Phylogenetics

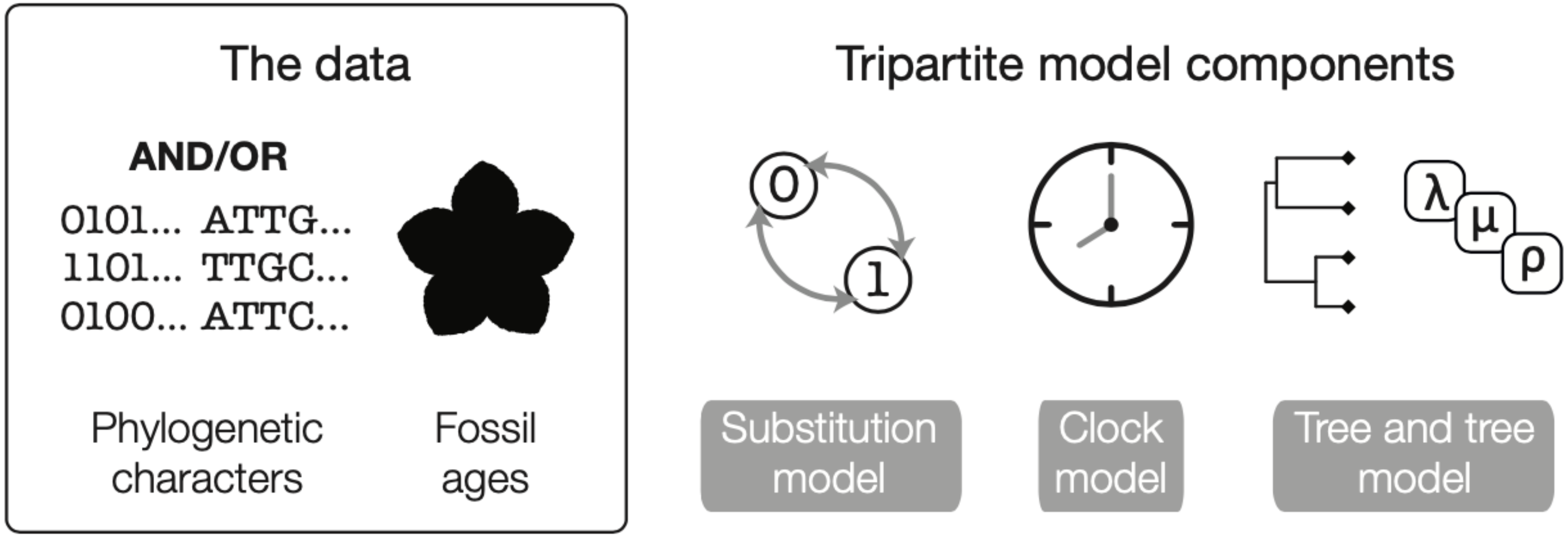
Morphological Substitution models

Rachel Warnock, Tim Brandler, Laura Mulvey

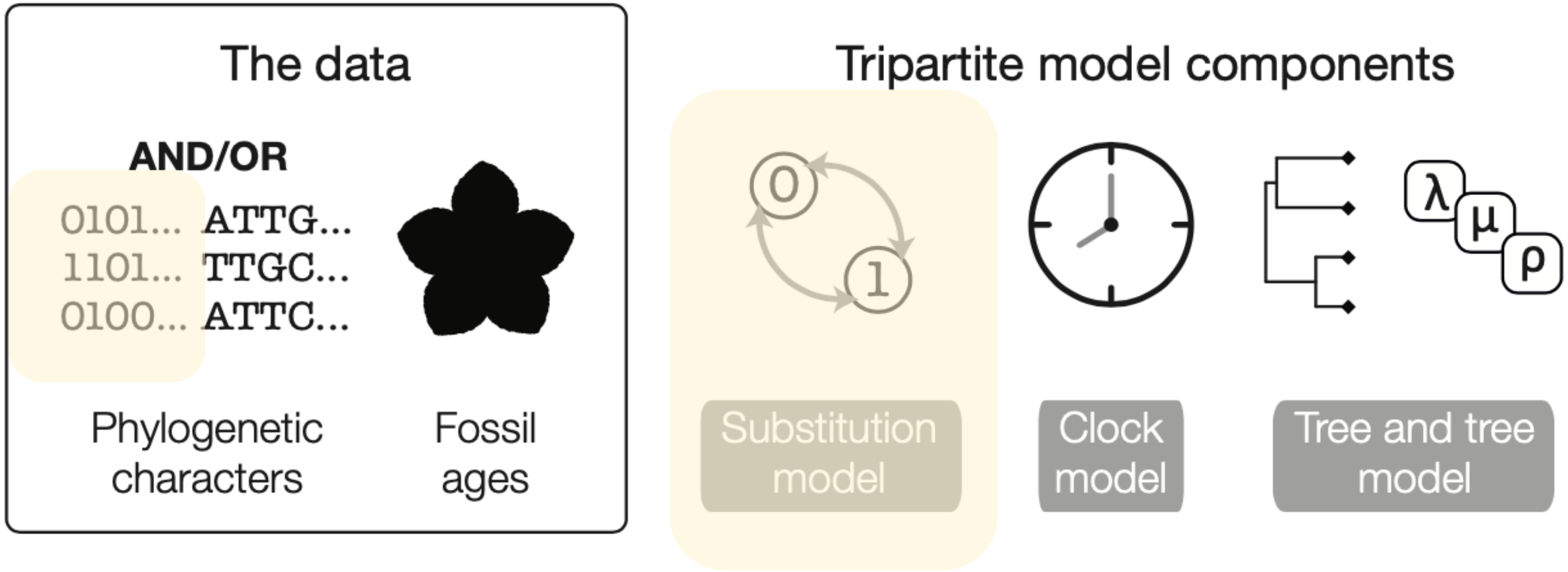
laura.l.mulvey@fau.de

June 13 2023

Bayesian Phylogenetic Analysis Components from last week



Bayesian Phylogenetic Analysis Components from last week



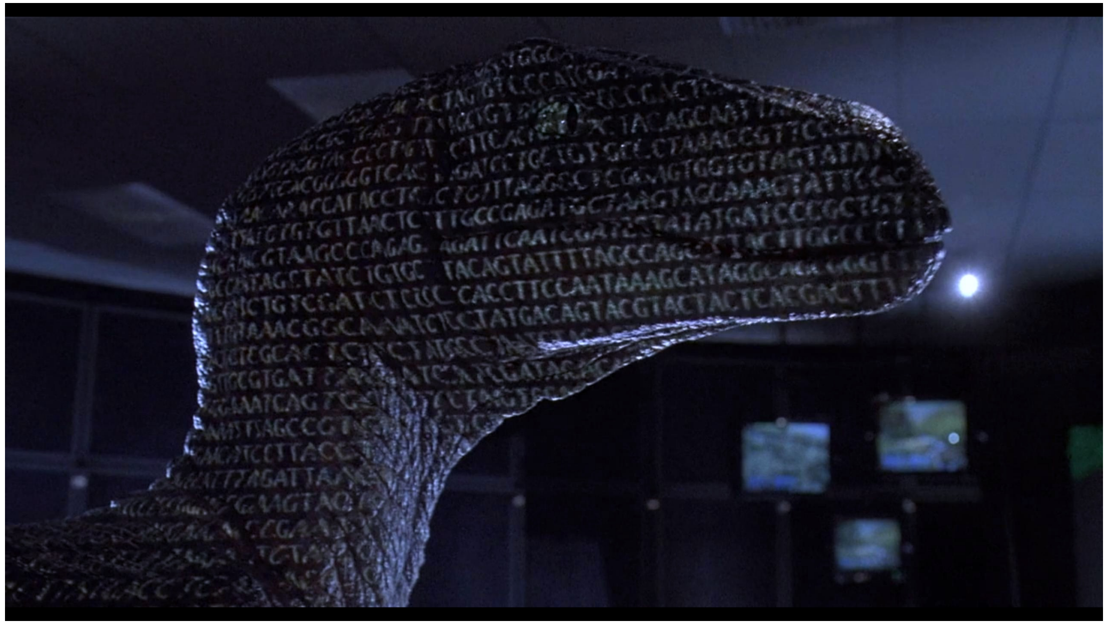
