Phylogenetics

Morphological Models & Model Adequacy

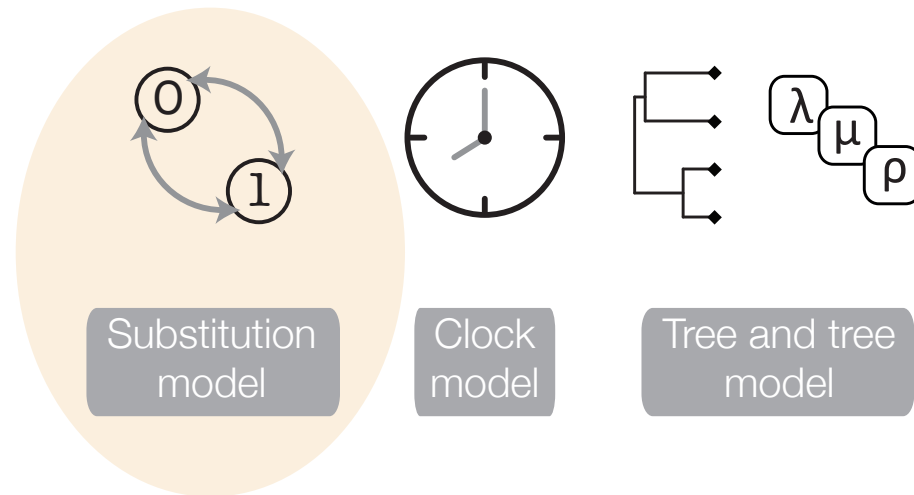Laura Mulvey, Rachel Warnock
June 09, 2022

Today's Objectives

Morphological substitution models used in Palaeobiology
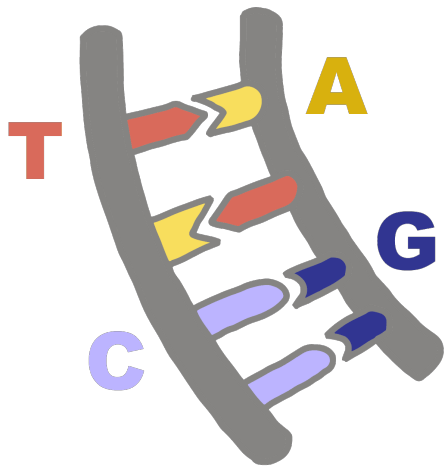
How to choose what model to use

Exercise using model adequacy

Phylogenetic Models

Substitution model

Clock model

Tree and tree model

Substitution models in Palaeobiology

- **Mk:** makes the same assumptions as the JC model. i.e. all transition probabilities are equal

- Only one model really used as morphological data is harder to generalise

Merychippus_primus_ORIG_M_primus
0000110000001001000100100102112200?001?0000011000100000
Merychippus_sejunctus_ORIG_M_nr_sejunctus
00011200001130120001000002020221110002???00012001100000
Merychippus_coloradense_ORIG_M_coloradense
00?00200?10371?2100321002014113011 0?102???00212001121011

Variations on the Mk model

- Discrete Gamma Model

- Variable Coding

- Multistate

Variations on the Mk model

- Discrete Gamma Model : allows for rate variation across sites

Variable Coding

	Lungs	jaws	Feathers	Gizzards	Fur
Lamprey	0	0	0	0	0
Sea bass	0	1	0	0	0
Antelope	1	1	0	0	1
Bald eagle	1	1	1	1	0
Alligator	1	1	0	1	0

- Causes ascertainment bias. Suggests that there is evolutionary distance between taxa than there is.

- Produces longer branches

Compared to molecular data

Sequence the whole gene. Not every character will be informative

Multistate

- Create different Q matrices for characters traits with different number of states.

