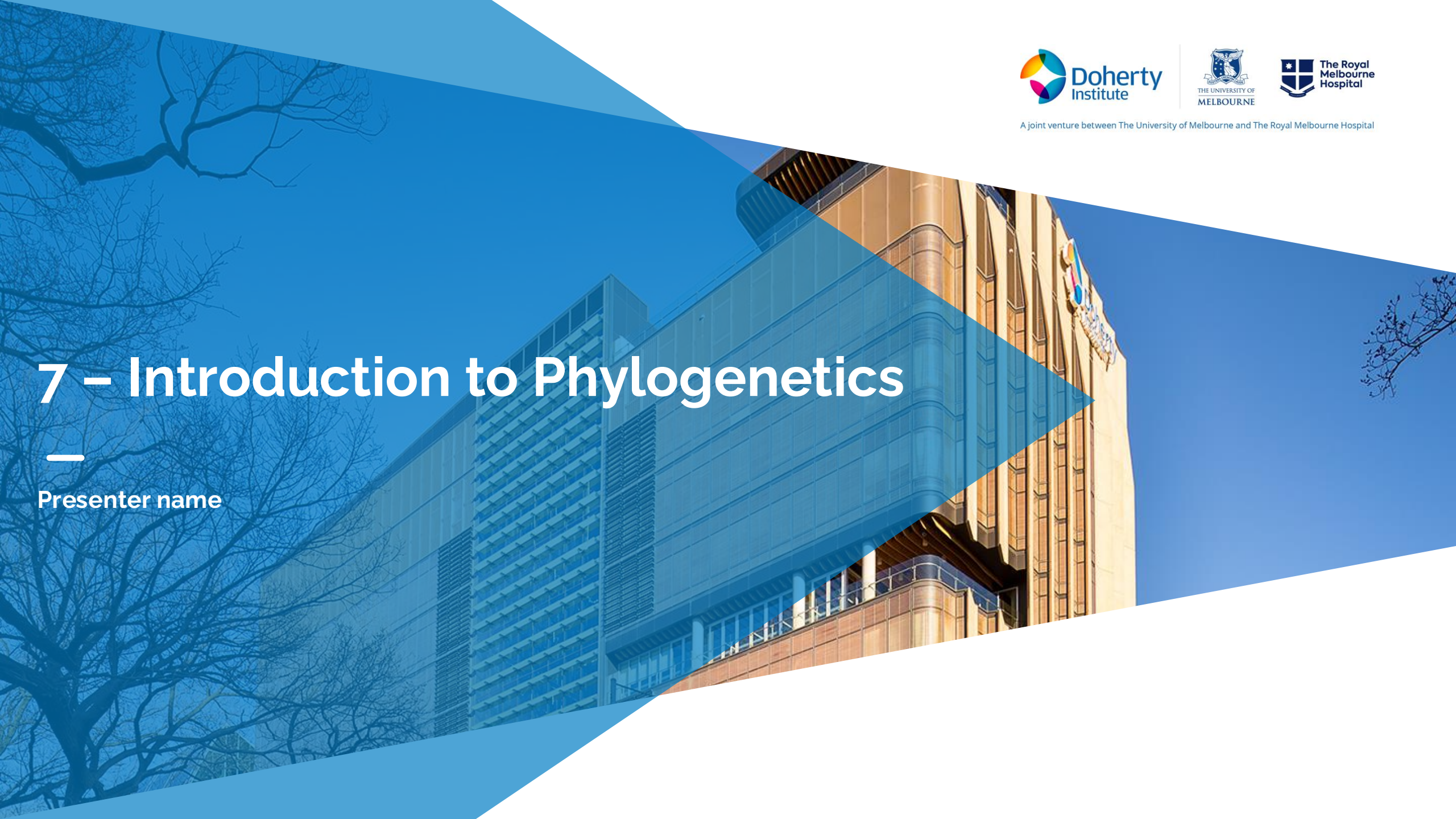


Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

7 – Introduction to Phylogenetics

Presenter name

Objectives

- Gain a basic understanding of phylogenetics
- To know the different methods of creating a phylogenetic tree

What is Phylogenetics?

