Phylogenetics

Introduction to likelihood

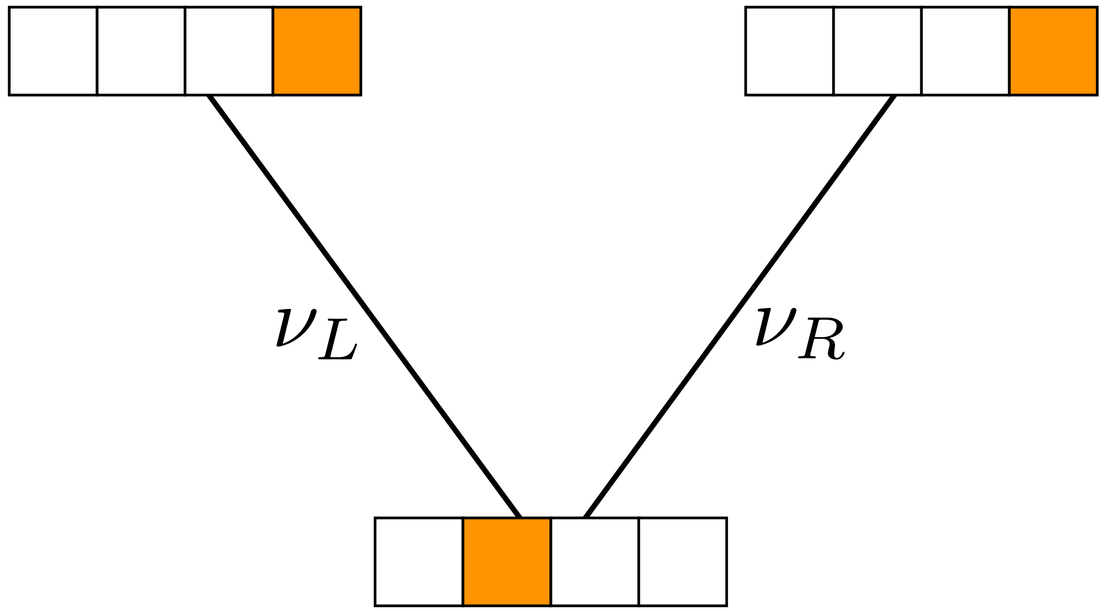
RL-V3 MPP

Rachel Warnock

14.04.25

Objectives

- Introduction to substitution models
- Gain an understanding of the maximum likelihood approach to tree-building



Recap

How do we find the ‘best’ tree?

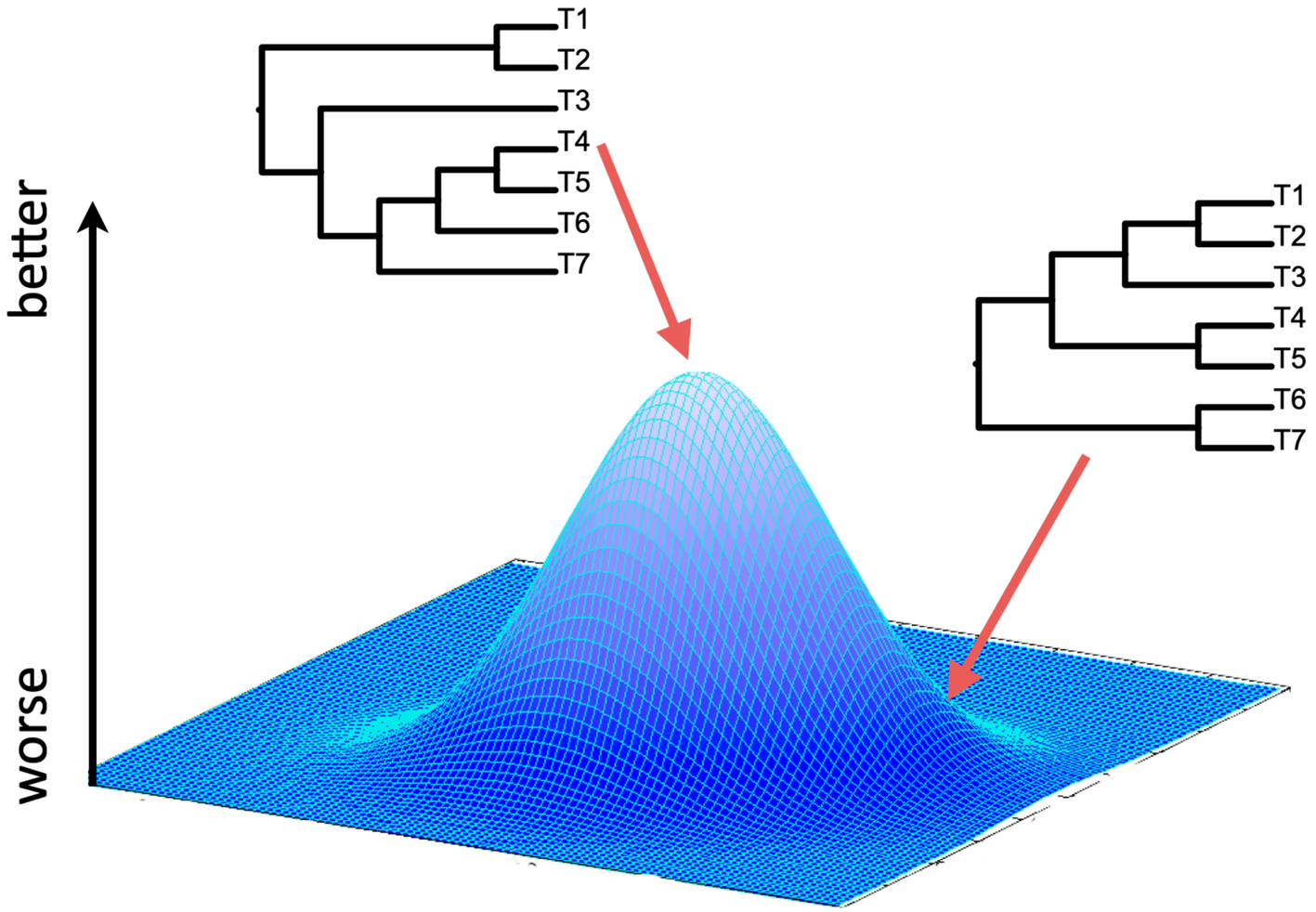


Image source Tracy Heath

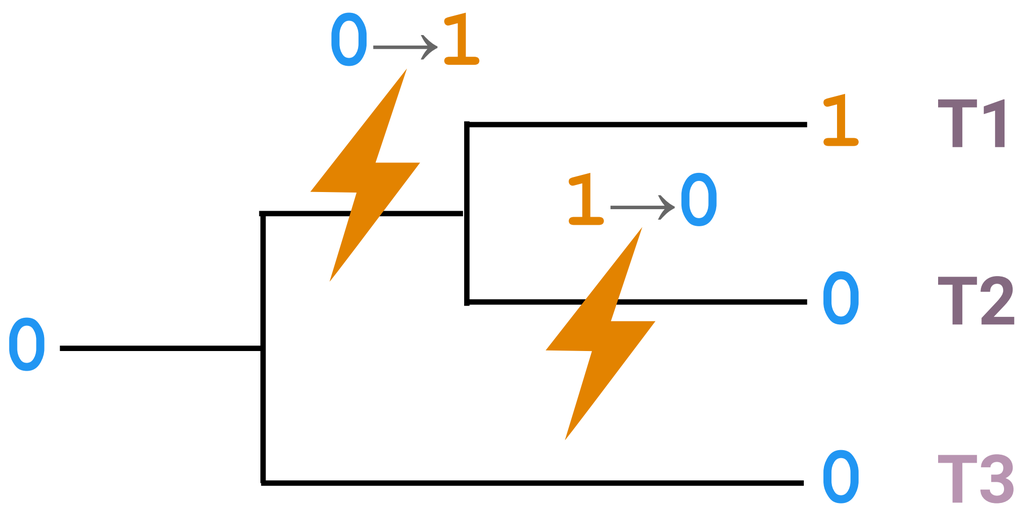
It depends how you measure ‘best’

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

Both maximum likelihood and Bayesian inference are model-based approaches

Note these are not the only approaches to tree-building but they are the most widely used

Convergence and parsimony



Hypothetical tree showing multiple transitions at the same character

Parsimony will always favour the tree with the smallest number of changes

The method does not account for multiple transitions (or “hits”), e.g., 0 → 1 → 0

We can only invoke convergent evolution *ad hoc* after inference

Recap: parsimony

Parsimony does not make explicit assumptions about the evolutionary process (although it makes implicit assumptions)

It has been demonstrated that in some scenarios parsimony is statistically inconsistent. The issue is known as long branch attraction

Model-based approaches on the other hand make explicit assumptions about evolutionary processes (+ have a wider ranger of applications in paleobio)

What do I mean by model?

(the following is my take on things — intended to be useful but not definitive)

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?

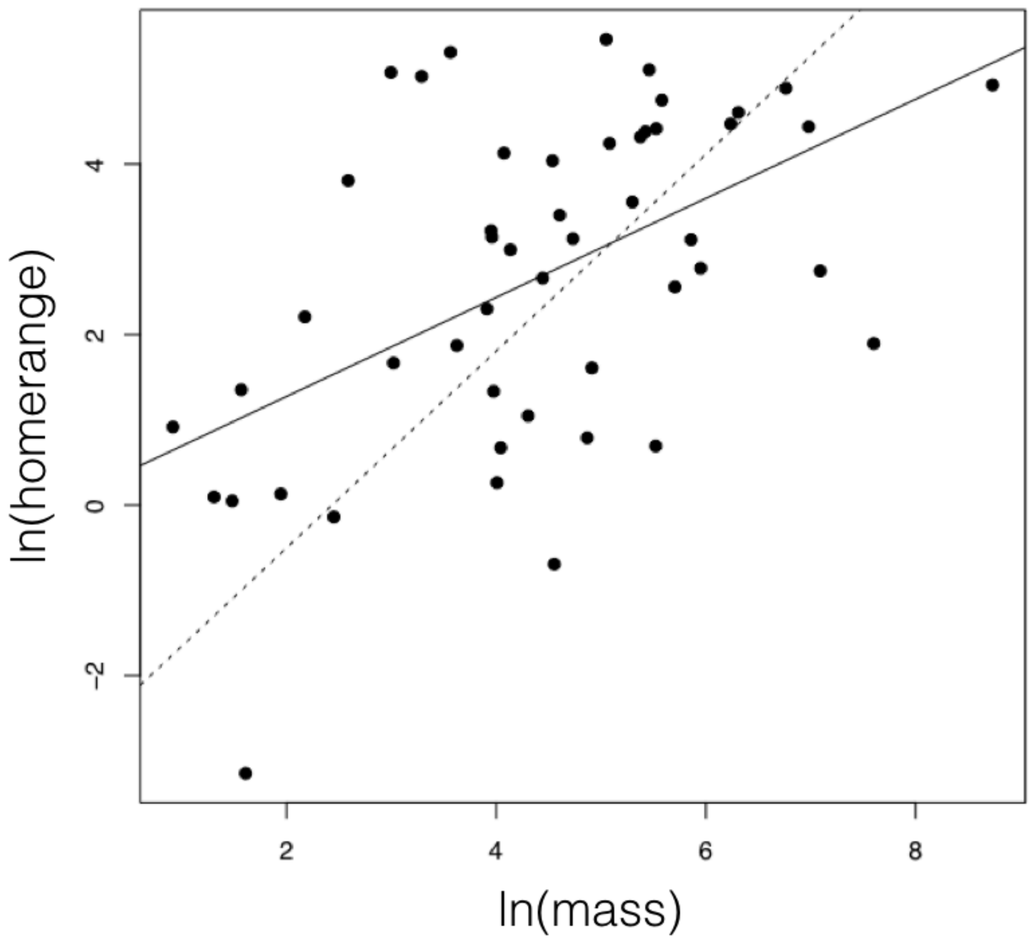
A statistical model is a type of model that includes a set of assumptions about the data-generating process

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



The solid black line is a linear regression line

We can estimate the parameters of the regression model

$y = X\beta + \varepsilon$

It's also straightforward to simulate data under this model

Image source *Harmon* (2019)

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 or geological 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 models used in phylogenetics are mechanistic, e.g., they might include parameters for origination, extinction, or sampling

