Phylogenetics

Introduction to statistical phylogenetics

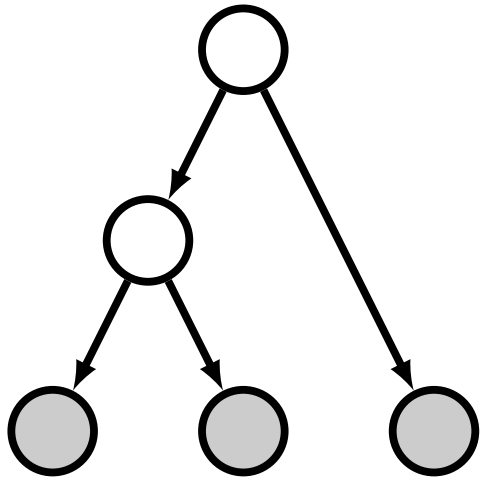
Rachel Warnock, Laura Mulvey
rachel.warnock@fau.de, laura.l.mulvey@fau.de

September 5, 2022

Analytical Paleobiology Workshop, Erlangen 2022

Part 2 objectives

- revisiting the definition of model
- maximum likelihood (briefly)
- substitution models



Caveat: the following is just my take on things!

What is a statistical model? When is an equation a model?

What is a mechanistic model?

What is the difference between an algorithm and a model?

A statistical model is a type of model that includes a set of assumptions about the data-generating process.

A statistical model is a type of model that includes a set of assumptions about the data-generating process.

It should be possible to simulate data under the assumptions of the model.

A statistical model is a type of model that includes a set of assumptions about the data-generating process.

It should be possible to simulate data under the assumptions of the model.

If we're lucky, we might also be able to estimate parameters under the model*. This isn't always possible because some models are too complex.

*A fancy way of saying this is "we can perform inference under the model".

An example

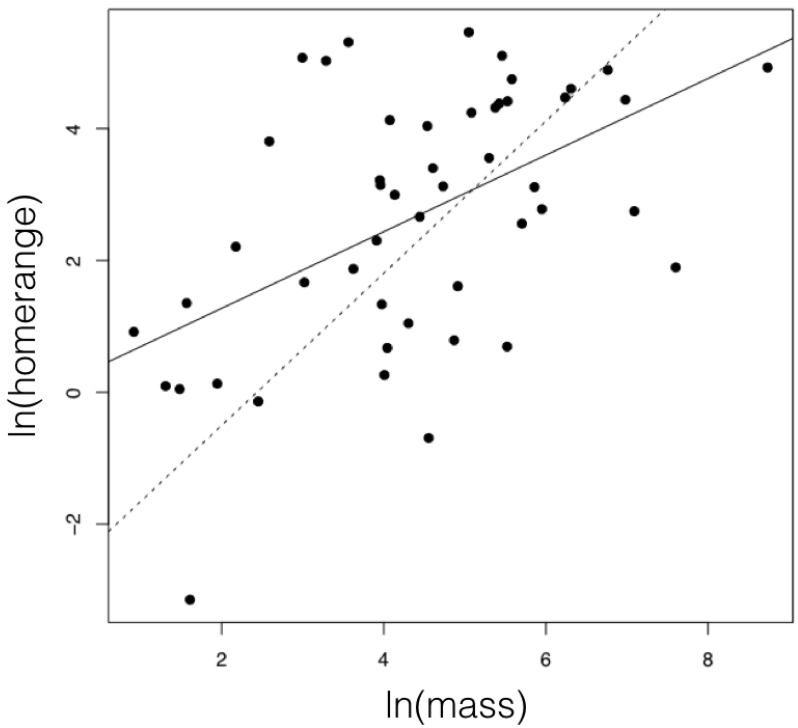


Image source: Harmon (2019)