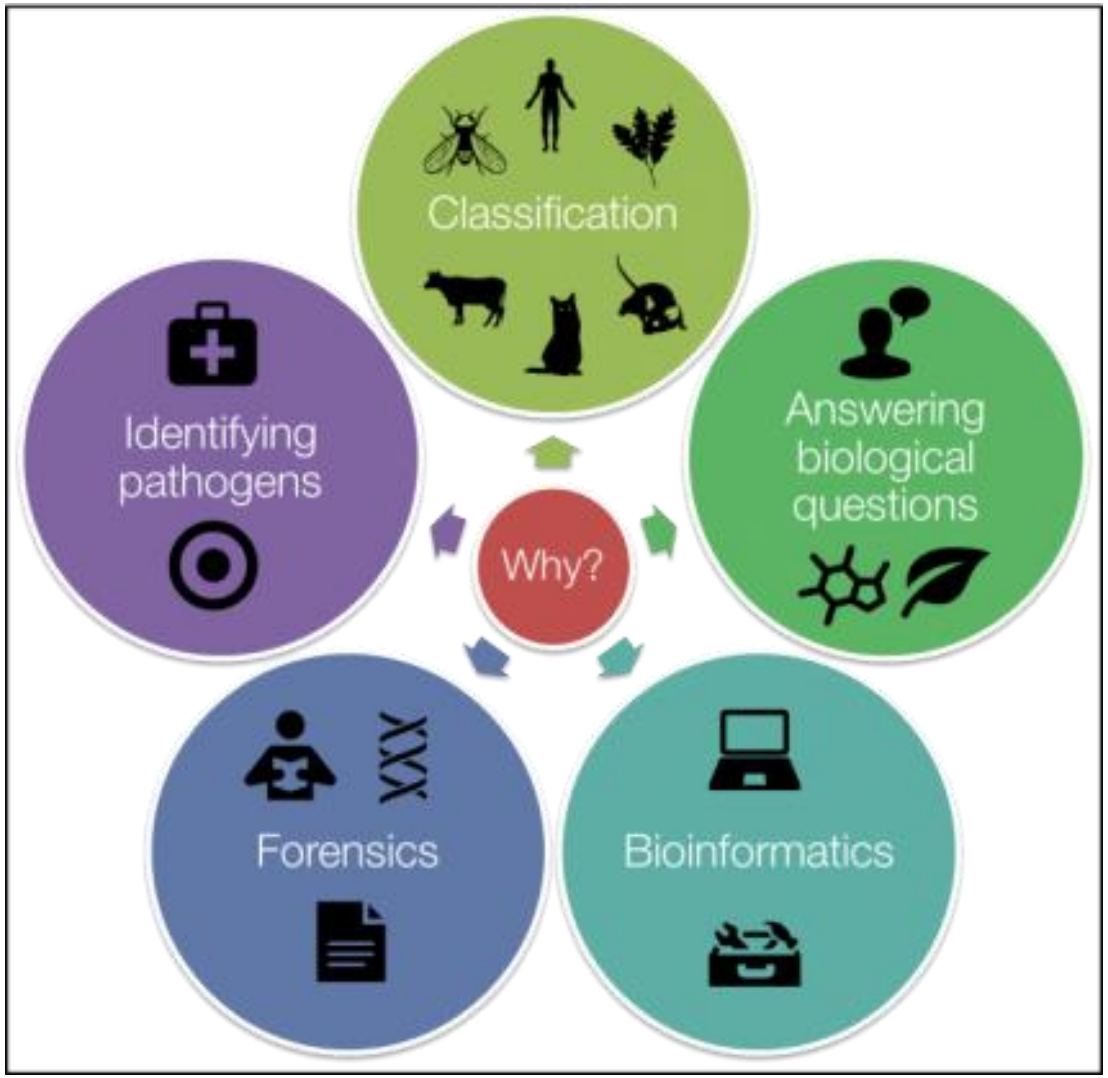
Defintion: Phylogenetics is the study of evolutionary relationships among biological entities – often species, individuals, genes, genomes. A phylogenetic tree is created (constructed) by looking at the nucleotide or protein sequences and combining with our understanding of sequence evolution.

Typical questions answered using phlyogenetics:

- What are the evolutionary relationships or histories among species/individuals/genes of interest?
- How do sequences evolve?
- Can I better describe processes of sequence evolution with a mathematical model?

This enables us to infer evolutionary events that happened in the past and provides more information about the evolutionary processes operating on sequences.

Why perform phylogenetics?



Why perform phylogenetics?

Classification: Phylogenetics based on sequence data provides us with more accurate descriptions of patterns of relatedness than was available before the advent of molecular sequencing. Phylogenetics now informs the Linnaean classification of new species.

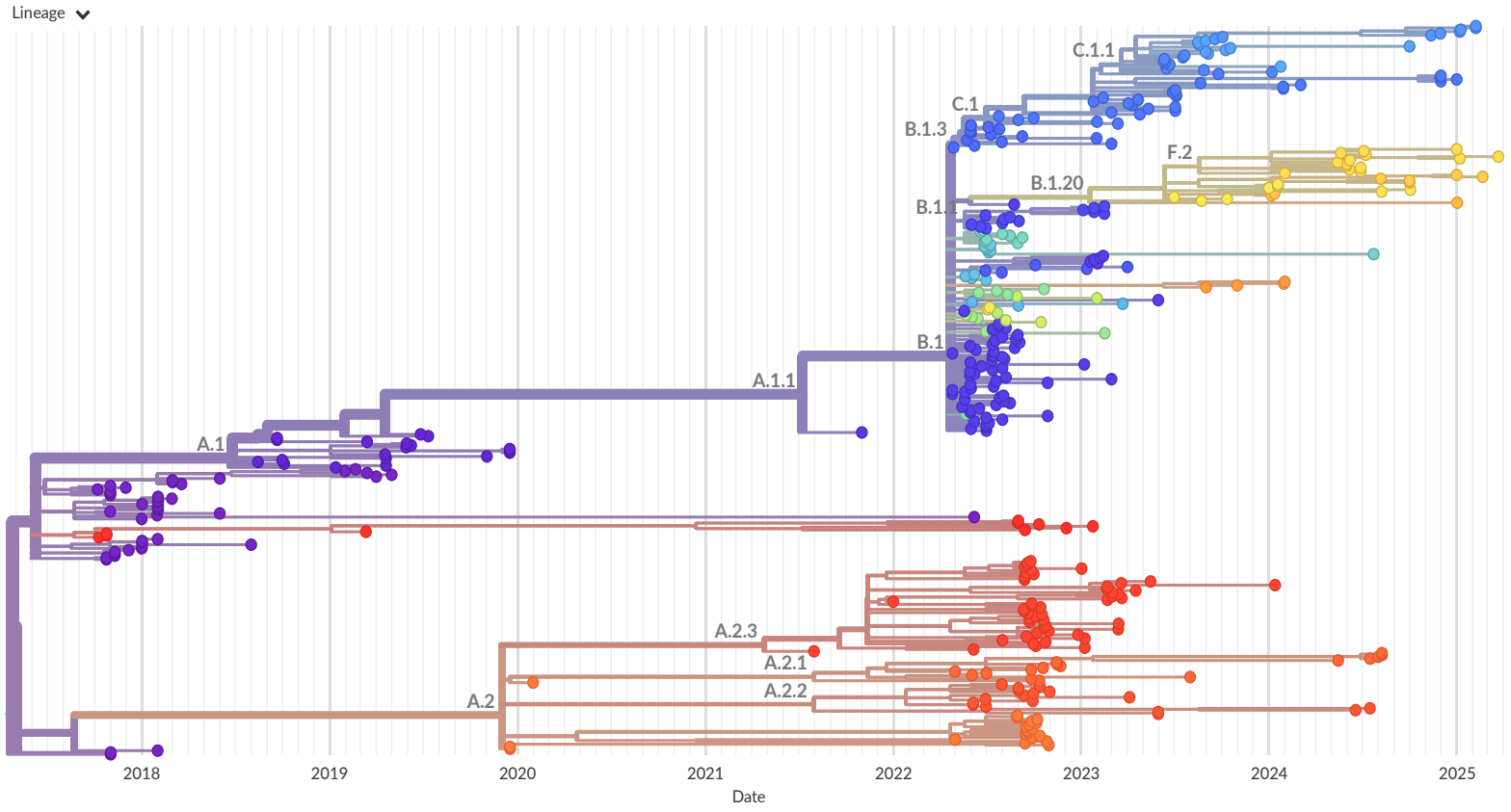
Forensics: Phylogenetics is used to assess DNA evidence presented in court cases to inform situations, e.g. where someone has committed a crime, when food is contaminated, or where the father of a child is unknown.