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)
```

Used in phylogenetics for all sorts of tasks, e.g., traversing tree space

Molecular evolution

Character evolution

along species trees

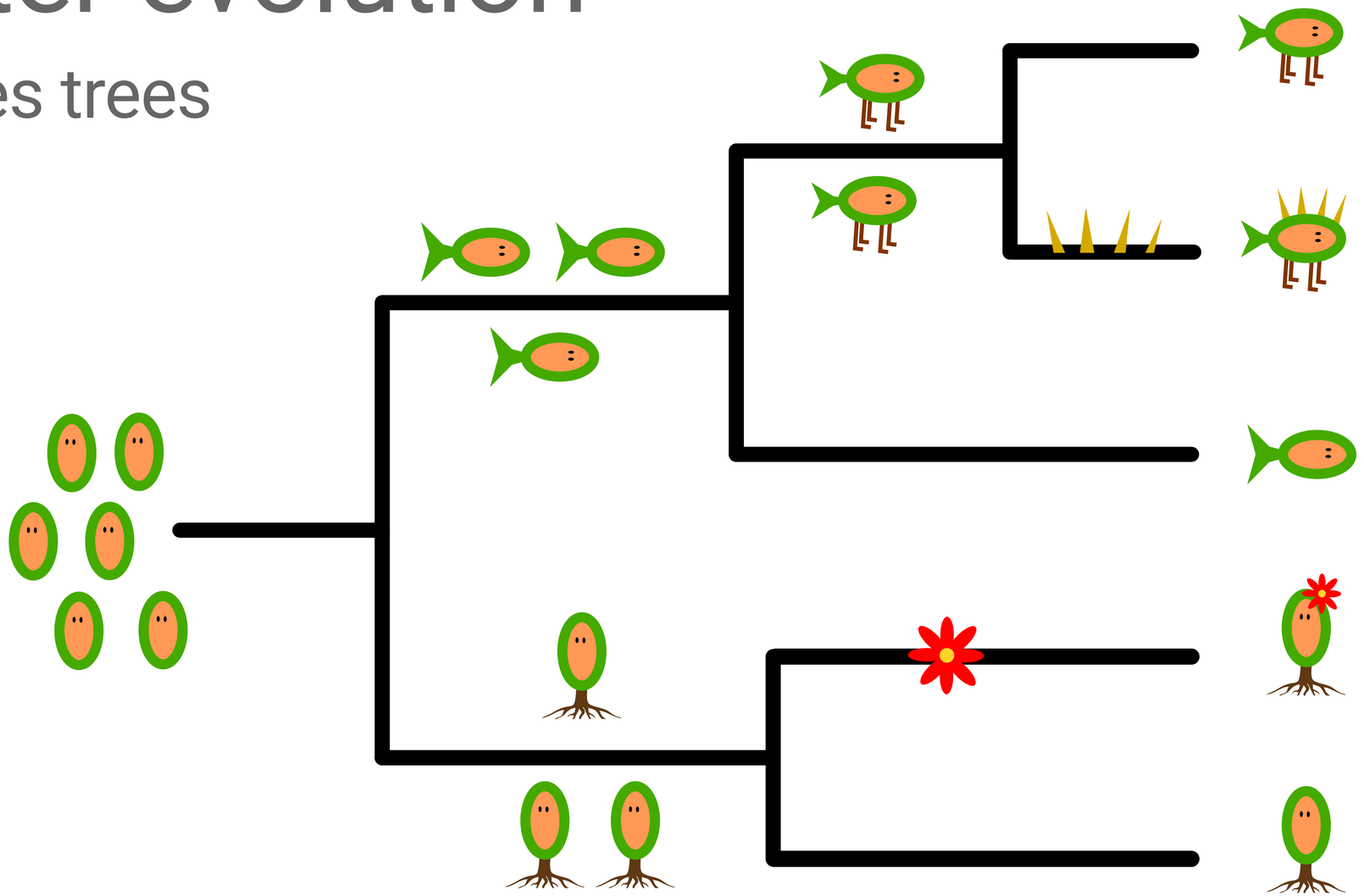
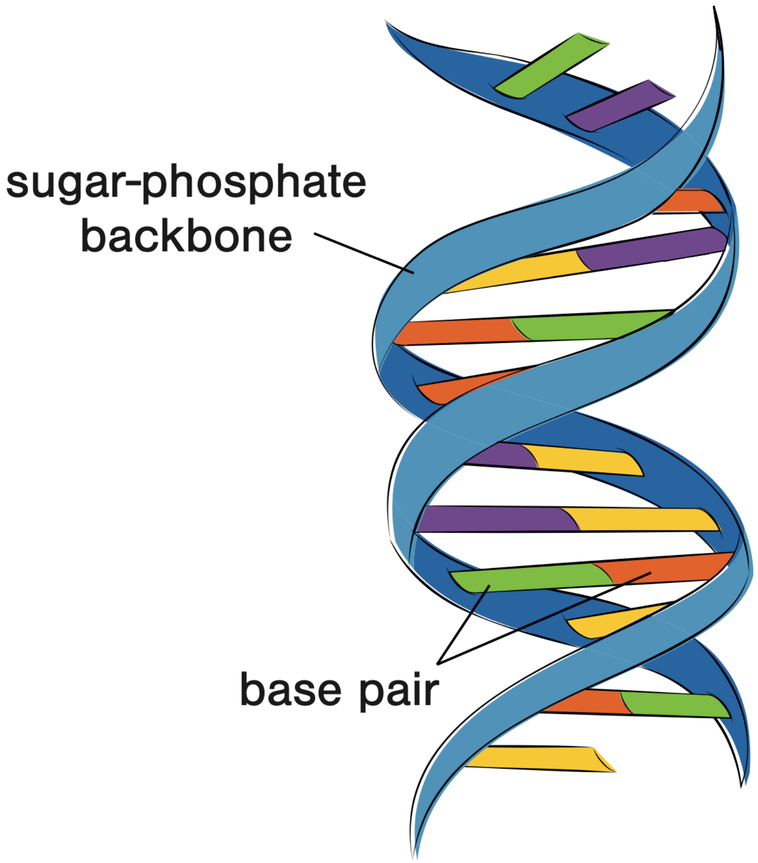
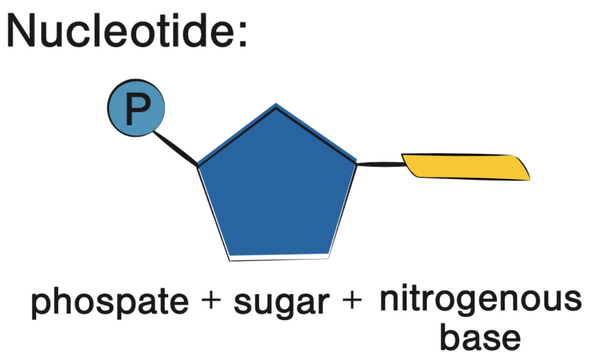
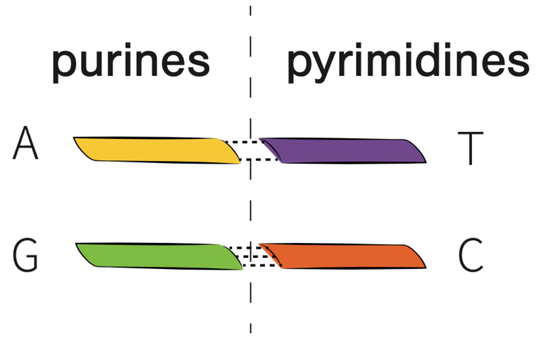


Image: Joëlle Barido-Sottani

Deoxyribonucleic acid (DNA)







Purines

- Adenine (**A**)
- Guanine (**G**)

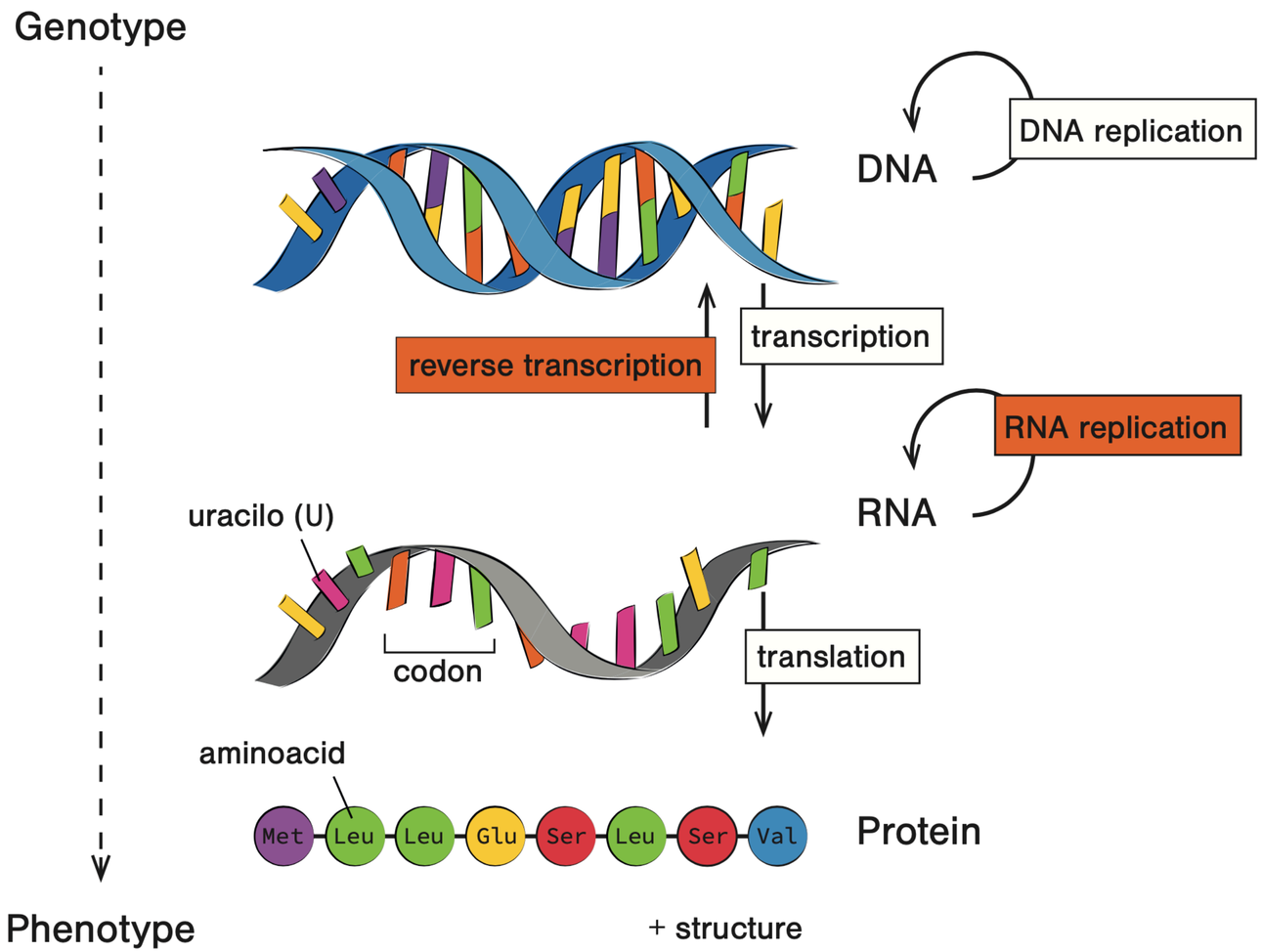
Pyrimidines

- Cytosine (**C**)
- Thymine (**T**)*

Image source Decoding Genomes *Stadler et al.* (2024)

*Replaced by uracil (U) in RNA

The central dogma of biology

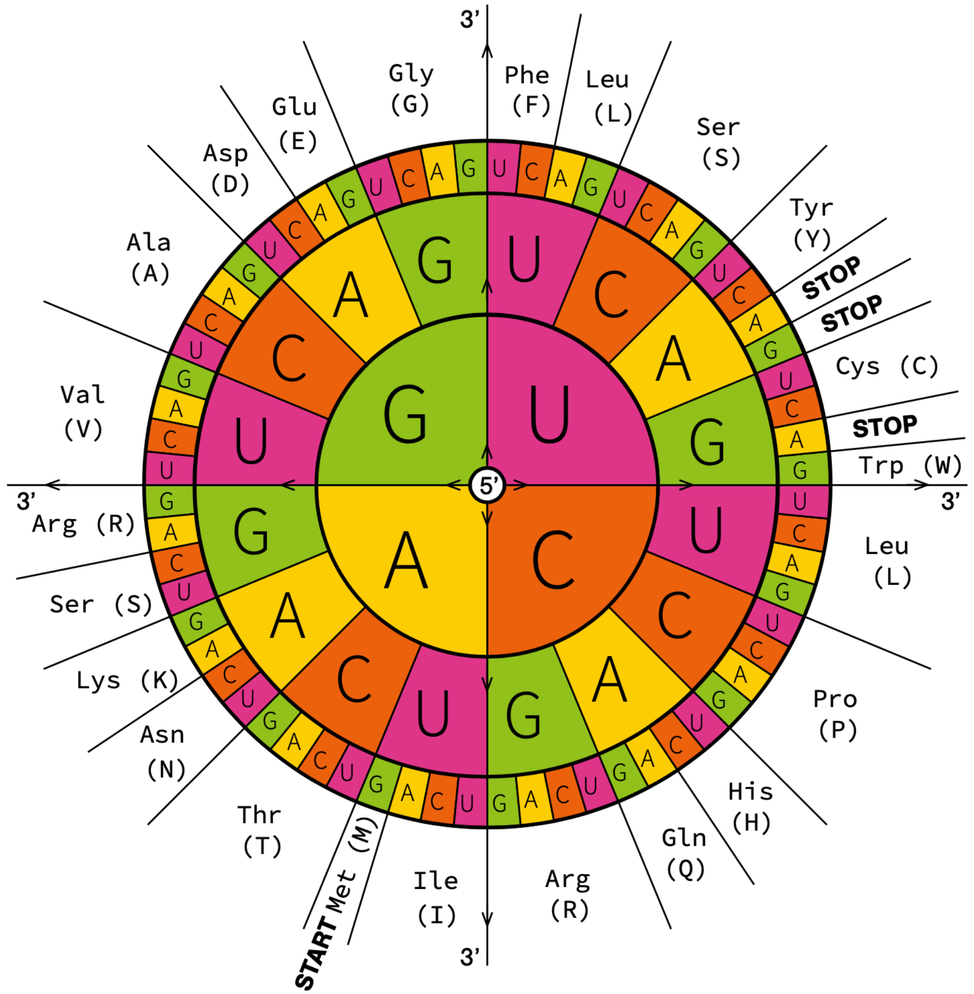


DNA → RNA → protein

Each group of 3 successive nucleotides in a gene is a codon that encodes an amino acid (or terminate translation)

Image source Decoding Genomes *Stadler et al.* (2024)

The universal genetic code



Amino acid	3-letter code	1-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

4^3 = 64 combinations

3 terminate translation

21 amino acids

Image source Decoding Genomes *Stadler et al.* (2024)

Mutation vs. substitution

Variation in genotypes (and in phenotypes) is due to errors that arise during DNA replication, termed mutations

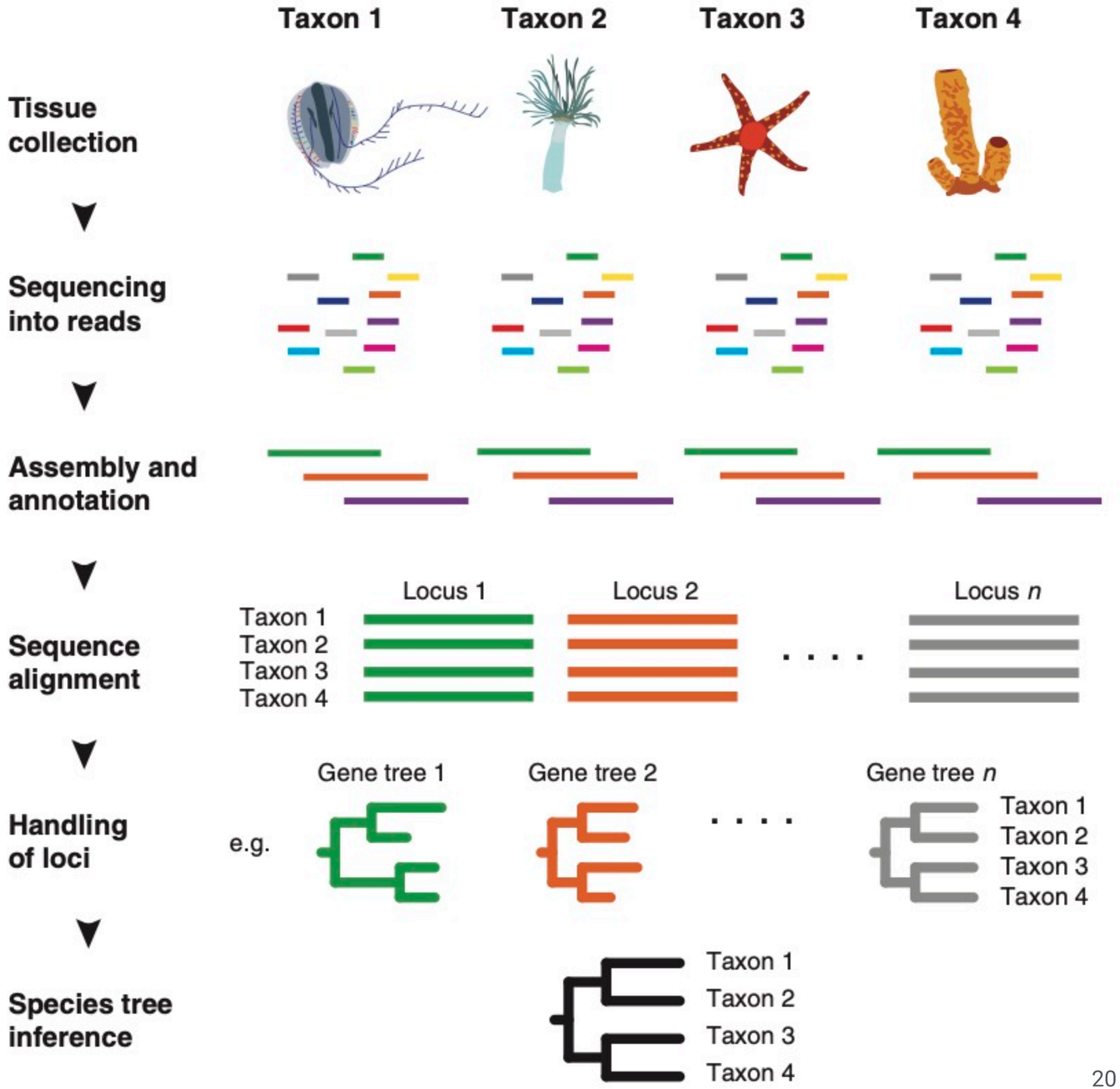
Individuals of the same species have identical characters at *most* positions in their genome (only 0.1% vary among humans)

Most mutations are repaired but can persist across generations

Mutations that spread throughout a population and become ‘fixed’ are called substitutions

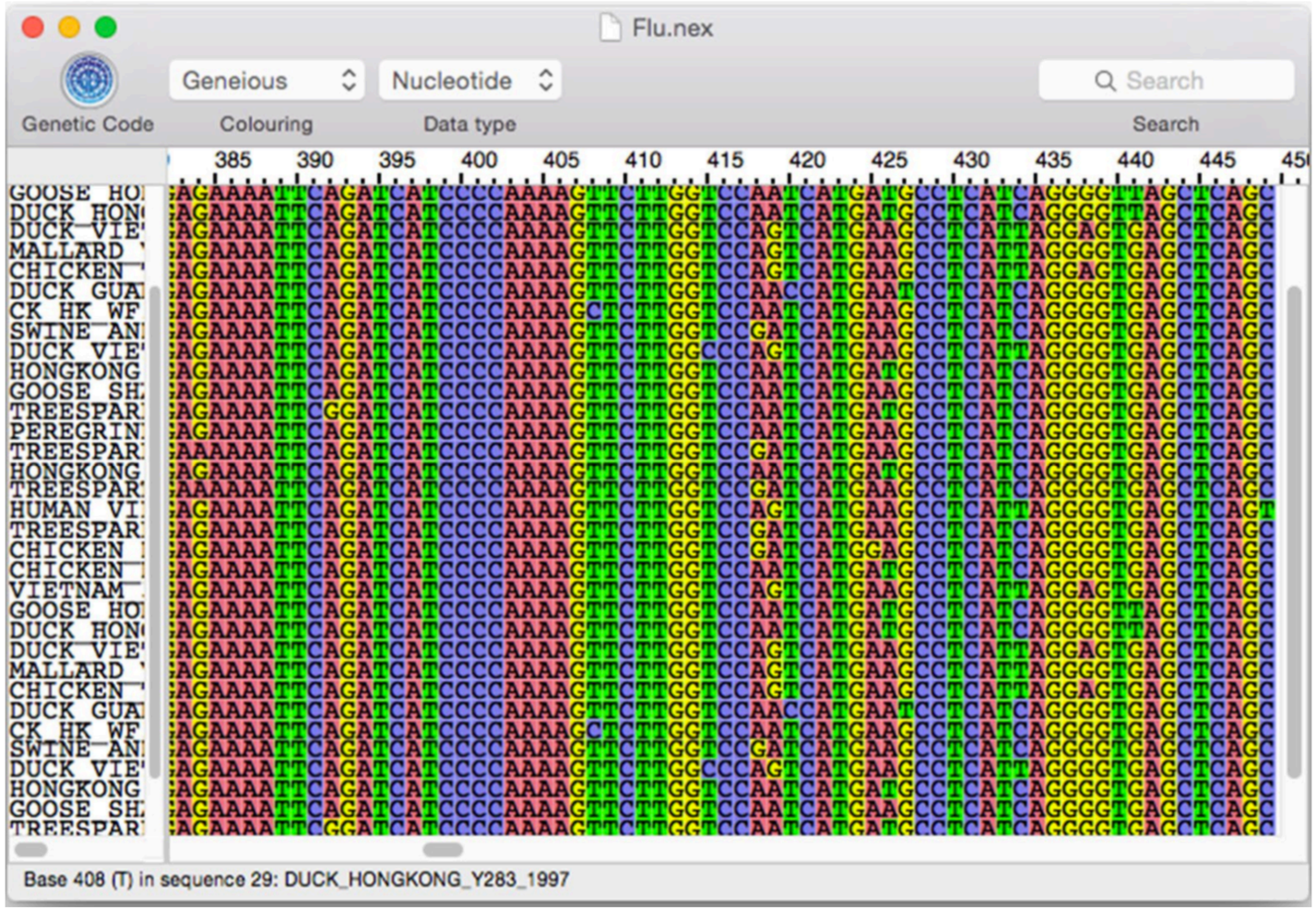
DNA sequencing

Multiple sequence alignment software establishes homology across sites from different species



Duchêne (2021) Phylogenomics Primer

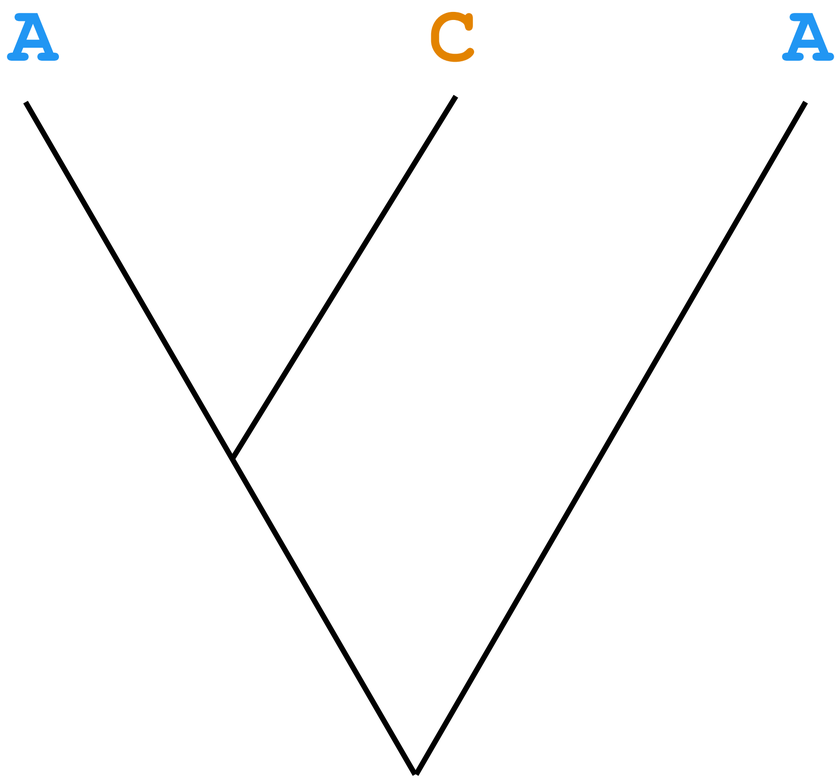
Multiple sequence alignment



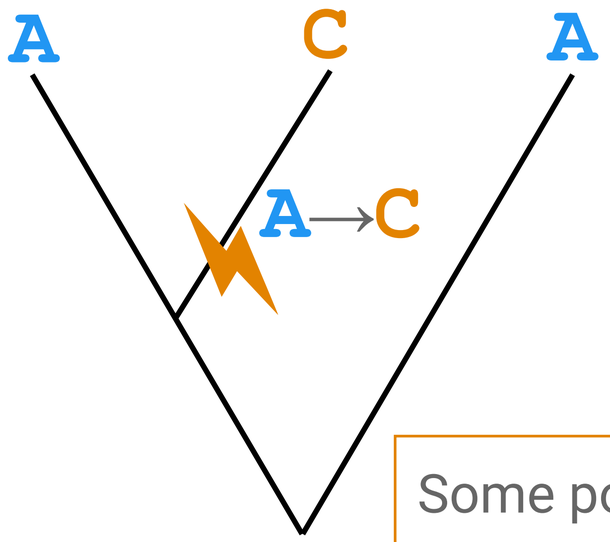
```
#NEXUS

