Traits, phylogenies, evolutionary models and divergence time between sequences

Adapted from

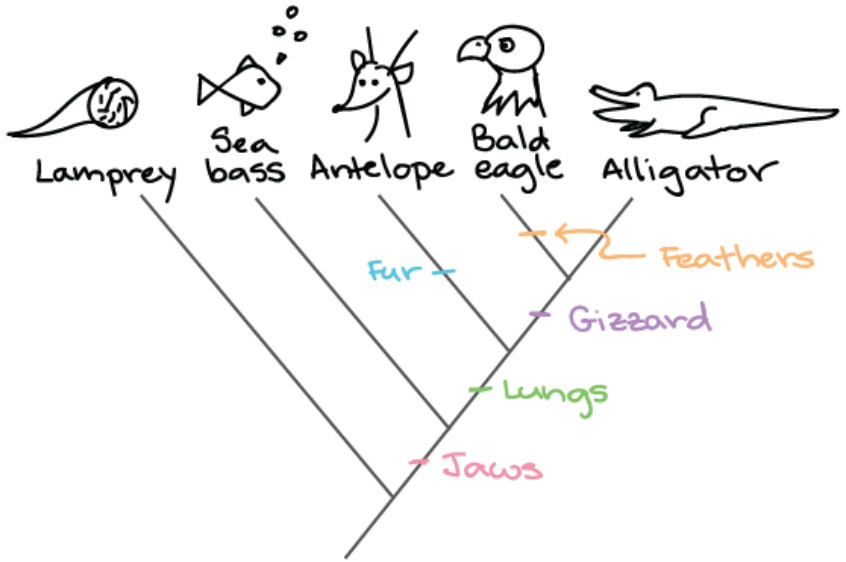
Traits

- A **trait** or **character** is any biological feature that can be compared across organisms, such as physical characteristics (morphology), genetic sequences, and behavioral traits
- Traits can be **qualitative/categorical** variables (e.g. aligned nucleotides or amino acids) 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.)
- Traits are used to construct phylogenetic trees that represent patterns of ancestry

Phylogeny

- Phylogeny is a way of classifying organisms (taxonomy) using evolutionary distance, or evolutionary relationship
- Phylogenetic relationship between organisms is given by the degree and kind of evolutionary distance
- Phylogenic relationships have been traditionally studied based on morphological data

Feature	Lamprey	Antelope	Bald eagle	Alligator	Sea bass
Lungs	**0**	+	+	+	0
Jaws	**0**	+	+	+	+
Feathers	**0**	0	+	0	0
Gizzard	**0**	0	+	+	0
Fur	**0**	+	0	0	0

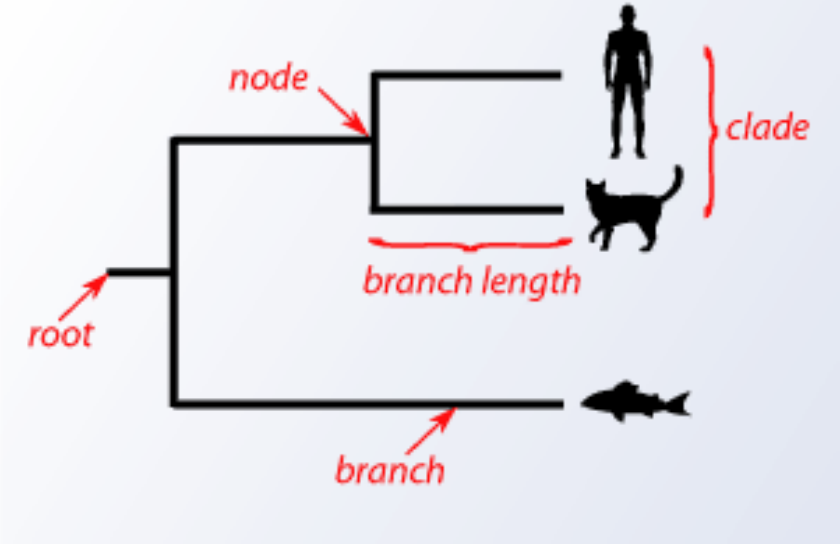


Phylogeny

- Phylogenetic relationships are currently studied using molecular data in order to classify organisms. Molecular methods are based on studies of gene sequences
- The relationships between species can be represented by a phylogenetic tree

Phylogenetic tree

- A **phylogenetic tree** or **evolutionary tree** is a branching diagram or "**tree**" showing the **evolutionary** relationships among various biological species
- A phylogenetic tree has;
  - Branches
  - Nodes
  - Branch length
  - Root

