Phylogenetics

Substitution models and maximum likelihood

Rachel Warnock, Tim Brandler, Laura Mulvey
rachel.warnock@fau.de

Updated May 9, 2023

Next week May 16 there will be no class!

Tasks

We'd like you to watch 4 video lectures by Paul Lewis and answer a set of questions.

The videos are part of the phyloseminar series and provide a foundation for understanding statistical phylogenetics.

We'll go over the answers all together in week 5 (May 23).

Things to note

The goal is not to understand *everything*!

Read the questions before you begin watching each one.

It will take 3.5–4 hours to watch the videos, so don't leave it until the last minute!

A brief introduction to maximum likelihood

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

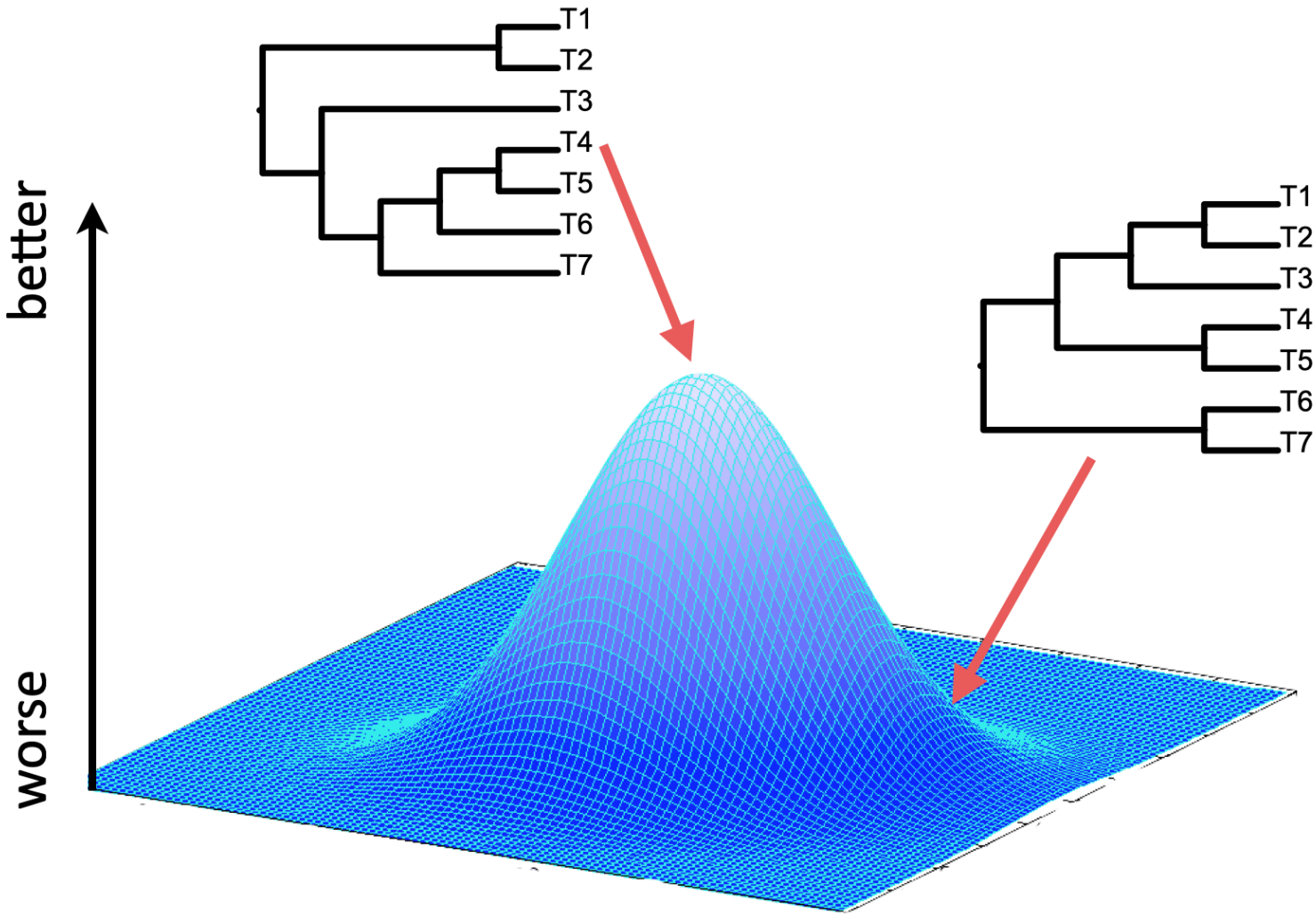


Image source: Tracy Heath

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.

