Part 1:
Morphological models

Laura Mulvey & Rachel Warnock

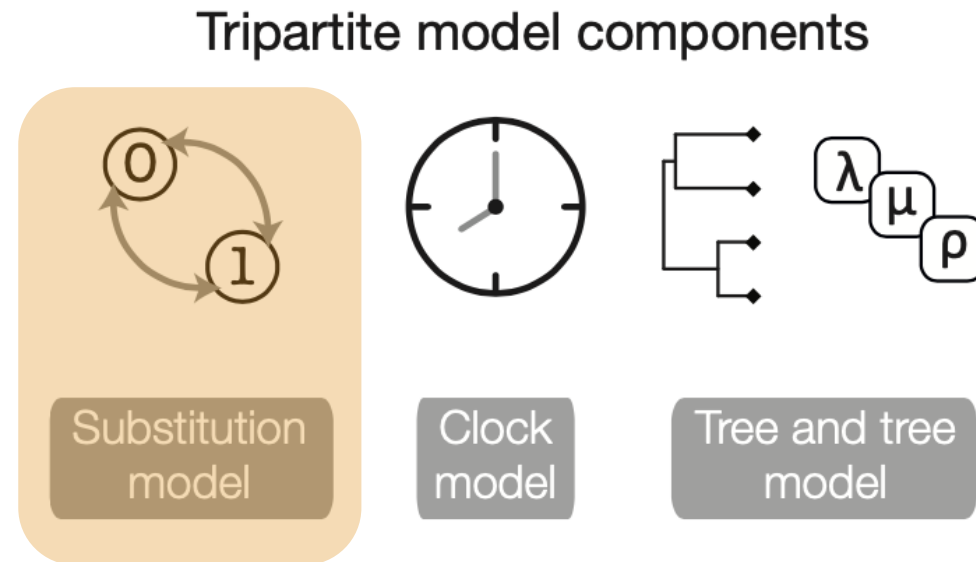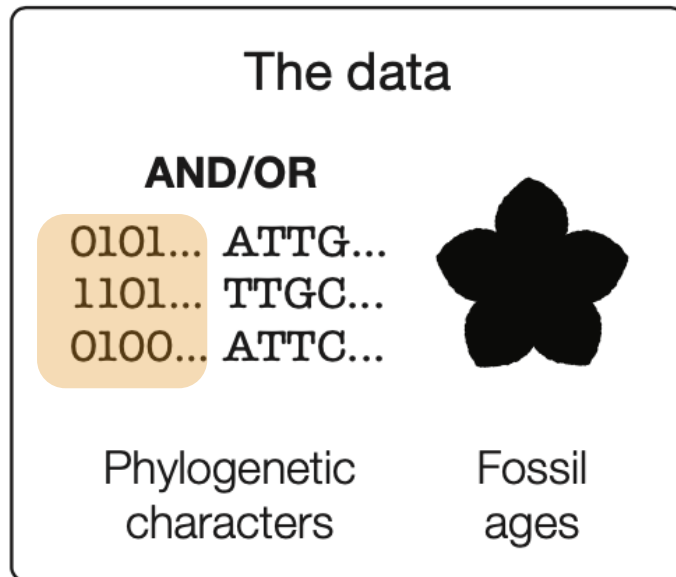FAU Erlangen-Nürnberg