Molecular Substitution models

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

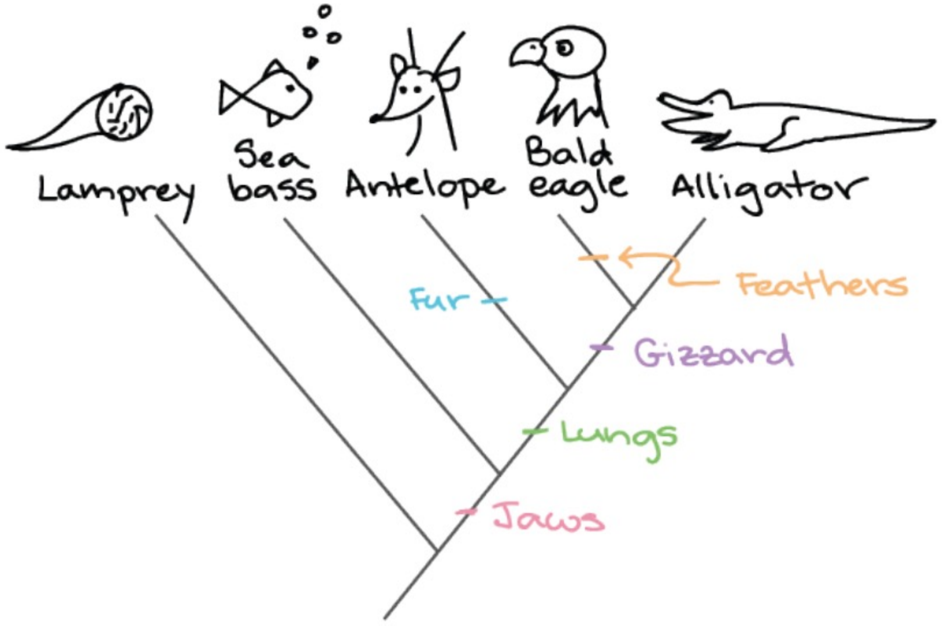
$$Q = \begin{pmatrix} * & \mu_{AG}\pi_G & \mu_{AC}\pi_C & \mu_{AT}\pi_T \\ \mu_{GA}\pi_A & * & \mu_{GC}\pi_C & \mu_{GT}\pi_T \\ \mu_{CA}\pi_A & \mu_{CG}\pi_G & * & \mu_{CT}\pi_T \\ \mu_{TA}\pi_A & \mu_{TG}\pi_G & \mu_{TC}\pi_C & * \end{pmatrix}$$



