

Molecular Evolution & Phylogenetics

Traits, phylogenies, evolutionary models
and divergence time between sequences

Jean-Baka Domelevo Entfellner
Bioinformatics Community of Practice,
BecA-ILRI Hub, September 2018

Learning Objectives

- understand the concepts and the vocabulary pertaining to phylogenetic trees
- know from what type of biological data it is possible to build phylogenies
- understand the concept of evolutionary divergence between sequences
- understand what evolutionary models are and know how to use them

Learning Outcomes

- be comfortable in discussing with a colleague about the parameters and details involved in running a phylogenetic analysis
- be able to prepare a phylogenetic analysis (choice of the data and of an appropriate model of evolution), upstream of the phylogenetic inference process itself

Molecular Evolution & Phylogenetics

**Traits, taxa,
phylogenetic trees**

Some vocabulary

- a phylogeny, a.k.a. phylogenetic tree, is a **tree** (connex acyclic graph) whose edges represent direct evolutionary links and nodes represent past or present taxa
- a **taxon** (pl. **taxa**) is a member of a taxonomic representation (species, virus strain, variety, identified subpopulation, etc)
- terminal nodes (**leaves** of a tree) are the extant taxa, a.k.a OTU (Operational Taxonomic Unit)
- a **trait** or **character** is any biological feature that can be compared across taxa
- traits can be **qualitative/categorical** variables (e.g. aligned nucleotides or aminoacids) or **quantitative**, in which case they can be discrete, semi-continuous or continuous (e.g. number of repeats of a microsatellite, frequency of an allele, diameter of the skull, etc)

Phylogenetic trees

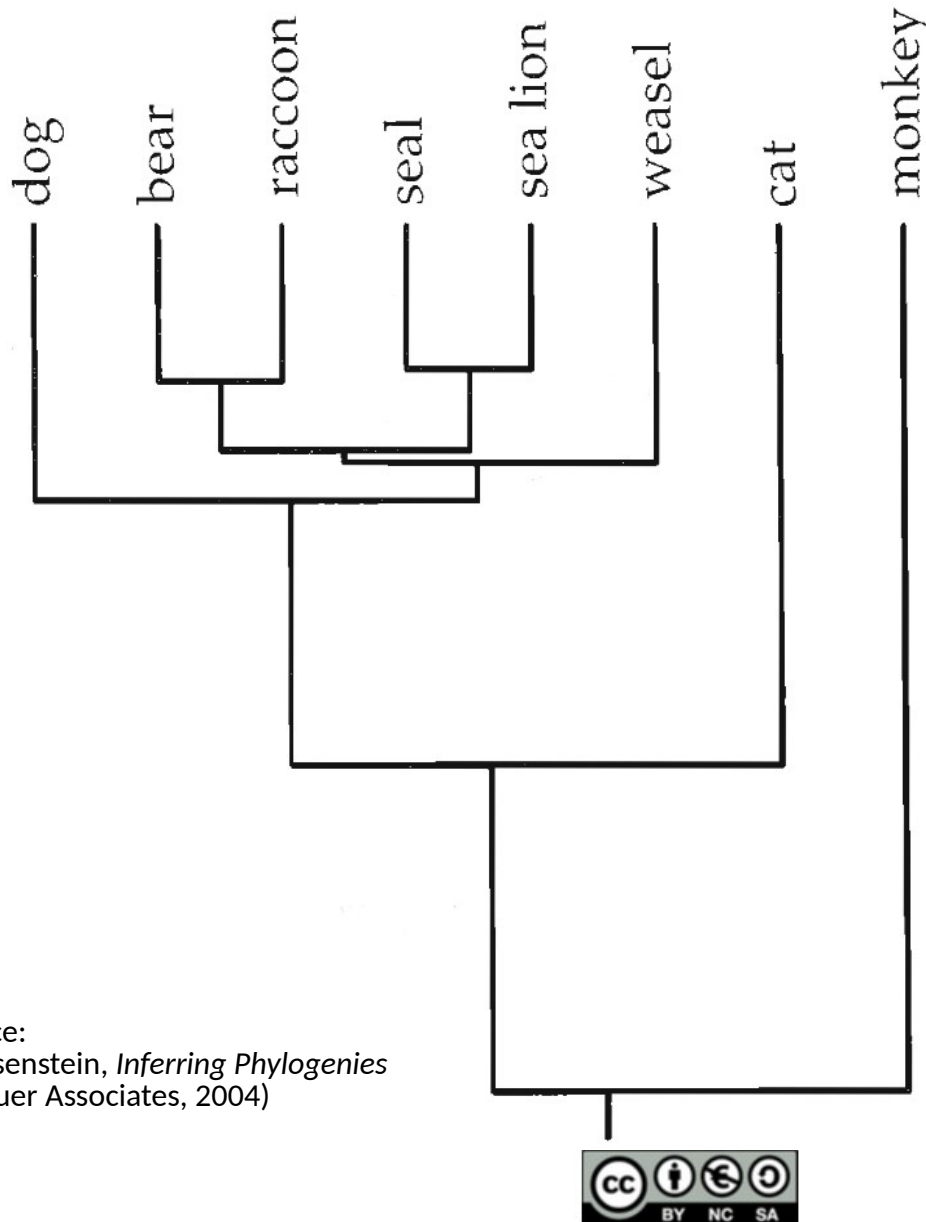
A mathematician's definition:

- tree **topology**: connex acyclic graph $G = (V, E)$
 - V : set of vertices or nodes (e.g. species, virus strains, genes)
 - E : set of edges or branches (materialising evolution)
- acyclicity and connexity impose that there is **exactly one path between any two nodes of the tree**
- the **degree** of a node is the number of edges it is connected to
- **branch lengths** are not always present. If present, they correlate with the amount of time elapsed (branch length unit: expected number of character substitutions per site)
- tree is **rooted** if we know where is the most ancient node (evolutionary origin), **unrooted** otherwise.

Some more vocabulary and remarks

- in a **binary tree**, all non-terminal nodes have 3 neighbours (or one parent node and two children in case the tree is rooted)
- a node with degree > 3 is called a **multifurcation**, aka. **polytomy**
- multifurcations represent well geological “moments” of sudden adaptive radiation, e.g. the Cambrian explosion
- trees can be built based on a single trait (e.g. one phenotypic characteristic) or (most commonly) on a set of characters (e.g. 1000 **aligned** sites)

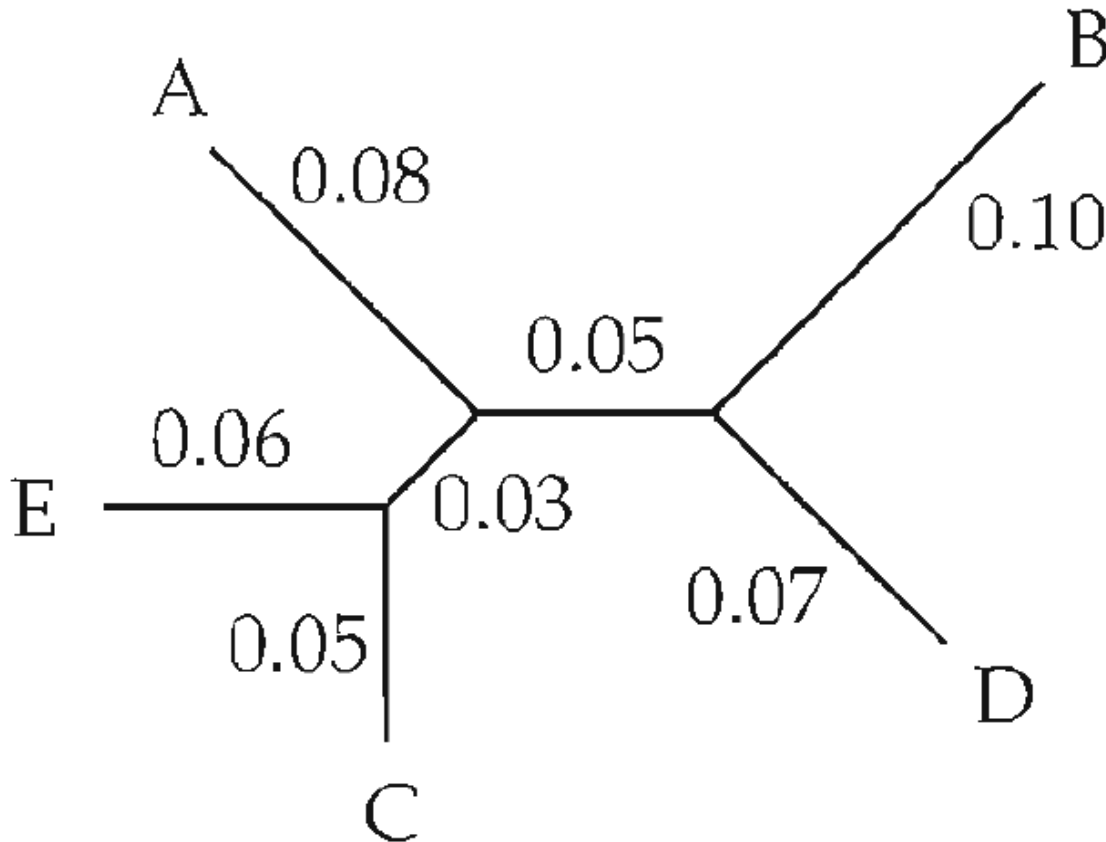
A rooted binary tree



- in a **rooted binary tree**, every internal node (except the root) has one father and two descendants (sons)
- n taxa
- $2n-1$ vertices
- $2n-2$ edges