[Cytochrome oxidase B genes - bears]
[Data source: https://revbayes.github.io/tutorials/dating/]


BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=1000;
    FORMAT DATATYPE = DNA MISSING=? GAP=- ;


MATRIX

Ailuropoda_melanoleuca     ATGATCAACATCCGAAAAACTCATCCATTAGTTAAAATTATCAACAACTCATTCATTGACCT...
Arctodus_simus             ATGACCAACATCCGAAAGACTCACCCACTGGCCAAAATTATCAATAACTCATTCATCGACCT...
Helarctos_malayanus        ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATTAACAACTCACTTATTGACCT...
Melursus_ursinus           ATGACCAACATCCGAAAAACCCACCCACTAGCTAAAATCATTAACAACTCACTCATTGACCT...
Ursus_americanus           ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCACTTATTGATCT...
Ursus_arctos               ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCATTTATTGACCT...
Ursus_maritimus            ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCATTTATTGATCT...
Ursus_thibetanus           ATGACCAACATCCGAAAAACCCATCCATTAGCCAAAATCATCAACAACTCACTCATTGATCT...
Ursus_spelaeus             ATGACCAACATCCGAAAAACCCATCCACTAGCTAAAATCATCAACAACTCATTCATTGACCT...
Tremarctos_ornatus         ATGACCAACATCCGAAAAACTCACCCACTAGCTAAAATCATCAACAACTCATTCATCGACCT...

;