	A	B	C
taxa 1	0	0	1
taxa 2	1	1	1
taxa 3	0	2	0
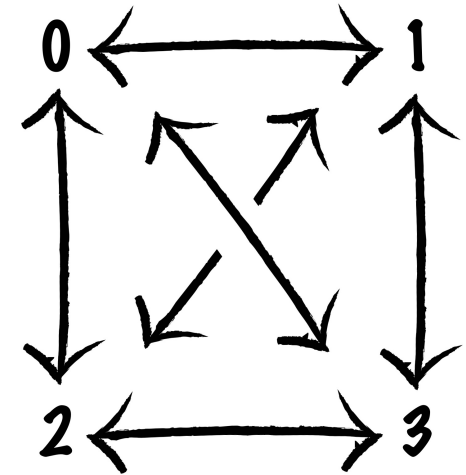
Variations on the Mk model

- Discrete Gamma Model

- Variable Coding

- Multistate

Effects different models have on the inference

2_Demar
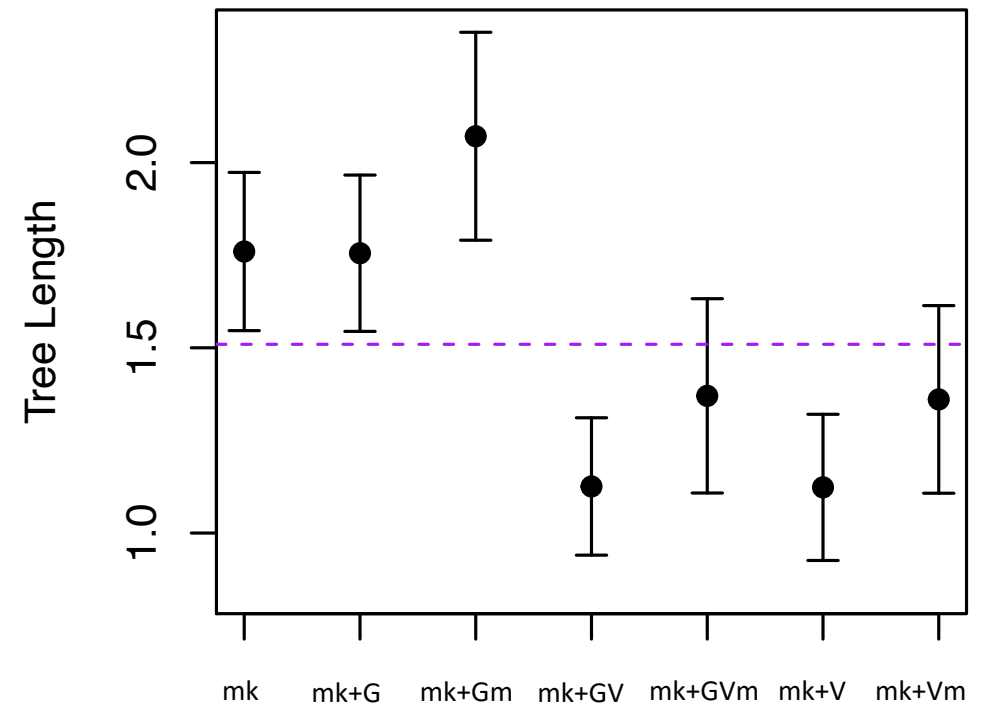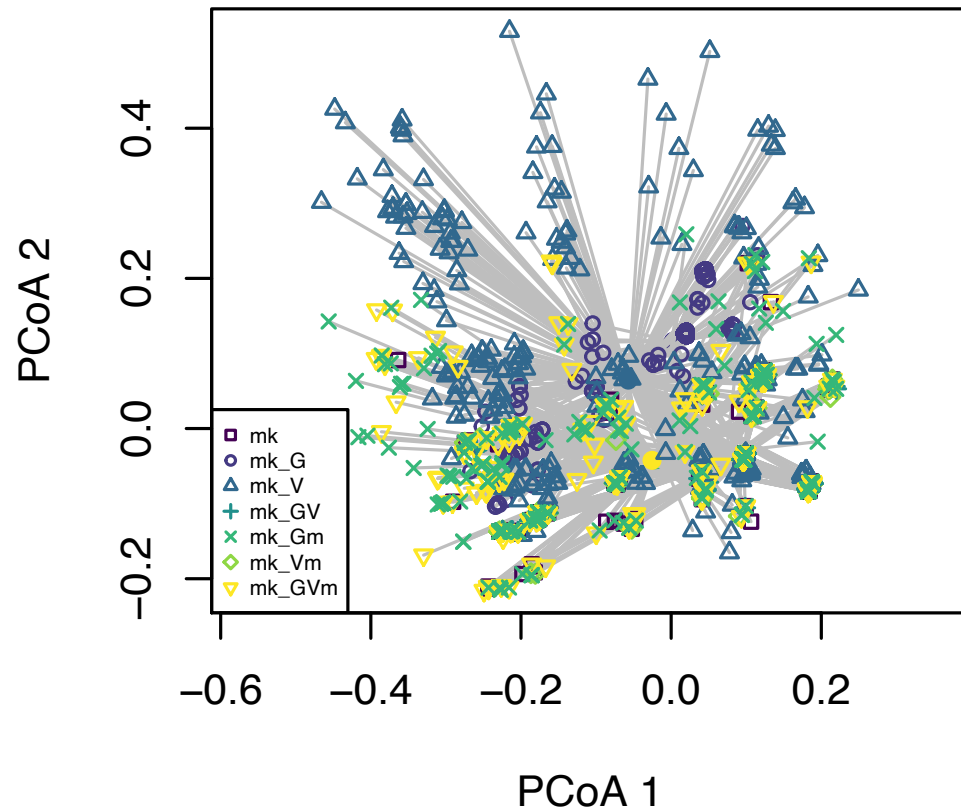
(Mulvey et al in prep.)