Source:
J. Felsenstein, *Inferring Phylogenies*
(Sinauer Associates, 2004)

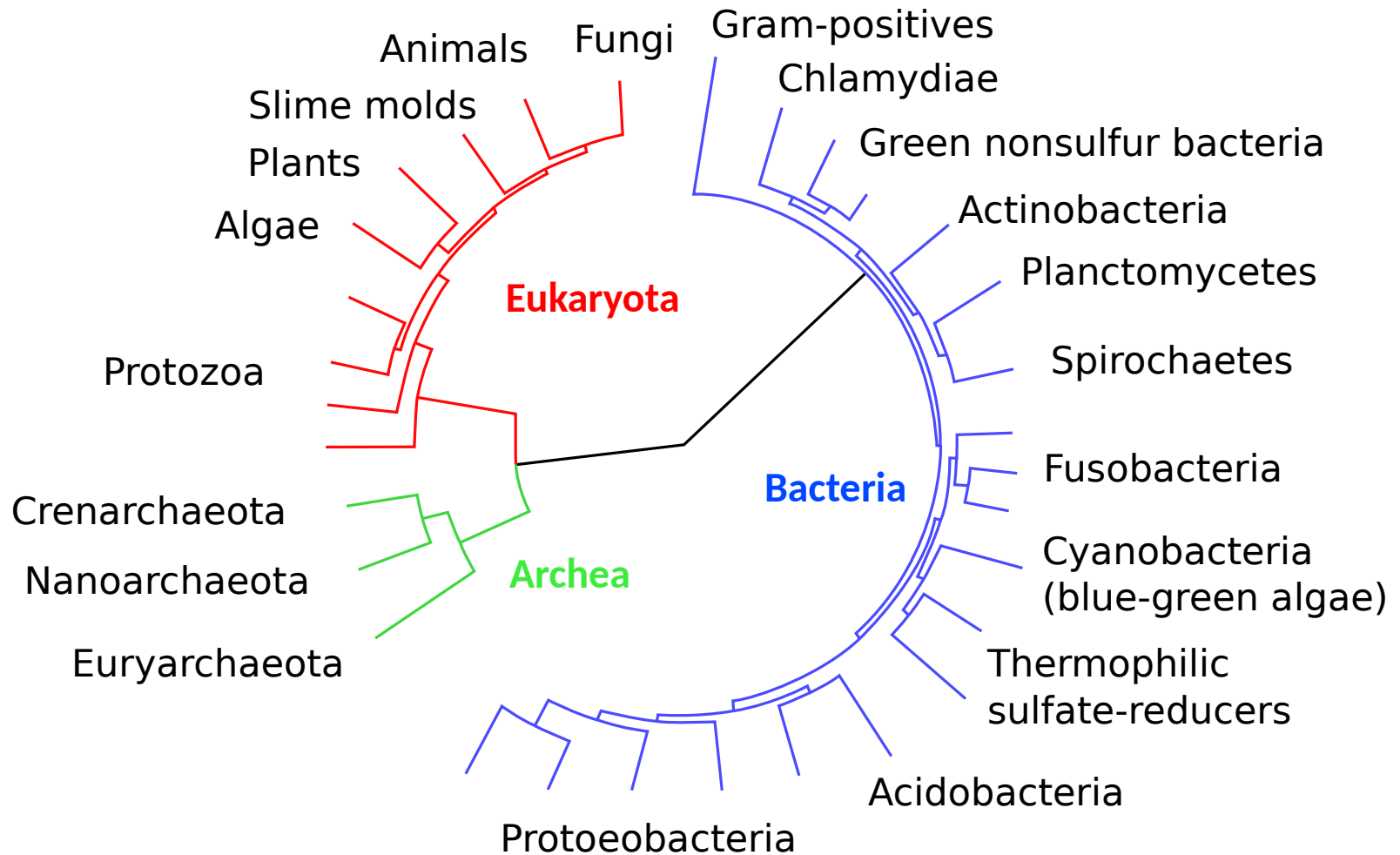
An unrooted binary tree



- in an **unrooted binary tree**, every internal node has three neighbours
- n taxa
- $2n-2$ vertices
- $2n-3$ edges
- this tree has branch lengths

Source:
J. Felsenstein, *Inferring Phylogenies*
(Sinauer Associates, 2004)