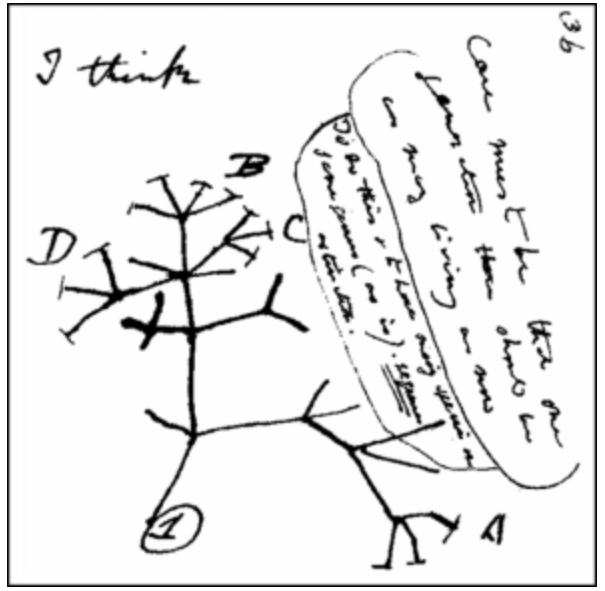
Identifying the origin of pathogens: Molecular sequencing technologies and phylogenetic approaches can be used to learn more about a new pathogen outbreak. This includes finding out about which species the pathogen is related to and subsequently the likely source of transmission. This can lead to new recommendations for public health policy.

Conservation: Phylogenetics can help to inform conservation policy when conservation biologists have to make tough decisions about which species they try to prevent from becoming extinct.

What is a phylogeny?

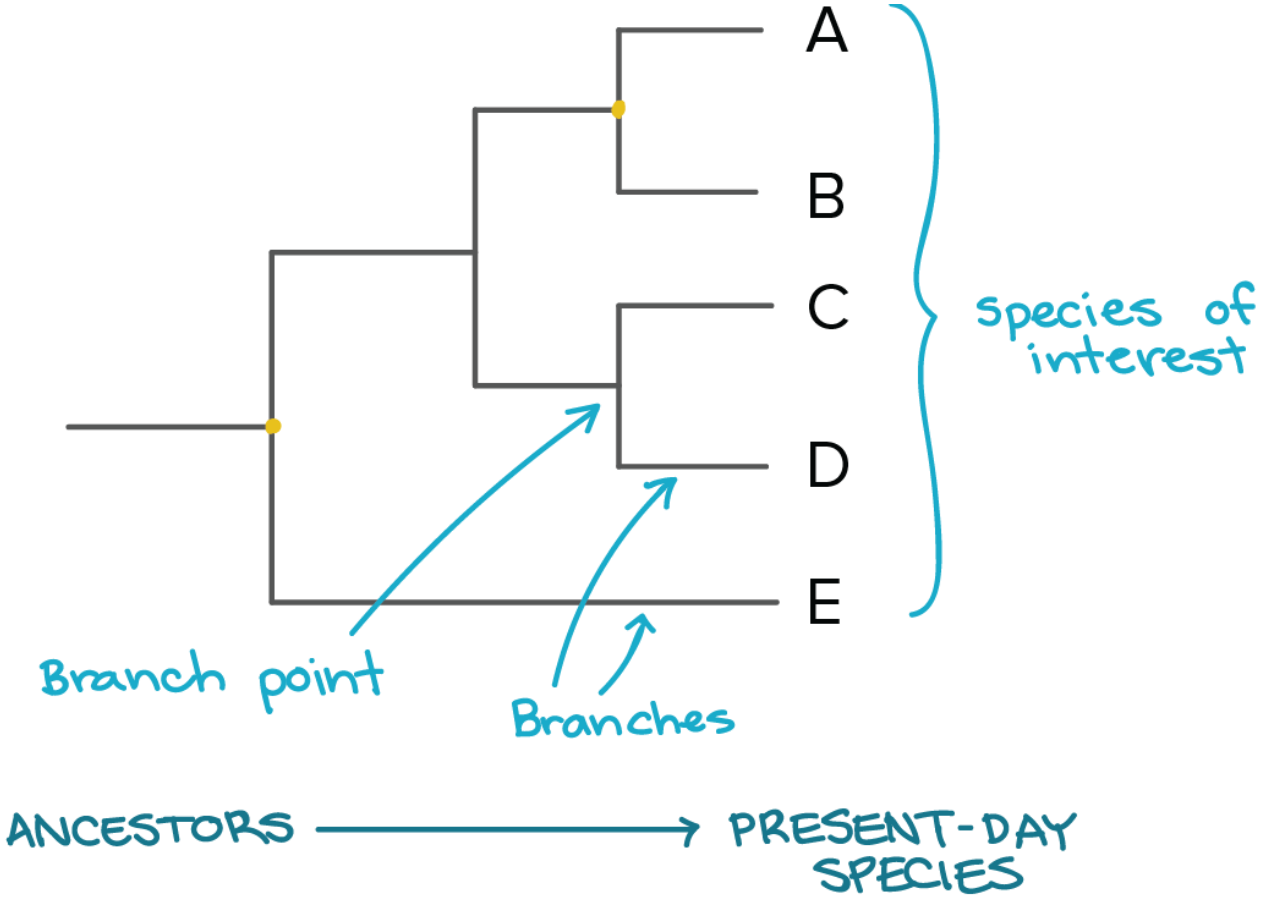
A phylogeny, also known as a tree, is an explanation of how sequences evolved, their genealogical relationships, and therefore how they came to be the way they are today.





