Phylogenetics

Introduction to statistical phylogenetics

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

May 5, 2022

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

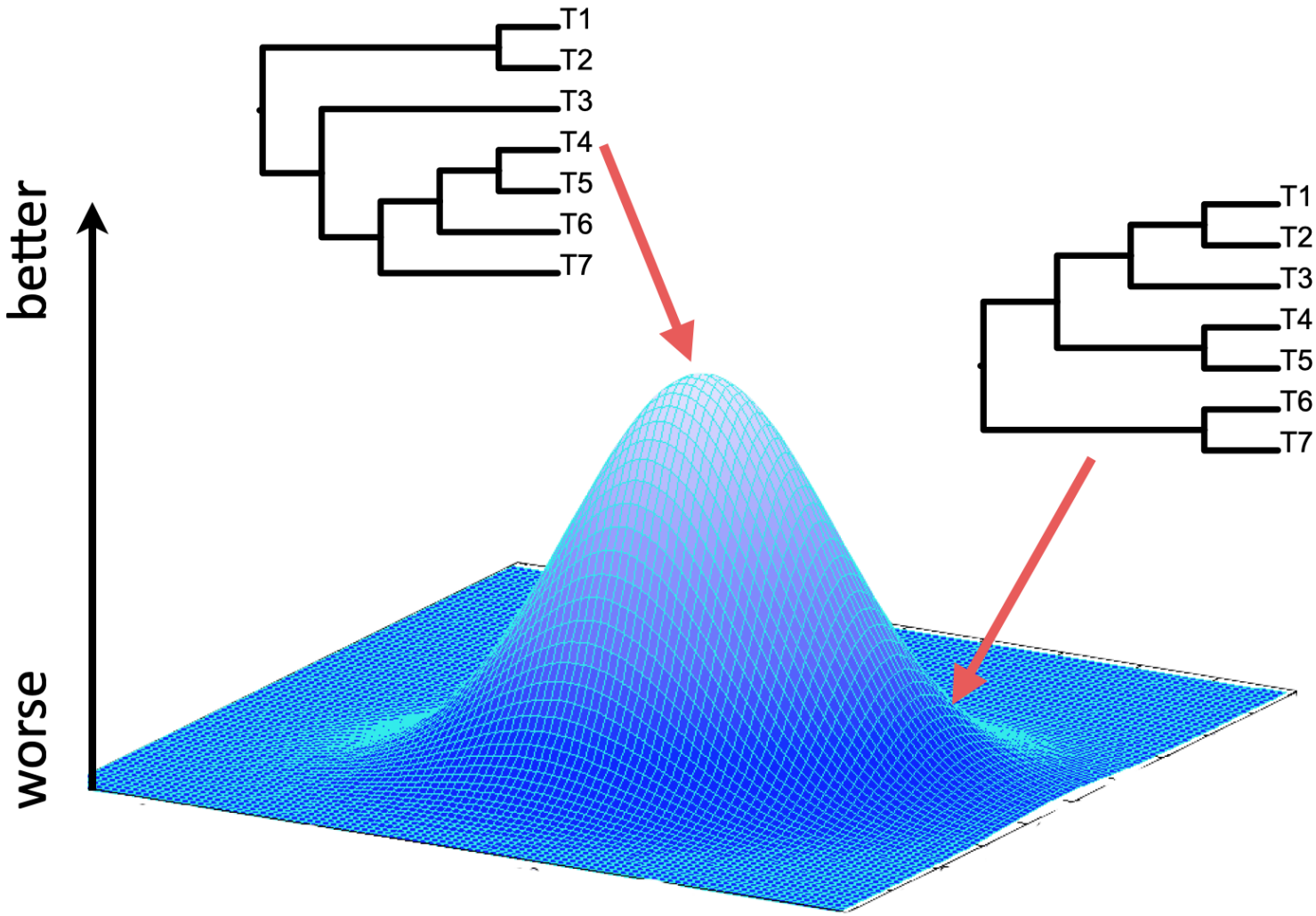


Image source: Tracy Heath

Recap — It depends how you measure "best"

Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes
Maximum likelihood	Log likelihood score, optimised over branch lengths and model parameters
Bayesian	Posterior probability, integrating over branch lengths and model parameters

Both maximum likelihood and Bayesian inference are model based approaches.

Recap — Parsimony: advantages and disadvantages

The greatest advantage of parsimony is its beautiful simplicity

Computationally fast

Often produces sensible results

Some argue that parsimony is assumption free

Others argue that parsimony does make assumptions, even if we don't know what they are

Yang (2014) Molecular Evolution: A Statistical Approach

Today

A brief introduction to:

maximum likelihood

substitution models

RevBayes and the Rev language

Next week May 12 there will be no class!