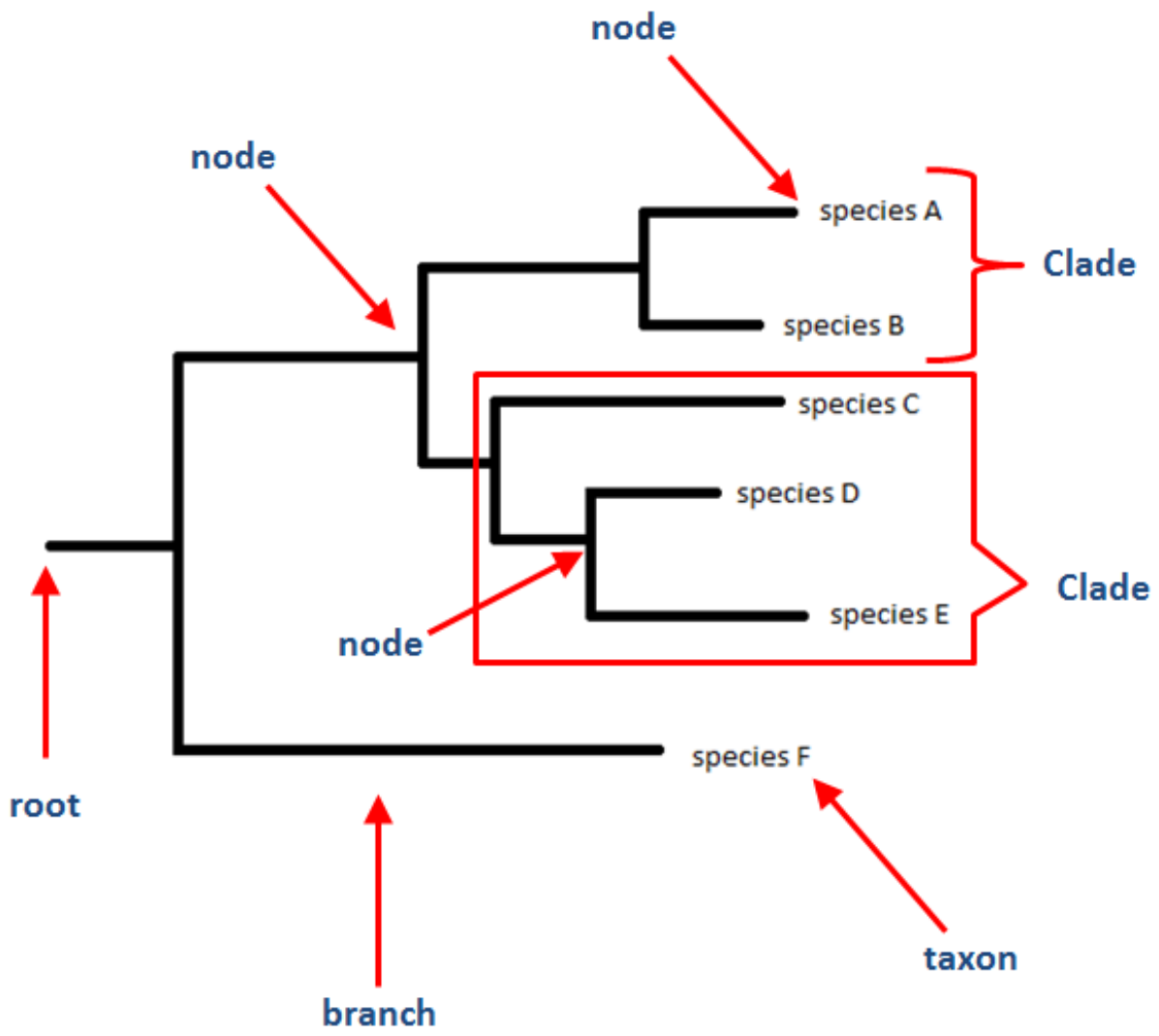


- The nodes represent taxonomic units. Branches reflect the relationships of these nodes in terms of descendants

Phylogenetic tree

- The branch length is not always present, if present it usually indicates some form of evolutionary distance
- The tree is **rooted** if we know where is the most ancient node (evolutionary origin), **unrooted** otherwise
- 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)
- Phylogenetic trees are hypotheses, not definitive facts

Parts of a phylogenetic tree



Terminologies

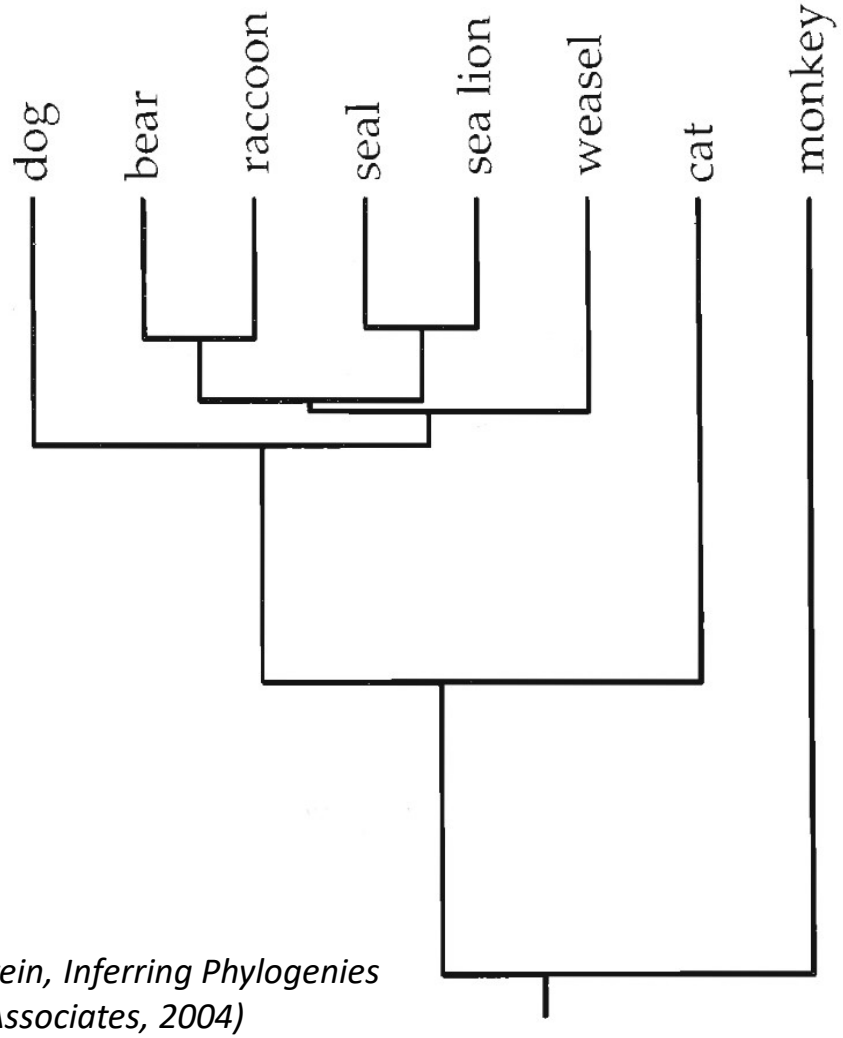
- A mathematical definition of tree **topology** is a connex acyclic graph