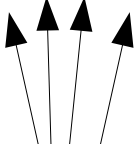
Circular trees: a fancy representation



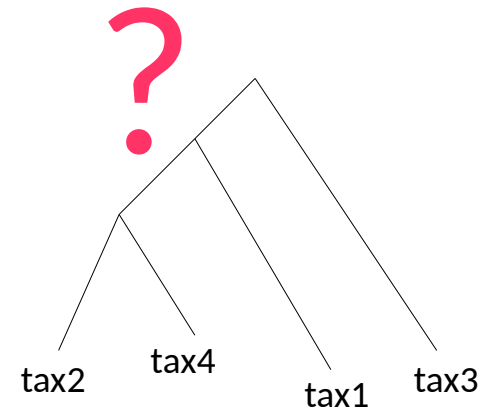
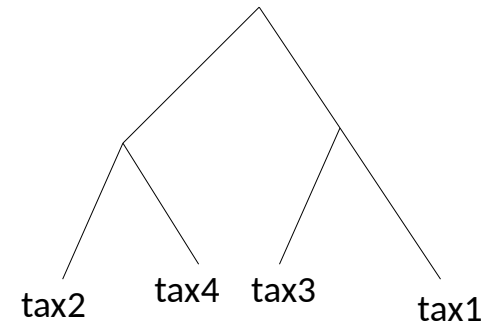
The tree inference problem

Problem: Assuming common descent, how to derive the “most probably correct” tree from the knowledge of the traits in the extant taxa (leaves)?

tax1: ACGG
tax2: AAGG
tax3: AAGT
tax4: GAGG


input data =
aligned nucl.

phylogenetic inference

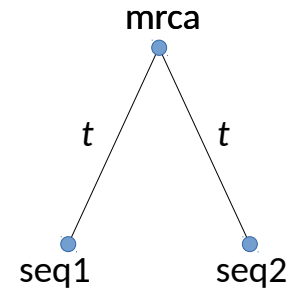


Simpler problem: evolutionary distance between two sequences

Let us consider two aligned sequences and the elementary tree linking them:

seq1: ACGGGTATTG

seq2: ACGATTATTT



We want to know the evolutionary time ($= 2t$) separating seq1 and seq2 from their MRCA (Most Recent Common Ancestor).

Naive distance = edit distance = $3/10 = 0.3$

Shortcomings of the edit distance between two sequences

- **all mutations** (transitions or transversions) yield **same** distance, while from a biochemical standpoint, transitions (within purines, $A \leftrightarrow G$, or within pyrimidines, $C \leftrightarrow T$) are “less costly” than transversions.
- temporal chains of successive mutations on the same site result in **underestimation of the true evolutionary distance** by the observed (edit) distance, e.g. $A \rightarrow G \rightarrow T$ (edit distance 1, true distance 2) or $A \rightarrow G \rightarrow A$ (edit distance 0, true distance 2)
- we have no means to translate edit distances into actual biological time, and date the MRCA in Mya (= “Million years ago”): problem of **calibration**

Molecular Evolution & Phylogenetics

Modeling evolution

What is a model, in general?

- **simplification of the reality** based on current human understanding
 - must be **simple enough** to be computationally **tractable**
 - must be **complex enough** to lead to measures, findings or simulations in approximate **agreement with our observations** (empirical truth)
-
- tradeoff between the complexity of a model and its tractability (a.k.a. “computability”, which implies practical usefulness)
 - Einstein: “It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.”
or: “Everything should be made as simple as possible, but not simpler.”

What is a model of evolution, specifically?

- deals with **aligned sequences** (nucleotides or amino acids)
- all evolutionary models are stochastic: they predict **probabilities of change**, without yielding any certainty
- models evolution in terms of **character substitutions**: enables to calculate e.g. $P_t(A \rightarrow C)$
- is defined with a certain number of **parameters** to be estimated from some training data
- essentially all models in use are **Markovian** (memory-less): the fate of a character depends only on its **present state**, not on its previous history of mutations.
- some models are **time-reversible**, some others are not.

Stochasticity

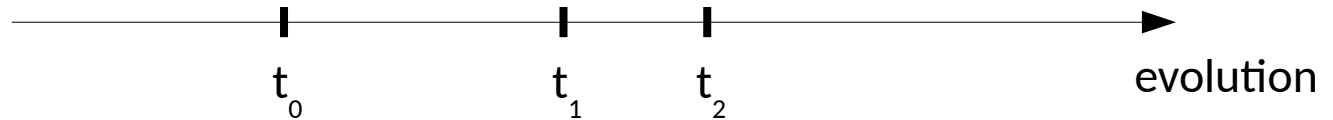
Stochastic (probabilistic) models are based on stochastic processes, i.e. processes defining the trajectory of some **random variables**.