How to choose a model

P (parameters | data, model) =

Posterior

Likelihood

Priors

P (data | parameters, model) P (parameters | model)

P (data | model)

Marginal
probability

Model selection: Stepping Stone
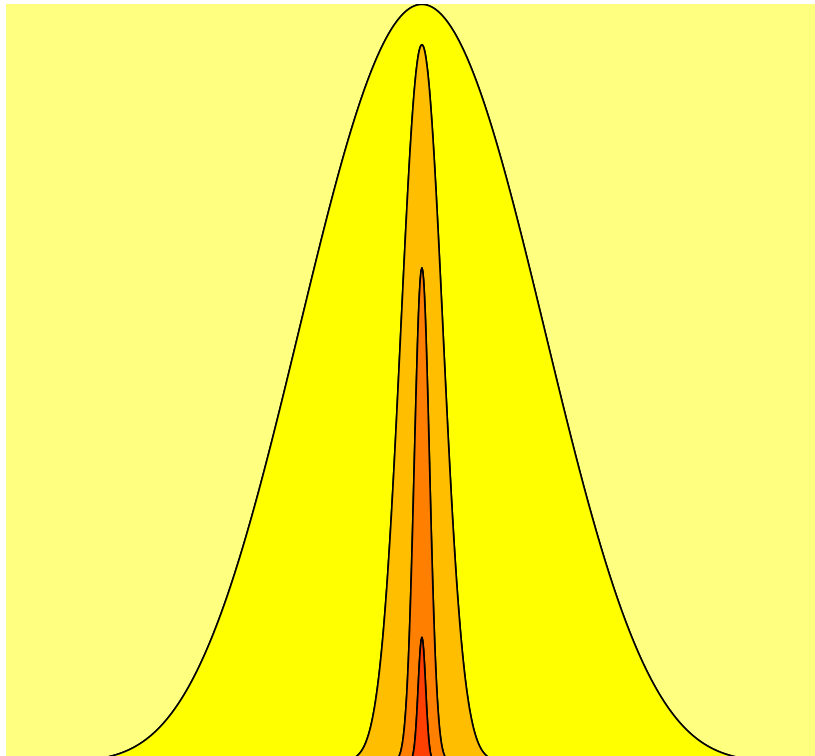
- Estimates the marginal likelihood by estimating the area under nested curves. Using smaller and smaller curves until you get down to the size of the marginal likelihood curve

- Simulated data sets show this approach seems to always choose the most complex model

Image from Paul Lewis Phyloseminar

Model selection vs. Model adequacy

Take a bunch of different models and test which is the *best*

Gives the **relative** fit

0102
0110
2001

?
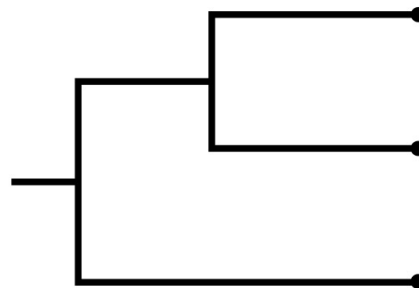
Assess whether a model is capturing the evolutionary dynamics that generated the data

Gives the **absolute** fit

Model Adequacy

1. Carry out inference under the chosen model, while sampling parameters from the posterior

2. Simulate new data sets using the parameters from step 1

3. Carry out the same inference as in step 1 using the new simulated data sets

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.