  $G = (V,E)$

  - 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**

  (Hage J, Harju T. Acyclicity of switching classes. European Journal of Combinatorics. 1998 Apr 1;19(3):321-7.)
- In a **binary tree**, all non-terminal nodes have 3 neighbours (or one parent node and two children in case the tree is rooted) - **bifurcation**
- A node with degree > 3 is called a **multifurcation**, aka. **polytomy**

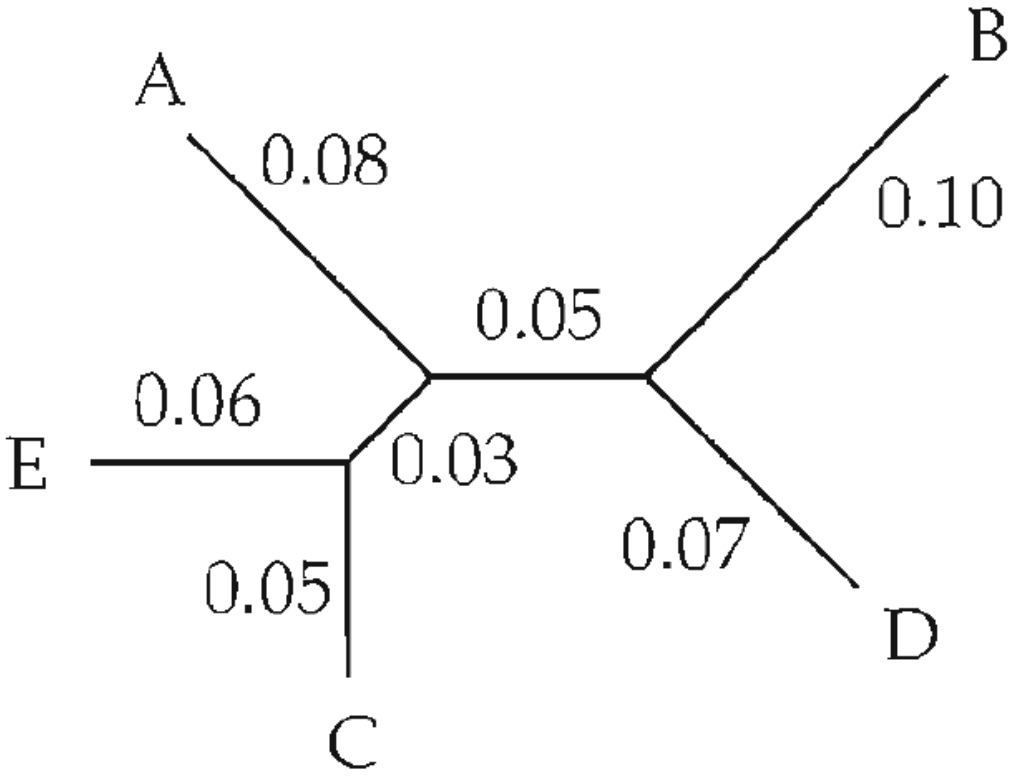
A rooted binary tree



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

- 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

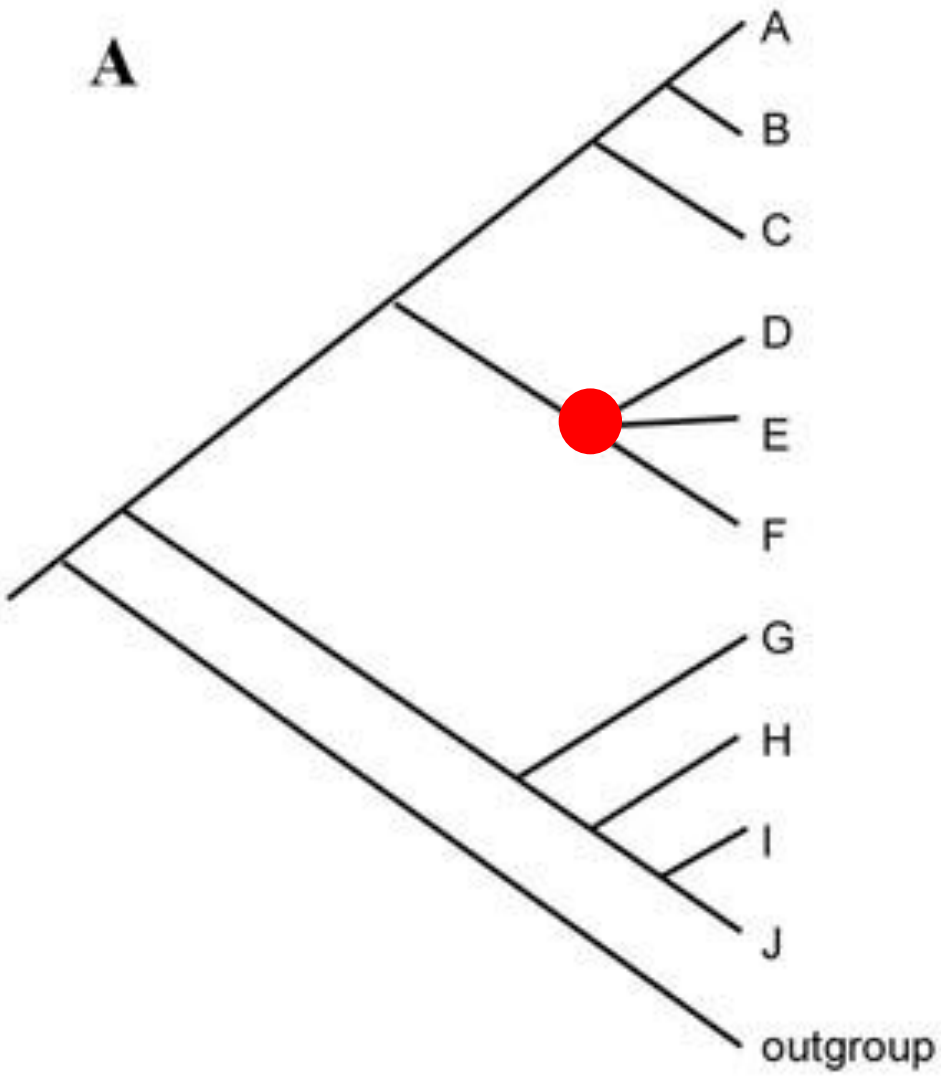
An unrooted binary tree

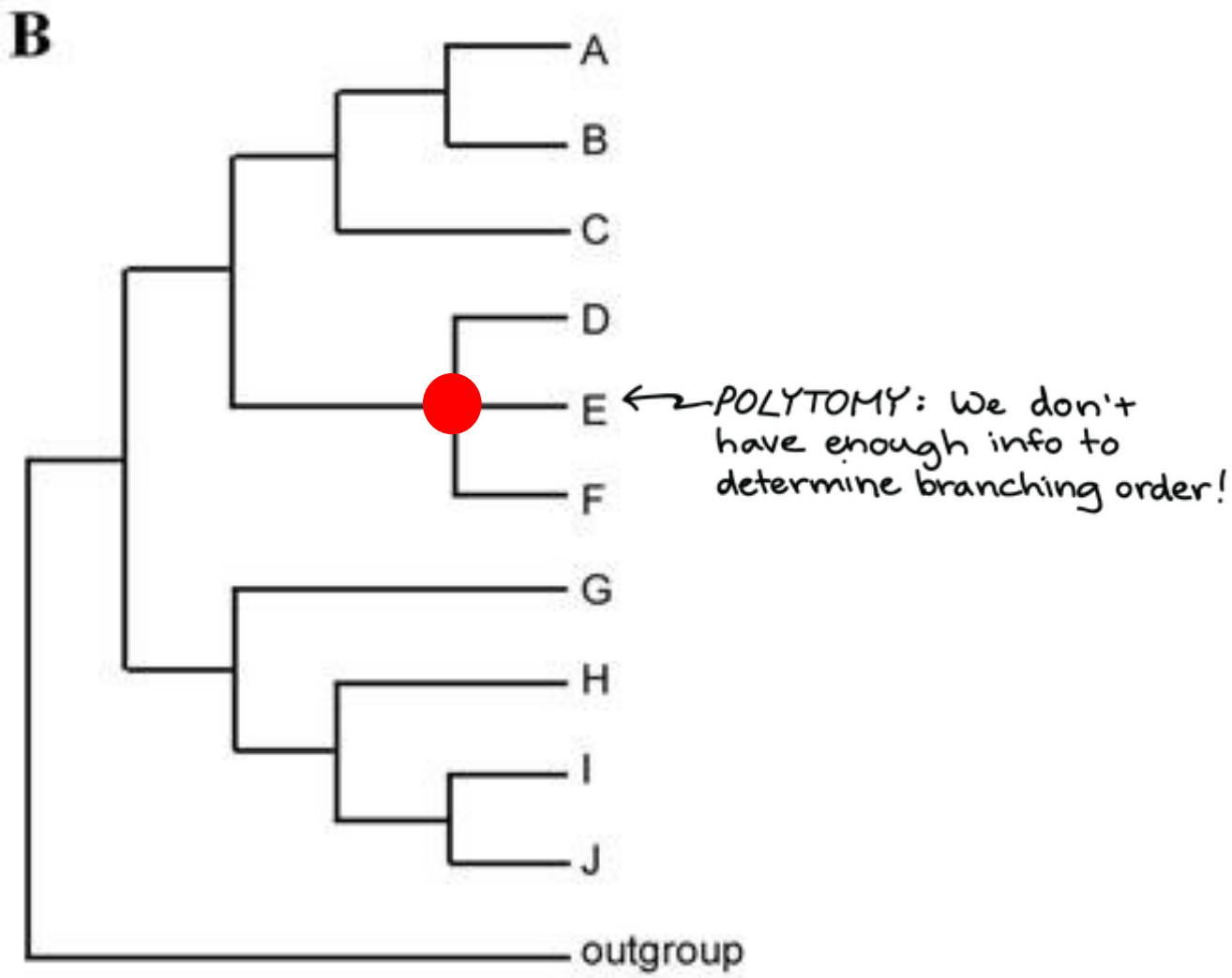


- in an **unrooted binary tree**, every internal node has three neighbors
- *n* taxa
- *2n-2* vertices
- *2n-3* edges
- this tree has branch lengths

