher rates** and visa versa

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 (±349,100)	0.206 (±0.060)
Branch B	0.05	0.43210 (±0.13756)	0.050 (±0.018)
Branch X	0.05	54.646 (±1,725.3)	0.052 (±0.023)
Branch C	0.2	143,950 (±228,910)	0.206 (±0.059)
Branch D	0.05	0.022 (±0.054)	0.051 (±0.019)

Lewis 2001

Partitioning the data

Researchers have argued that it is reasonable partition a morphological matrix by the number of character states

Taxa A	010023
Taxa B	201102
Taxa C	112131

Exercise 1:

This exercise will walk you through an unconstrained (no temporal information) inference using a morphological data set of terror birds

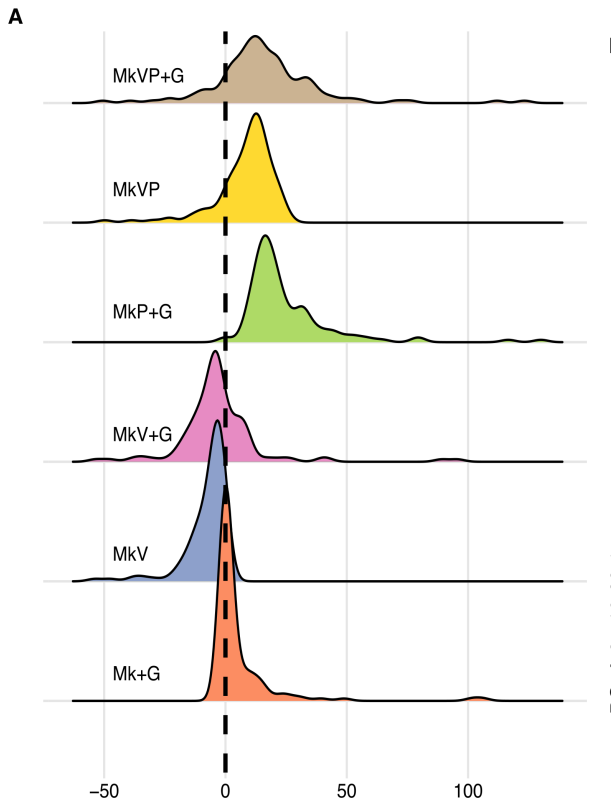
Interpreting the output

Did you get a different topology from the different models?

Are the tree lengths different from each other? What does it mean in terms of evolutionary distance if the tree lengths are different?

Which model should we use?

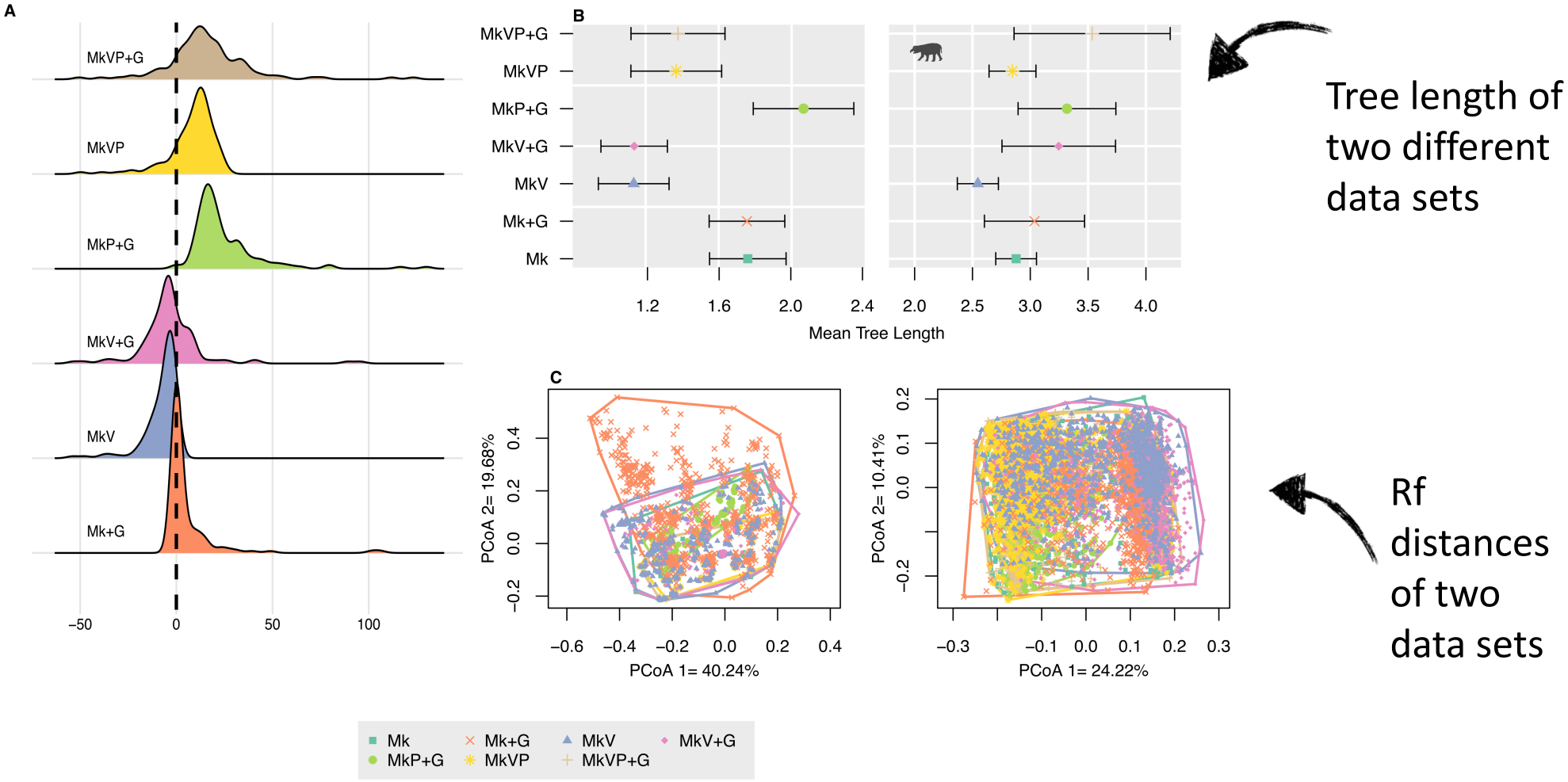
Models do impact key parameter estimates