Tasks

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

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 3 (May 19).

Things to remember

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!

Model-based phylogenetics

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)

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

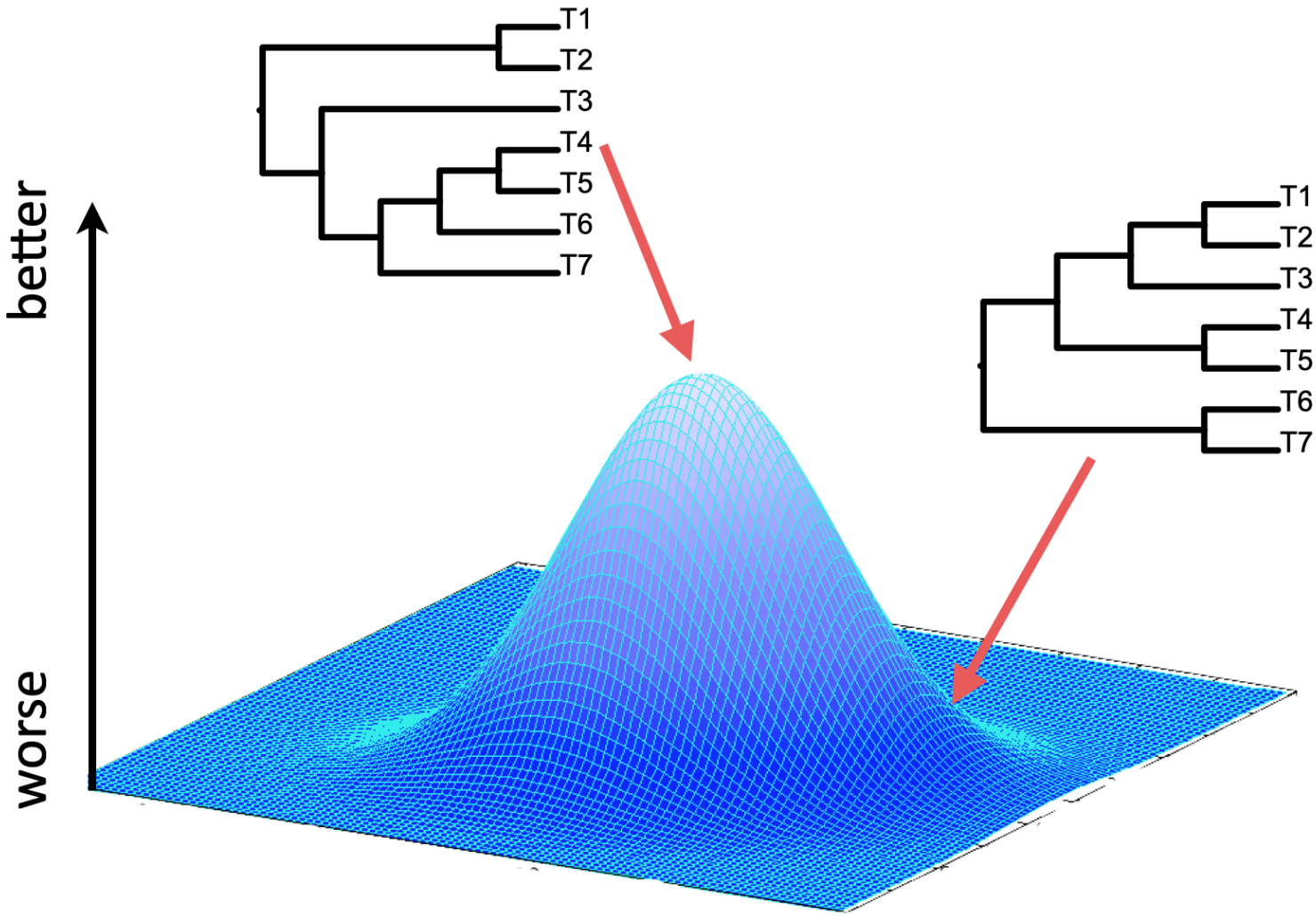


Image source: Tracy Heath