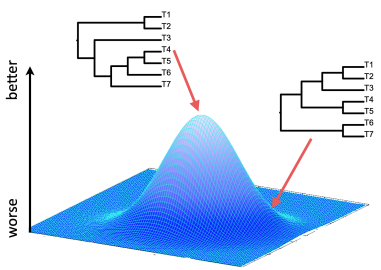
Probability (data | model, tree)

Maximum likelihood algorithm 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. Proceed until we can't improve the likelihood any further.



Phylogenetic character data

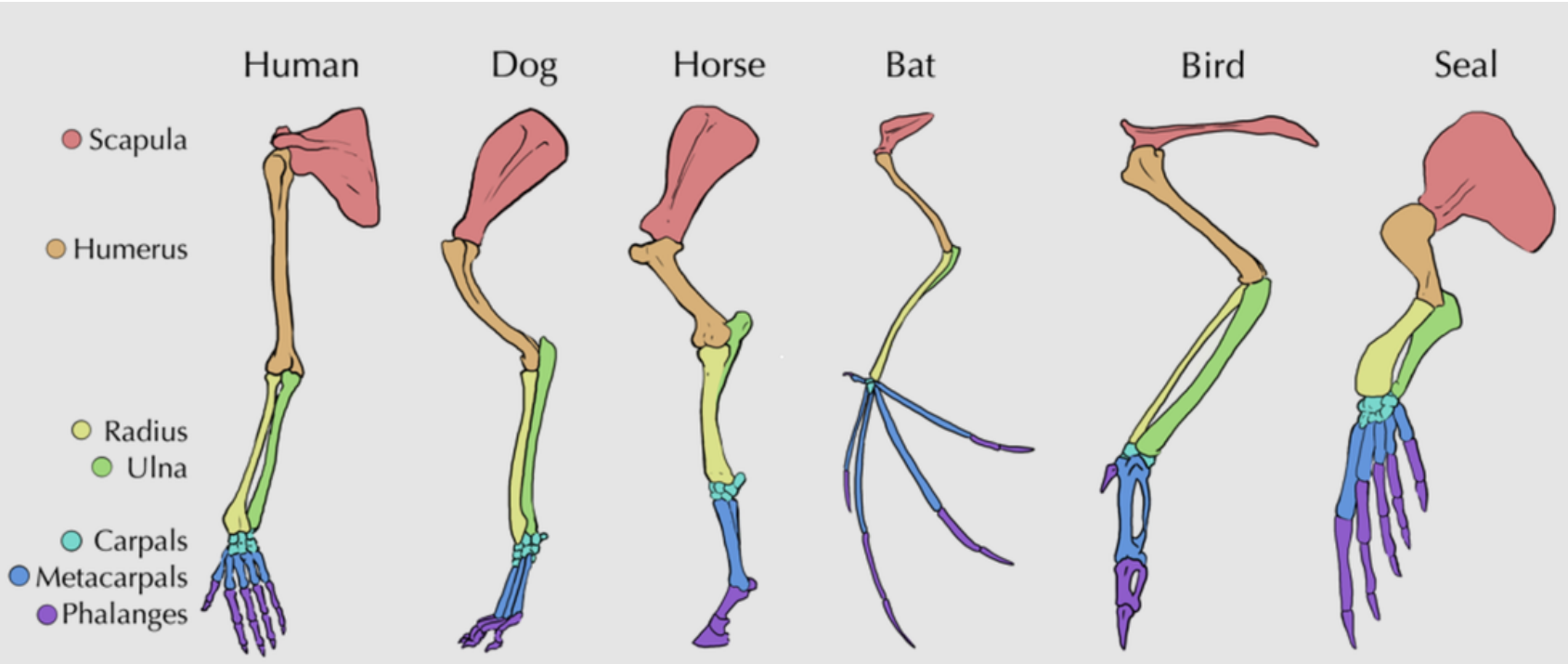
Two main sources of data for building trees:

1. Molecular sequences (nucleotides or proteins)

2. Morphological characters (discrete or continuous)

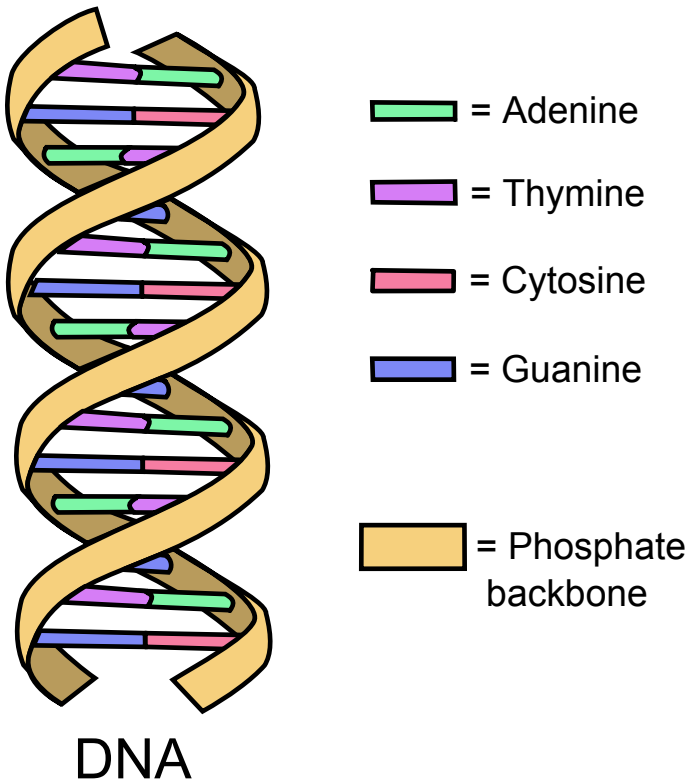
First we need to collect the data and establish homology.

Homology - similarity due to shared ancestry



Each coloured bone is a homologous structure.

Molecular sequence data



DNA

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

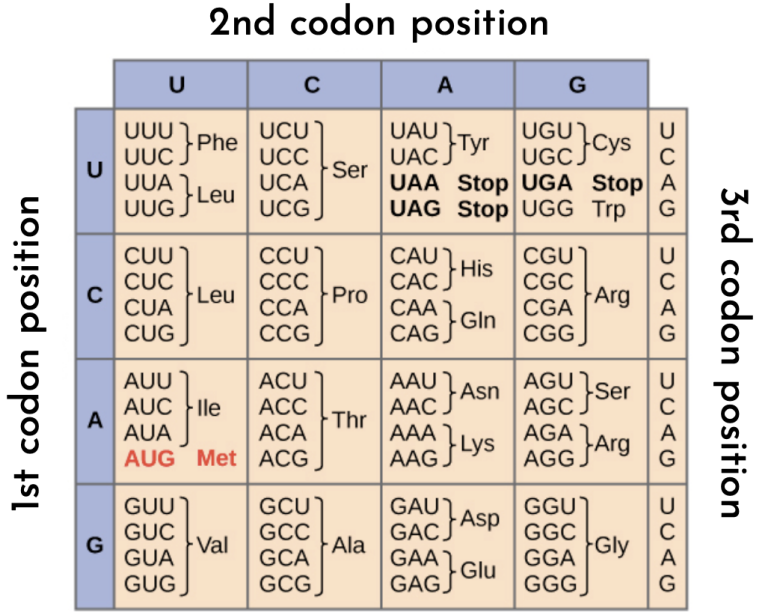
Genes encode amino acids (proteins) that in turn provide a 20 letter alphabet.

Protein sequences are typically used for more distant evolutionary relationships.