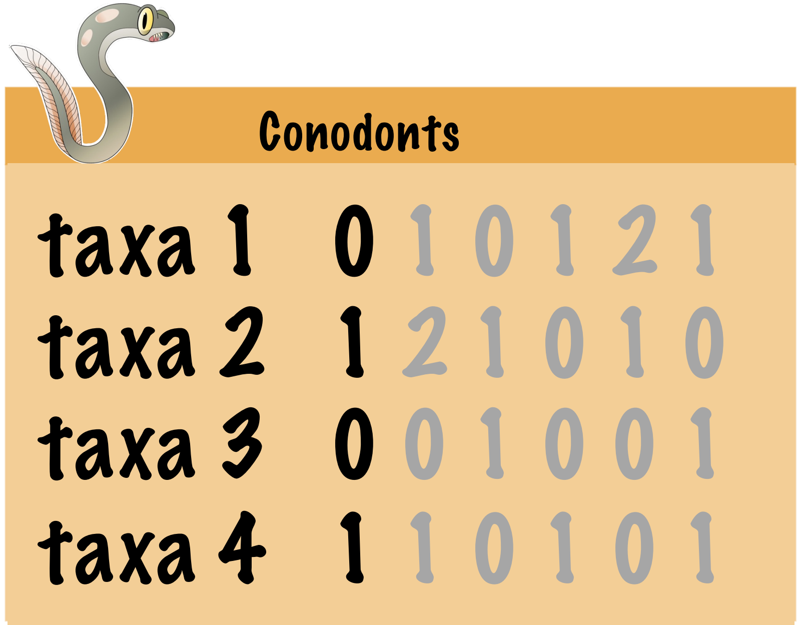
µ = substitution rate
Π = stationary frequency

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



Issues with Morphological data

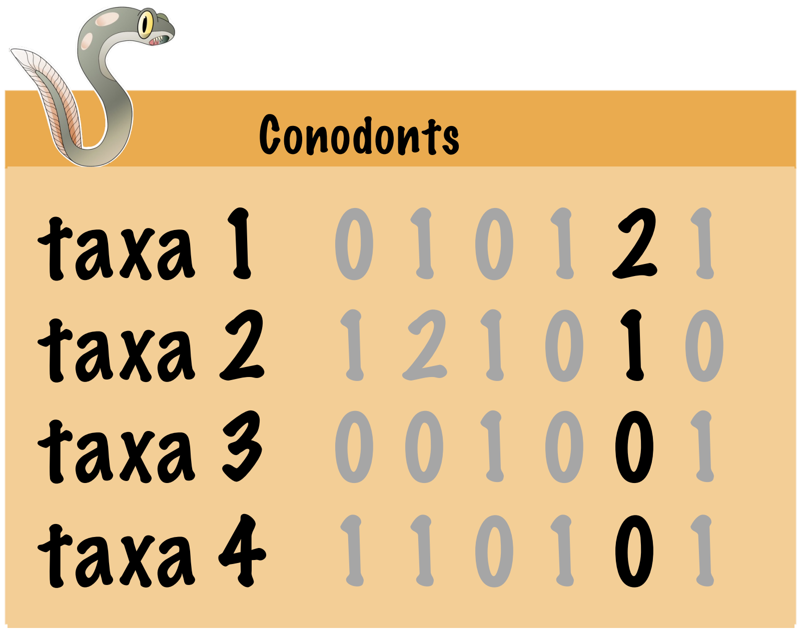


Conodonts

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

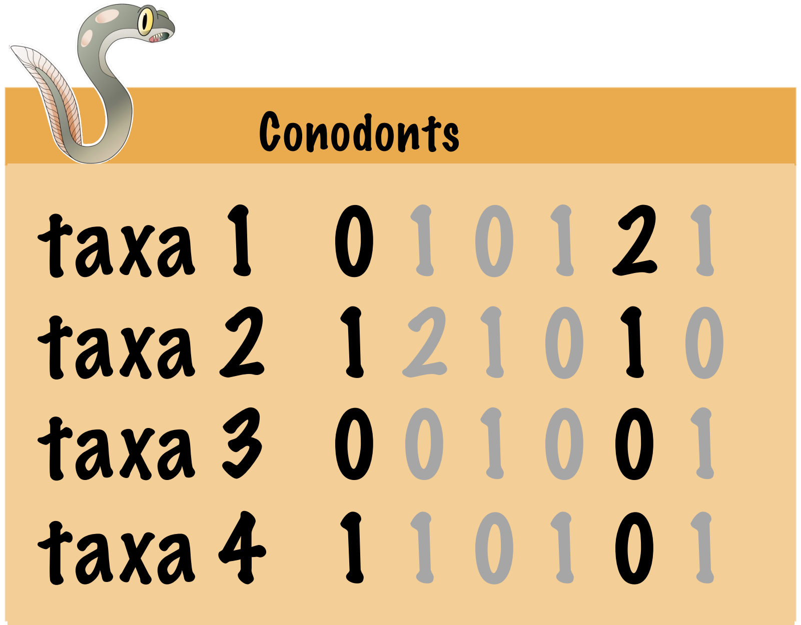
Issues with Morphological data



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

Issues 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

Trait 1		Trait 2
0	≠	0
1	≠	1

Generalising morphological data is much more difficult that molecular

Differences between molecular and morphological data to consider when modelling

Molecular data has a similar biological meaning throughout the alignment.

A T in one part of the alignment represents the same biological unit as a T somewhere else in the alignment.

This is not the same for morphological data.