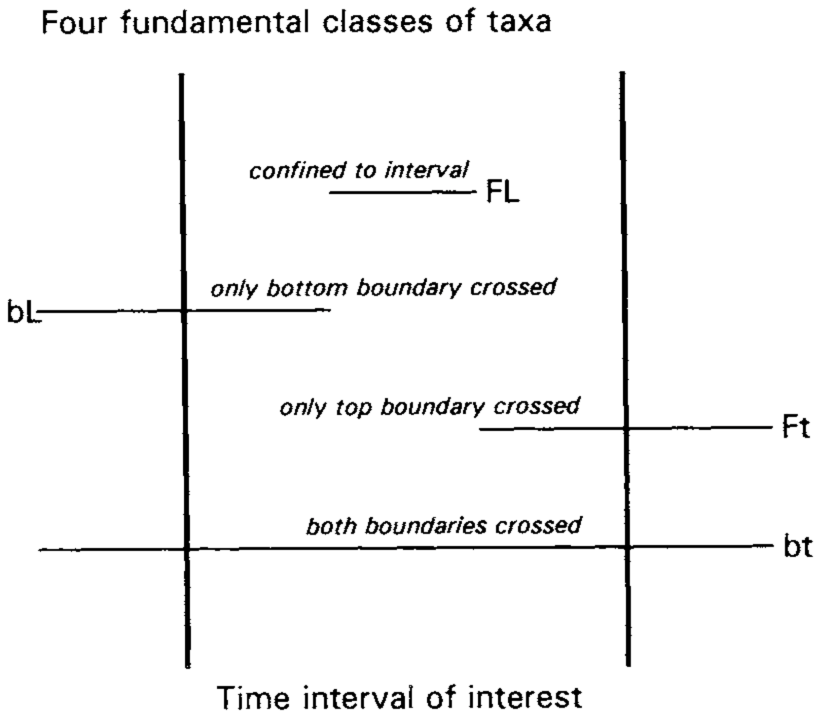
The solid black line is a linear regression line.

We can estimate the parameters of the regression model.

$$y = X\beta + \epsilon$$

It's also straightforward to simulate data under this model.

Some approaches that are not model based



Foote (2000)

The boundary-crosser and three-timer metrics are not models.

They provide a clever way of approximating origination and extinction rates (and often perform well), but they don't describe the data generating processes.

Mechanistic or process based models *are based on "physical principles". They describe the data as a function of a set of parameters that have a tangible biological meaning.*

Mechanistic or process based models *are based on "physical principles". They describe the data as a function of a set of parameters that have a tangible biological meaning.*

A regression model is not mechanistic – it describes the relationship between x and y but the parameters don't have a biological meaning.

Mechanistic or process based models *are based on "physical principles". They describe the data as a function of a set of parameters that have a tangible biological meaning.*

A regression model is not mechanistic – it describes the relationship between x and y but the parameters don't have a biological meaning.

Many of the models we use in phylogenetics are mechanistic models, e.g. they might include origination, extinction and sampling parameters explicitly.

Mechanistic or process based models *are based on "physical principles". They describe the data as a function of a set of parameters that have a tangible biological meaning.*

A regression model is not mechanistic – it describes the relationship between x and y but the parameters don't have a biological meaning.

Many of the models we use in phylogenetics are mechanistic models, e.g. they might include origination, extinction and sampling parameters explicitly.

Note the definition of different model types varies a lot. The above is just my take on things from a very phylogenetics perspective.

An algorithm is a precise rule (or set of rules) specifying how to solve some problem.

```
i = 1
while  i < 11:
    print(i)
    i = i + 1
```

```
for i in range(1,11):
  print(i)
```

Algorithms are used in phylogenetics for all sorts of tasks, inc. searching tree space or traversing trees.

A brief introduction to maximum likelihood

Recap — How do we find the "best" tree?