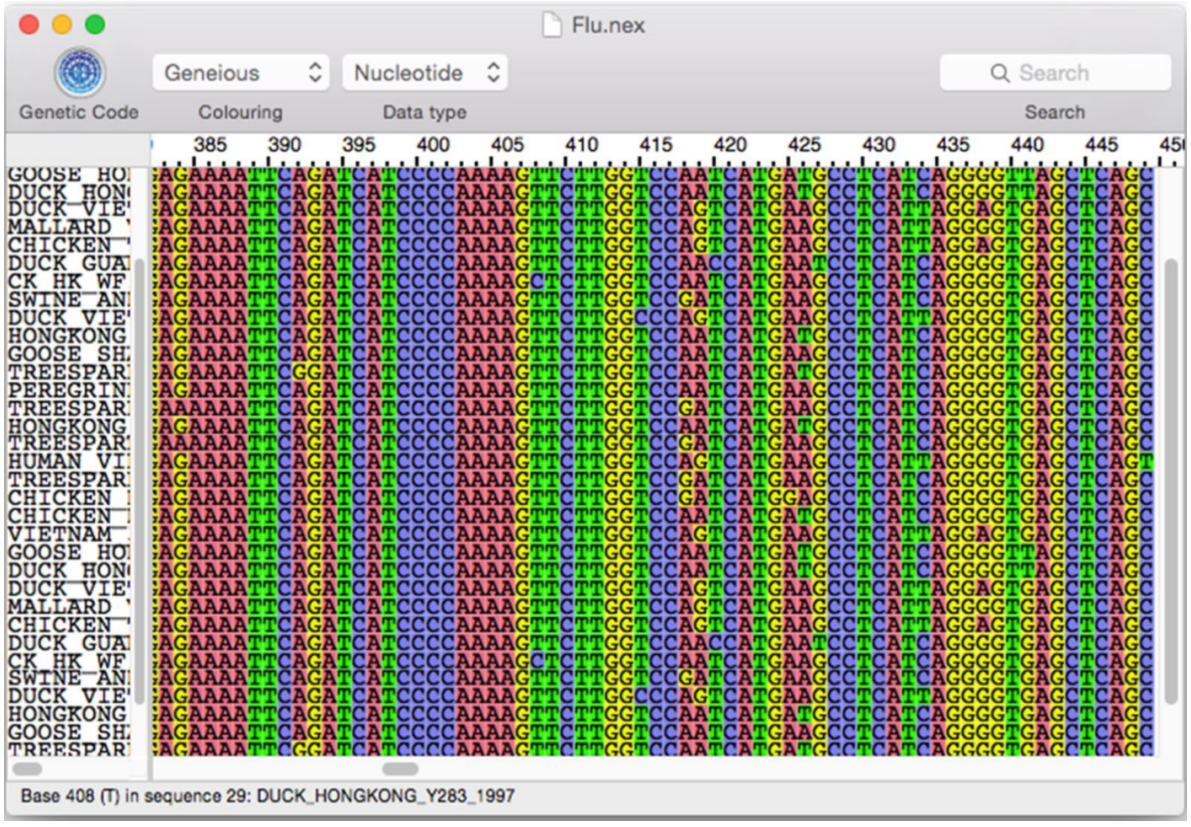
Check out the following resources to recap DNA, RNA and protein sequences and the Universal Genetic Code.