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

Multifurcation/Polytomy

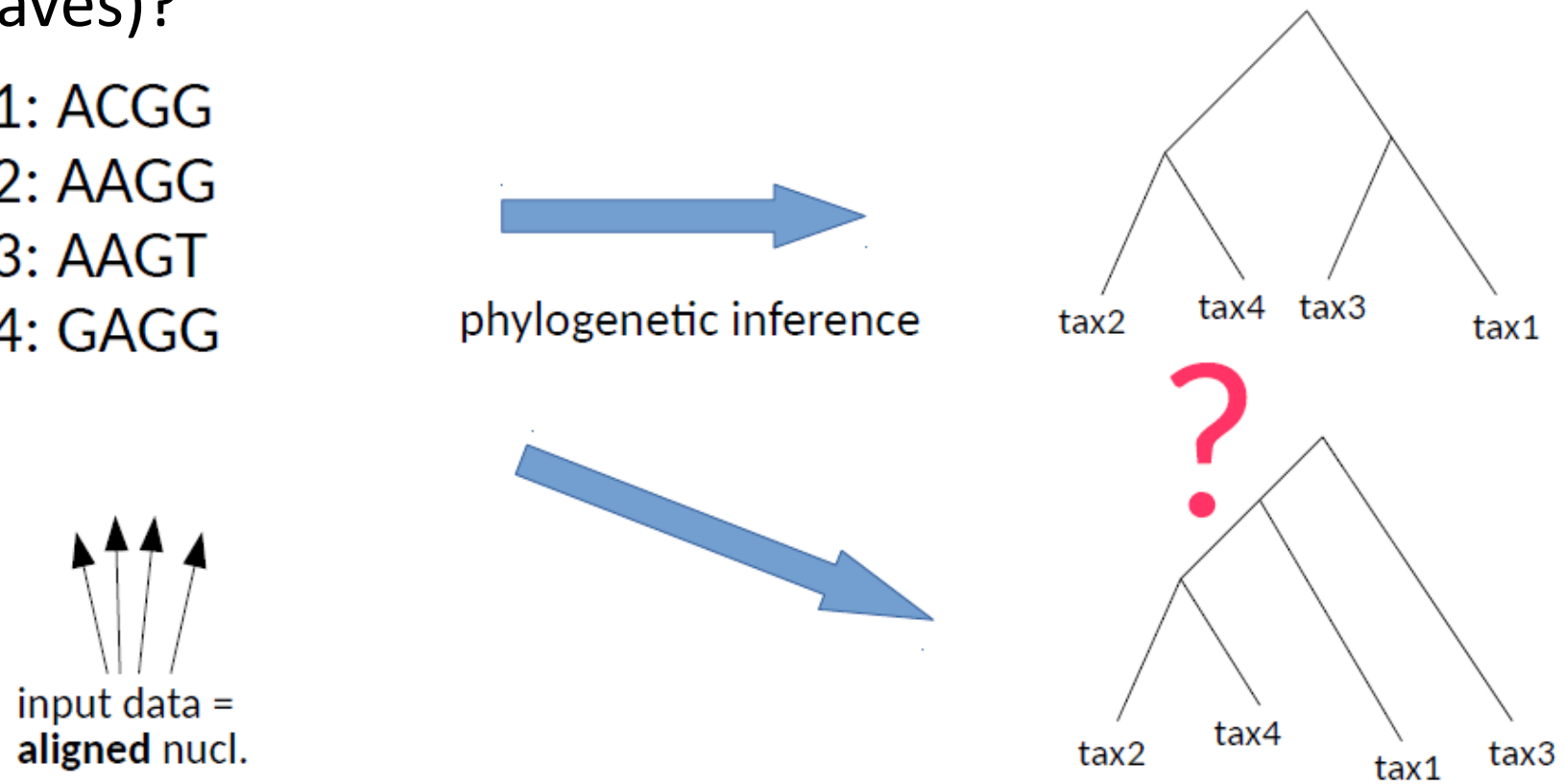




The tree inference problem

- 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

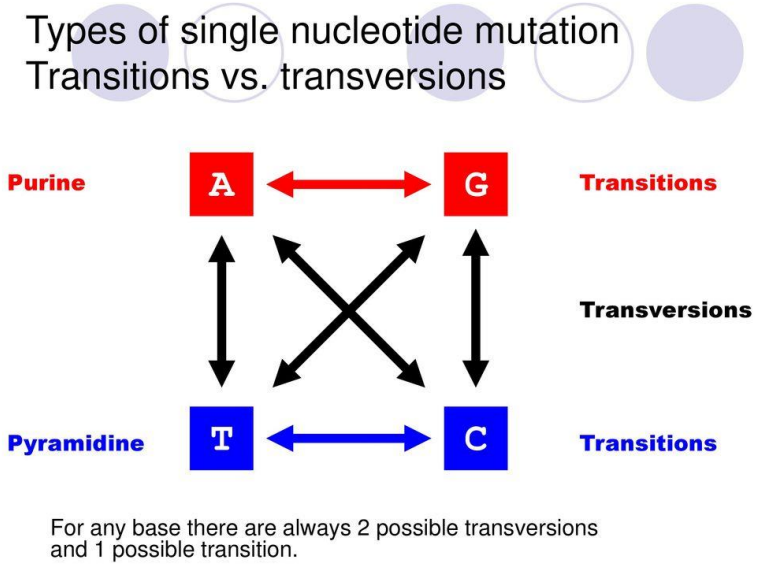


Evolutionary models

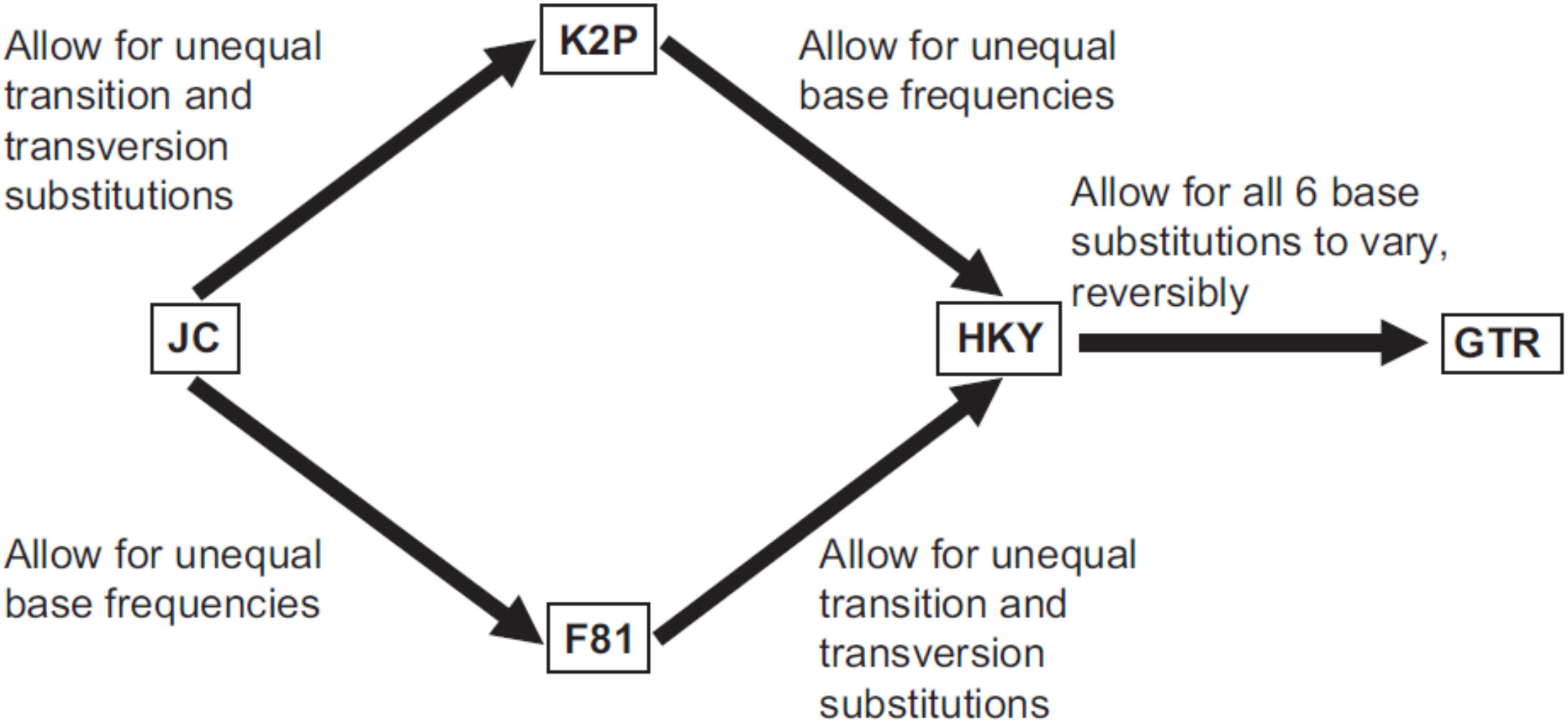
- An evolutionary model deals with **aligned sequences** (nucleotides or amino acids)
- All evolutionary models are stochastic: they predict **probabilities of change**, without yielding any certainty
- Models evolution in terms of **character substitutions**: enables to calculate e.g. $Pt\ (A \rightarrow C)$
- A model 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

Evolutionary models

- The model of nucleotide substitution chosen for the data set is important and affects the final tree selected
- The model makes basic assumptions about the base composition, rate and frequency of base substitutions among different sites and the nature of base substitutions, *i.e.* transitions versus transversions
- Some well known evolutionary models are;
  - Jukes-Cantor (JC, 1969): the earliest substitution model
  - Kimura-2-parameter (K2P, 1980)
  - Felsenstein (F81, 1981)
  - Hasegawa, Kishino and Yano (HKY85, 1985)
  - General time reversible model (GTR/REV, 1984, 1986)