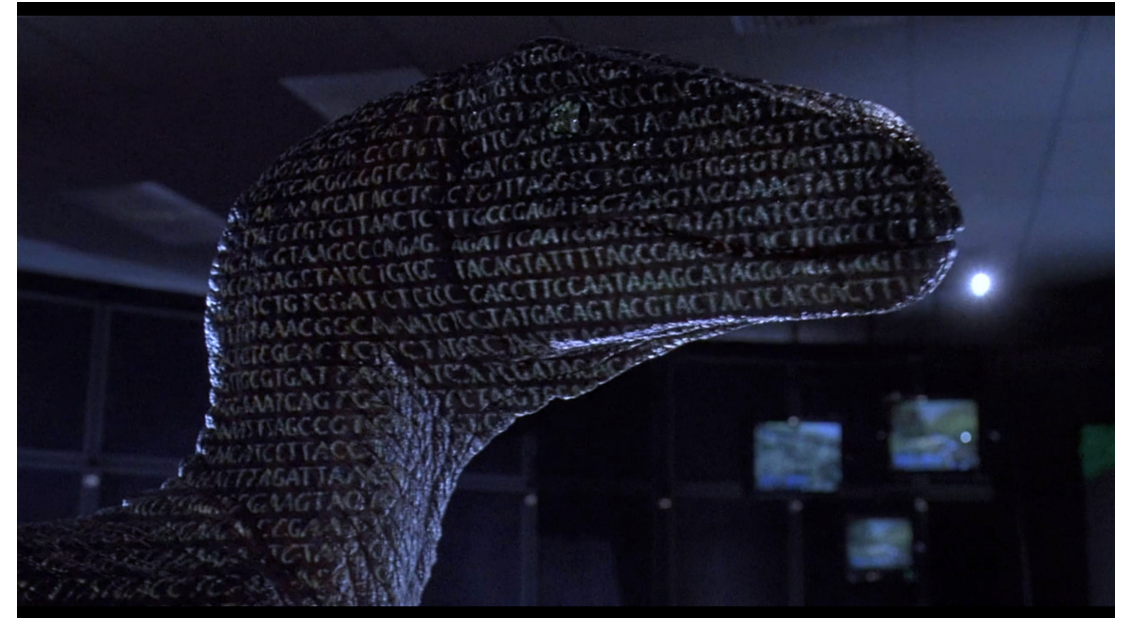Wednesday 07.09.22

Warnock & Wright 2020

Already know about substitution models for molecular data

JC substitution model

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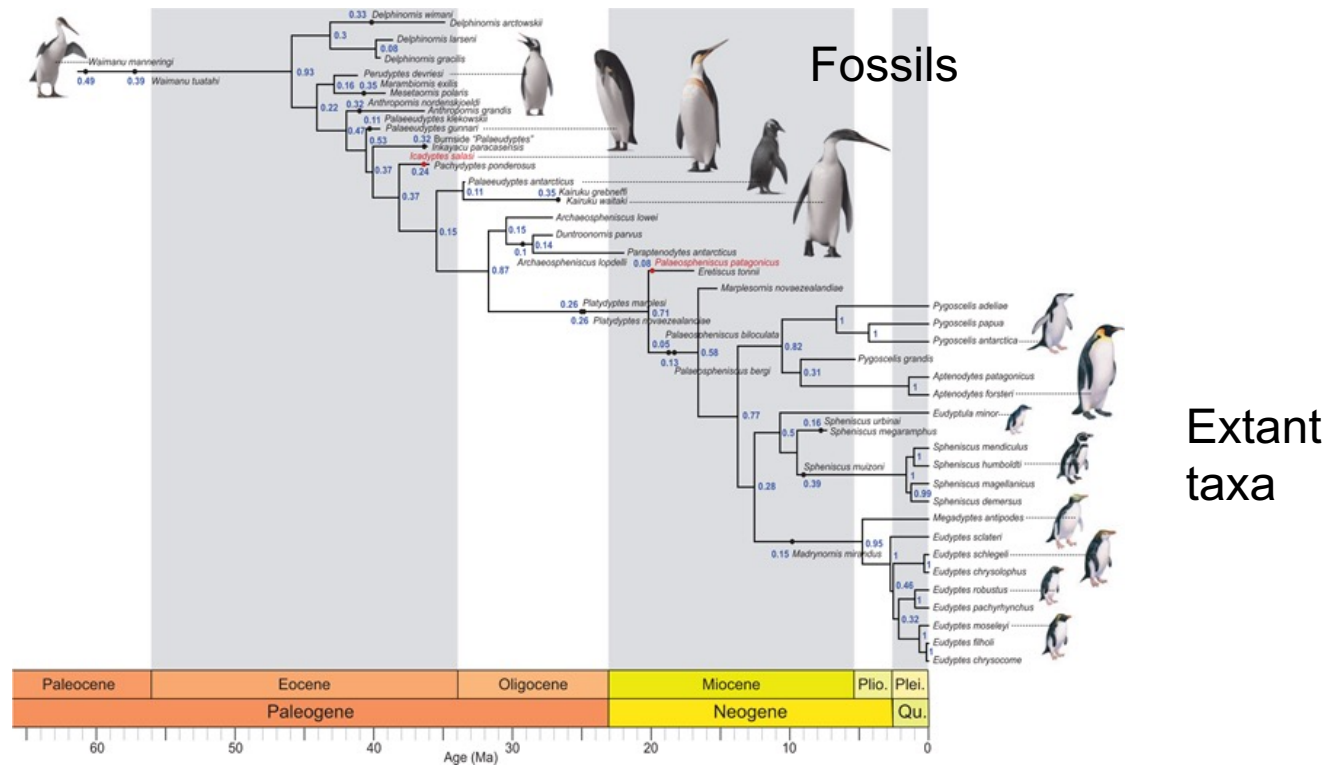
GTR substitution model