Percentage difference in tree length relative to Mk model

Mk Mk+G MkV MkV+G
MkP+G MkVP MkVP+G

Models do impact key parameter estimates



Tree length of two different data sets

Rf distances of two data sets

Percentage difference in tree length relative to Mk model

Comparing topologies

Generalised Robinsons Foulds distance is the most common method used for comparing phylogenetic trees

It calculates the number of difference in terms of nodes across a tree

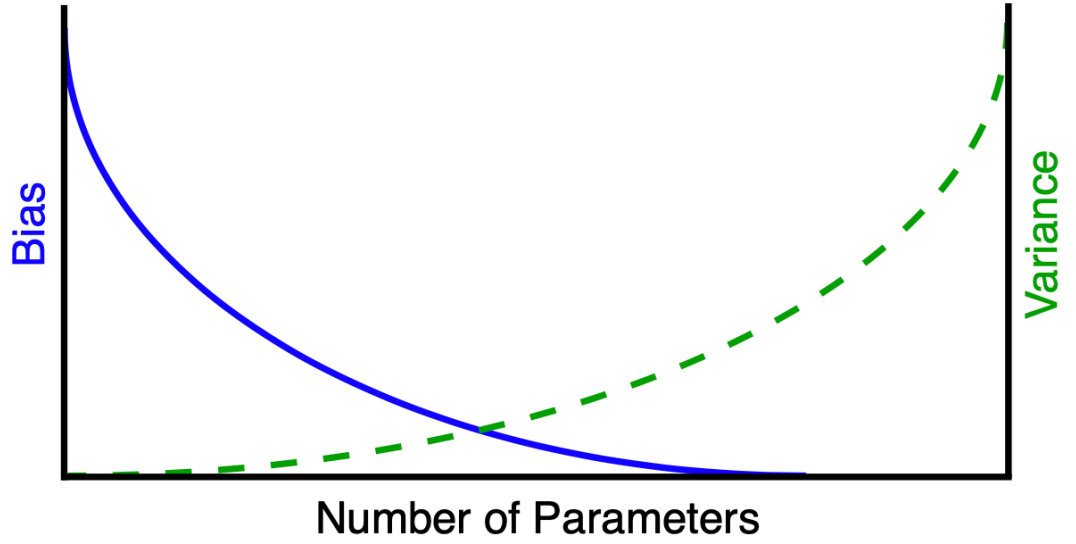
It can be quite conservative though and makes trees appear to be extremely different from one another

Another approach which aims to be less conservative is **Quartets**

What does a good model look like?