END;
```

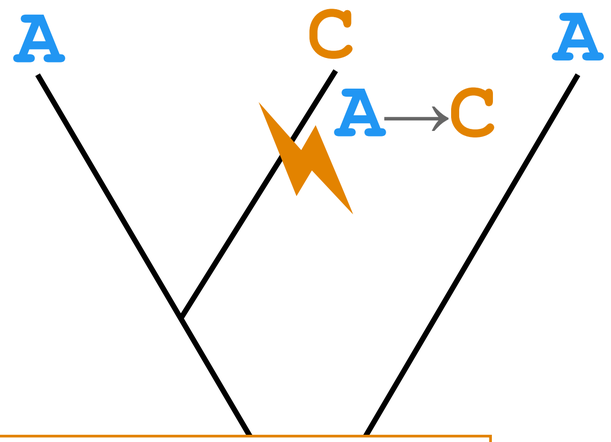
The data are the observed states at the tips

How probable is our data, given my tree?

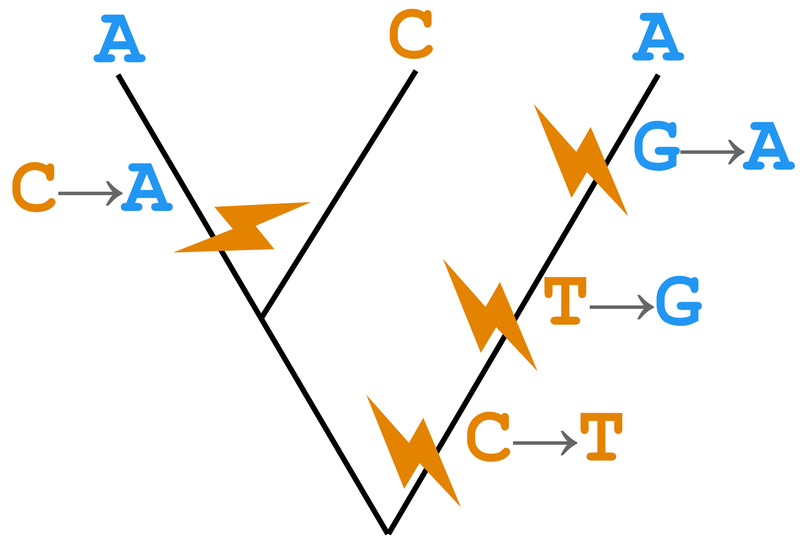


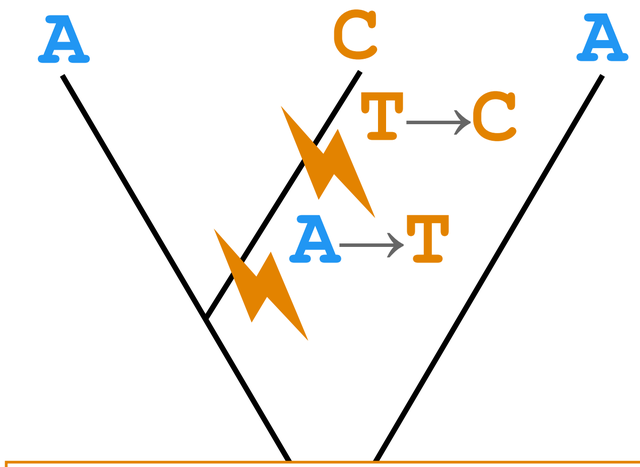
To apply a model based approach we need to be able to compute the probability of our sequence alignment (or character matrix)

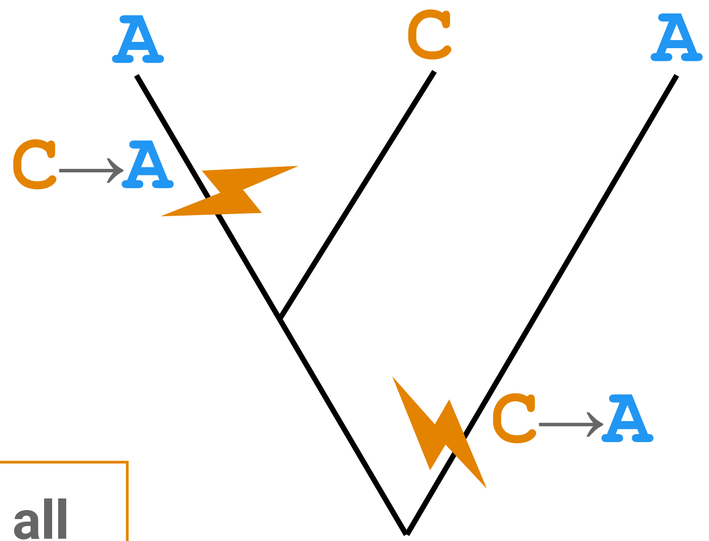




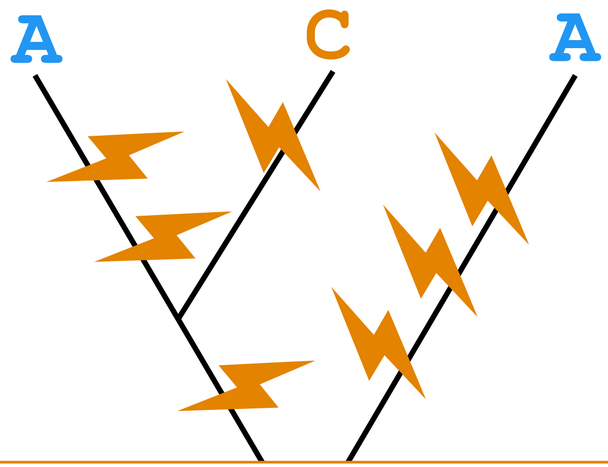
Some possible character histories...







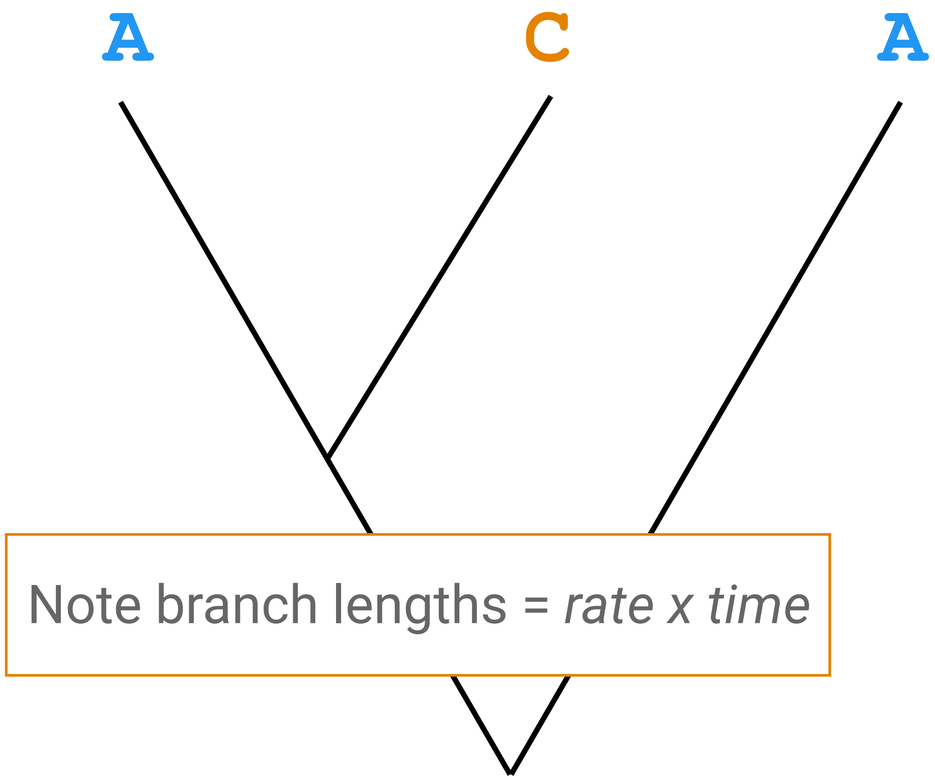
Model based approaches take **all possible histories** into account!



There are **infinitely** many histories leading to the same observed data

The data are the observed states at the tips

How probable is our data, given my tree?



To compute *P*, we need:

- A model of sequence (or character) evolution
- A way of calculating the probability for given a phylogeny (tree topology + branch lengths)

Substitution models

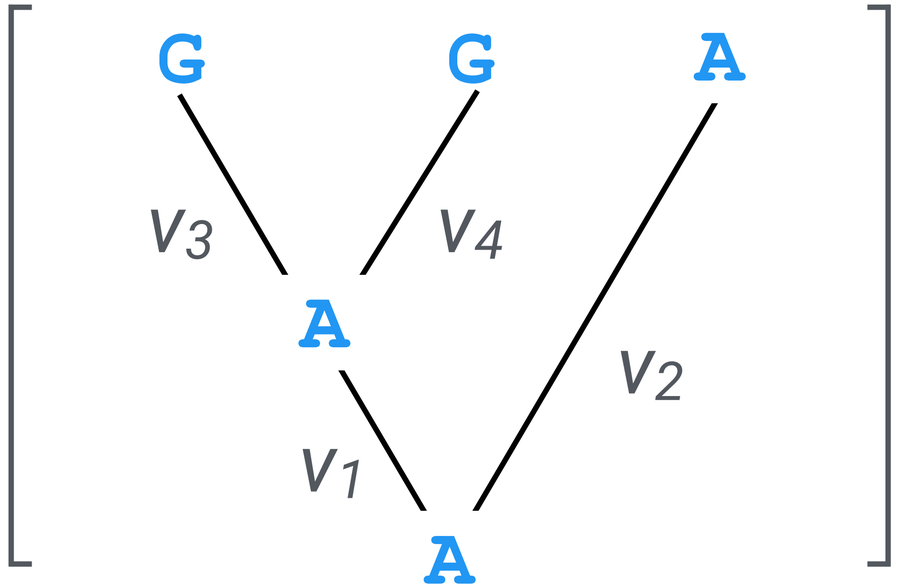
Models of molecular sequence evolution

Also known as substitution / site / character models

They allow us to compute the probability of changing from one state to another over branch length *v*

Computing the probability of the observed data

$P =$



Just suppose for now we know the ancestral states at internal nodes

$P_{AA}(v_1) \times P_{AA}(v_2) \times P_{AG}(v_3) \times P_{AG}(v_4)$

$P_{ij}(v)$ — transition probabilities

Rate matrix