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

Models of discrete character data

For most organisms (essentially of extinct taxa) all we have is morphological data

Important for dating phylogenies, ancestral state reconstructing and testing evolutionary hypothesis.

Fossils

Extant taxa

Gavryushkina et al 2016

Morphological data

```
001510010?00-100--0000000000
000500010?200100--0010010000
002500010?200100--0?10010000
00?5?0010?200100?-0???010110
001500010120100043010001111
001500010120101044011011111
??050????201000440?11011111
01050?010-210000?501??010110
000200100210100?3-1110010110
000200100211001441121011111
000201111-210010?-??11011121
?103?0?11?1001104-0000010000
100500211010001000--0?00110?20
100500200010101054010?00110020
```

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

Challenges with Morphological data

Conodonts						
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

Image from Bryan Shirley

Challenges with Morphological data

Conodonts

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

Often used to indicate presence absence data

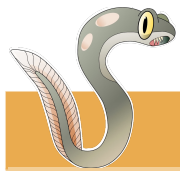Image from Bryan Shirley

Challenges with Morphological data

Conodonts

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

Multistate characters can be used to represent types of a trait

Image from Bryan Shirley

Challenges with Morphological data

Image from Bryan Shirley

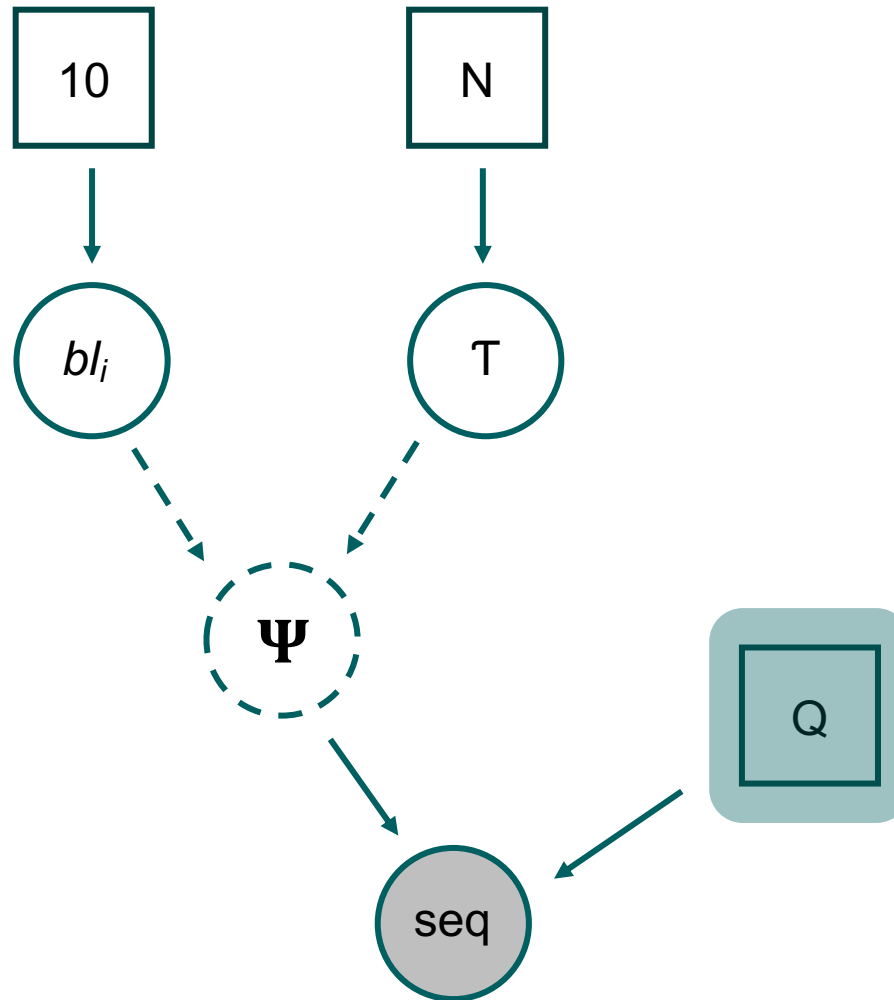| Conodonts						
taxa 1	0	1	0	1	2	1
taxa 2	1	2	1	0	1	0
taxa 3	0	0	1	0	0	1
taxa 4	1	1	0	1	0	1