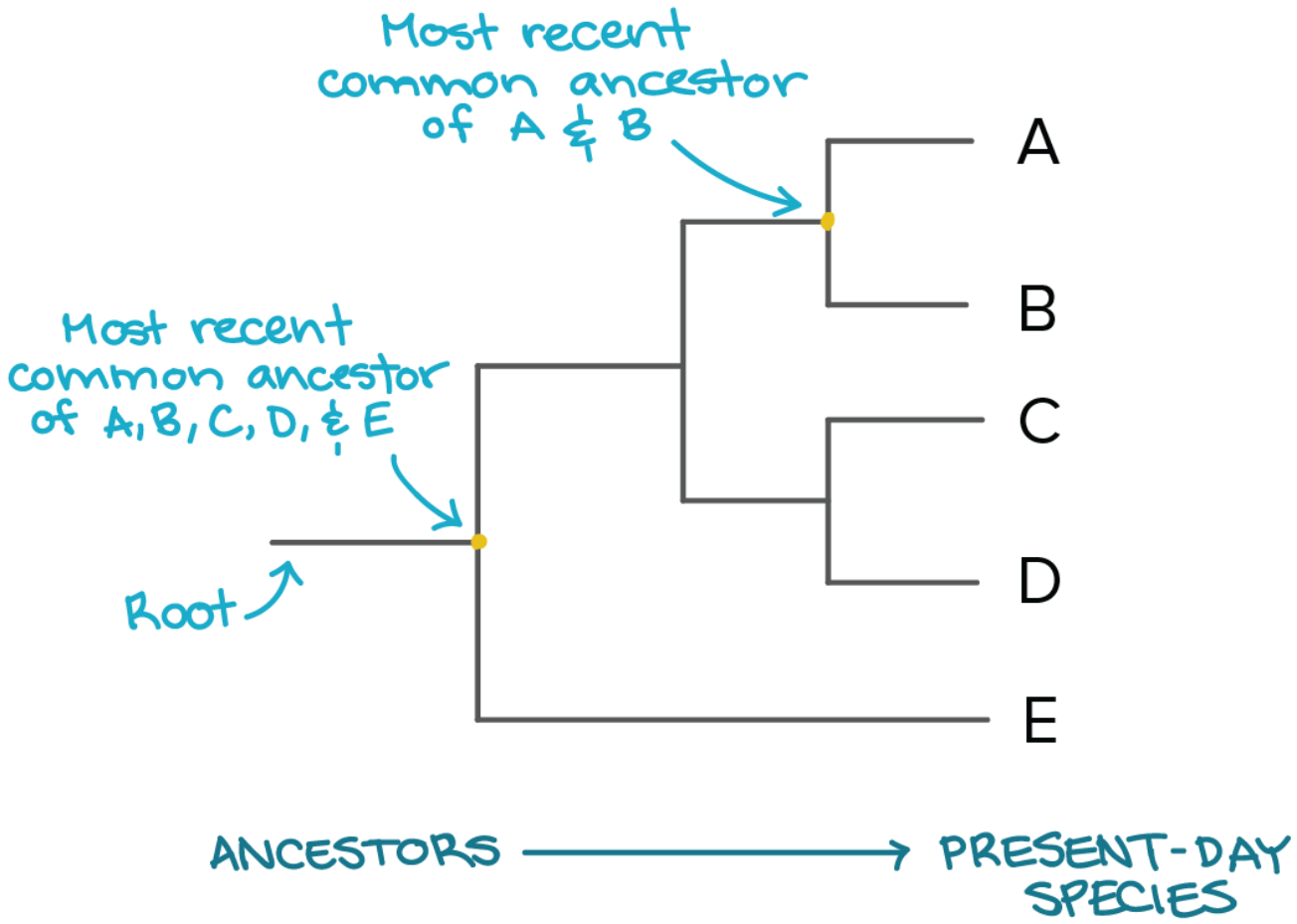
One of the first drawings of a phylogeny by Charles Darwin.

Parts of Phylogeny (or Tree)