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{T} \\ \mathtt{G} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{T} & \mathtt{G} & \mathtt{C} \end{array} \\ \begin{bmatrix} - & \lambda & \lambda & \lambda \\ \lambda & - & \lambda & \lambda \\ \lambda & \lambda & - & \lambda \\ \lambda & \lambda & \lambda & - \end{bmatrix} \end{array}$$

In this model, we only have one parameter, substitution rate parameter λ

This is the Jukes-Cantor (1969) or JC69 model

Rate matrix

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{T} \\ \mathtt{G} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{T} & \mathtt{G} & \mathtt{C} \end{array} \\ \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix} \end{array}$$

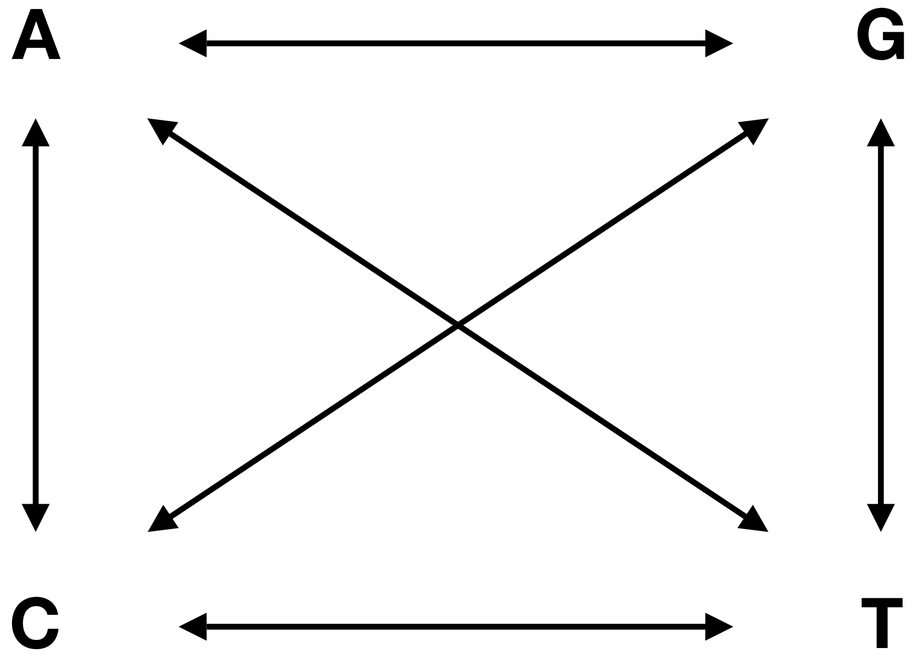
In this model, we only have one parameter, substitution rate parameter λ

This is the Jukes-Cantor (1969) or JC69 model

Rate matrix

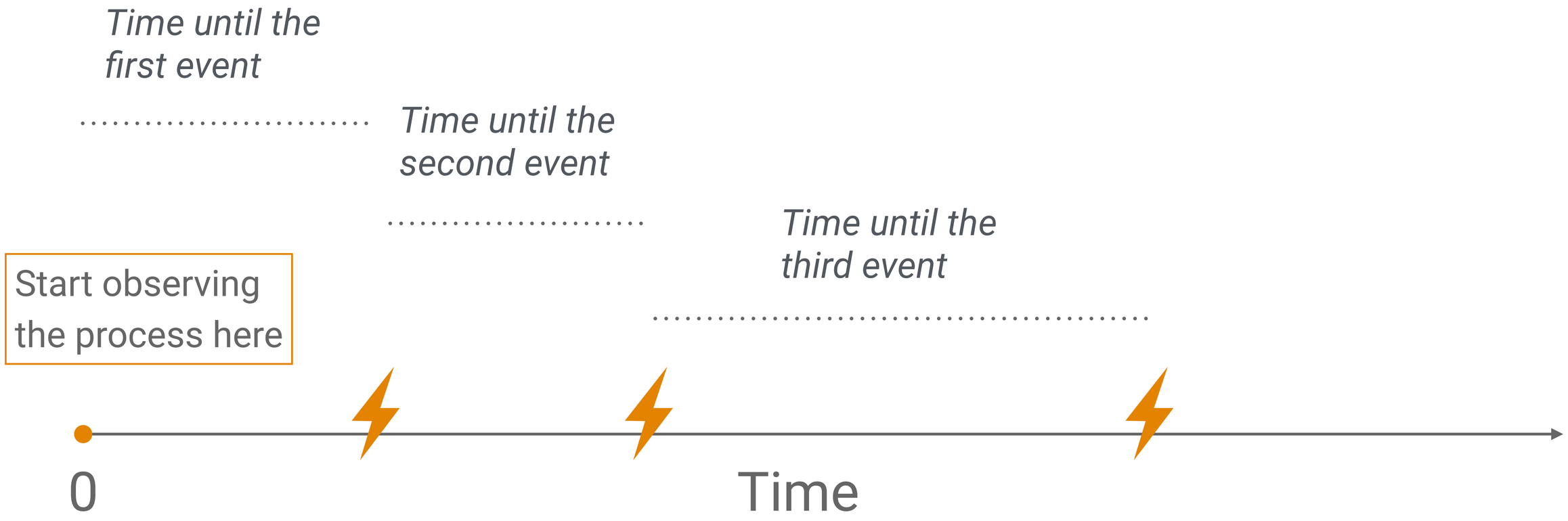
An alternative representation of the JC model

Arrows represent (symmetric) rates of change between states

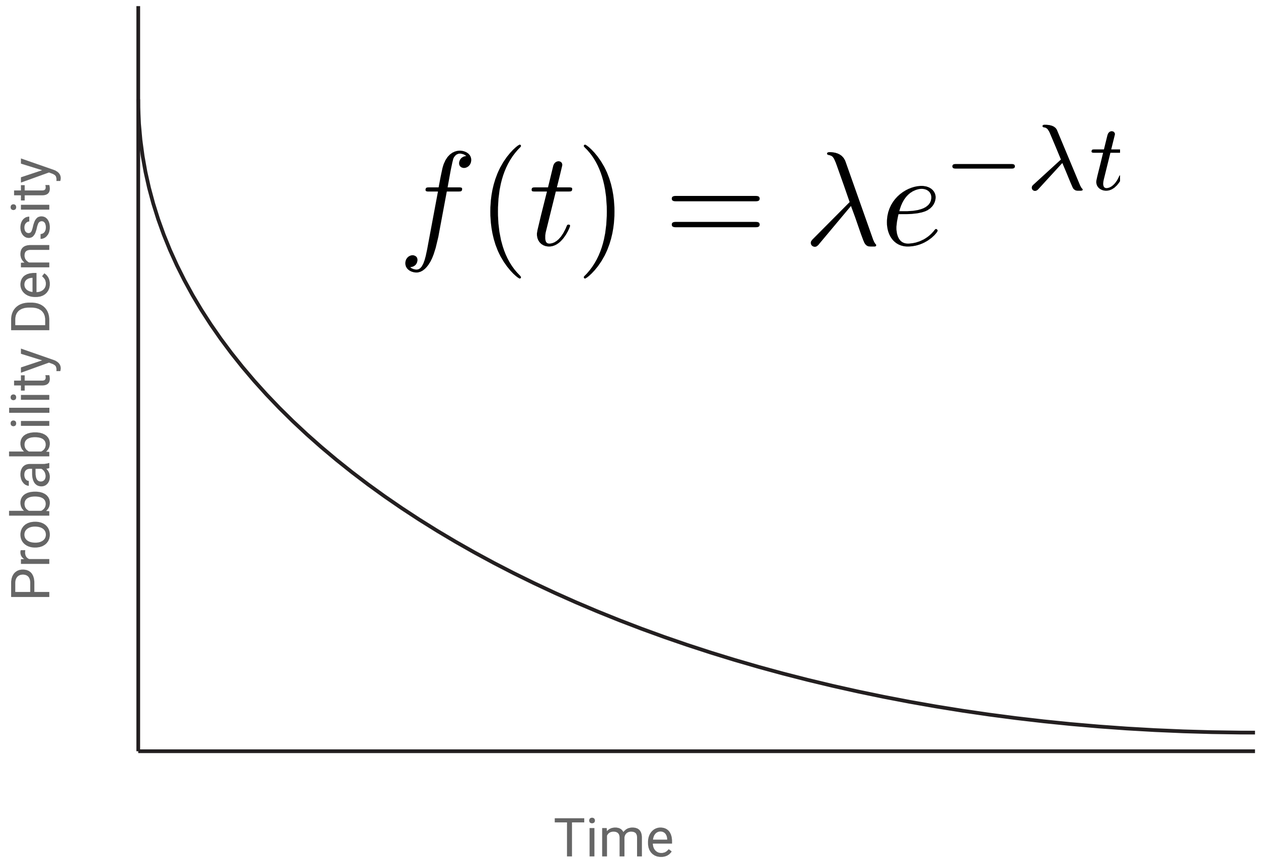


Continuous time Markov chain

Nucleotide substitutions (events) occur at a constant rate (λ)



The poisson process



The waiting times are exponentially distributed random variables

We can use this to calculate the probability of change over time (or branch length *v*)

The longer the interval of time, the more likely we are to observe change