To do statistical inference we need a model
What model should that be?
Our goal should be to have a model that is **complex enough** to capture the “important” variation in the data, but **not be more complex** than it needs to be

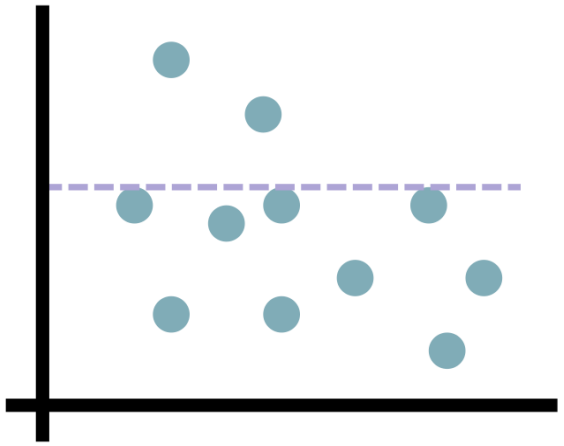
Too simple, misinterpreting the data



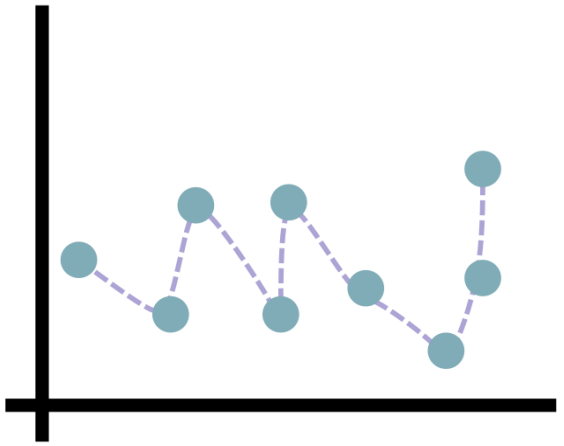
Too complicated, not enough information

What does a good model look like?

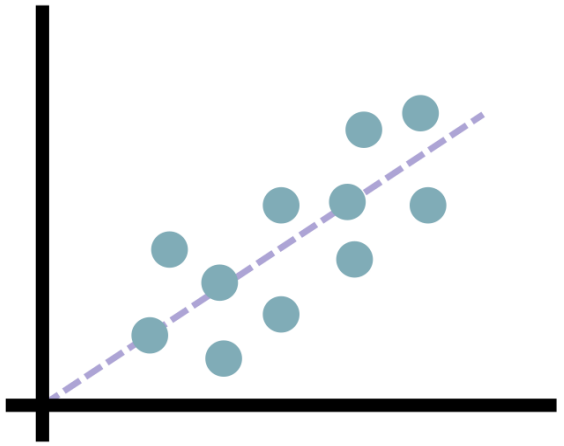
To do statistical inference we need a model
What model should that be?
Our goal should be to have a model that is **complex enough** to capture the “important” variation in the data, but **not be more complex** than it needs to be



Underfitting



Overfitting



Proper fit

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 mdoels

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 mdoels

$$P(\text{parameters} \mid \text{data}, \text{model}) = \frac{P(\text{data} \mid \text{parameters}, \text{model})\, P(\text{parameters} \mid \text{model})}{P(\text{data} \mid \text{model})}$$

Posterior

Likelihood

Priors

Marginal probability

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 mdoels

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)

Höhna et al 2016

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

Empirical Data	
taxa 1	0 1 0 1 2 1
taxa 2	1 2 1 0 1 0
taxa 3	0 0 1 0 0 1
taxa 4	1 1 0 1 0 1

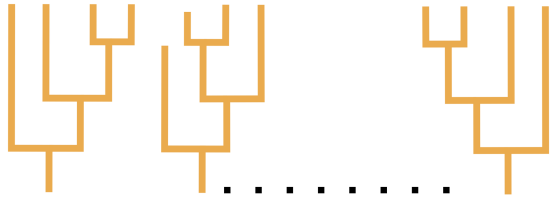
Höhna et al 2017

Posterior Predictive Simulations

Empirical Data						
taxa 1	0	1	0	1	2	1
taxa 2	1	2	1	0	1	0
taxa 3	0	0	1	0	0	1
taxa 4	1	1	0	1	0	1

1)

Standard MCMC inference while sampling from the posterior

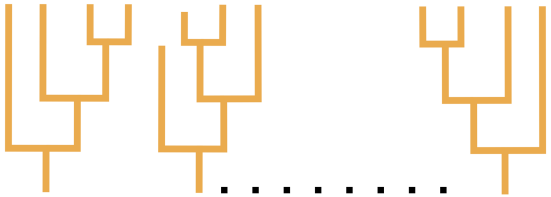


Posterior Predictive Simulations

Empirical Data	
taxa 1	0 1 0 1 2 1
taxa 2	1 2 1 0 1 0
taxa 3	0 0 1 0 0 1
taxa 4	1 1 0 1 0 1