Phylogenetic character data

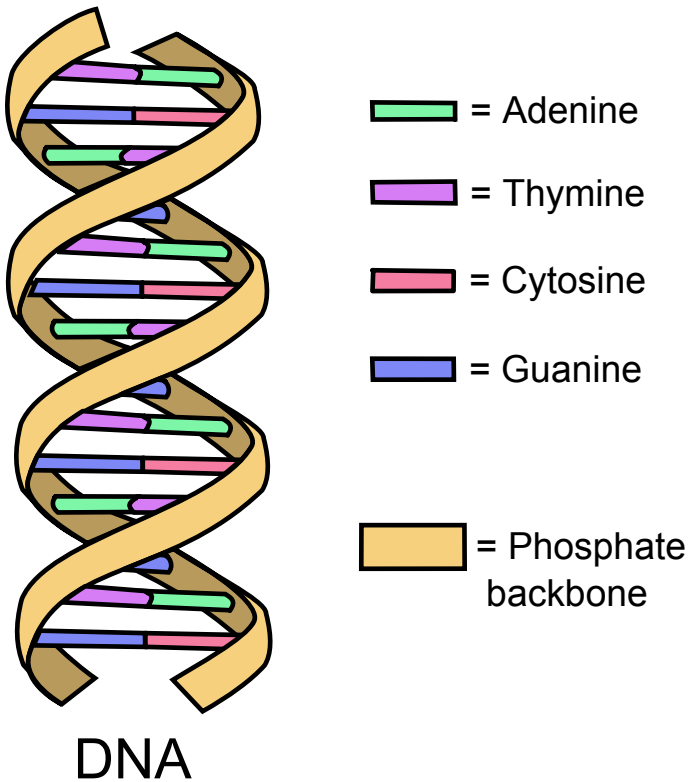
Two main sources of data for building trees:

1. Molecular sequences (nucleotides or proteins)

2. Morphological characters (discrete or continuous)

First we need to collect the data $\rightarrow$ then we need to establish homology.

Molecular sequence data



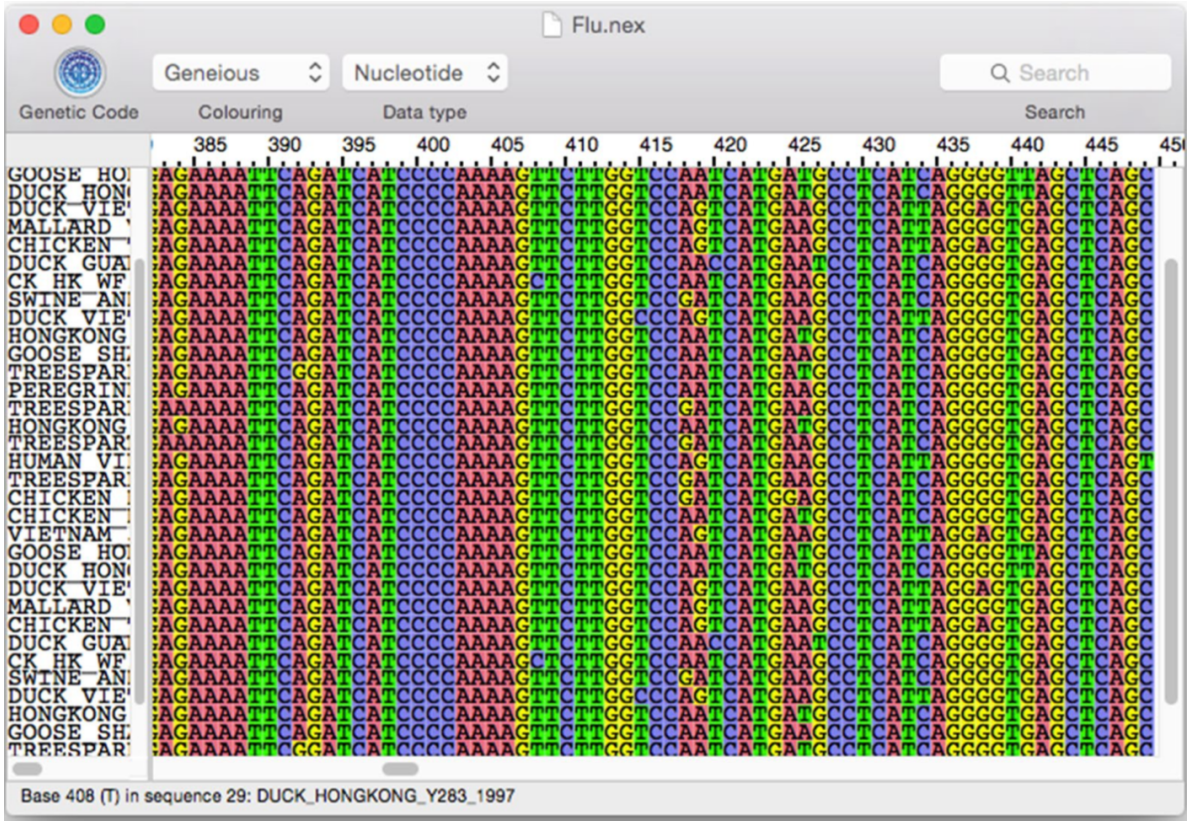
DNA

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.