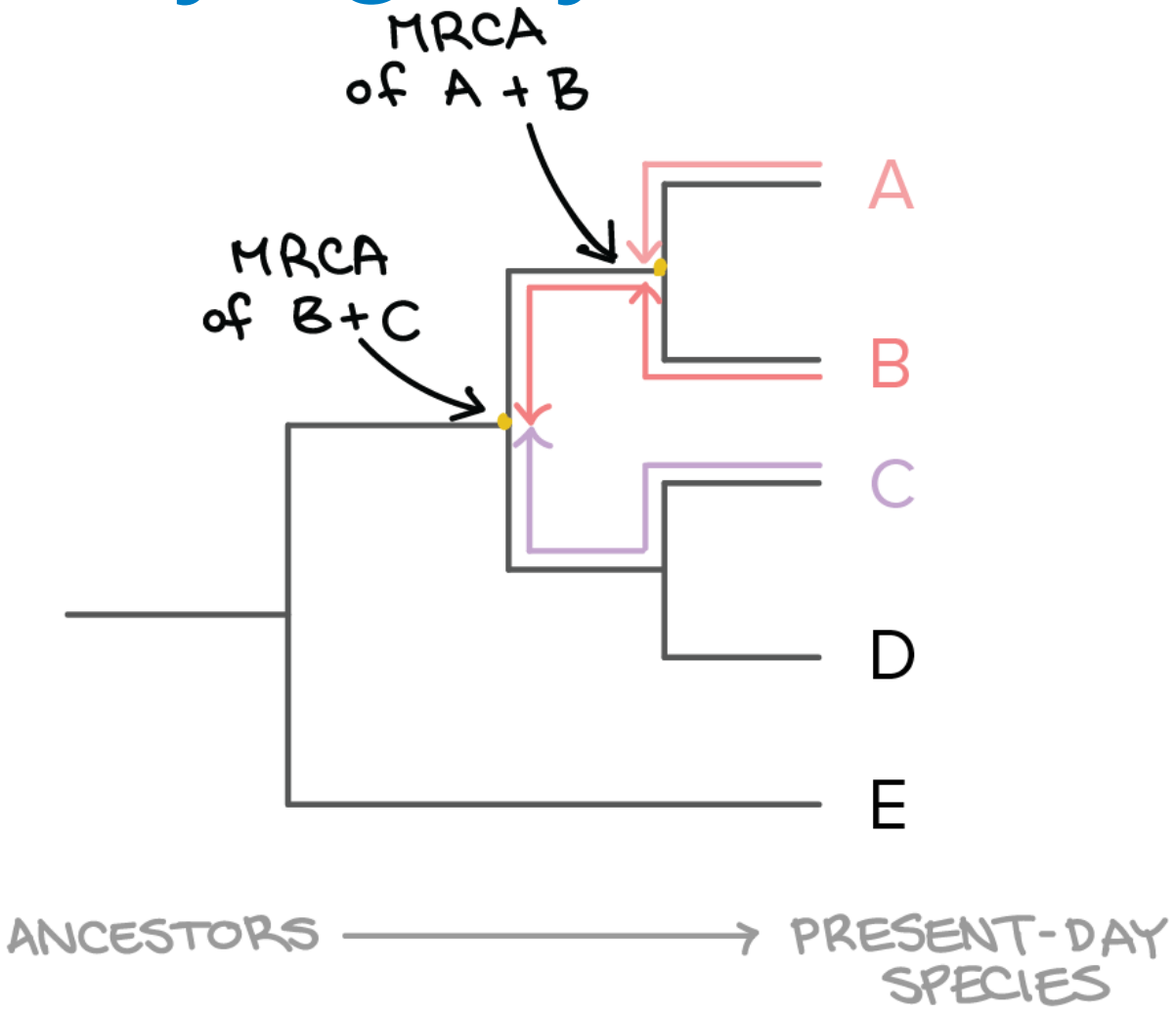
https://www.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/phylogenetic-trees

Parts of Phylogeny (or Tree)



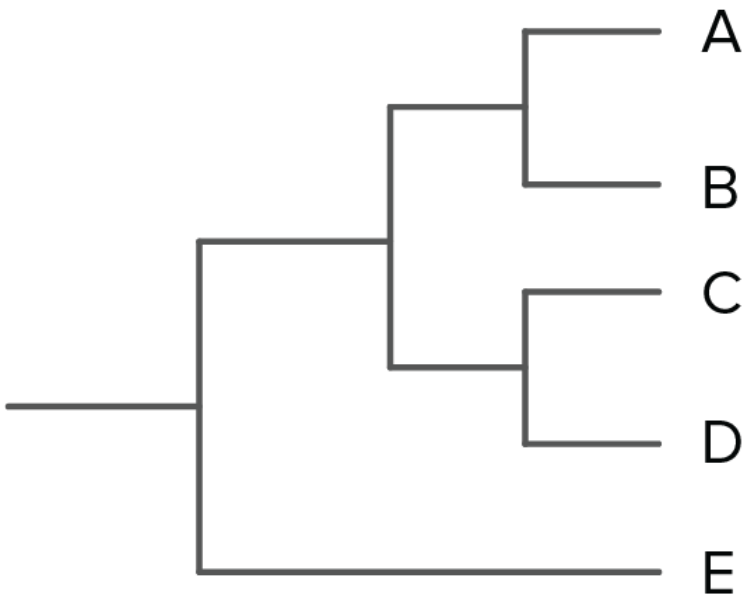
https://www.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/phylogenetic-trees

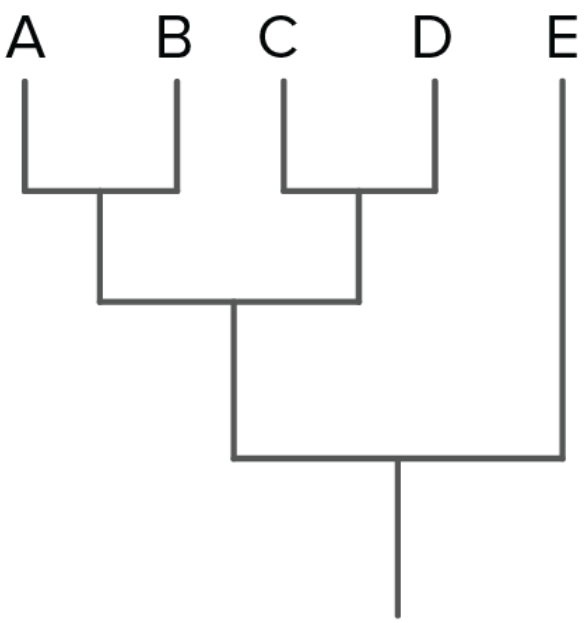
Parts of Phylogeny (or Tree)