1)

Standard MCMC inference while sampling from the posterior



2)

Using the information sampled in 1) generate new data sets

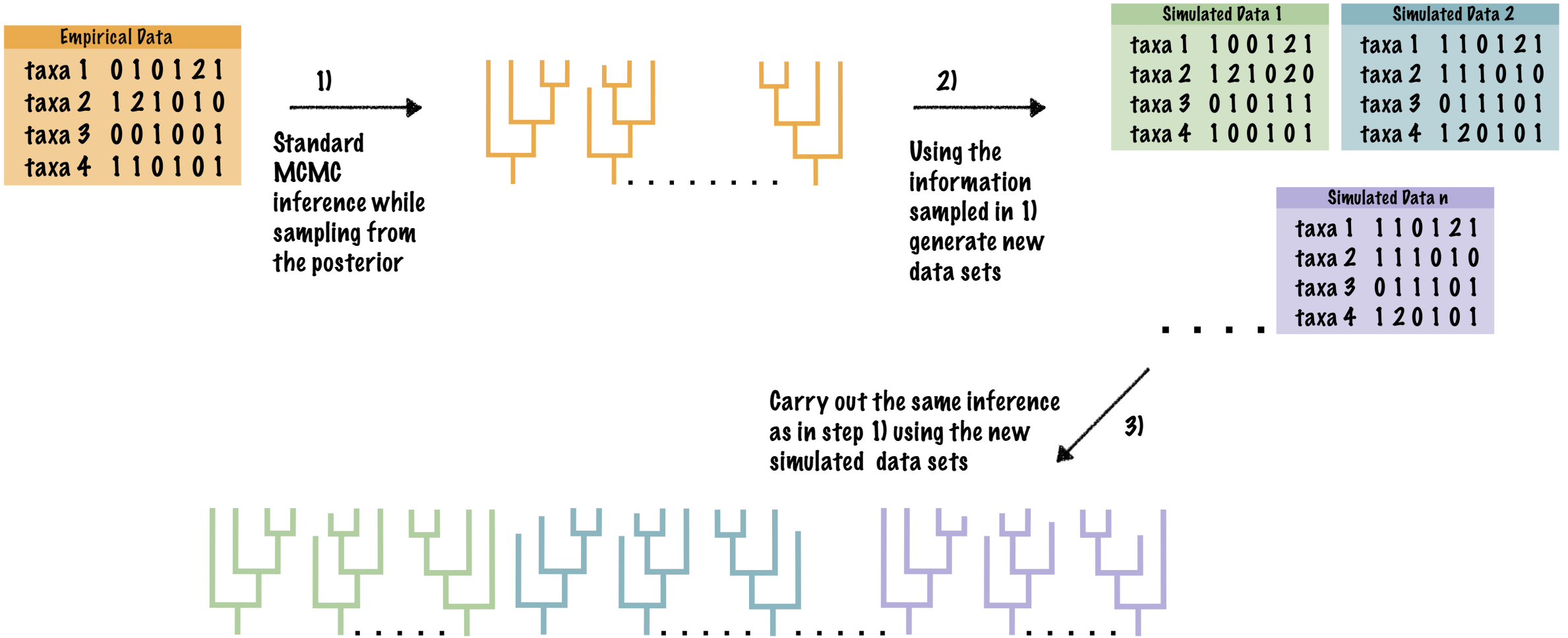
Simulated Data 1	
taxa 1	1 0 0 1 2 1
taxa 2	1 2 1 0 2 0
taxa 3	0 1 0 1 1 1
taxa 4	1 0 0 1 0 1