Genetic Code Quick Guide

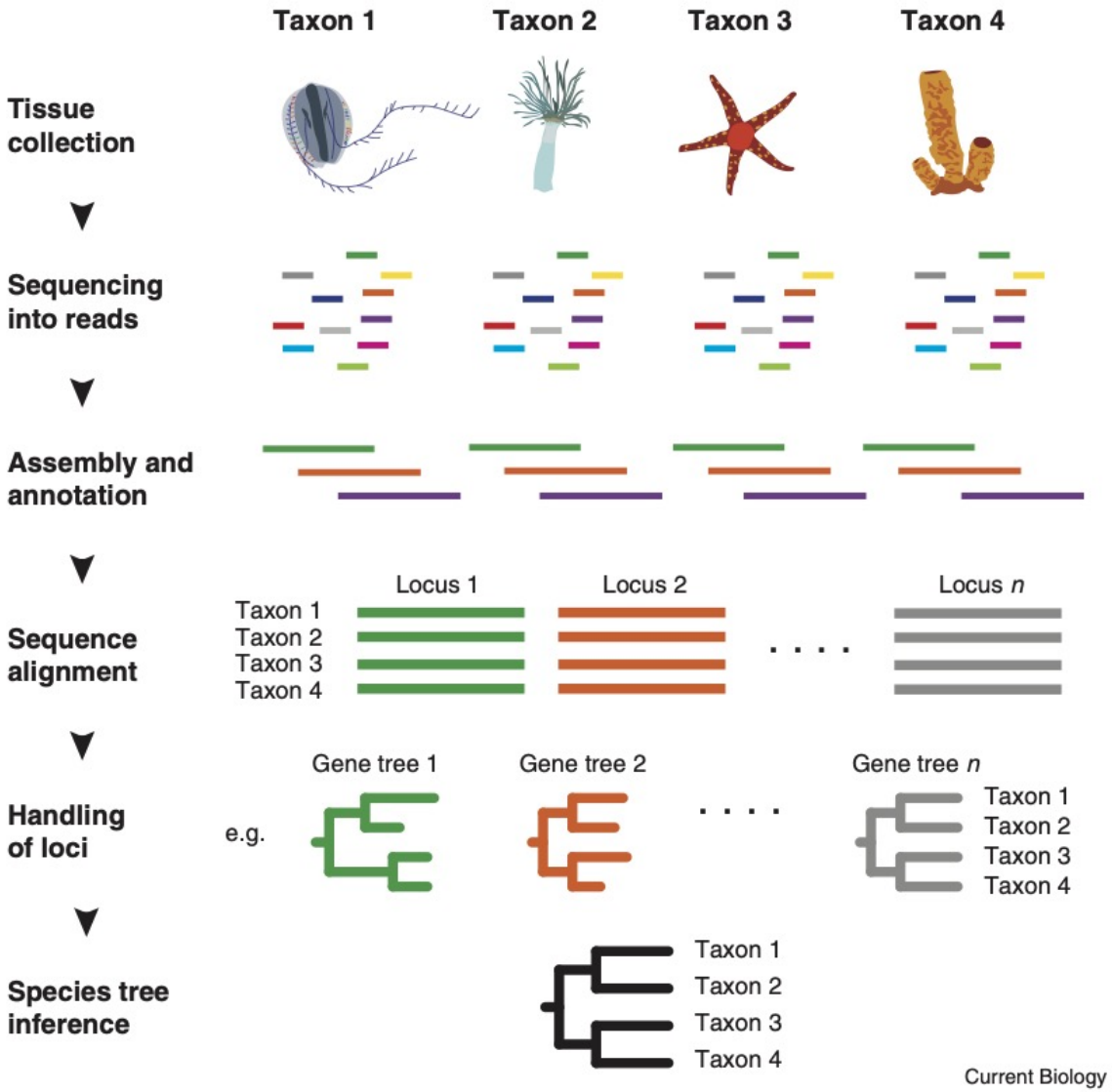
Khan Academy



Multiple sequence alignments are the primary input for molecular phylogenetic analysis



Phylogenomics pipeline



Duchêne (2021) *Phylogenomics Primer*

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?

What assumptions might you want to incorporate into a model of sequence evolution? e.g., would all sites evolve at the same rate?

Models of nucleotide evolution: rate matrix

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

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

We can calculate the the probability of changing between two states over a given 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

The simplest model of sequence evolution.

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

Assumptions: equal mutation rates and equal base frequencies.

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

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.

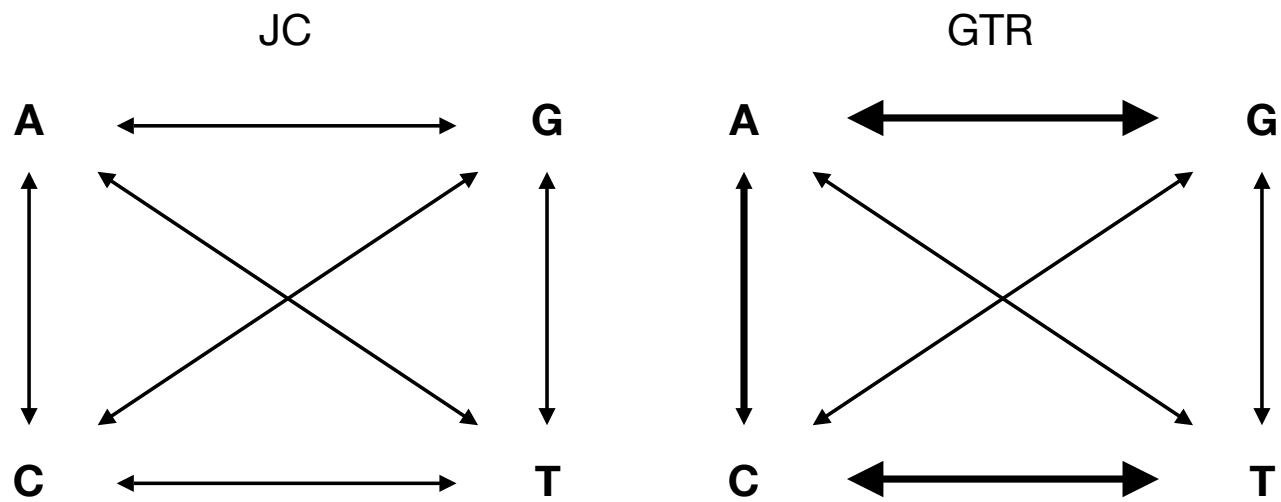
GTR assumptions: unequal mutation rates AND unequal base frequencies.