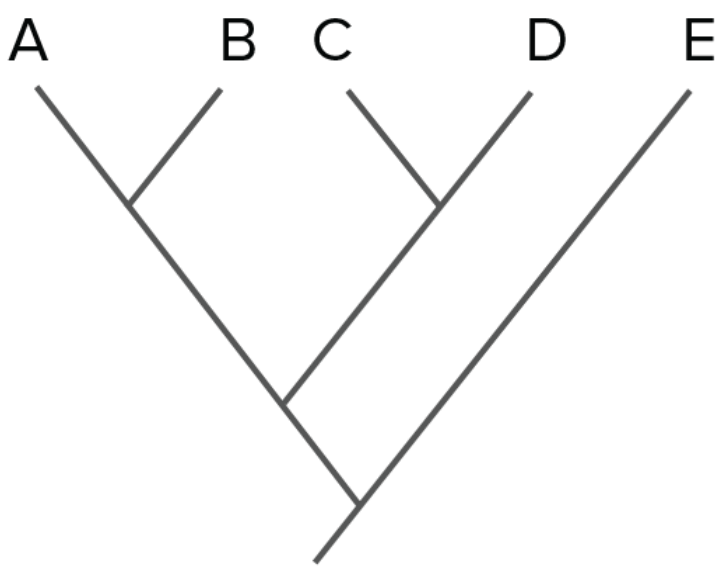


https://www.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/phylogenetic-trees

Tips for reading trees



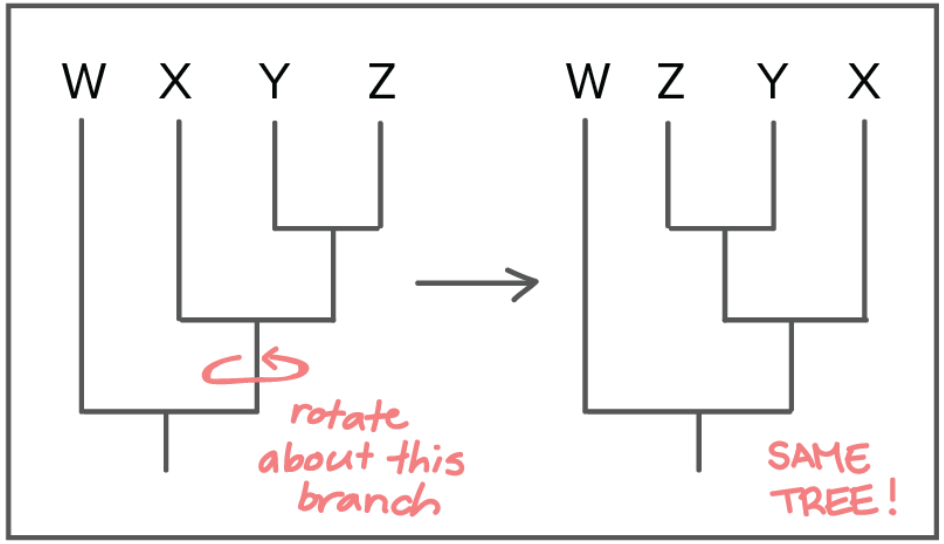




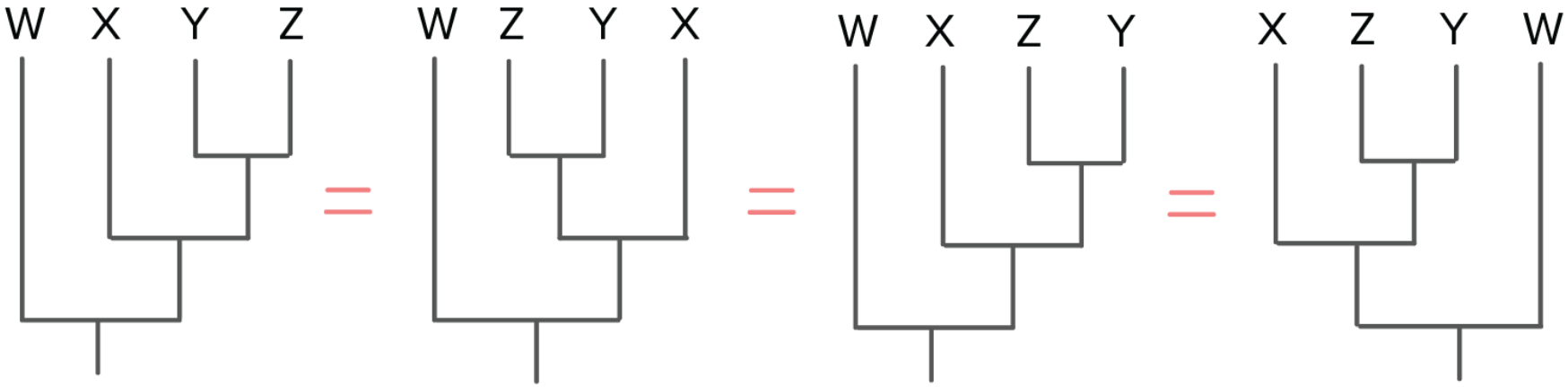
The three trees above represent identical relationships among species A, B, C, D, and E.

https://www.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/phylogenetic-trees