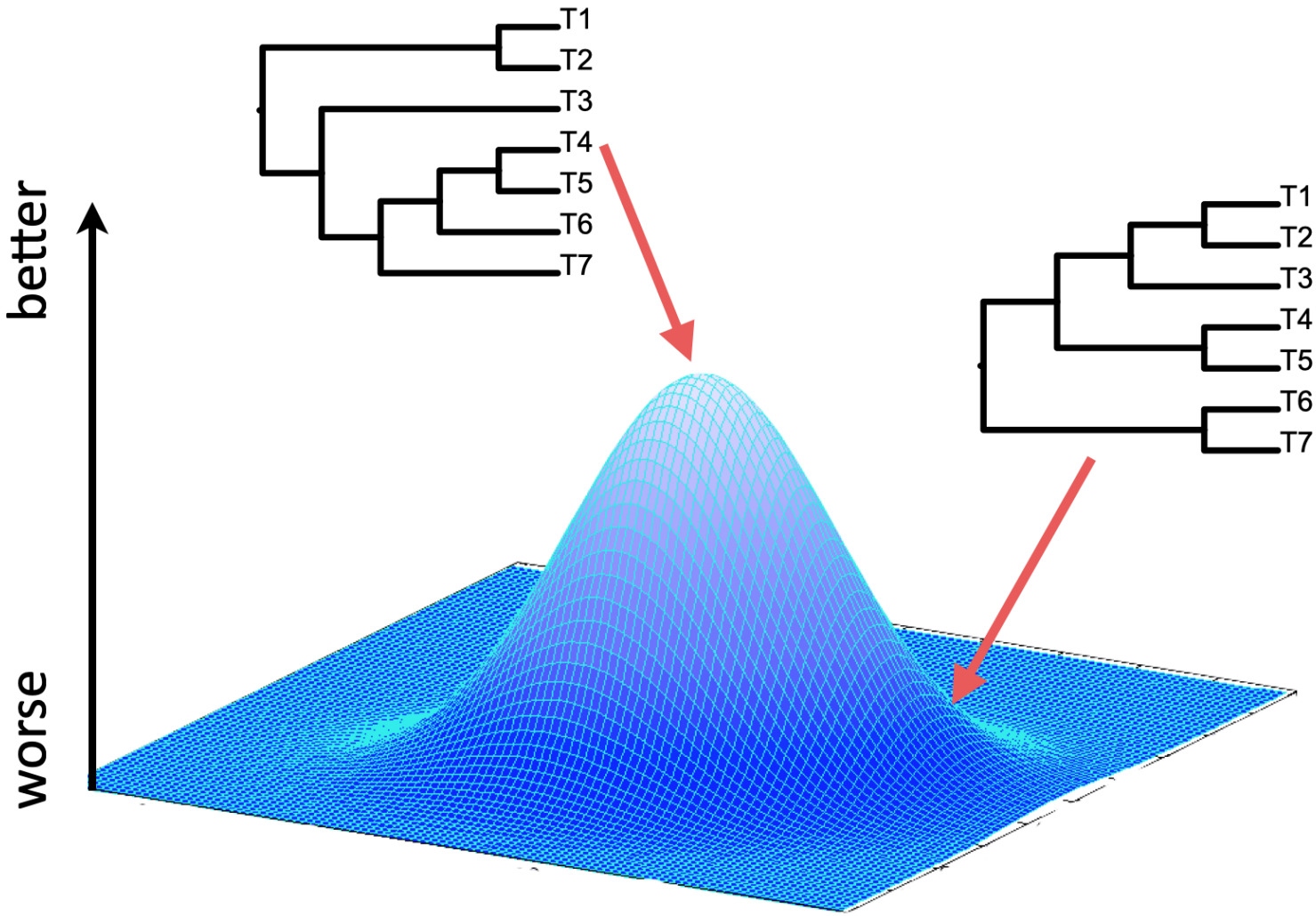


Image source: Tracy Heath

Model-based phylogenetics

Assume an explicit model of character evolution.

Maximum likelihood is a method for estimating unknown parameters in a model. The tree that maximises the likelihood is the best one.

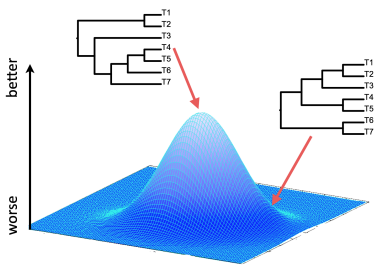
Probability (data | model, tree)

Maximum likelihood simplified

1. We first propose a topology with branch lengths and then calculate the likelihood (taking into account all sites).

2. We then propose a new tree or set of branch lengths and recalculate the likelihood. If the likelihood is $>$, we accept this tree as being better.