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

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.

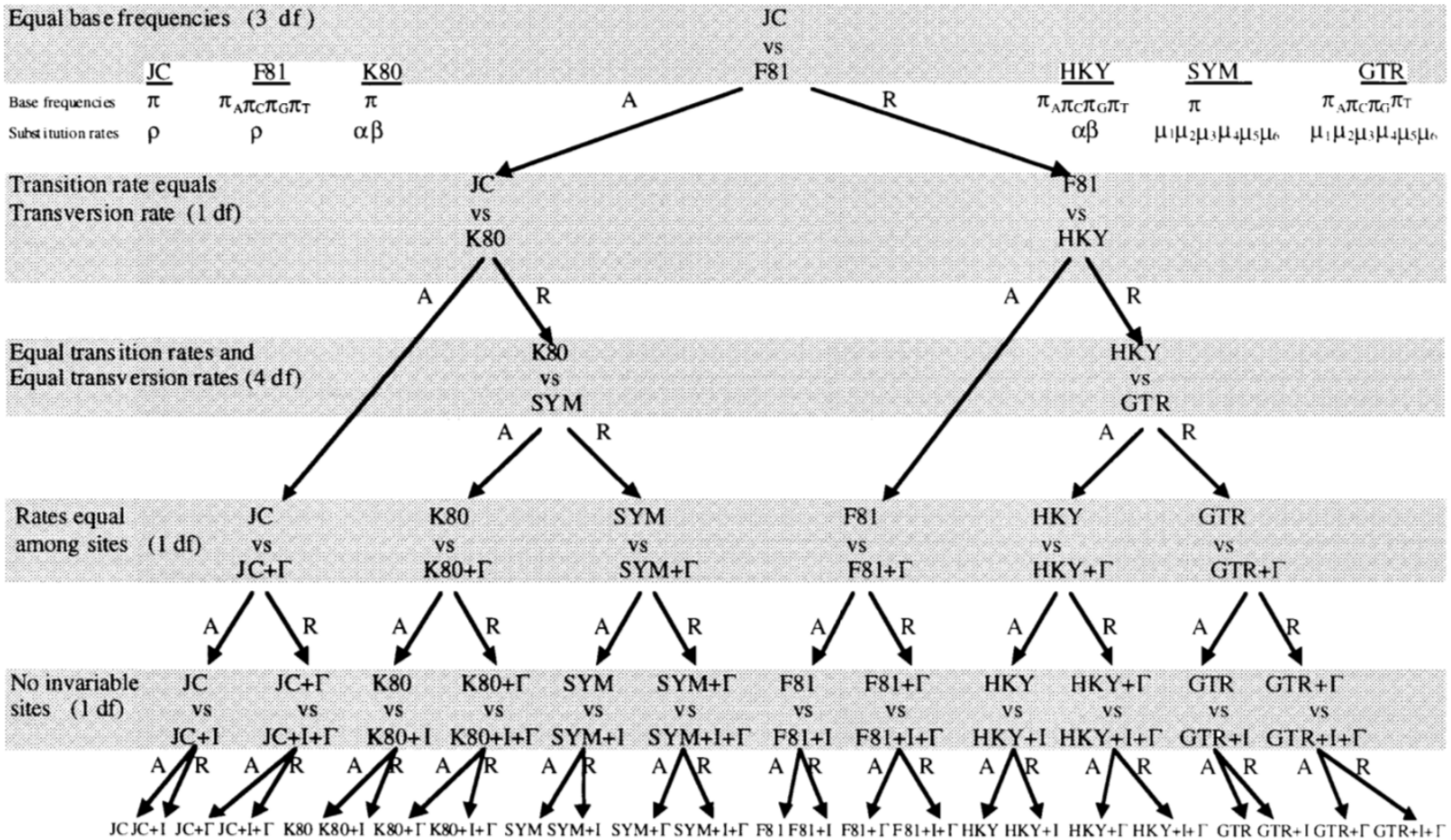
The JC versus GTR models

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*

Exercise