Trait 1   Trait 2
0   ≠   0
1   ≠   1

Generalising morphological data is much more difficult that molecular

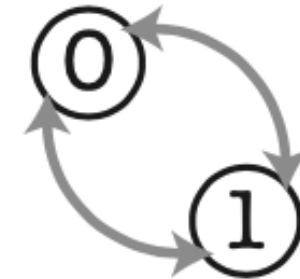April Wright has an awesome paper about this stuff!

Mk model – Lewis 2001

Generalised JC 69 model

Mk model assumptions

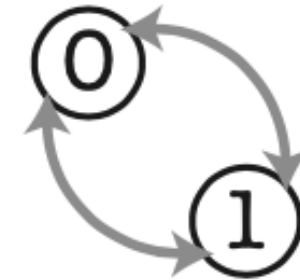It is a symmetric model : probability of changing from 0 to 1 same as 1 to 0

Mk model assumptions

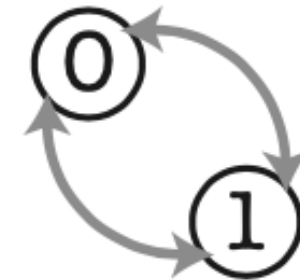It is a symmetric model : probability of changing from 0 to 1 same as 1 to 0

Characters are always in one of k states

Mk model assumptions

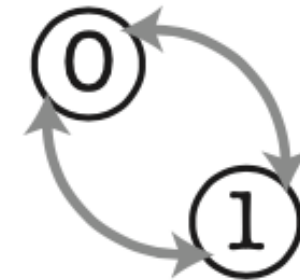It is a symmetric model : probability of changing from 0 to 1 same as 1 to 0

Characters are always in one of k states
Character change from one state to another is instantaneous along a branch

Mk model assumptions

It is a symmetric model : probability of changing from 0 to 1 same as 1 to 0

Characters are always in one of k states
Character change from one state to another is instantaneous along a branch
Changes are independent of each other
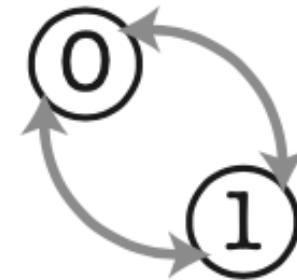
Mk model assumptions

It is a symmetric model : probability of changing from 0 to 1 same as 1 to 0

Characters are always in one of k states
Character change from one state to another is instantaneous along a branch
Changes are independent of each other
No state is a priori ancestral or derived

Extensions of the Mk model

It is possible to relax some of the assumptions of the MK model.