3. The algorithm proceeds until we can't improve the likelihood any further.



Phylogenetic character data

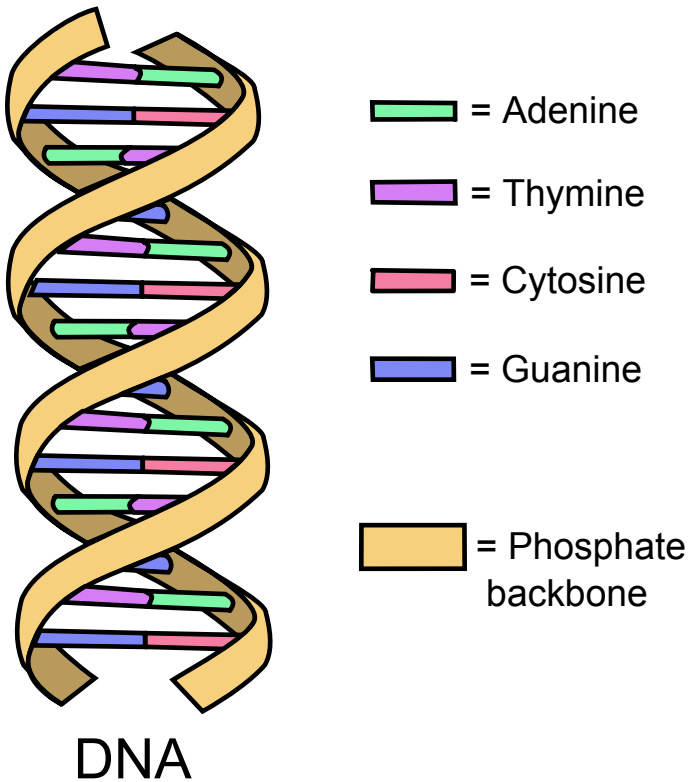
Two main sources of data for building trees:

1. Molecular sequences (nucleotides or proteins)

2. Morphological characters (discrete or continuous)

First we need to collect the data and establish homology.