Exercise

Jukes-Cantor model transition probability applet

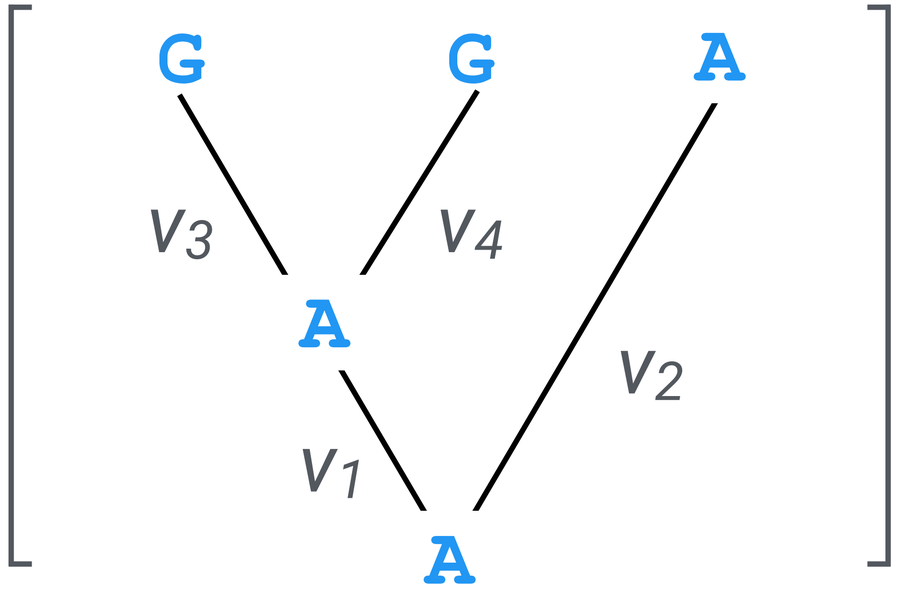
Written by Paul Lewis

Felsenstein’s pruning algorithm

The following slides are adapted from John Huelsenbeck (c/o Sebastian Höhna)

Computing the probability of the observed data

$P =$

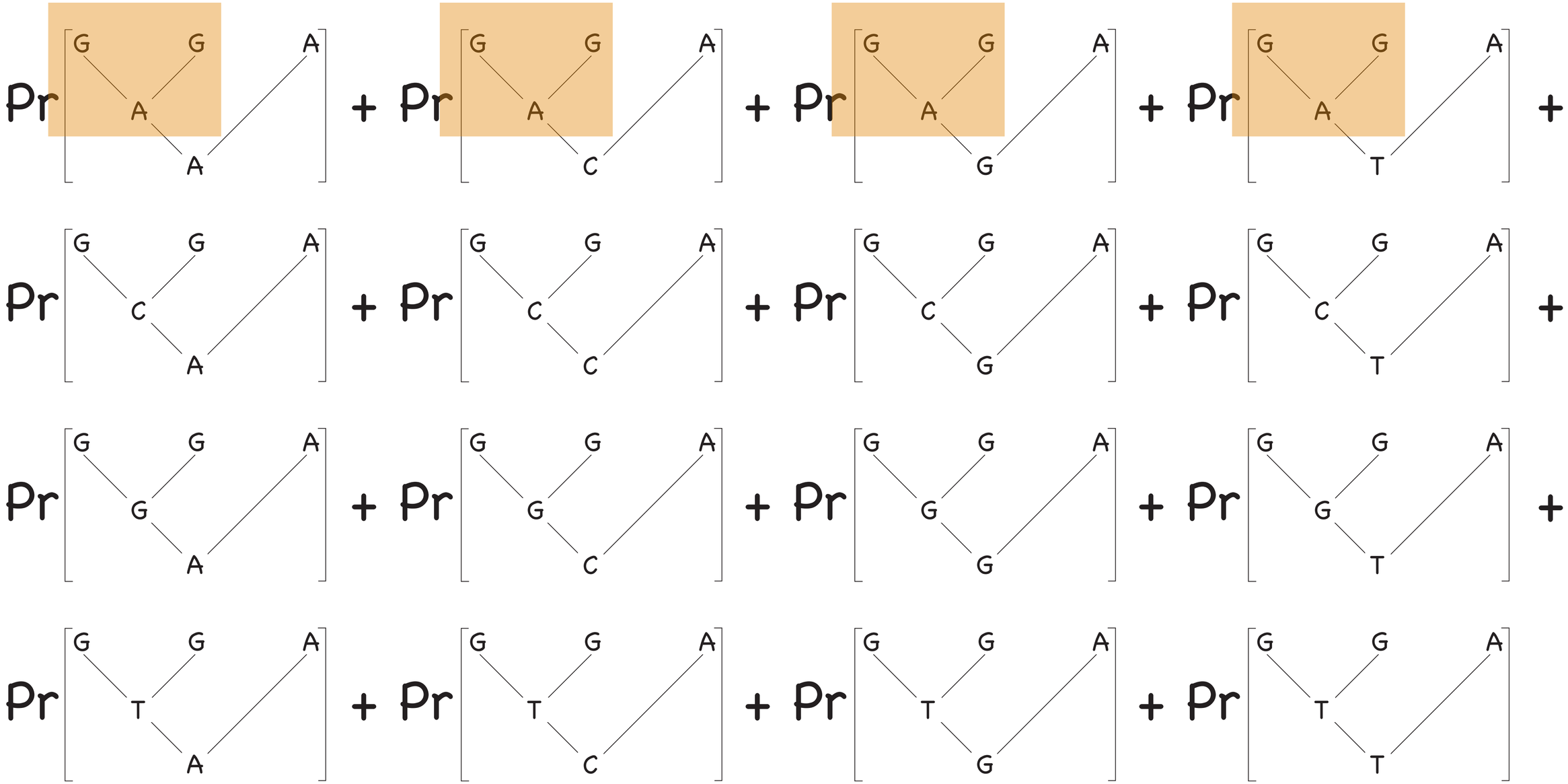


Just suppose for now we know the ancestral states at internal nodes

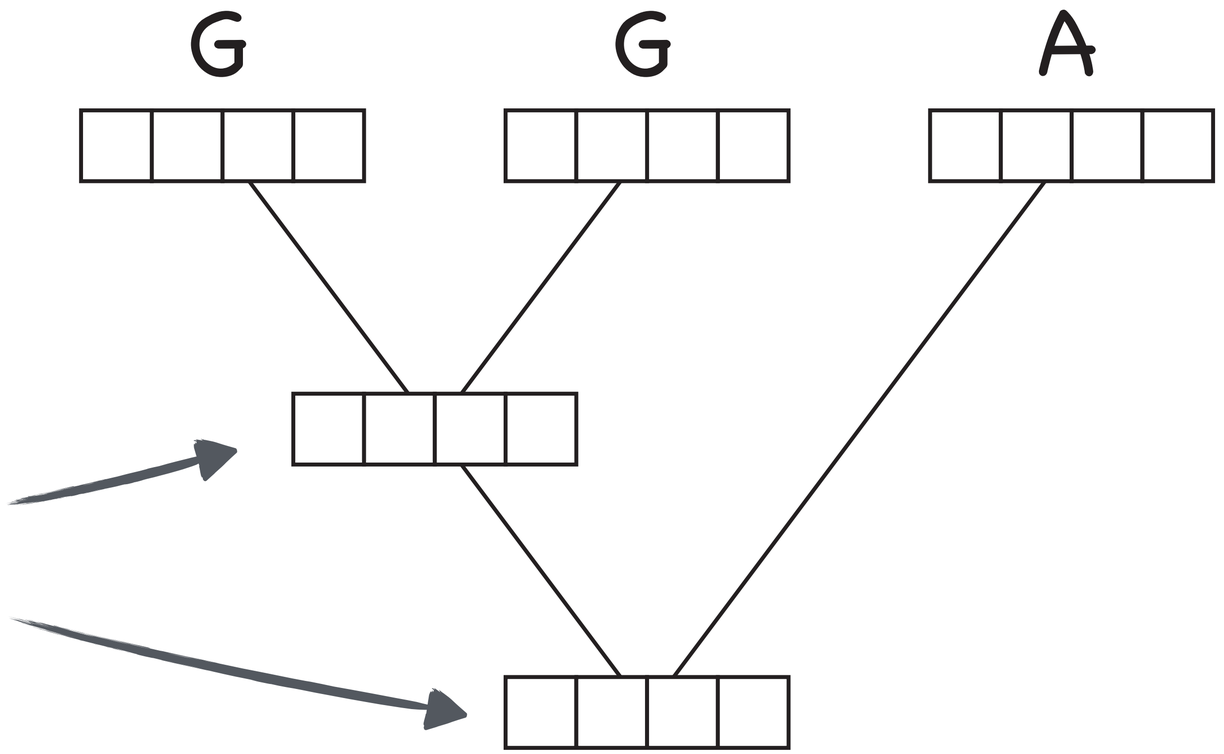
$\pi_A \times P_{AA}(v_1) \times P_{AA}(v_2) \times P_{AG}(v_3) \times P_{AG}(v_4)$

$P_{ij}(v)$ — transition probabilities

π_i — stationary frequencies

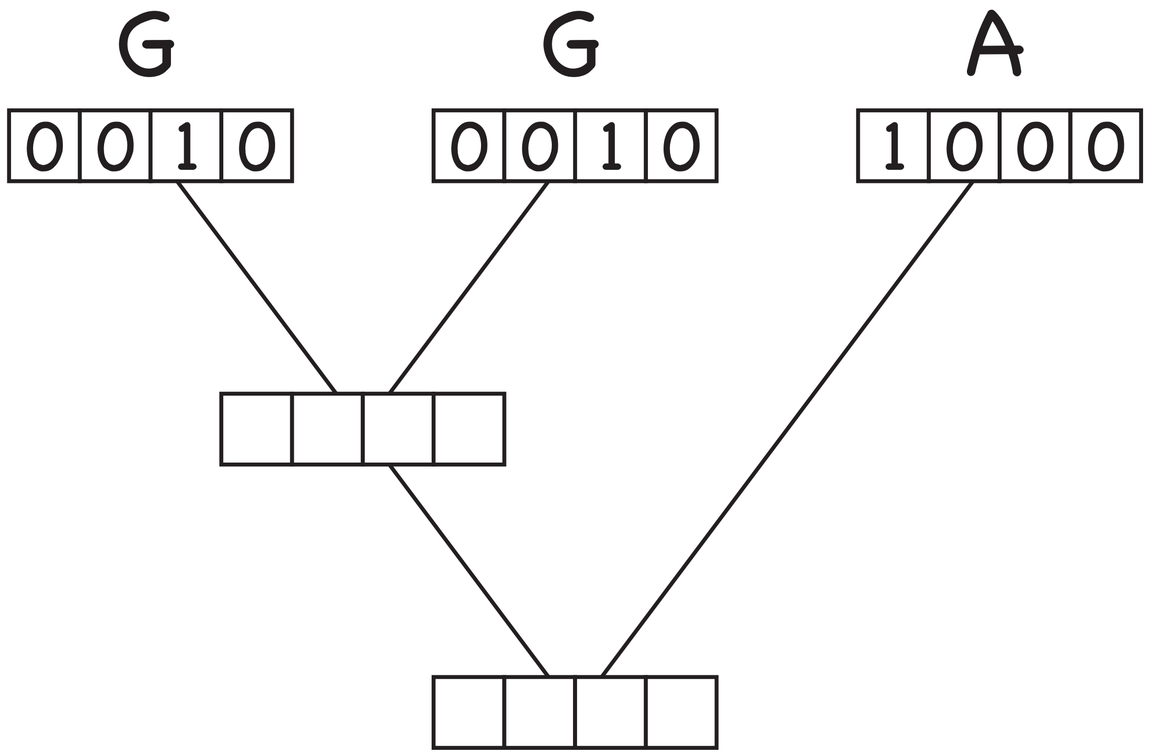

Pr [G G A / A / A] + Pr [G G A / A / C] + Pr [G G A / A / G] + Pr [G G A / A / T] +
Pr [G G A / C / A] + Pr [G G A / C / C] + Pr [G G A / C / G] + Pr [G G A / C / T] +
Pr [G G A / G / A] + Pr [G G A / G / C] + Pr [G G A / G / G] + Pr [G G A / G / T] +
Pr [G G A / T / A] + Pr [G G A / T / C] + Pr [G G A / T / G] + Pr [G G A / T / T]

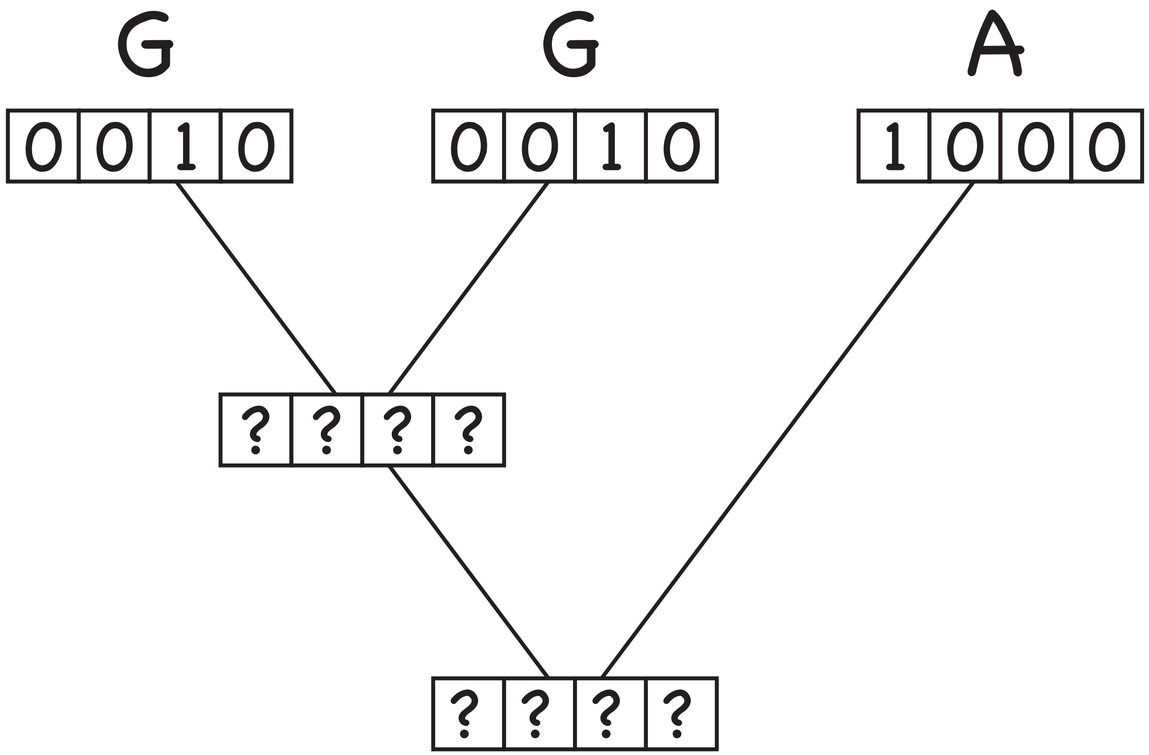
We’re going to calculate these probabilities based on the Pr calculated for the tips / nodes below



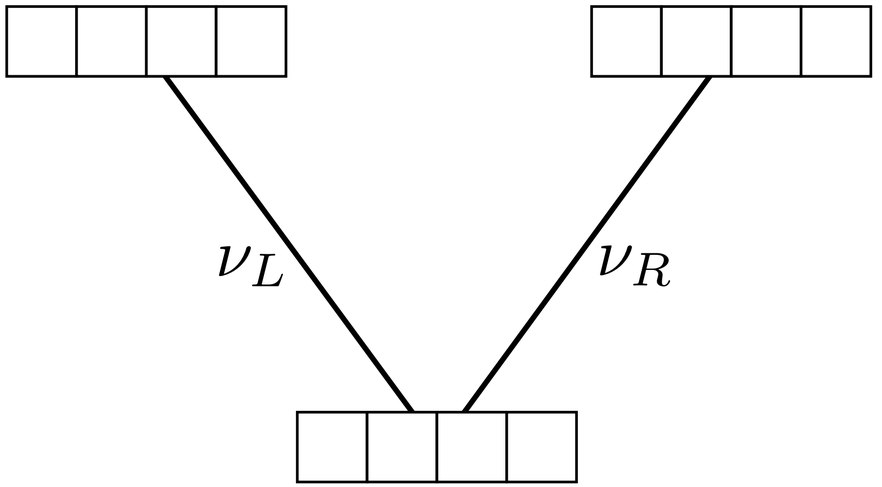
Felsenstein. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach.

First, we initialise the probabilities at the tips



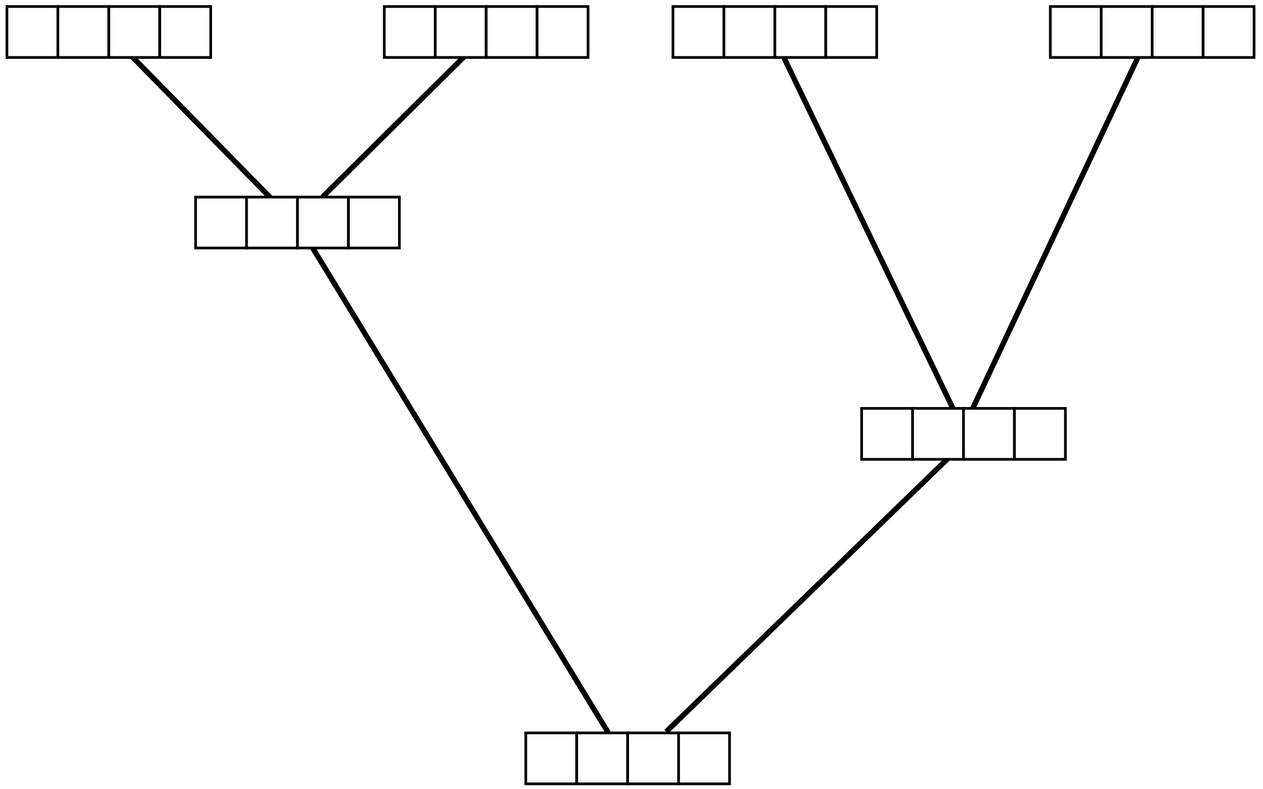


Next, we calculate the probabilities at each internal node



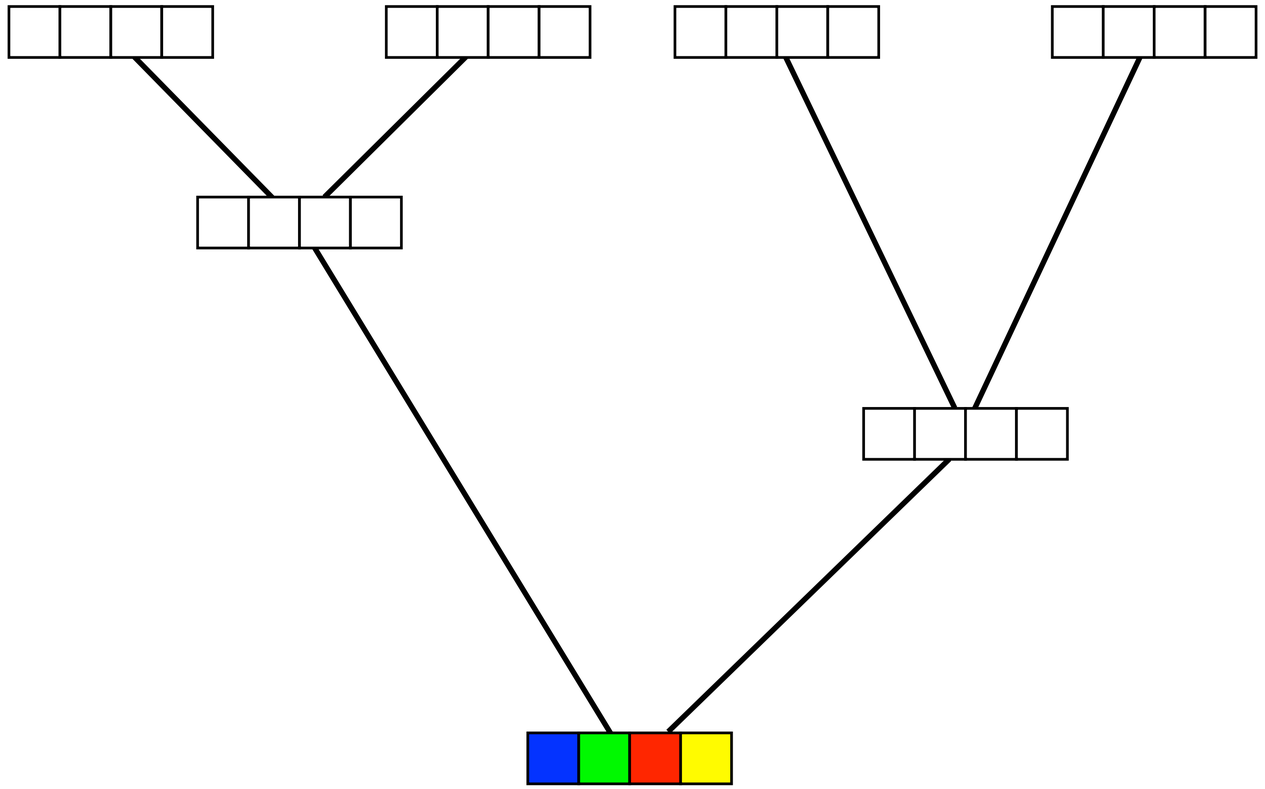