Extensions are included to account for factors that are considered important in morphological evolution.
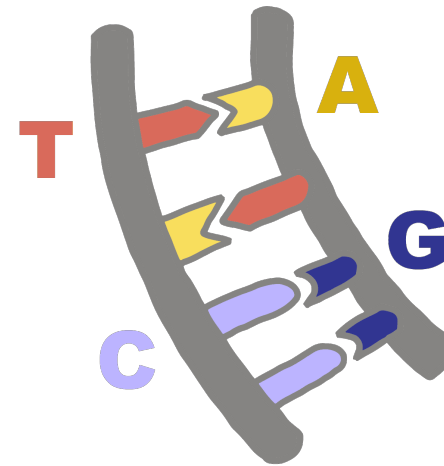
Aim to make the models closer to reality.

MkV model

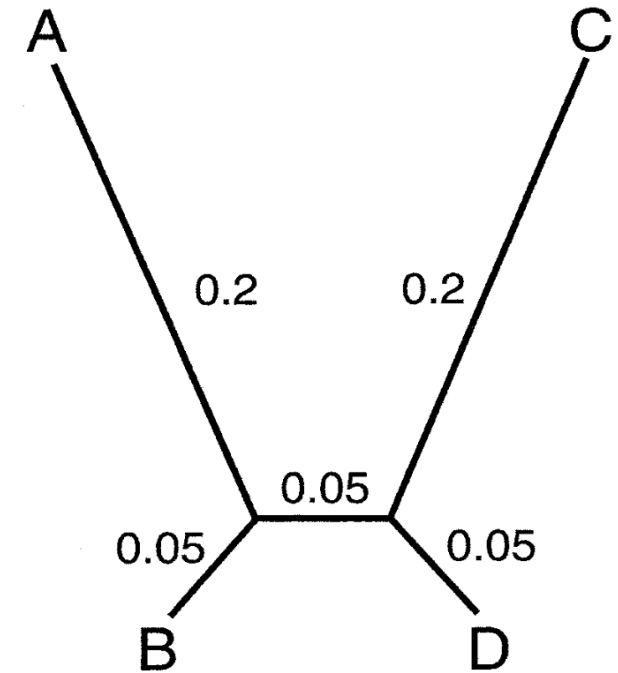$$P(D \mid T, V) = \frac{Pr(D, V \mid T)}{Pr(V \mid T)}$$

Corrects for the acquisition / ascertainment bias in your data

Lewis 2001

MkV model

	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

Rates of morphological evolution

Conodonts												
taxa 1	0	1	0	1	2	1	0	1	0	1	2	1
taxa 2	1	2	1	0	1	0	0	0	1	0	0	1
taxa 3	0	0	1	0	0	1	1	1	0	1	0	1
taxa 4	1	1	0	1	0	1	0	1	0	1	2	1