Molecular sequence data



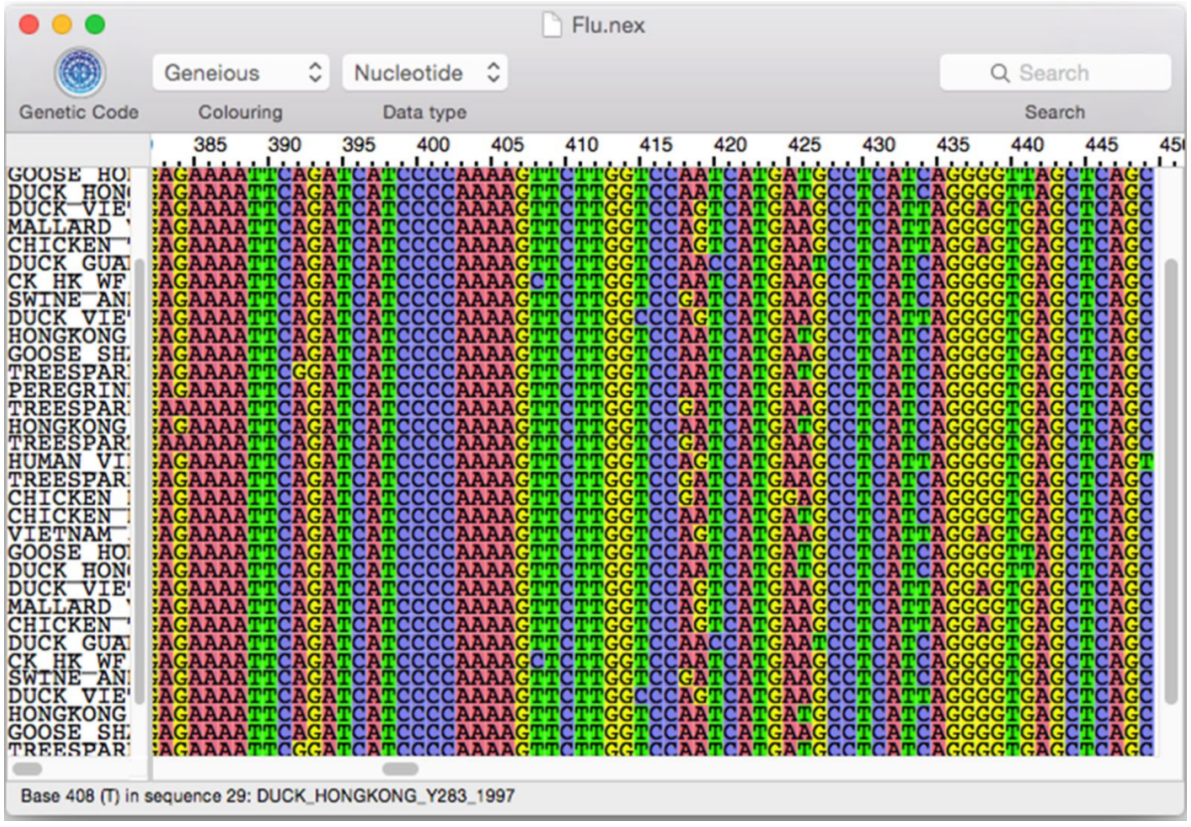
DNA

Nucleotides provide a four letter alphabet we can use to generate trees.

Genes encode amino acids (proteins) that in turn provide a 20 letter alphabet.

Protein sequences are typically used for more distant evolutionary relationships.

Multiple sequence alignments are the primary input for molecular phylogenetic analysis



Models of molecular sequence evolution

Also known as substitution / site / character models.

They capture the process of character evolution.

Allow us to ask, what is the probability of transitioning from one state to another over time?

Thinking about the snacks (or your favourite group of species), what assumptions would you want to incorporate into a model of character evolution?

Models of nucleotide evolution: rate matrix

Using the model we can calculate the probability of transitioning between different nucleotides.

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

Probability of changing between two states over the branch lengths.

μ is the substitution rate.

The longer the interval of time has past, the more likely we are to observe a change.

You can explore this principle via this app by Paul Lewis.

The Jukes-Cantor model of sequence evolution

This is the simplest model of sequence evolution.

Assumptions: equal mutation rates and base frequencies.

Base frequencies are the proportion of each nucleotide within the dataset.

$$Q = \begin{pmatrix} * & \mu & \mu & \mu \\ \mu & * & \mu & \mu \\ \mu & \mu & * & \mu \\ \mu & \mu & \mu & * \end{pmatrix}$$

The GTR model of sequence evolution

Nucleotides (ATCG) occur at different frequencies depending on the group of species or gene.
If a given nucleotide appears in our dataset at a low frequency, we are less likely to observe a transition to that state.