For each node, we take into account all possible changes, along both branches

$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$

n.b. This is the probability of a **single site** in your alignment!



Finally, we need to calculate & multiply this Pr across all sites in our alignment

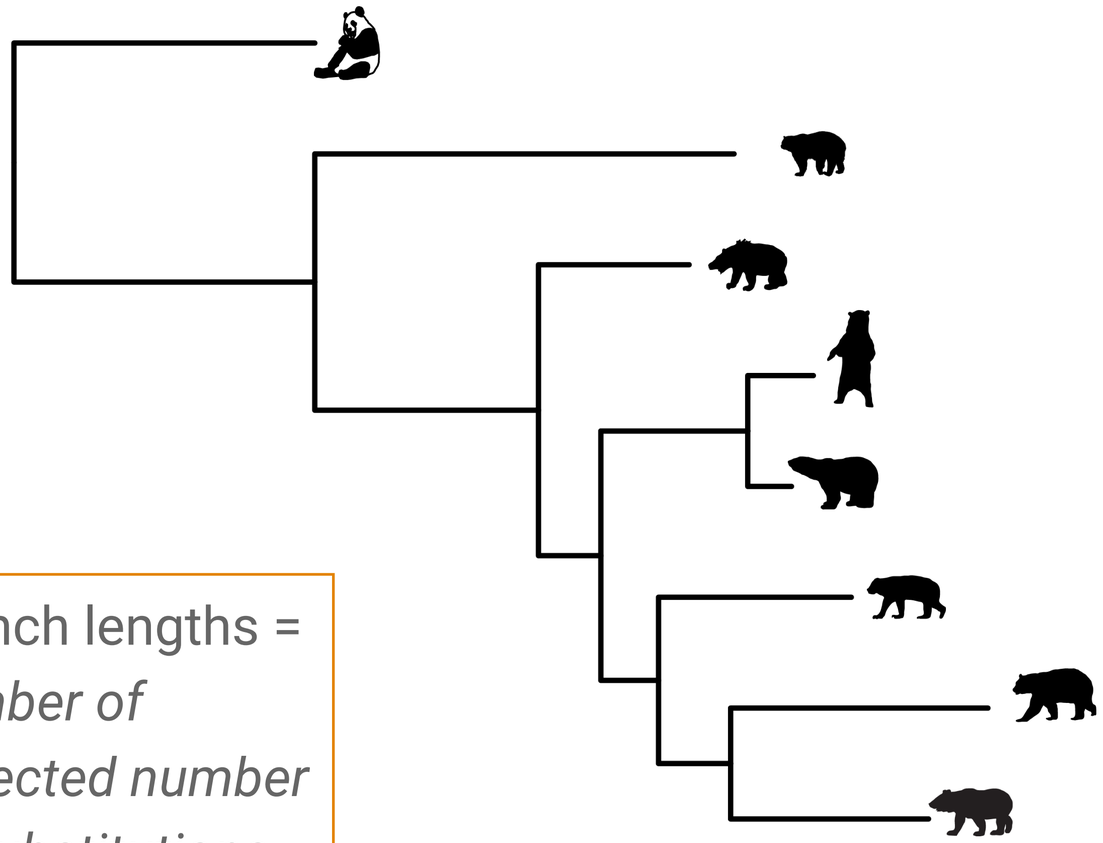
$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$

$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$

$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$

Another nice description of the pruning algorithm: *Harmon* (2019) Phylogenetic Comparative Methods, Chapter 8

Branch lengths



Branch lengths = *number of expected number of substitutions per site*

Branch lengths are a product of rate and time

Without temporal information we can only measure relative genetic distance

Maximum likelihood

How do we find the ‘best’ tree?

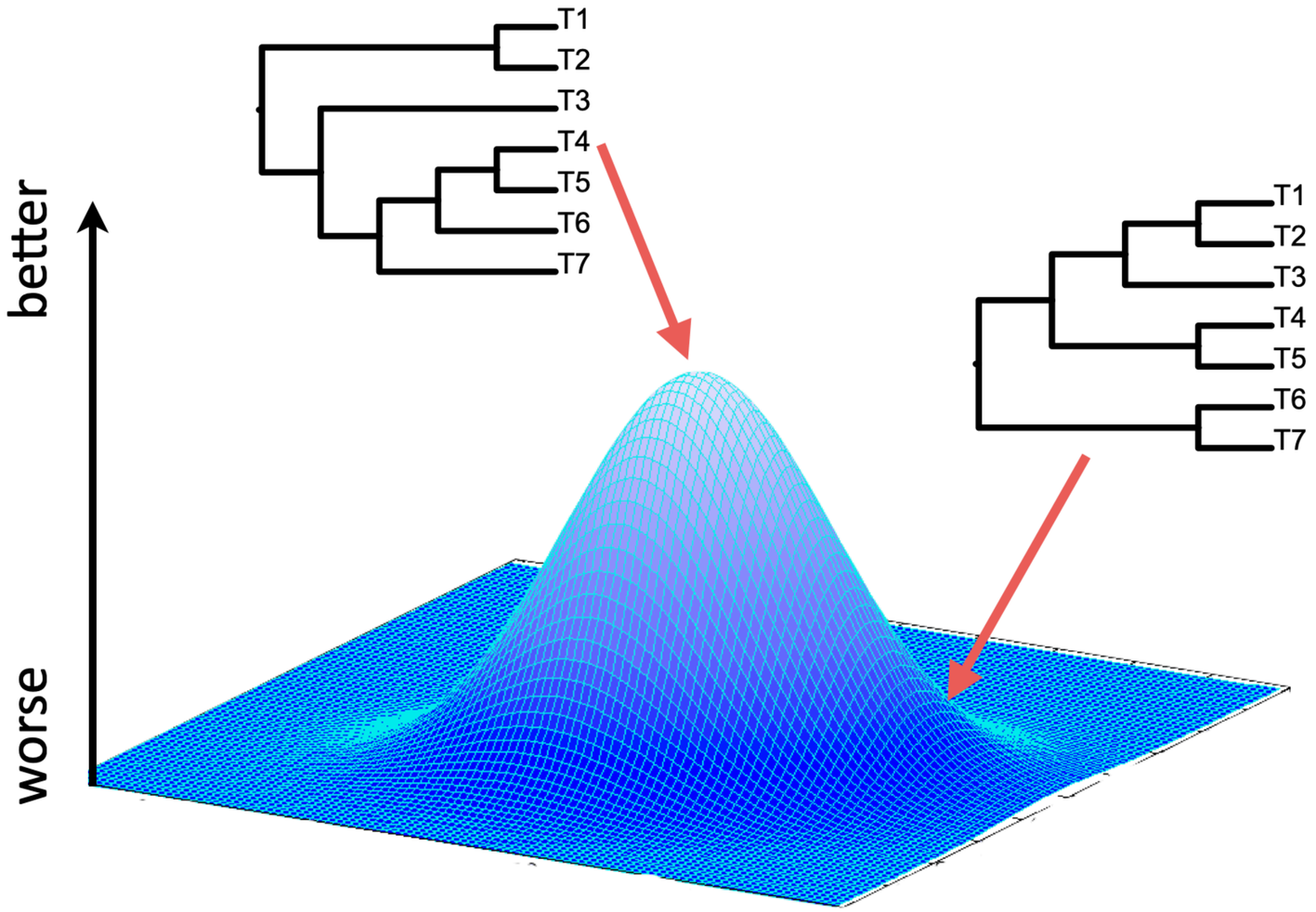


Image source Tracy Heath

It depends how you measure ‘best’

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

Both maximum likelihood and Bayesian inference are model-based approaches

Note these are not the only approaches to tree-building but they are the most widely used

Exercise

Tree building using likelihood

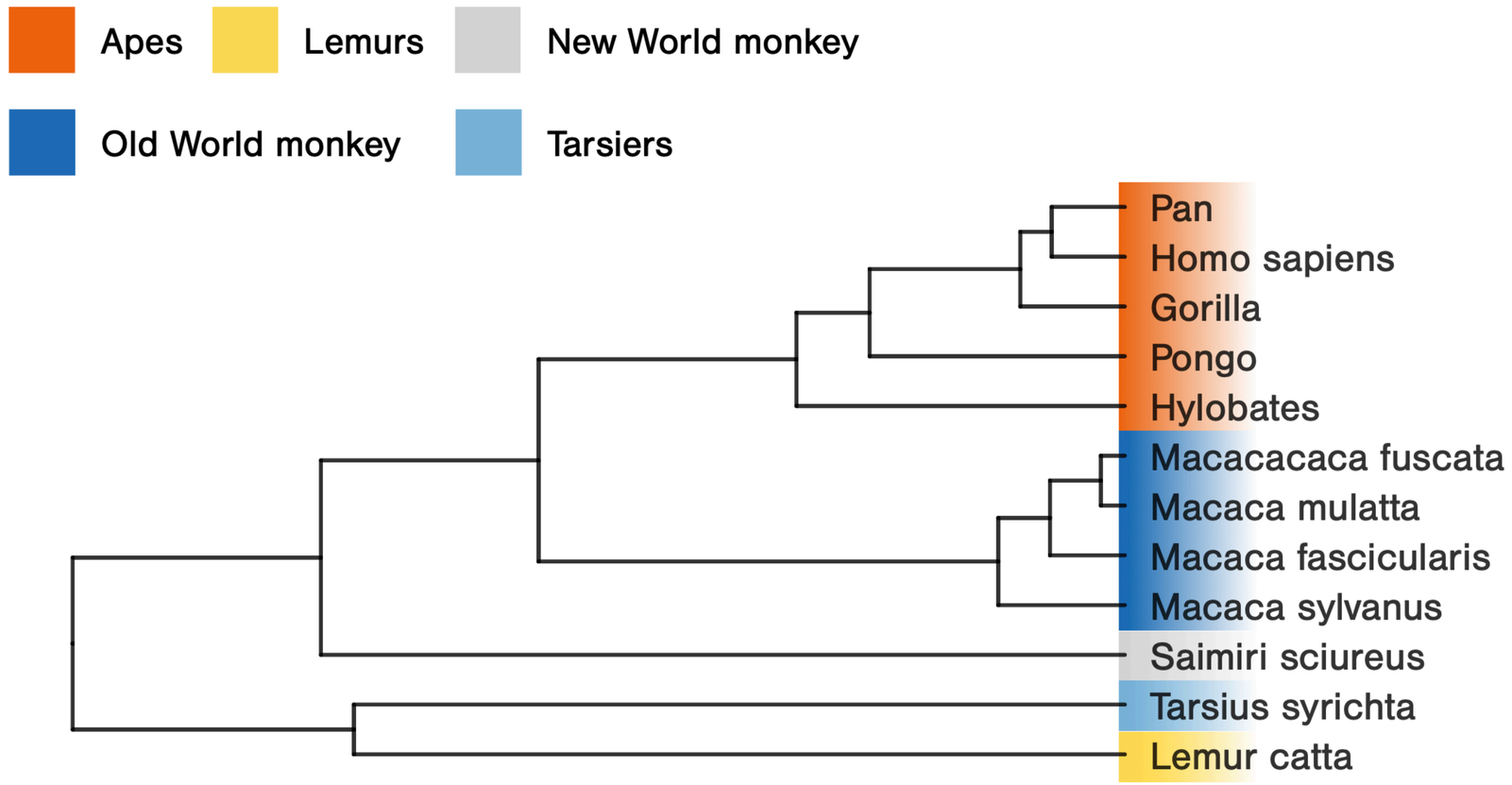


Image source Decoding Genomes *Stadler et al.* (2024)

How does your chosen tree compare to [published results]?