Tips for reading trees

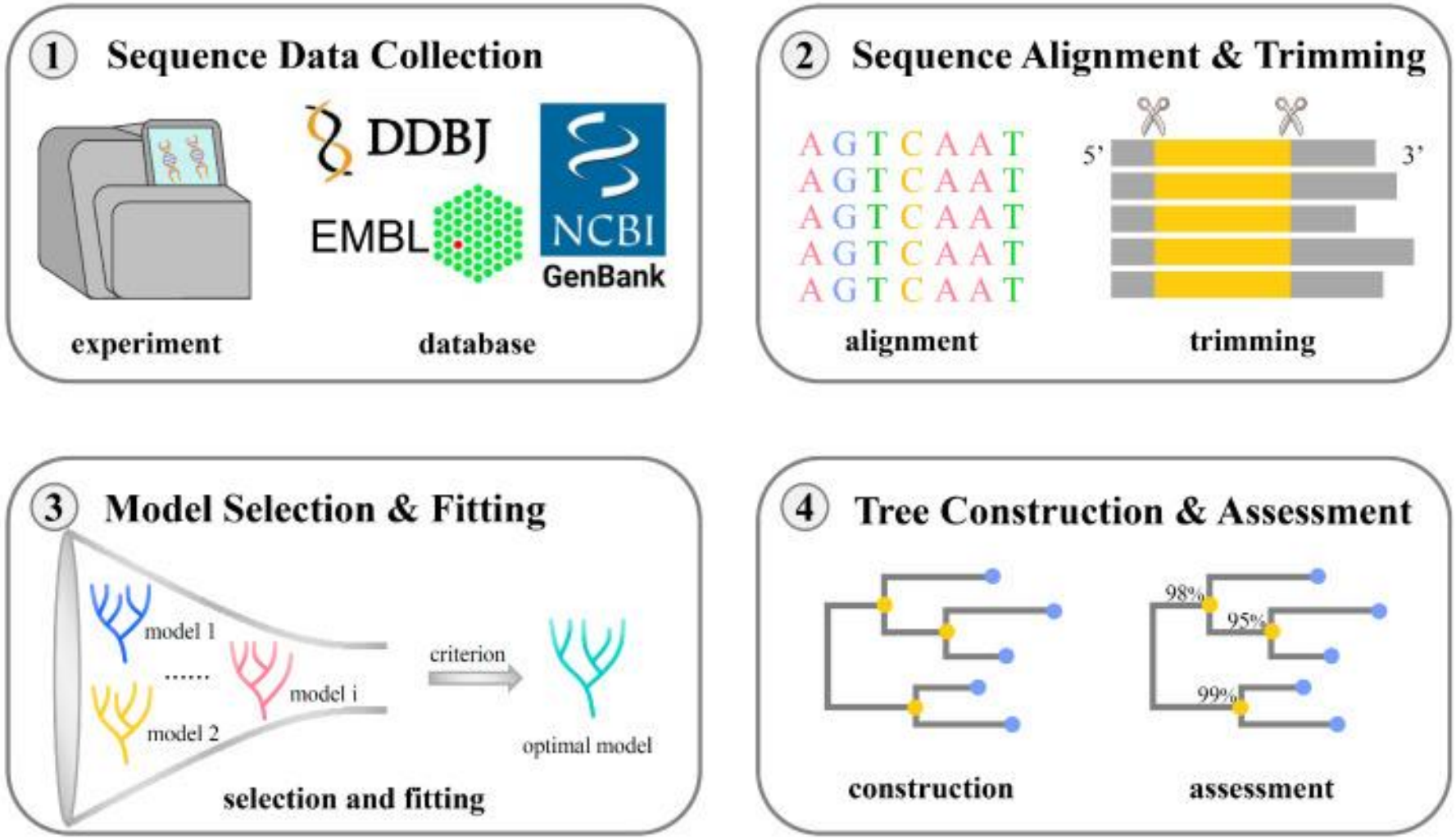


Another critical point about these trees is that if you rotate the structures, using one of the branch points as a pivot, you don't change the relationships.



https://www.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/phylogenetic-trees

Major stages in phylogenetic tree analysis

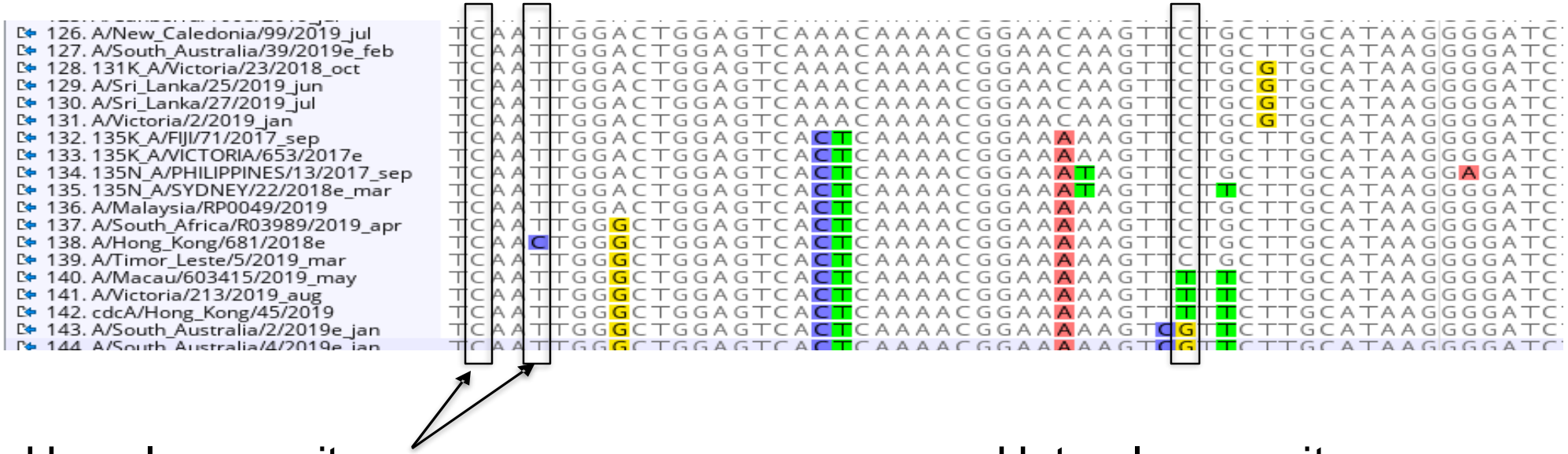


https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Creating a (phylogenetic) tree

1. Retrieve sequence data
2. Align sequence data
3. Identify informative sites
4. Model selection
5. Tree construction
6. Tree visualisation