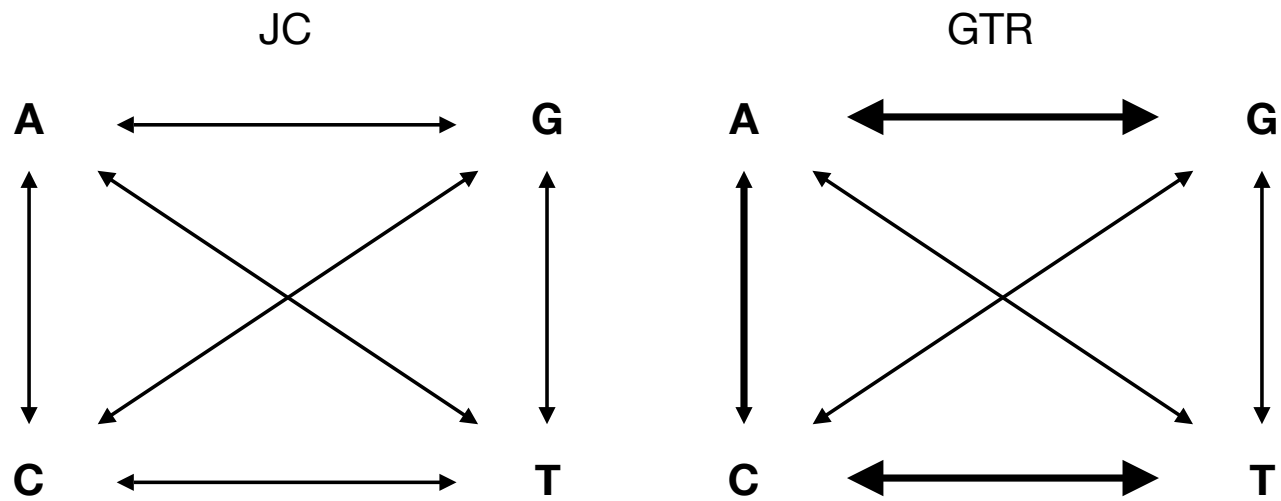
GTR assumptions: unequal mutation rates AND unequal base frequencies.

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

Note the rates are symmetric – e.g., the rate of change between A and T, is the same in both directions – but the proportion of each character state also affects the probability of change.