Becomes more **difficult to generalise** morphological data in any biologically meaningful way

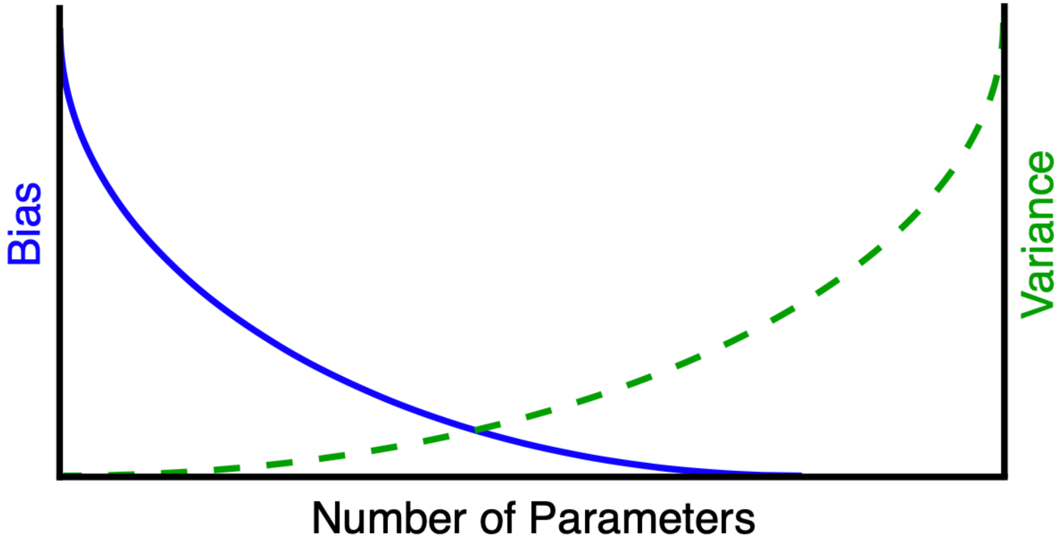
What does a good model look like?

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



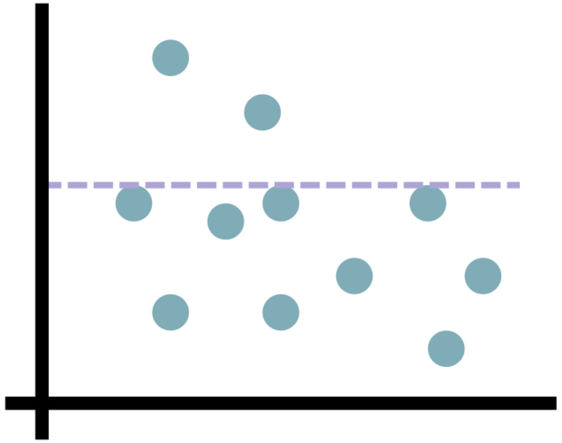
Too complicated, not enough information

What does a good model look like?

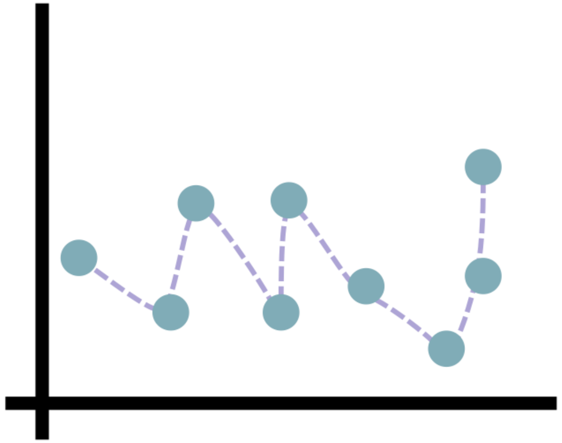
To do statistical inference we need a model

What model should that be?

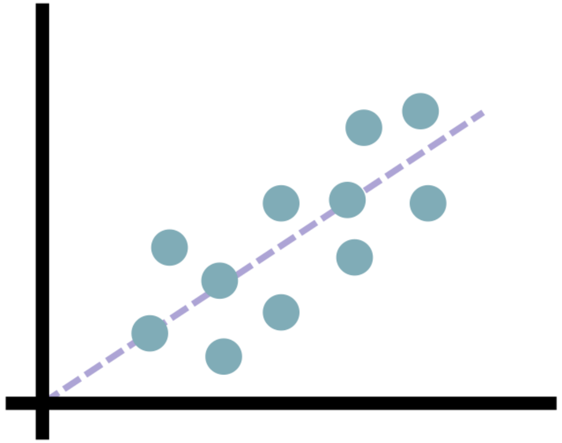
Our goal should be to have a model that is **complex enough** to capture the “important” variation in the data, but **not be more complex** than it needs to be