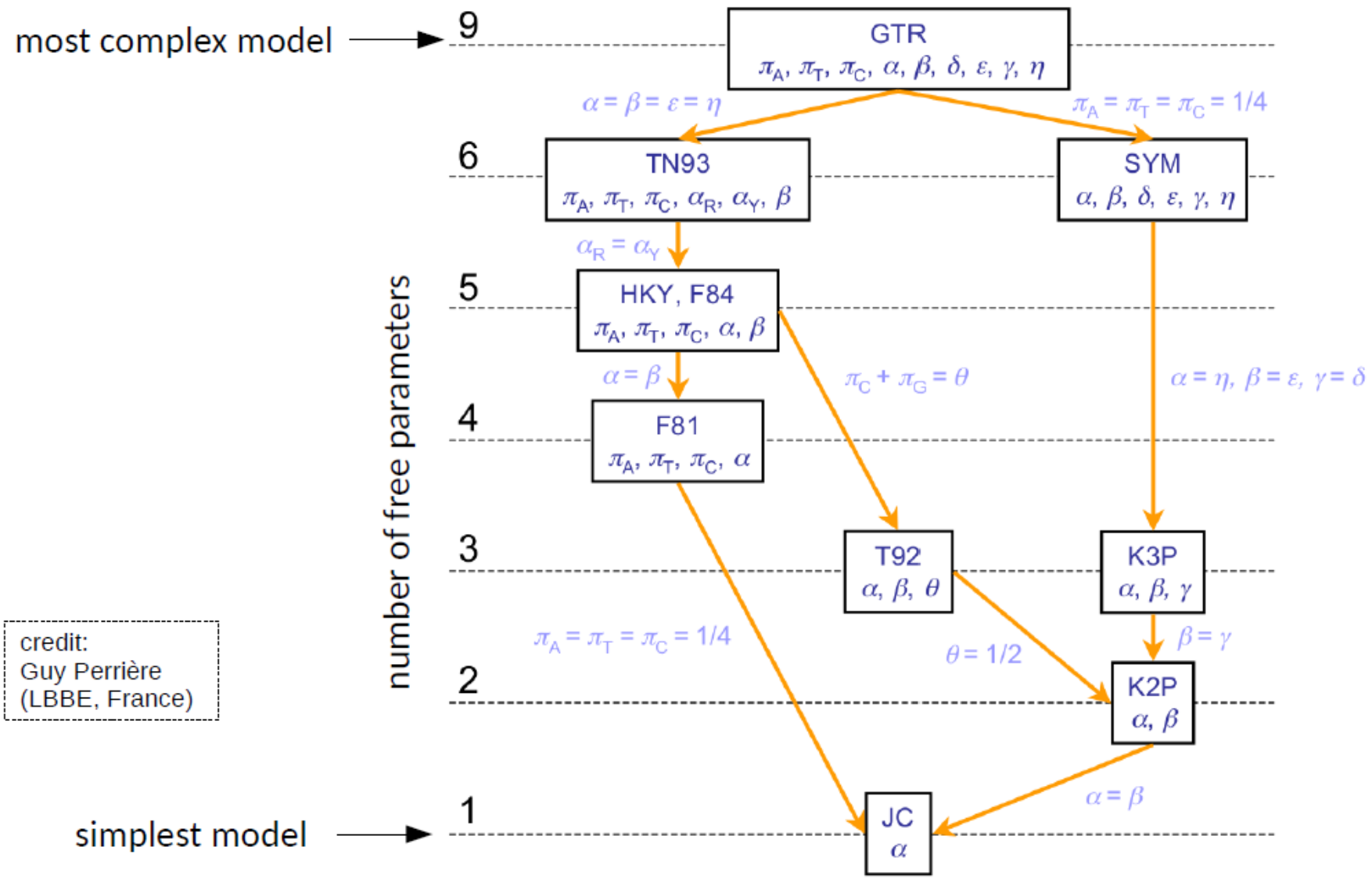


Evolutionary models for nucleotide data



Source:
Tsan-Yuk Lam et al. 2010

Hierarchy of models for nucleotide data



credit:
Guy Perrière
(LBBE, France)

Evolutionary models for amino acids

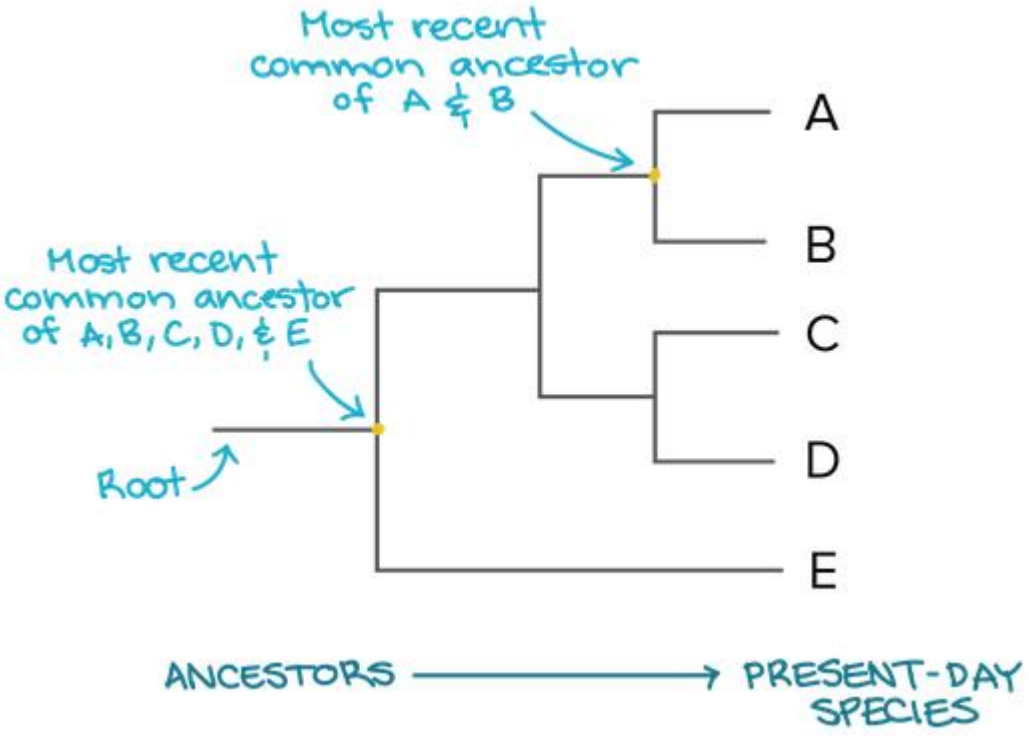
Matrices of amino-acid substitution rates have been developed empirically:

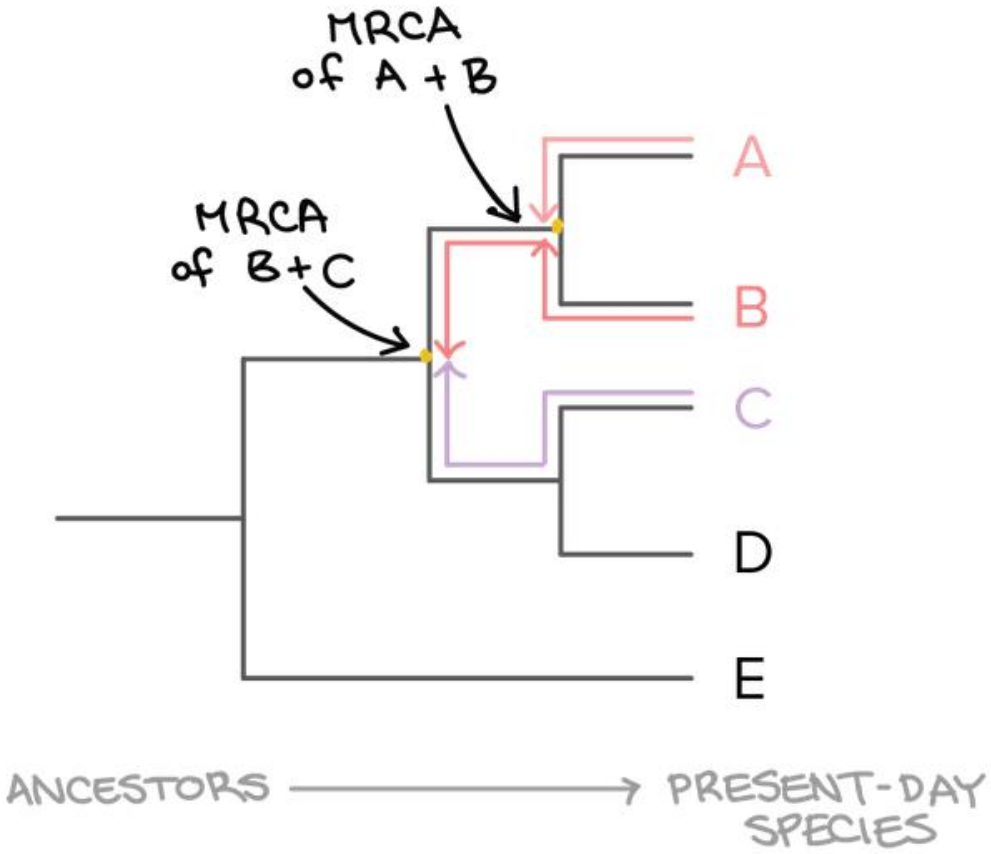
- JTT (Jones, Taylor, Thornton 1992): first matrix built from a large number of pairwise alignments from the Swissprot databank
- WAG (Whelan and Goldman 2001): derived from 3905 sequences in 182 protein families
- LG (Le & Gascuel 2008): estimated on 3,912 alignments from Pfam, comprising approximately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

Divergence times between sequences

- It is the time which two or more populations of an ancestral species accumulated independent genetic changes and diverged
- The ability to date the time of divergence between lineages using molecular clock provides the opportunity to answer important questions in evolutionary biology
- Methods used to estimate time of divergence require a large amount of computational time
- The divergence times help us to find the time for most common recent ancestor(MRCA) between two taxa

Divergence times





Molecular clocks

- Molecular clocks are used to estimate time of divergence between different species
- The molecular clock hypothesis states that DNA and protein sequences evolve at a rate that is relatively constant over time and among different organisms – **strict molecular clock**
- A direct consequence of this constancy is that the genetic difference between any two species is proportional to the time since these species last shared a common ancestor
- The limitation led to development of **"relaxed" molecular clocks**, which allow the molecular rate to vary among lineages

BEAST

BEAST is a cross-platform program for Bayesian analysis of molecular sequences using MCMC

Uses both strict or relaxed molecular clock models

BEAST

Bayesian Evolutionary Analysis Sampling Trees

THANK YOU