Goal: to model the process of character evolution across the tree.

What is the probability of transitioning from one state to another over time?

Models of nucleotide evolution: rate matrix

Our model needs to determine 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

Assumptions: unequal mutation rates AND unequal base frequencies.

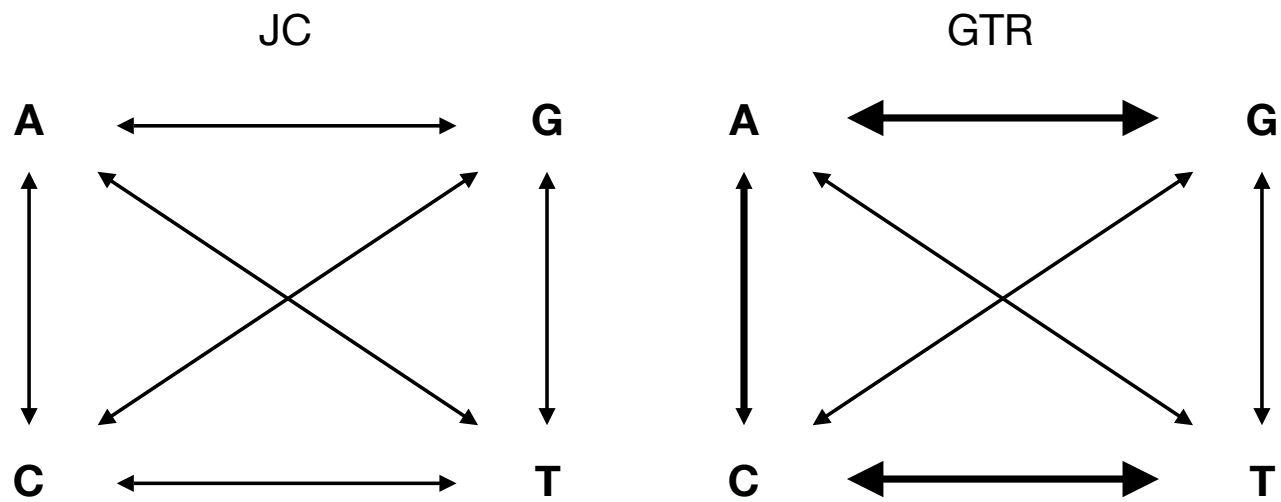
If a given nucleotide appears in our dataset at a low frequency, we are less likely to observe a transition to that state.

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

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.