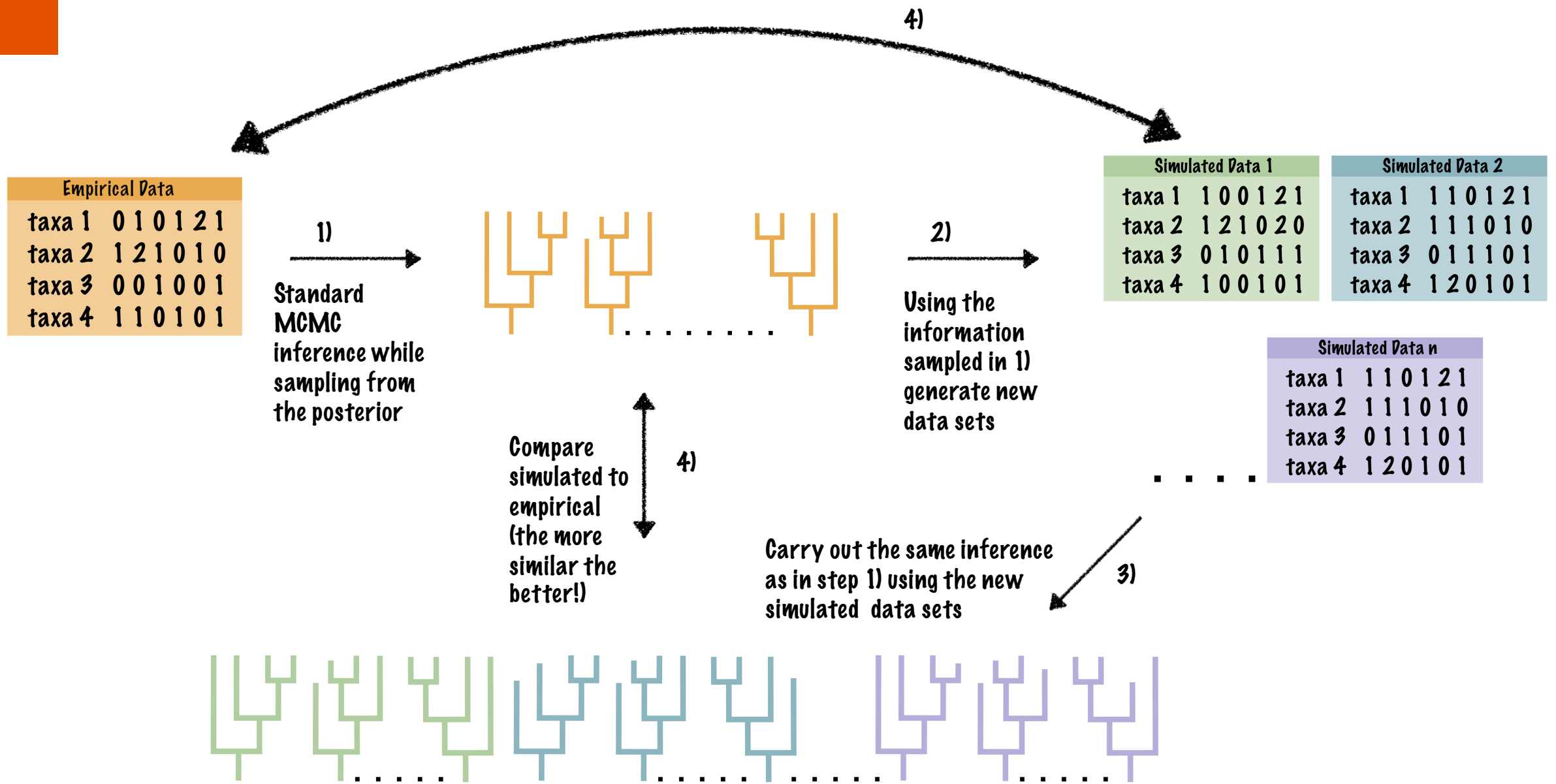
Simulated Data 2	
taxa 1	1 1 0 1 2 1
taxa 2	1 1 1 0 1 0
taxa 3	0 1 1 1 0 1
taxa 4	1 2 0 1 0 1

. . . .

Simulated Data n	
taxa 1	1 1 0 1 2 1
taxa 2	1 1 1 0 1 0
taxa 3	0 1 1 1 0 1
taxa 4	1 2 0 1 0 1

Posterior Predictive Simulations




4)
Empirical Data
taxa 1 0 1 0 1 2 1
taxa 2 1 2 1 0 1 0
taxa 3 0 0 1 0 0 1
taxa 4 1 1 0 1 0 1
1)
Standard MCMC inference while sampling from the posterior
2)
Using the information sampled in 1) generate new data sets
Simulated Data 1
taxa 1 1 0 0 1 2 1
taxa 2 1 2 1 0 2 0
taxa 3 0 1 0 1 1 1
taxa 4 1 0 0 1 0 1
Simulated Data 2
taxa 1 1 1 0 1 2 1
taxa 2 1 1 1 0 1 0
taxa 3 0 1 1 1 0 1
taxa 4 1 2 0 1 0 1
Simulated Data n
taxa 1 1 1 0 1 2 1
taxa 2 1 1 1 0 1 0
taxa 3 0 1 1 1 0 1
taxa 4 1 2 0 1 0 1
Compare simulated to empirical (the more similar the better!)
4)
Carry out the same inference as in step 1) using the new simulated data sets
3)

Test Statistics

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

A value that captures the characteristic of you 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

Test Statistics

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

A value that captures the characteristic of you 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?