Underfitting



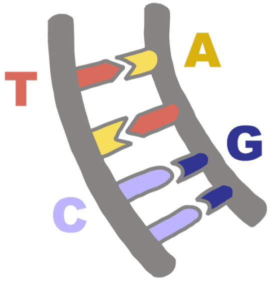
Overfitting

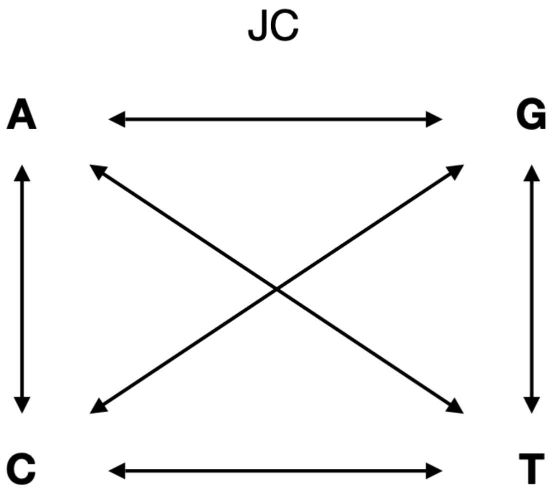


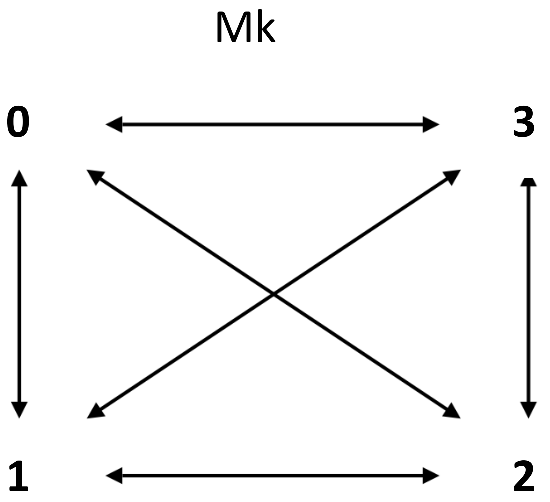
Proper fit

What assumptions might you want to incorporate into a model of morphological character evolution?

Substitution models for morphological data



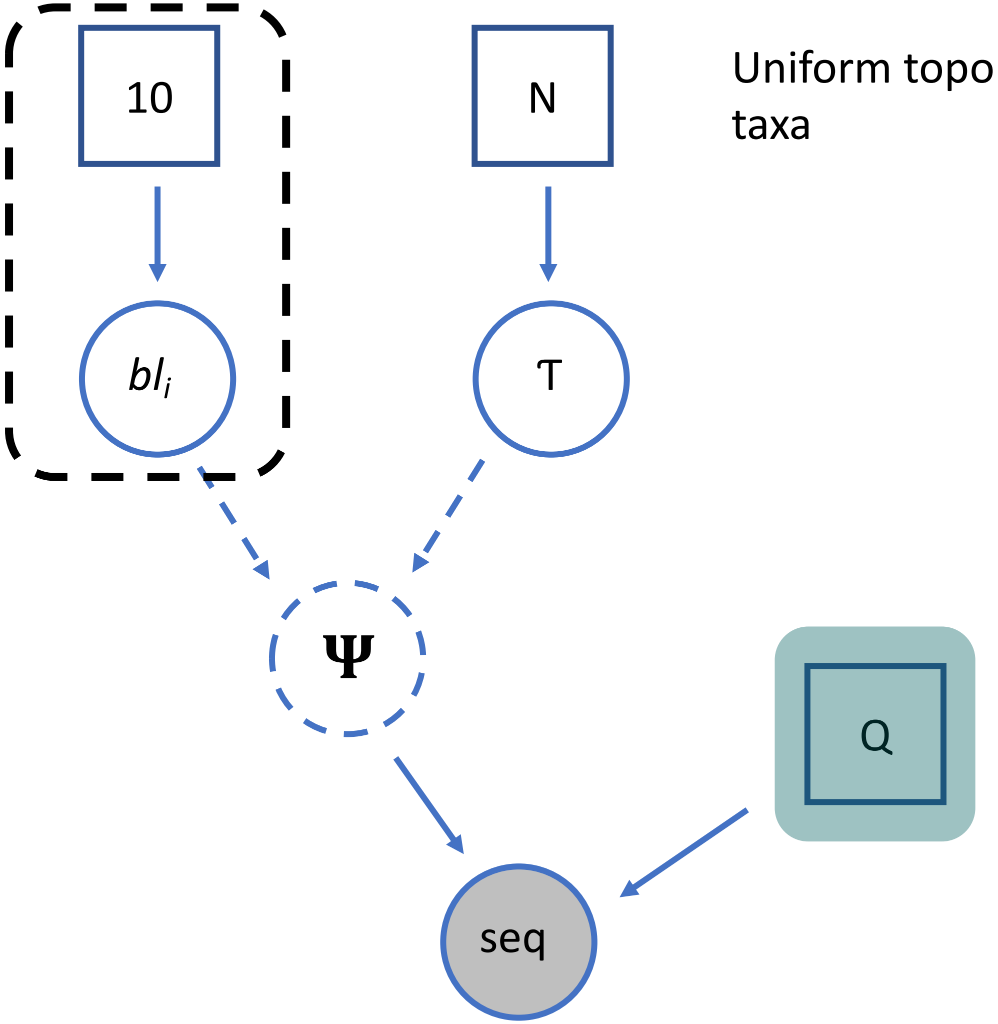




*4 state here as an example, can be any number from 2!

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

Mk Model



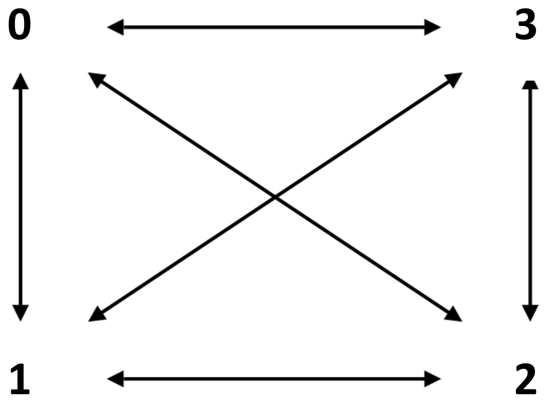
Exponential rate parameter of 10 on branch lengths.