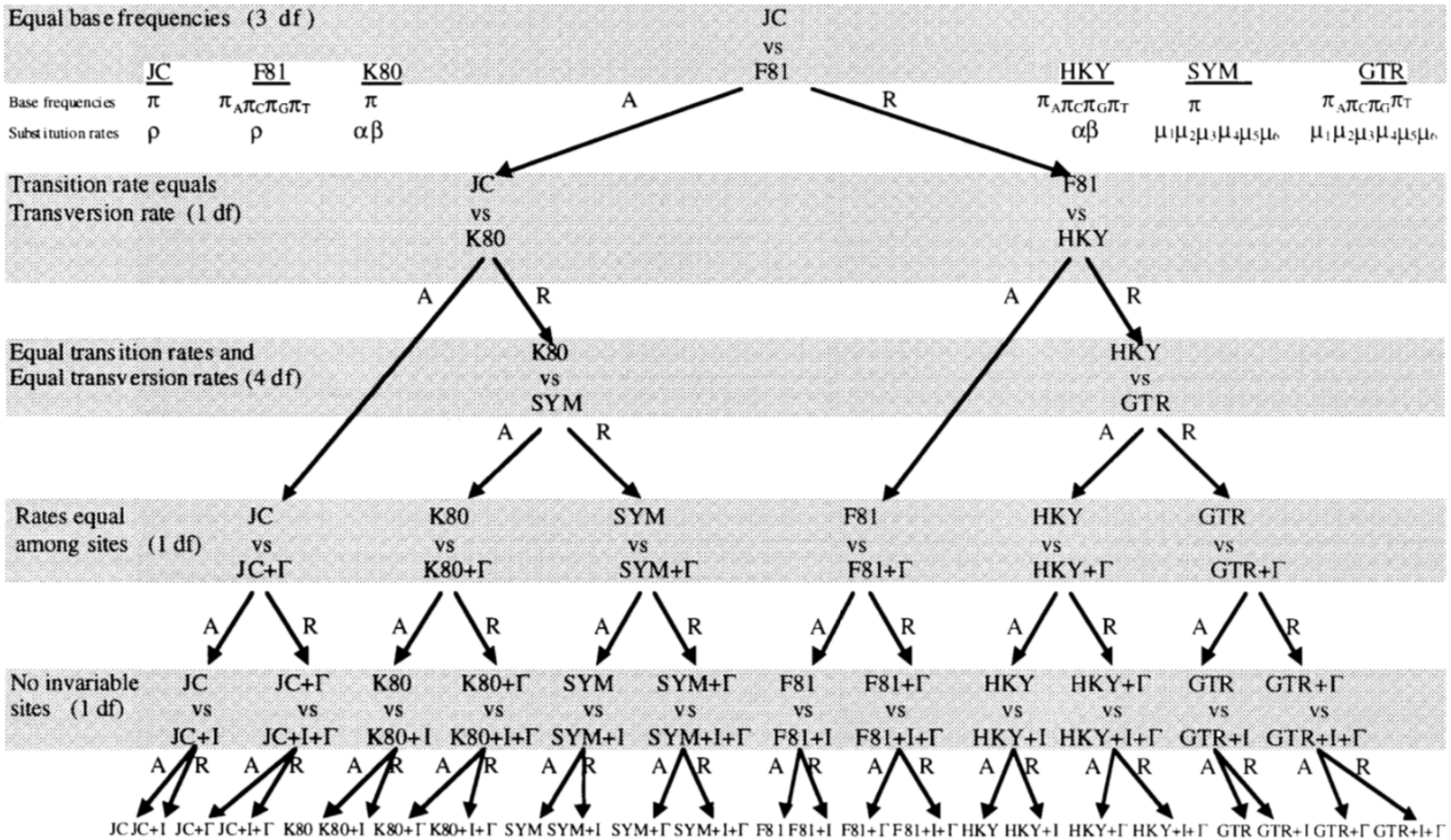
The JC versus GTR models

Another way of visualising substitution models.



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

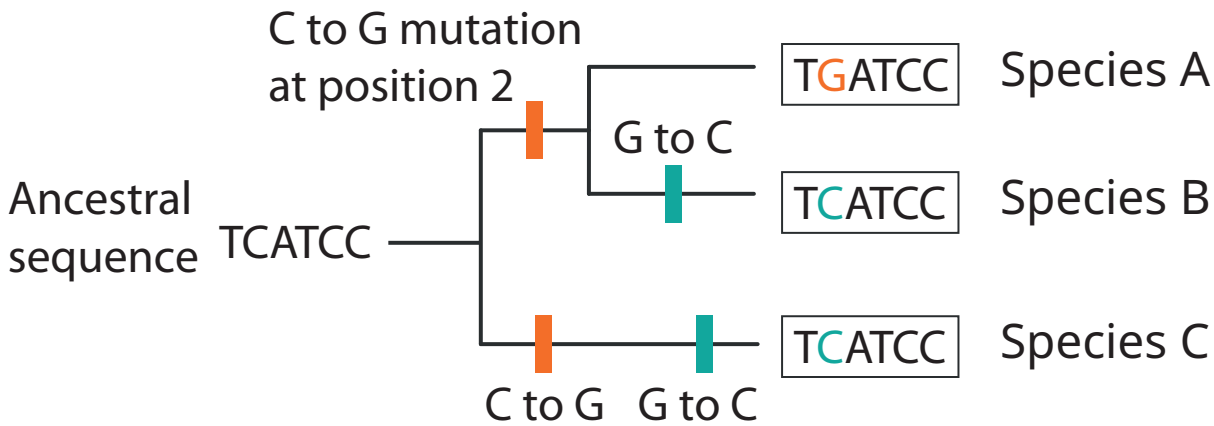
JC & GTR belong to a large family of substitution models



Posada & Crandall (1998) *Bioinformatics*

Model-based phylogenetics

Models can account for multiple changes at the same site.



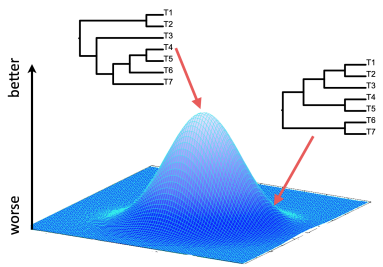
Branch lengths = *expected number of changes per site*

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.



Things to bare in mind

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

For molecular data, base pairs (ATCG) occur at different frequencies depending on the group of species or gene.

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

Exercises