Unlikely that all characters evolve under same process

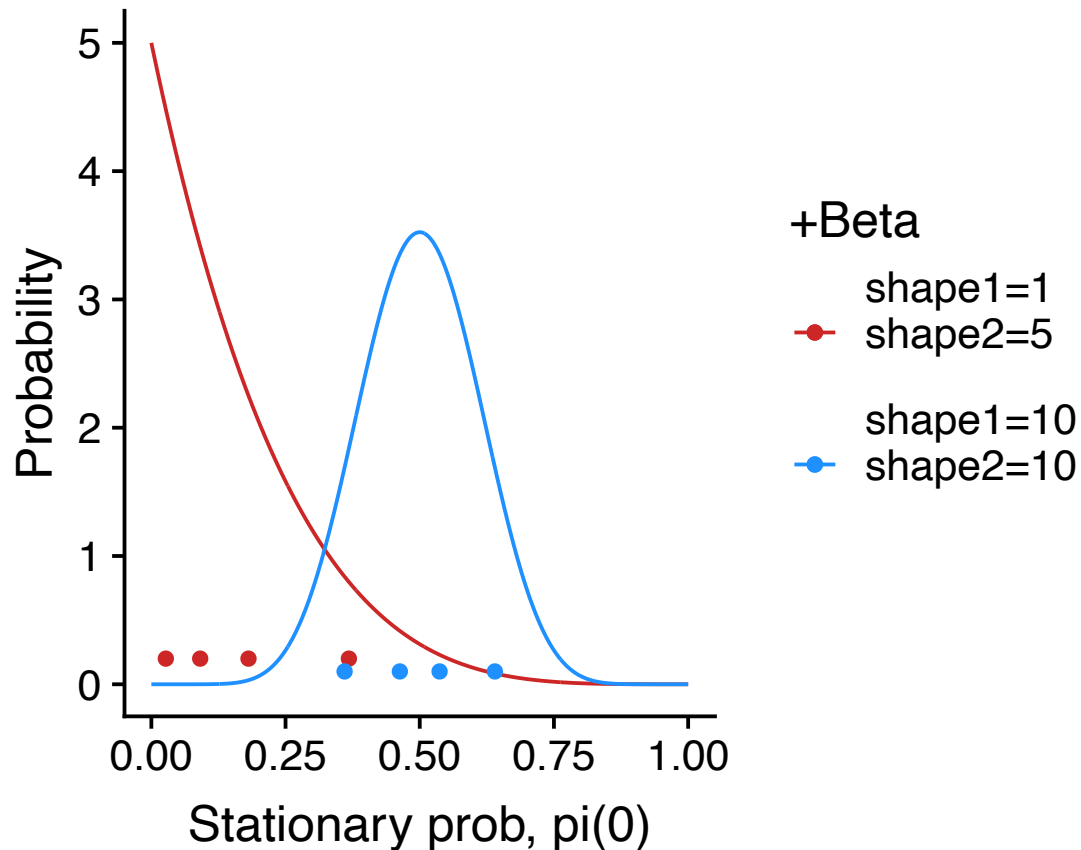# Across site rate variation

+Gamma

- alpha=10
- alpha=2

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

Yang 1994

Allowing heterogeneity in character change symmetry

Image from Michael Landis

This model allows for unequal transitions between states.

Relaxing the assumption that the probability of changing from 0 to 1 is the same as 1 to 0

Uses state frequencies for transition probabilities – i.e.It may be very likely, in a character, to change from 0 to 1. But if the frequency of 0 is very low, we will still seldom see this change.

Wright et al 2015

Partitioning the data

Partitioned analyses allow for different sets of homologous sites to evolve according to different sets of evolutionary parameters

Character sets might share

number of states

ancestral/derived polarity

structure/order

variability

ontogeny

function

model fit

Conodonts											
taxa 1	0	1	0	1	2	1	0	1	0	1	2 1
taxa 2	1	2	1	0	1	0	0	0	1	0	0 1
taxa 3	0	0	1	0	0	1	1	1	0	1	0 1
taxa 4	1	1	0	1	0	1	0	1	0	1	2 1