Uniform topology of N taxa

Mk model

Substitution models for morphological data

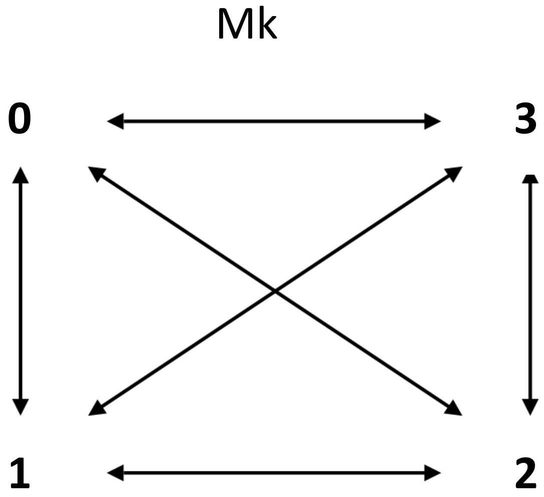
Mk



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

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

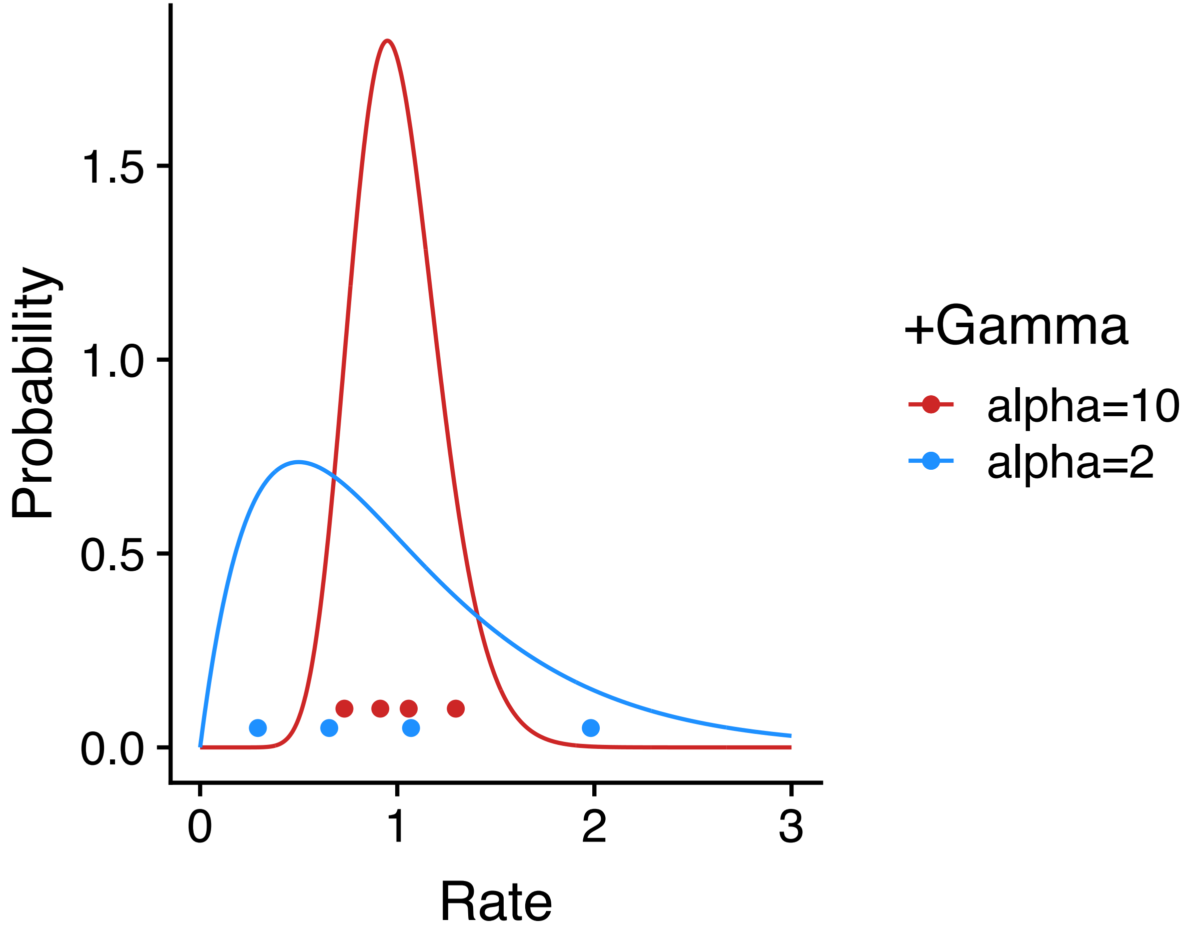
Substitution models for morphological data



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

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

Across Site Rate Variation (+G)