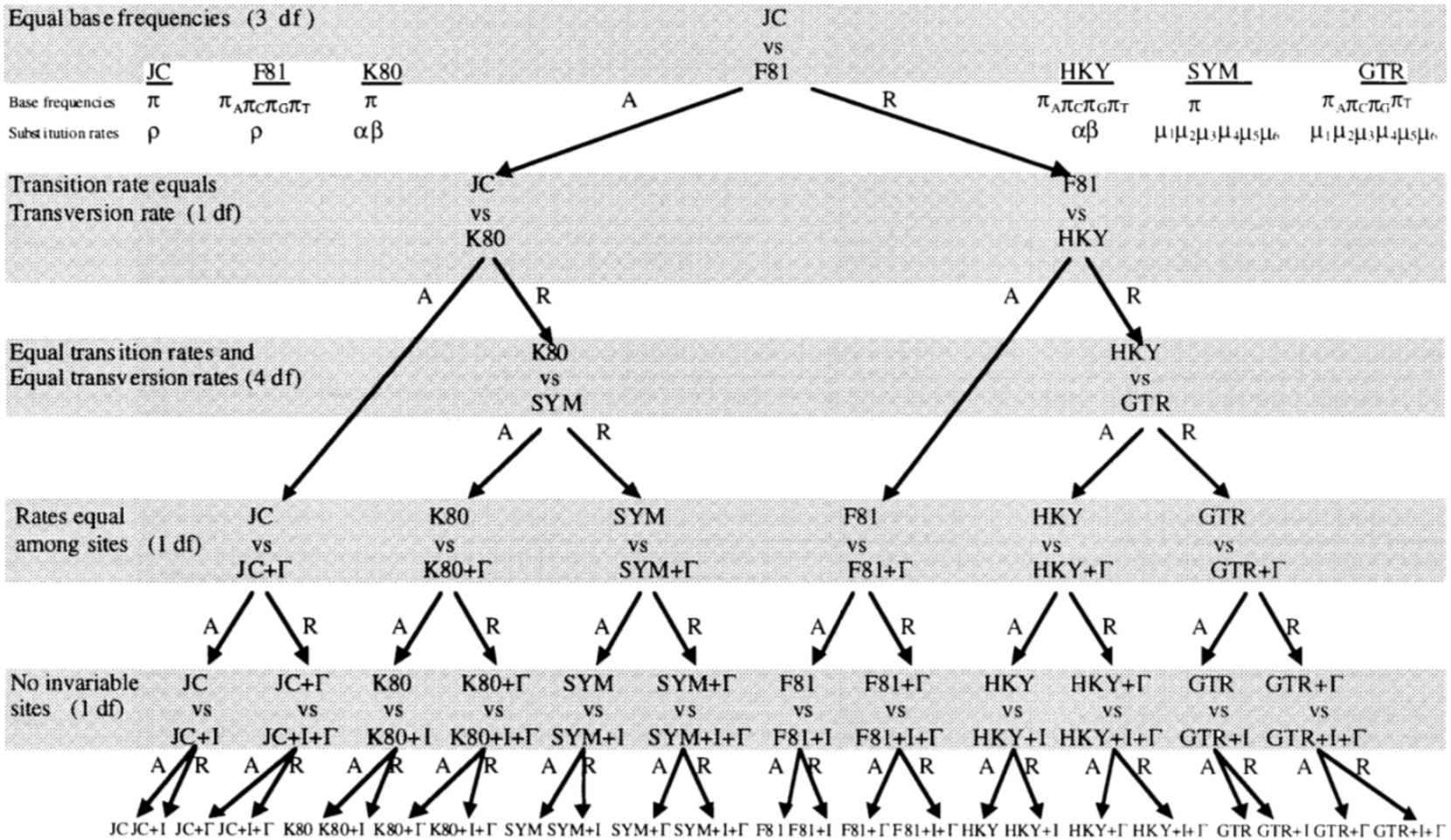
The JC versus GTR models

Another way of visualising substitution models.



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

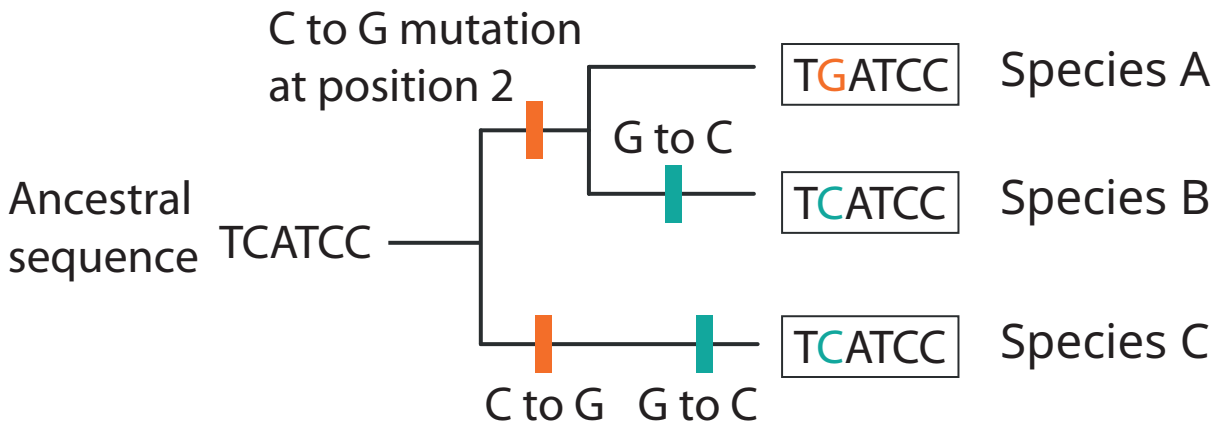
JC & GTR belong to a large family of substitution models



Posada & Crandall (1998) *Bioinformatics*

Model-based phylogenetics

Models can account for multiple changes at the same site.



Branch lengths = *expected number of changes per site*

Something to bare in mind

In the absence of any information about time, rates are *relative*, i.e. rates are expected substitutions per site, independent of any time unit.

Model-based methods: advantages and disadvantages

Statistically more sound

Can test and update explicit assumptions

Model-based methods: advantages and disadvantages

Statistically more sound

Can test and update explicit assumptions

Computationally slow (often)

Results are sensitive to model choice

Model-based methods: advantages and disadvantages

Statistically more sound

Can test and update explicit assumptions

Computationally slow (often)

Results are sensitive to model choice

There are many more things we can do with models in palaeobiology!

Yang (2014) Molecular Evolution: A Statistical Approach

Exercise