Partitioning Morphological data sets

Partitioning Morphological data sets
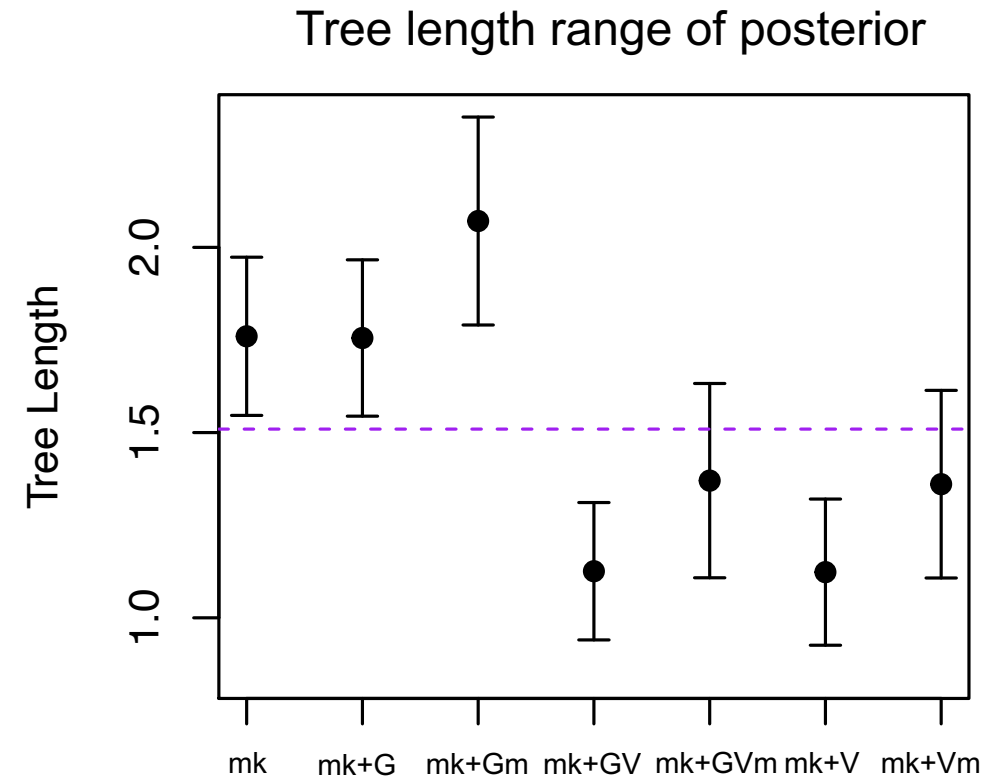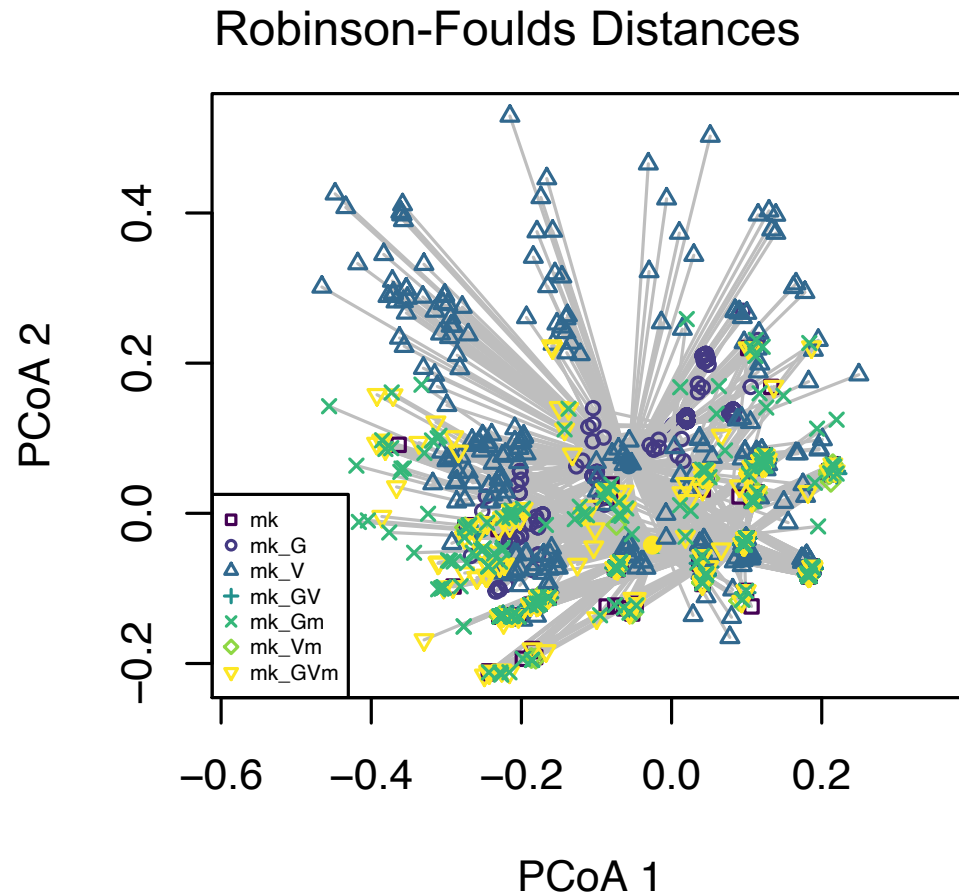
Does changing the substitution model really matter for empirical data?

Impacts of substitution model on inferred parameters

Robinson-Foulds Distances

2_Demar
Tree length range of posterior

(Mulvey et al in prep.)

Exercise 1

Run an MCMC inference using **two** "versions" of the Mk model