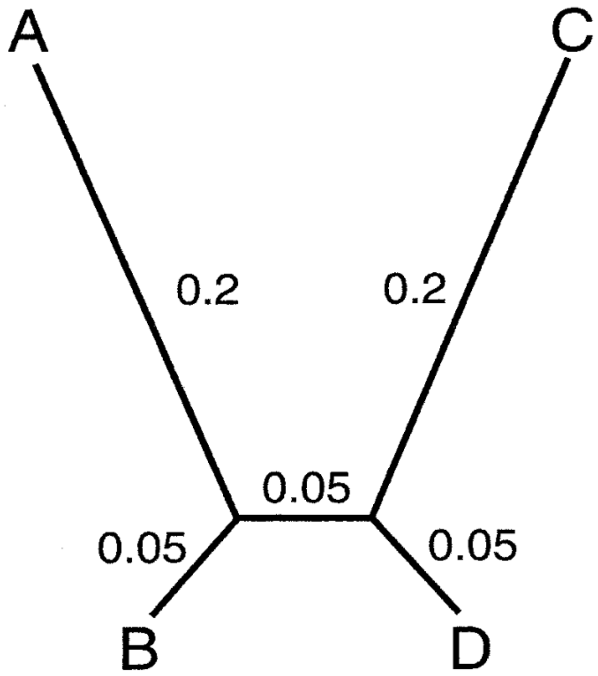


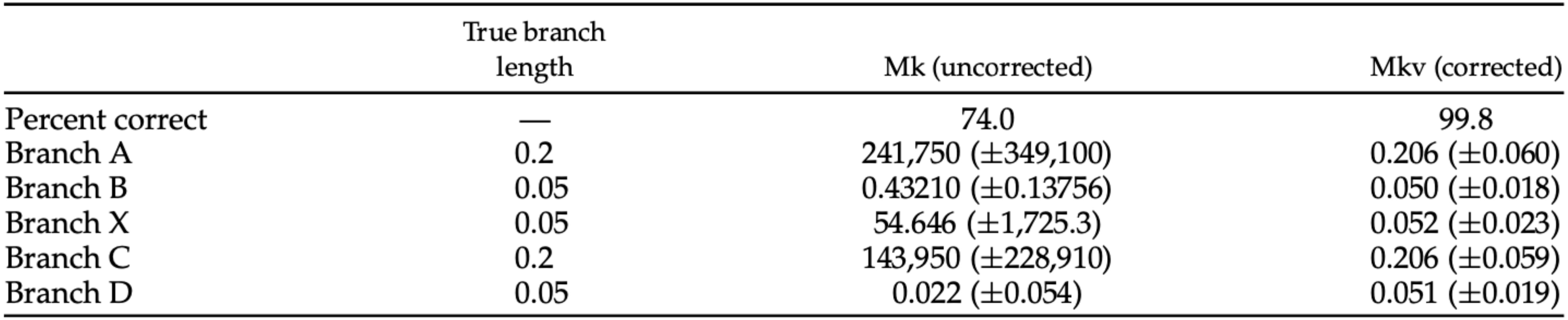
alpha = 10, the rates are similar
alpha = 2 the rates differ

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



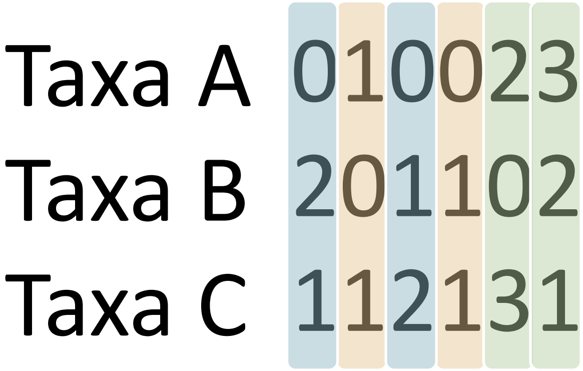


	True branch length	Mk (uncorrected)	Mkv (corrected)
Percent correct	—	74.0	99.8
Branch A	0.2	241,750 (±349,100)	0.206 (±0.060)
Branch B	0.05	0.43210 (±0.13756)	0.050 (±0.018)
Branch X	0.05	54.646 (±1,725.3)	0.052 (±0.023)
Branch C	0.2	143,950 (±228,910)	0.206 (±0.059)
Branch D	0.05	0.022 (±0.054)	0.051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