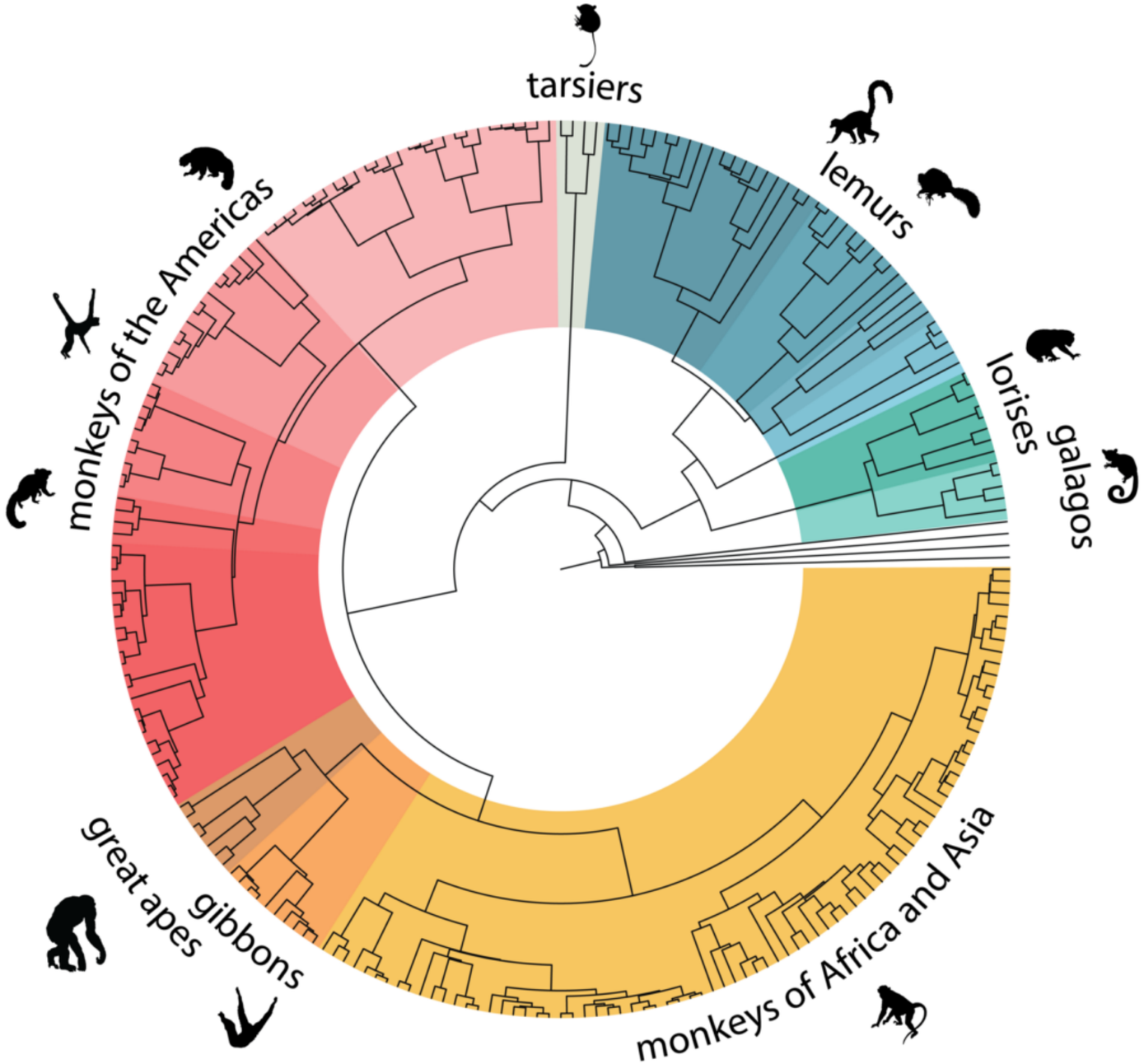
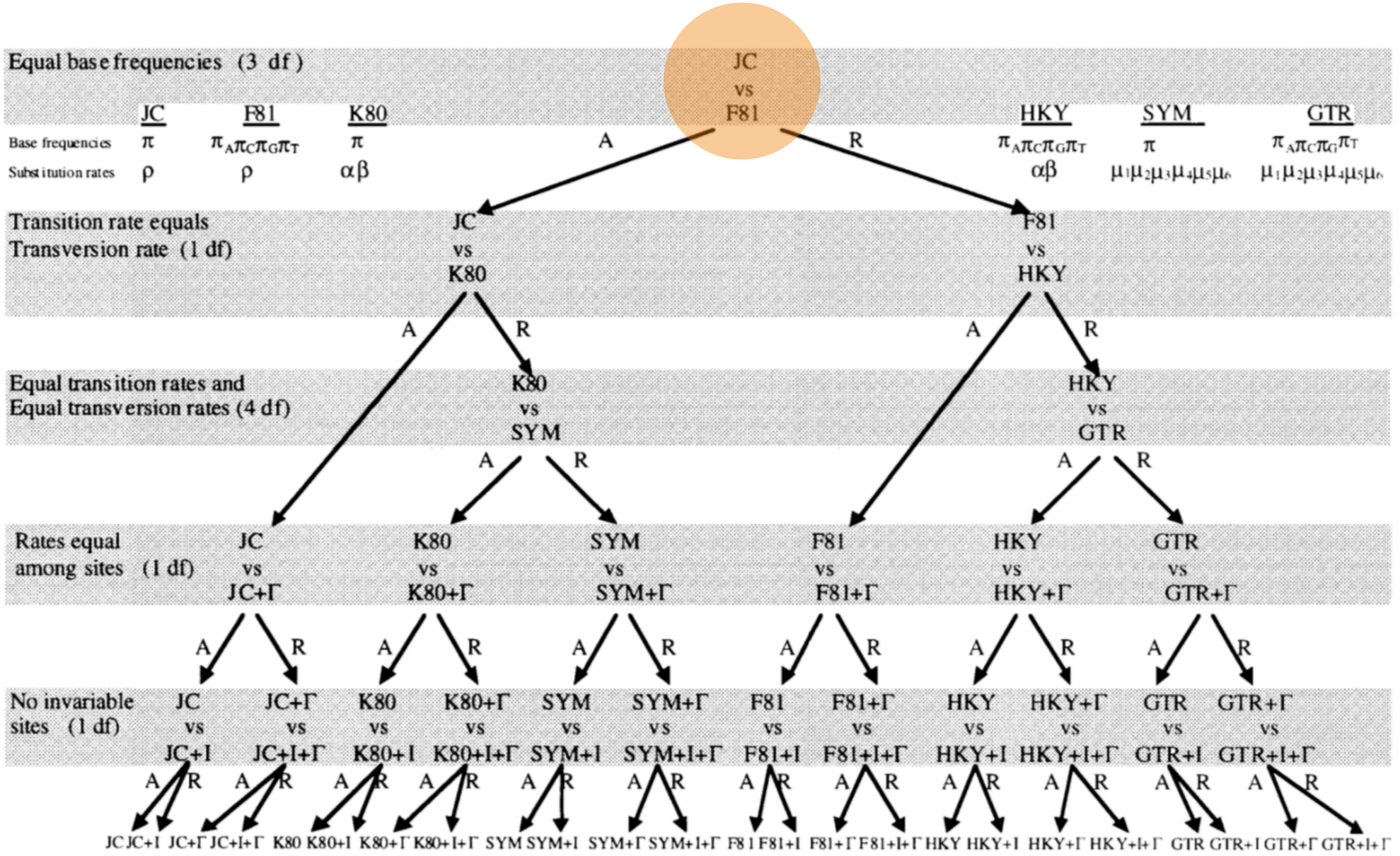


Image source genengnews.com

Other substitution models



Posada & Crandall (1998)

Base frequencies

The JC69 model assumes equal transition rates and equal base frequencies

Base frequencies are the proportion of each nucleotide in the dataset

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

The general time reversible model

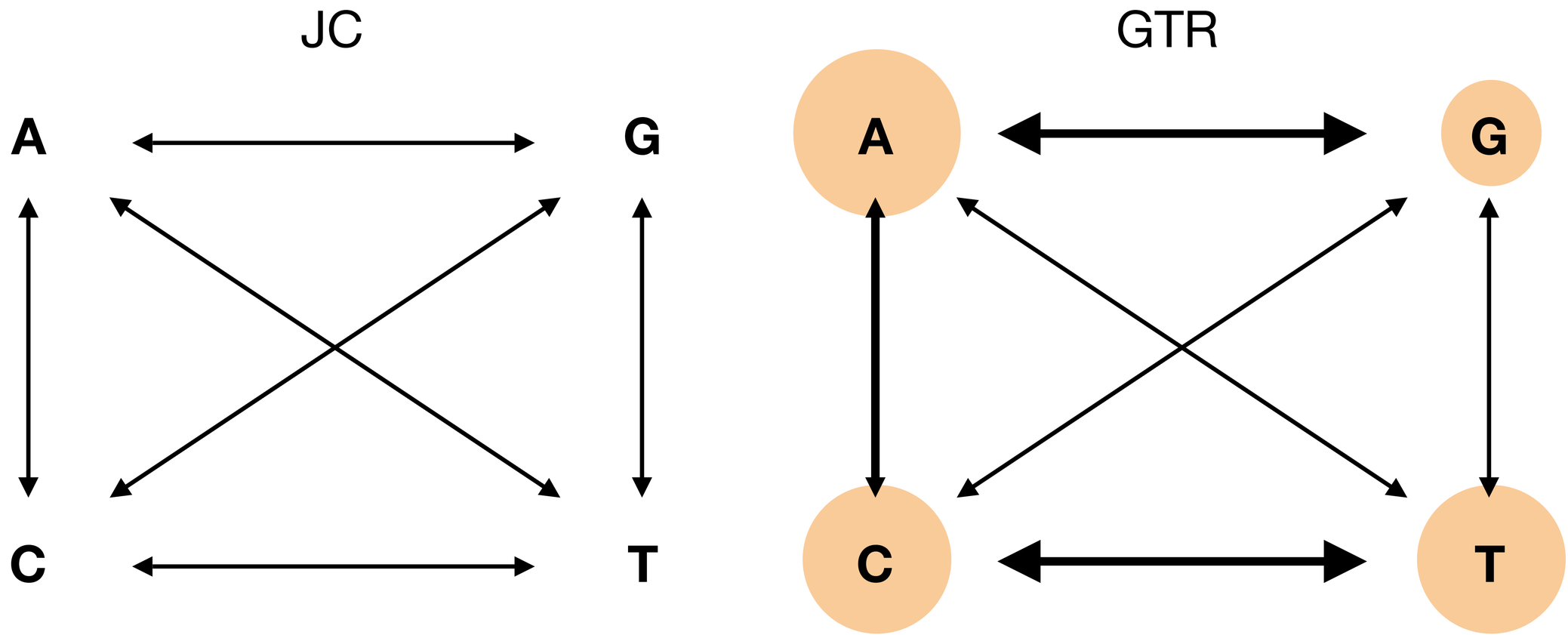
$$Q = \begin{pmatrix} * & \mu_{AG}\pi_G & \mu_{AC}\pi_C & \mu_{AT}\pi_T \\ \mu_{GA}\pi_A & * & \mu_{GC}\pi_C & \mu_{GT}\pi_T \\ \mu_{CA}\pi_A & \mu_{CG}\pi_G & * & \mu_{CT}\pi_T \\ \mu_{TA}\pi_A & \mu_{TG}\pi_G & \mu_{TC}\pi_C & * \end{pmatrix}$$

Allows for unequal transition rates (μ) *and* unequal base frequencies (π)

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

Tavaré (1986)

The JC versus GTR models

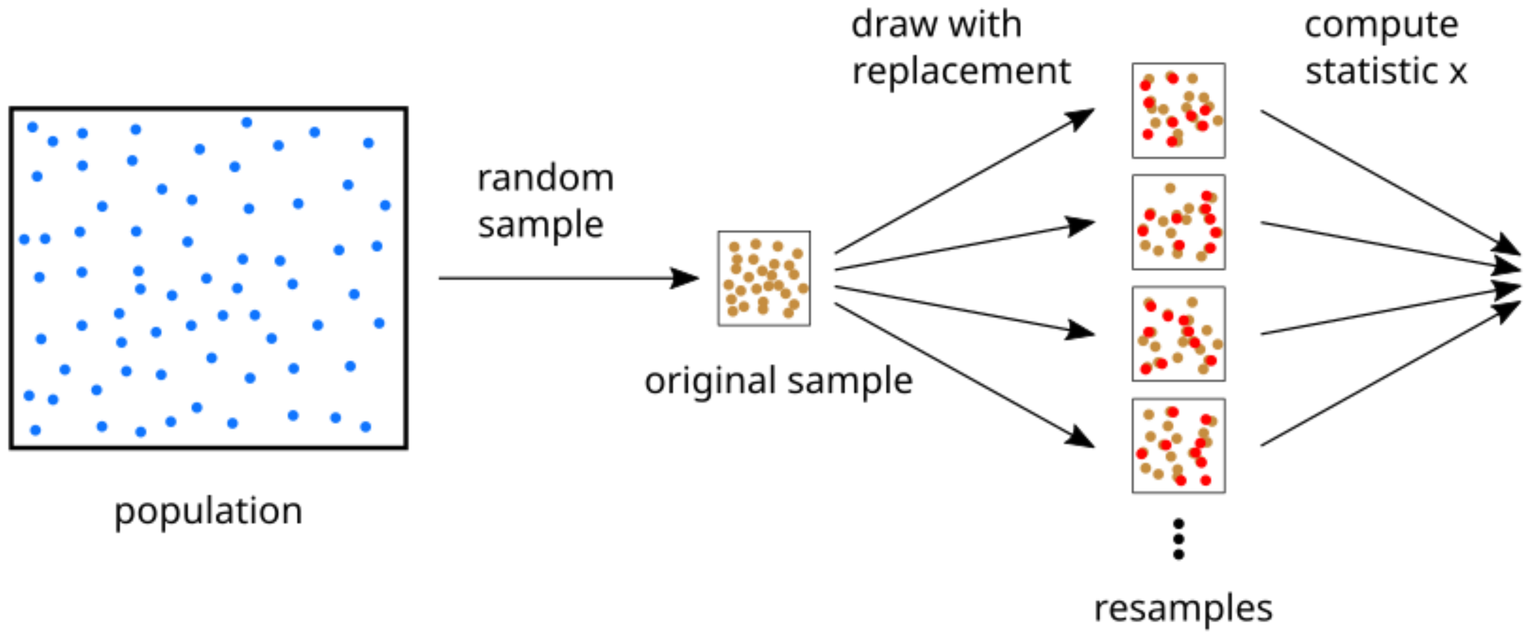


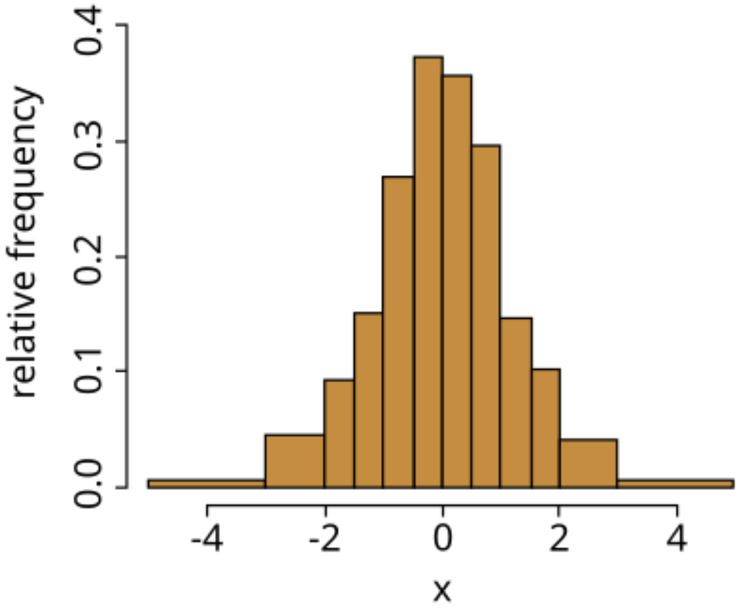
Bootstrapping

Bootstrapping

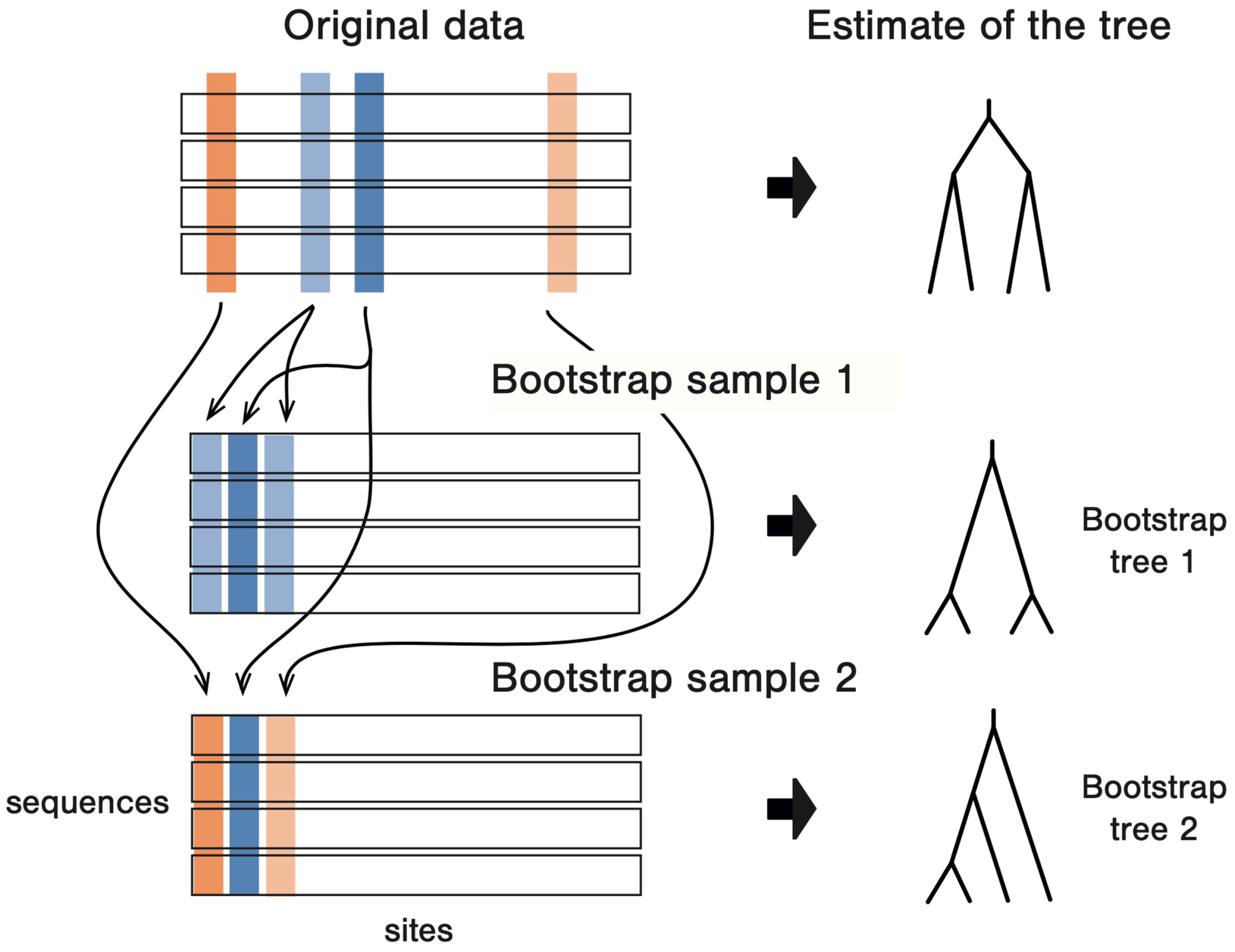
Resampling with replacement

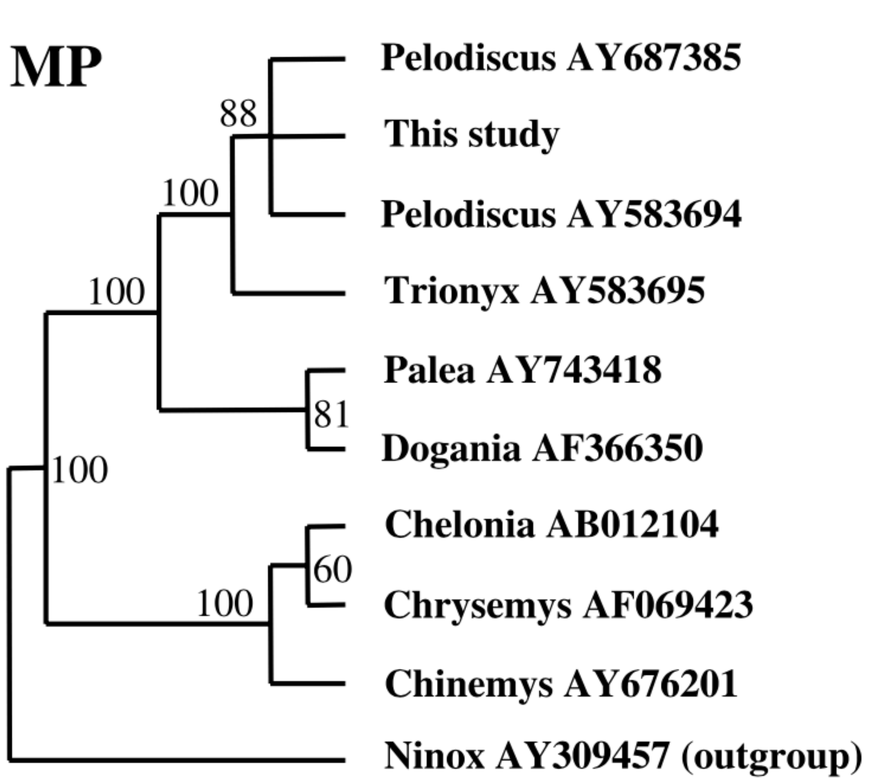
Image source Wiki

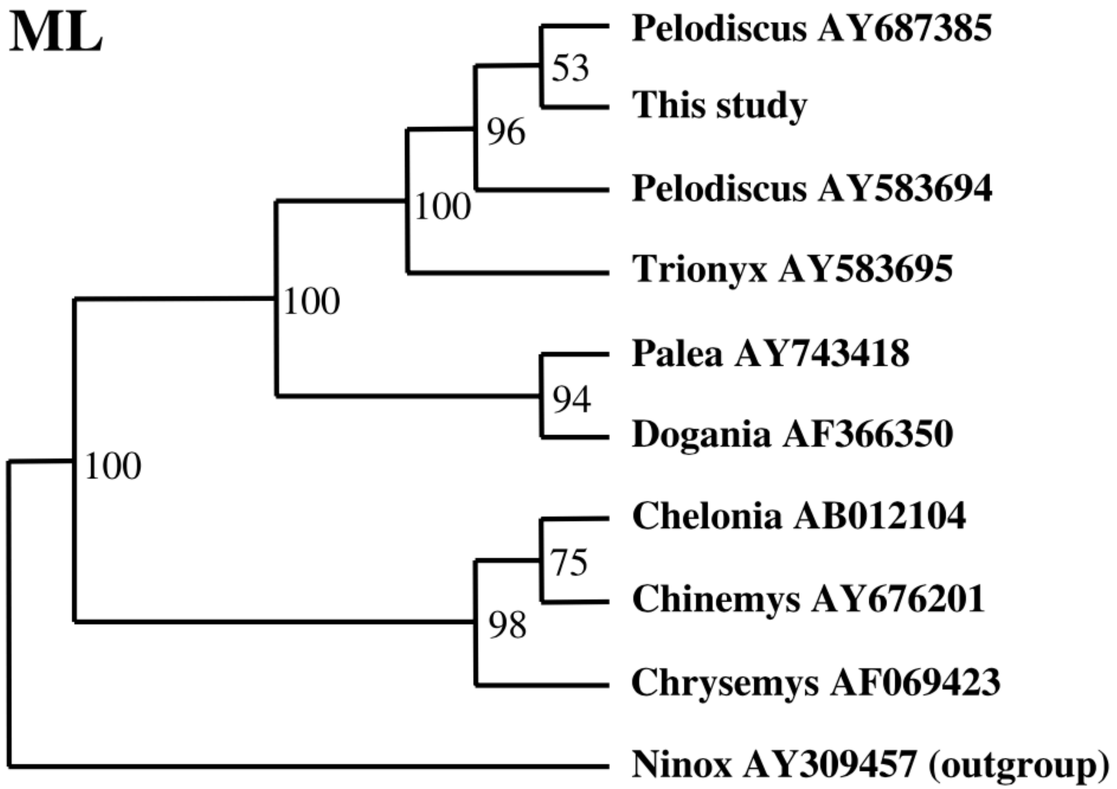




In our case, the random resamples are alignments *X* is the tree


Original data
Estimate of the tree
Bootstrap sample 1
Bootstrap sample 2
Bootstrap tree 1
Bootstrap tree 2
sequences
sites





Jung et al. (2006)

Next

Install the following software:

RevBayes

FigTree

Tracer