Examples of models:

- non stochastic: “nucleotides mutate once every thousand years according to the cycle: $A \rightarrow C, C \rightarrow G, G \rightarrow T, T \rightarrow A$ ”
- stochastic: “nucleotide substitutions occur randomly at a constant rate α (expected number of substitutions per site per time unit), equal for all substitutions.” (JC69)

Markovian property

The evolutionary history of a character X doesn't further impact its future trajectory, only its present state matters to determine its future (memory-less property):



$$Pr(X(t_2) | \{X(t_0), X(t_1)\}) = Pr(X(t_2) | X(t_1))$$

Stationarity

Stationary models accept limits for the frequencies of the characters, after the stochastic model has run for an “infinite” amount of time. For instance,

$$\lim_{t \rightarrow \infty} Pr_t(A \rightarrow C) = \pi_C$$

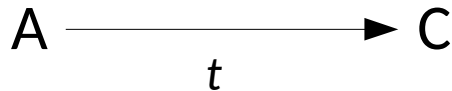
$$\lim_{t \rightarrow \infty} Pr_t(G \rightarrow C) = \pi_C$$

probability to have a C after an “infinite” amount of time **does not** depend on the original nucleotide



Time-reversibility

Time-reversible models give no preference to evolution “in one direction” compared to “in the opposite direction”, e.g. on some branch of a tree: they give no indication about the orientation of the time arrow.



$$Pr(A) Pr_t(A \rightarrow C) = Pr(C) Pr_t(C \rightarrow A)$$

Consequence: time-reversible models cannot lead by themselves to a rooted tree. Unrooted trees are built first, and then rooted by some extra information (outgroup).

Exchangeability rates

When dealing with **time-reversible models**, one has:

$$\pi_A Pr_t(A \rightarrow C) = \pi_C Pr_t(C \rightarrow A)$$

And using instantaneous rates:

$$\pi_A q_{AC} = \pi_C q_{CA}$$

so we can define symmetric **exchangeabilities**:

$$\frac{q_{AC}}{\pi_C} = \frac{q_{CA}}{\pi_A} = s_{AC} = s_{CA}$$

Molecular Evolution & Phylogenetics

**Simplest evolutionary model
on nucleotide data:**

JC69

JC69: Jukes & Cantor, 1969

- single constant rate for all substitutions: model has only one parameter $\alpha \in \mathbb{R}^+$ (odd situation for $\alpha = 0$: no evolution).
- only substitutions, no insertions and deletions
- \Rightarrow model explains or creates only ungapped alignments
- \Rightarrow model can be trained on relatively small datasets
- - oversimplification of the reality: observations give credit to varying substitution rates depending on the position in the genome or in the Tree of Life
- + very simple maths to calculate the probability of one sequence having evolved into another one

JC69 instantaneous rate matrix (Q)

		to:			
		A	C	G	T
from:	A	-3α	α	α	α ← q_{AT}
	C	α	-3α	α	α
	G	α	α → q_{GC}	-3α	α
	T	α	α	α	-3α

matrix Q of **instantaneous rates**: over a very short duration dt , probability $\alpha \cdot dt$ that e.g. a nucleotide A mutates into a C

Dynamics of nucleotide evolution under JC69

- We consider a population of N nucleotides evolving independently from $t = 0$ under the JC69 model.
- $n_A(t)$: count of adenines at time t
- Calculus leads to:
$$n_A(t) = \left[n_A(0) - \frac{N}{4} \right] e^{-4\alpha t} + \frac{N}{4}$$

(and symmetric expressions for C, G and T)
- JC69 is a **stationary** model: when $t \rightarrow \infty$, $n_A(t) \rightarrow N/4$

From instantaneous rates to probabilities of change

- we want to calculate from Q the probability of having two aligned nucleotides e.g. A and C separated by an evolutionary time t
- fundamental relation: $Pr_t(A \rightarrow C) = [e^{Qt}]_{A,C}$
- equilibrium probabilities for the JC69 model:

$$\pi_A = \pi_C = \pi_G = \pi_T = 0.25$$

Estimation of divergence times under JC69

- if p = proportion of observed differences between two nucleotide sequences (edit distance) and **d is the expected number of mutations per site** between the two sequences, we have:

$$d = -\frac{3}{4} \log\left(1 - \frac{4}{3} p\right) > p$$

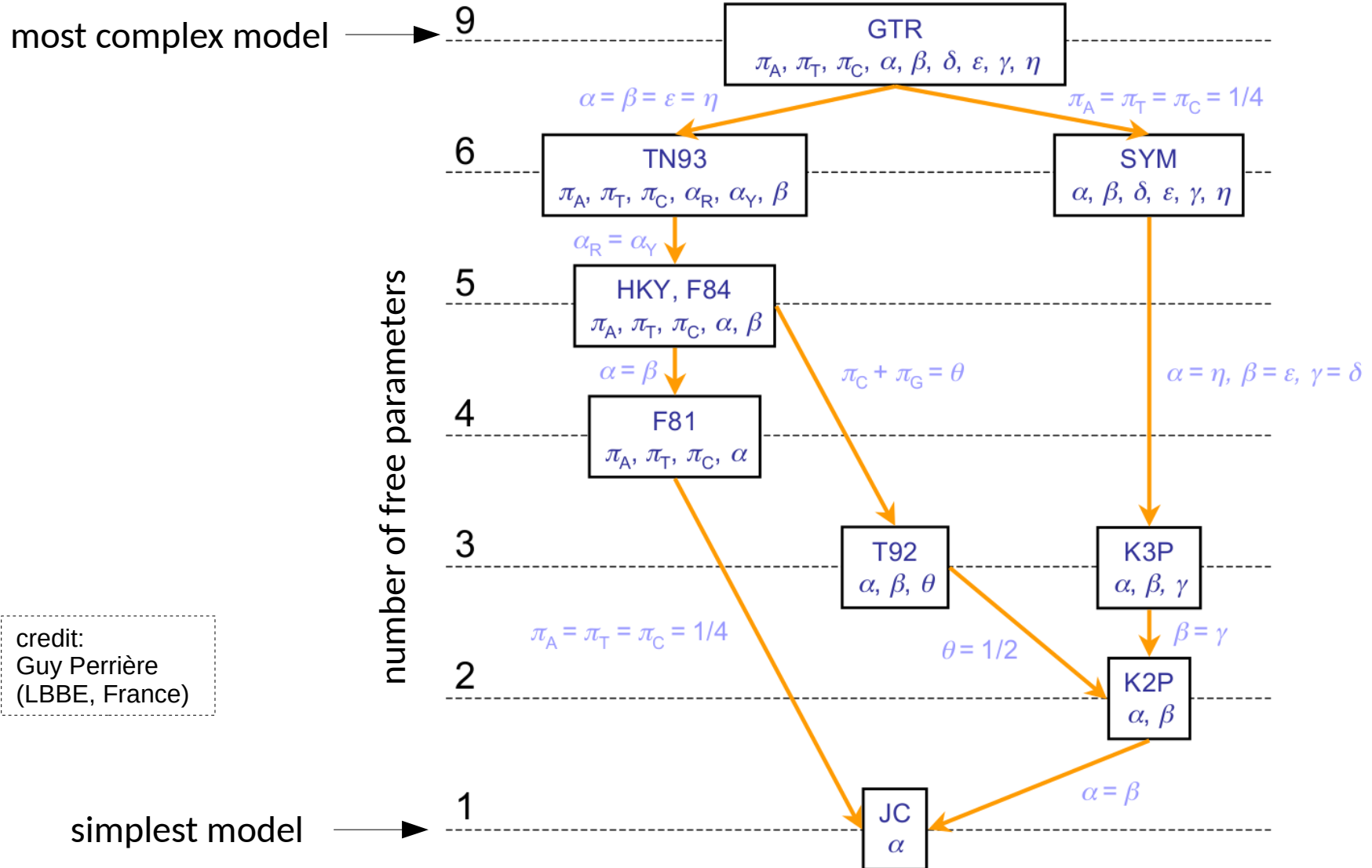
- in order to get to proper geological durations (millions of years), one needs to **calibrate** the rate of mutations (order of magnitude: 10^{-9} mutations per site and per year) using reliably dated data (e.g. fossils)

More complex models for nucleotide data

- K2P (2 parameters): equal equilibrium frequencies but two exchangeabilities (one for transitions and one for transversions)
- F81 (4 parameters): one single exchangeability rate but four different equilibrium frequencies π_A , π_C , π_G and π_T
- F84/HKY (5 parameters): different equilibrium frequencies and two exchangeabilities (one for transitions and one for transversions)
- GTR (9 parameters): general time-reversible, most complex time-reversible model on nucleotides