Lewis 2001

Partitioning the data

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

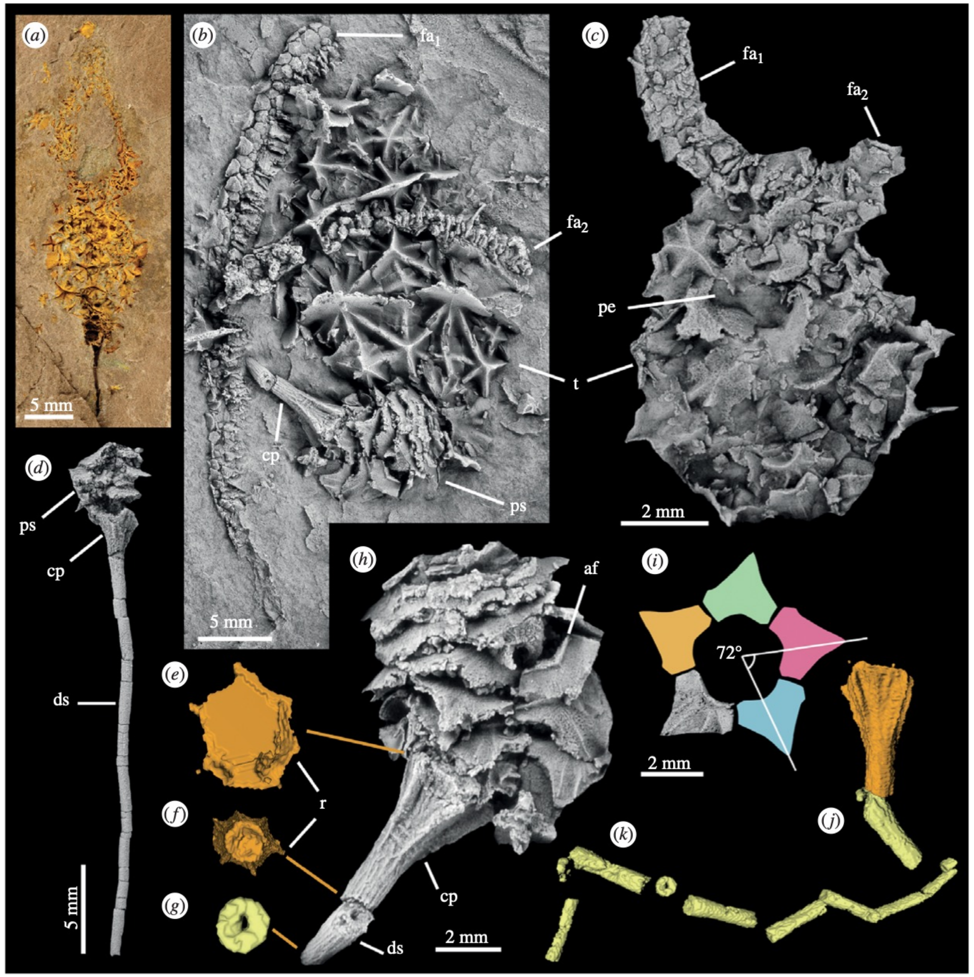


Taxa A 010023

Taxa B 201102

Taxa C 112131

```
001510010?00-100--0000000000
000500010?200100--0010010000
002500010?200100--0?10010000
00?5?0010?200100?-0???010110
0015000101201000430100011111
0015000101201010440111011111
??050?????201000440?11011111
01050?010-210000?501??010110
0002000100210100 3-1110010110
0002000100211001441121011111
000201111-210010?-??11011121
?103?0?11?1001104-0000010000
1005002110100010--0?00110?20
1005002000101010540?00110020
```

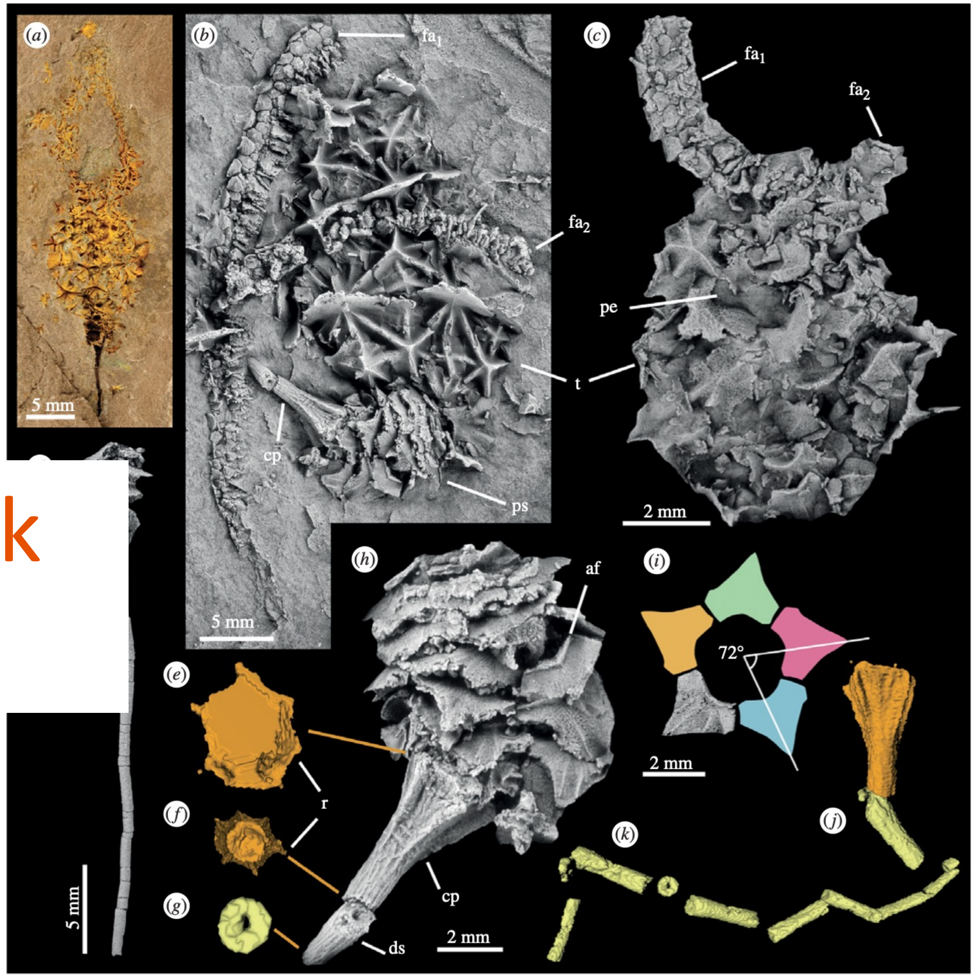


Cambrian stalked echinoderms show
unexpected plasticity of arm construction
Zamora & Smith. 2012 Proc B

```
0015100 10?00-100--0000000000
```

Can you draw the Q-matrix for an Mk model for this data set?

```
01050?010-210000?501??010110
0002000100210100 3-1110010110
0002000100211001441121011111
00020111 1-210010?-??11011121
?103?0?11?1001104-0000010000
1005002110100010--0?00110?20
1005002000101010540?00110020
```



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

Exercise