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

2.

3.
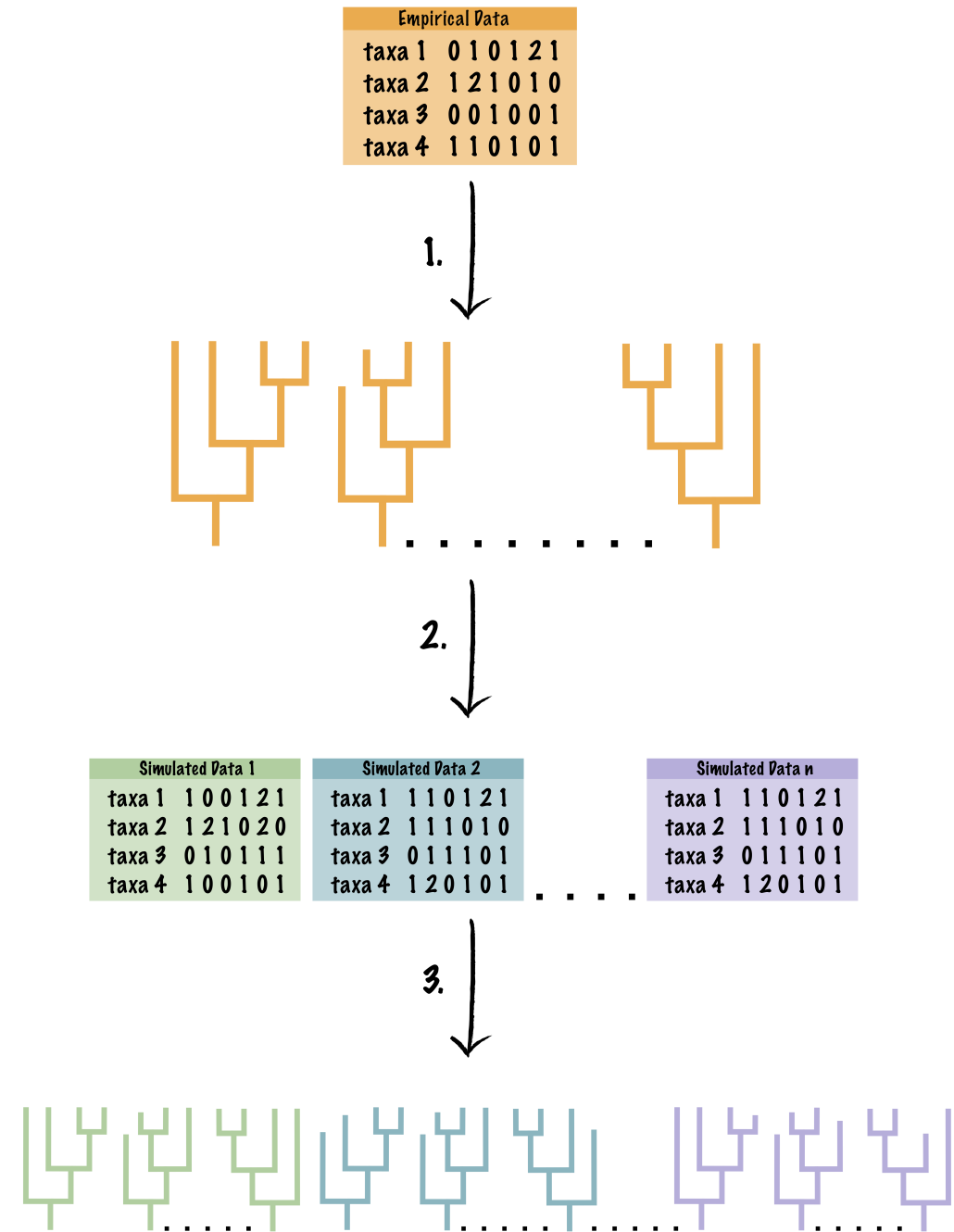
(Mulvey et al in prep.)

How can you compare different models using model adequacy?

1. Inference based

Tree length : Sum of all the branches
Robinson Foulds: the distance between phylogenetic trees

2. Mixed Inference

Consistency Index : measure of how well an individual character fits on a phylogenetic tree
Retention Index : measure of how well an individual character fits on a phylogenetic tree

Exercise 7