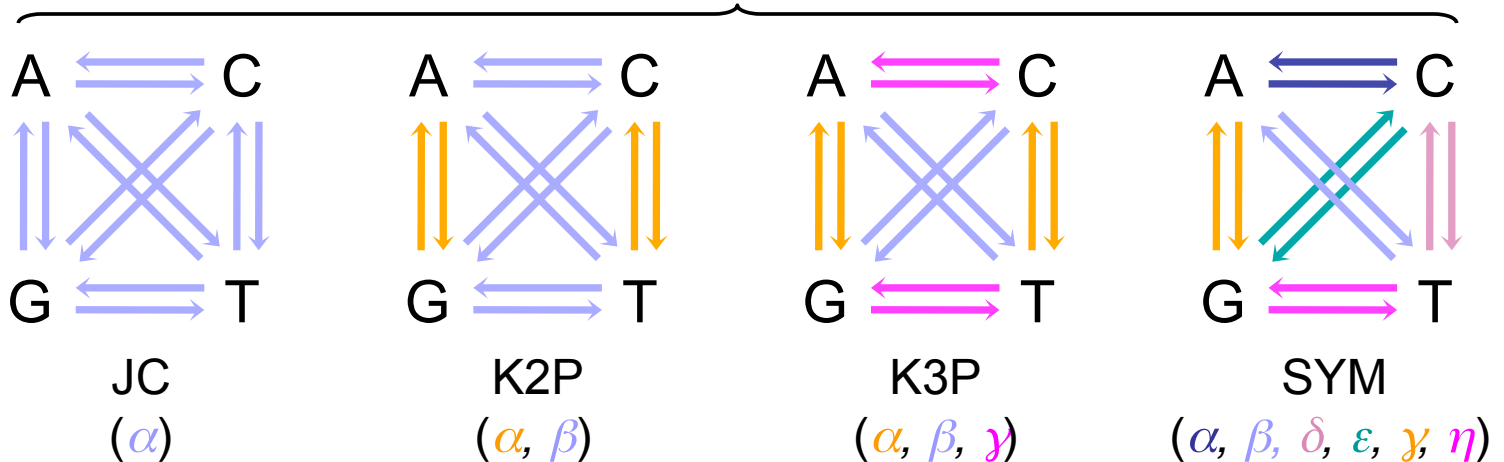
Beware! Using the most complex model on a small dataset is not the best strategy because of overfitting in parameter estimation!

Hierarchy of models for nucleotide data

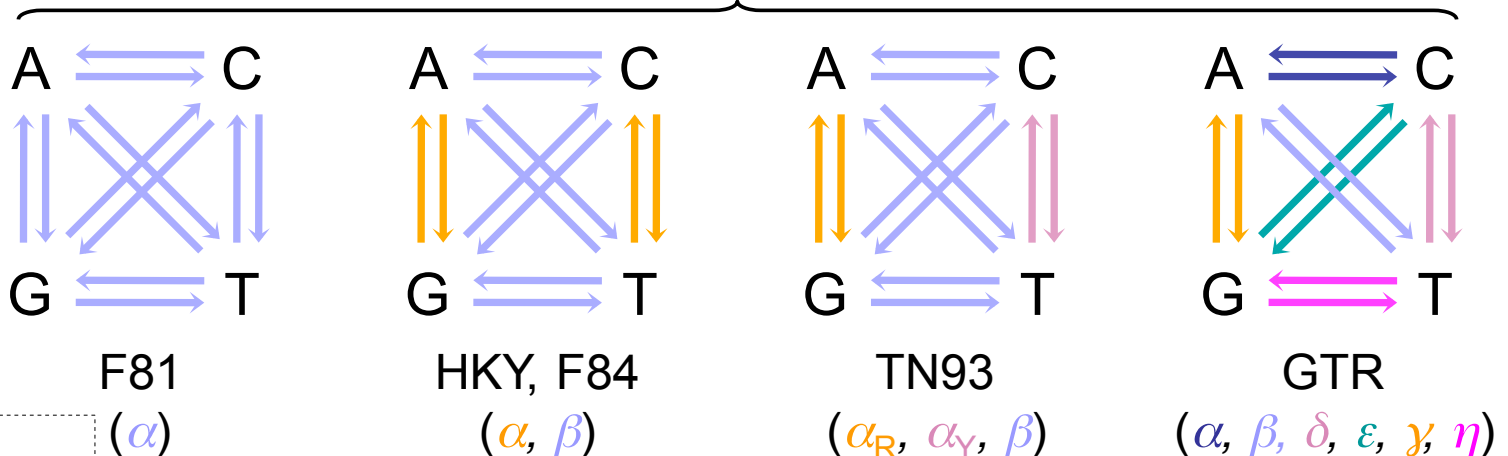


Different models, different parameters

$$\pi_A = \pi_C = \pi_T = \pi_G = 1/4$$



$$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$$



credit:
Guy Perrière
(LBBE, France)



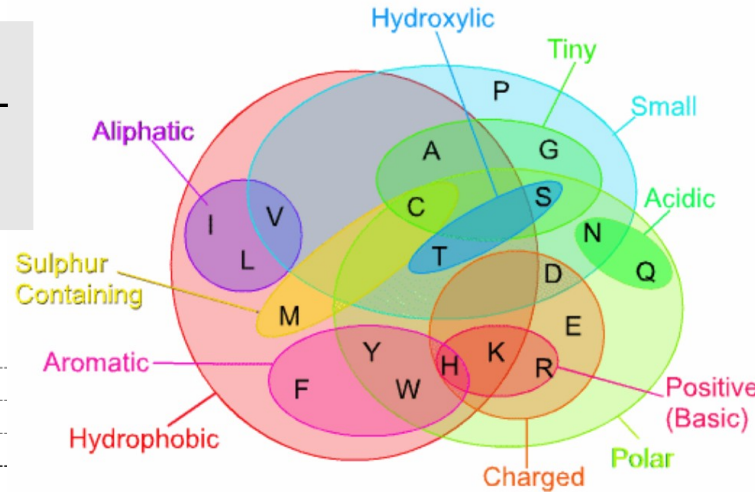
Substitution models for amino acid data

Matrices of amino-acid substitution rates have been developed empirically:

- JTT (Jones, Taylor, Thornton 1992): first matrix built from a large number of pairwise alignments from the Swissprot databank
- WAG (Whelan and Goldman 2001): derived from 3905 sequences in 182 protein families
- LG (Le & Gascuel 2008): estimated on 3,912 alignments from Pfam, comprising approximately 50,000 sequences and approximately 6.5 million residues overall
- mtREV: for mitochondrial protein data

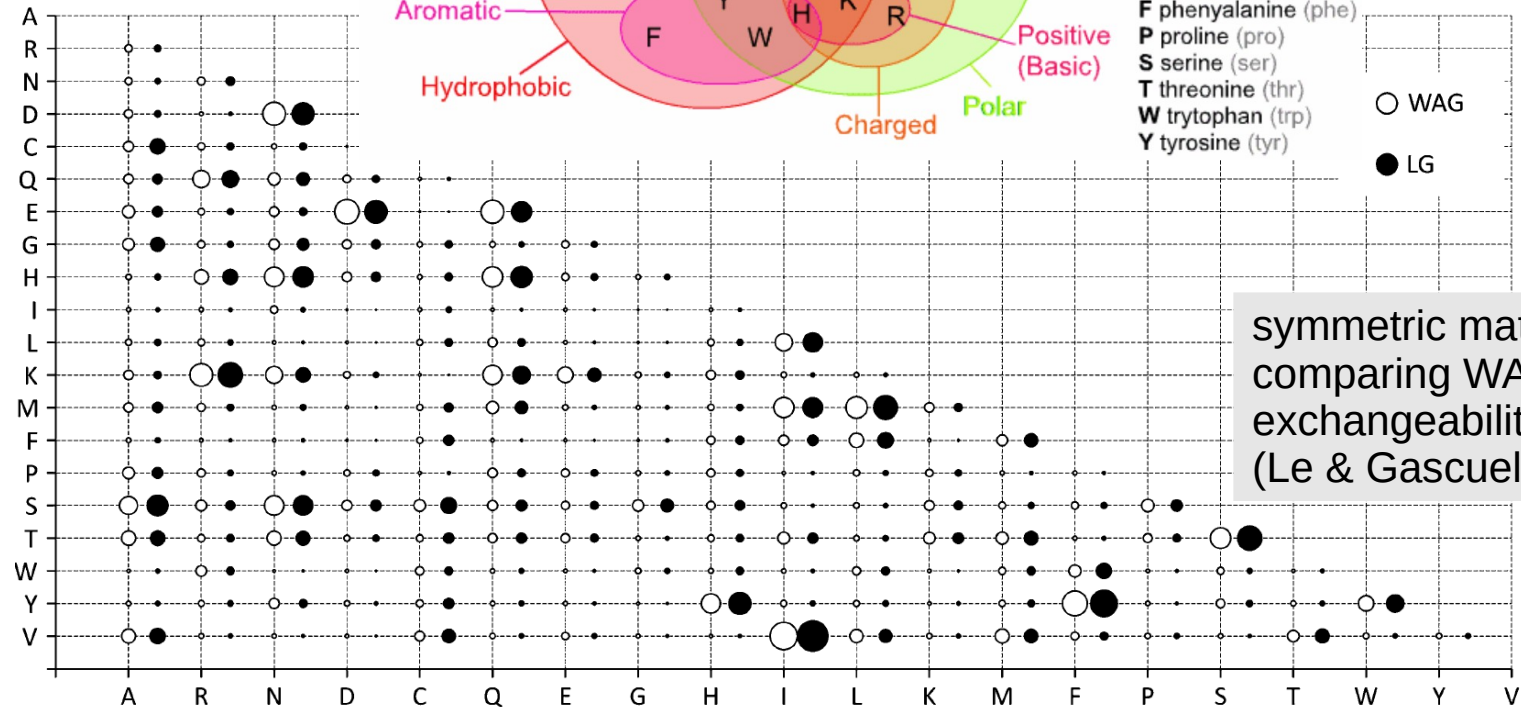
Amino acid exchangeabilities derive from their physico-chemical properties

amino acids grouped according to physico-chemical properties (Esquivel & al. 2013)



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M methionine (met)
F phenylalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W tryptophan (trp)
Y tyrosine (tyr)



symmetric matrix
comparing WAG & LG
exchangeabilities
(Le & Gascuel 2008)