Creating a (phylo)tree

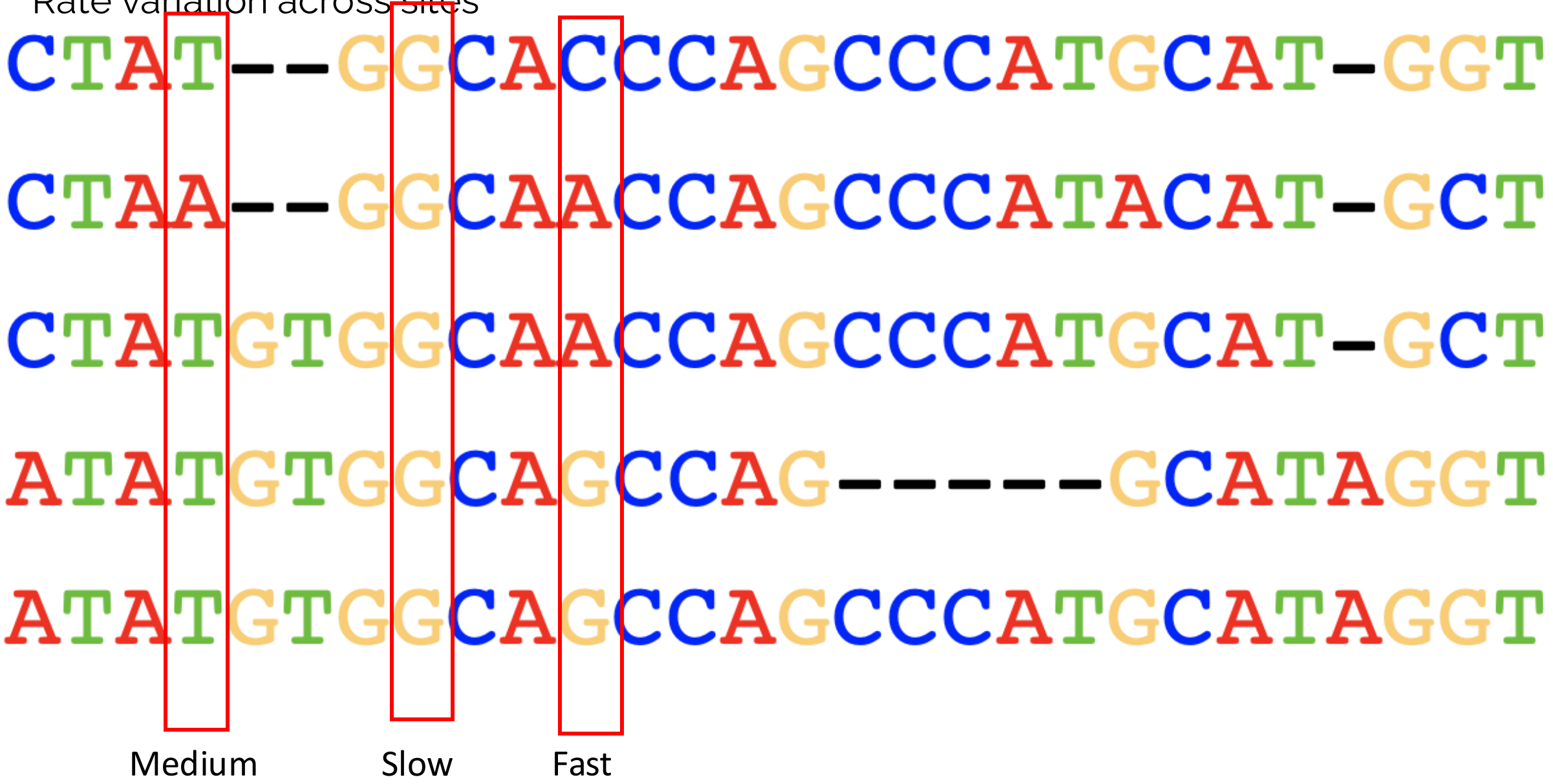


Homologous sites:
- Inherited from the common ancestor of all the species in the alignment
- Not informative for inferring information on the amount of evolution
i.e. branch lengths, times, evolutionary rates

Heterologous sites:
- Informative for inferring information on the amount of evolution

- Both type of sites are useful in phylogenetic studies as are insertions & deletions

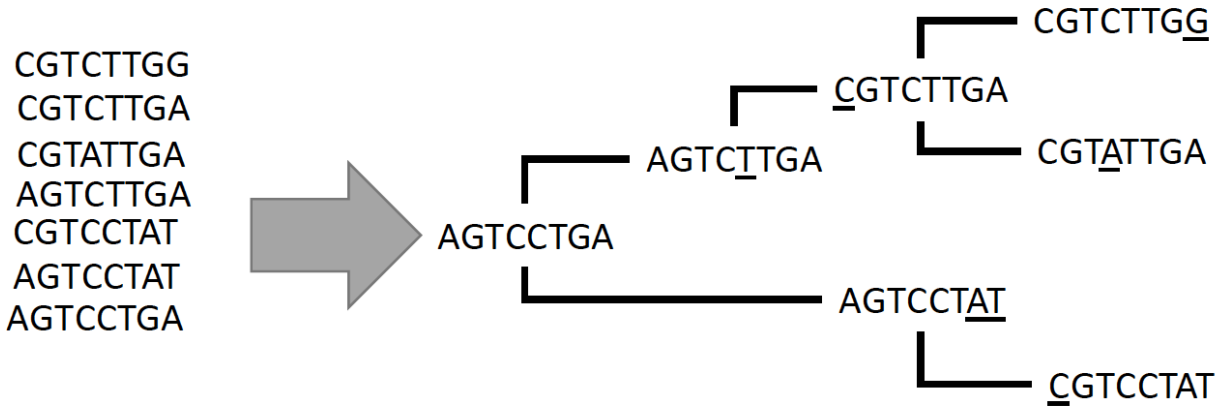
Rate variation across sites



Phylogenetic Methods

Fundamental assumptions are made in phylogenetic methods:

- Each aligned site represents a set of orthologous characters
  - i.e. a gene in different species that evolved from a common ancestor by speciation
- Sites and Lineages in alignment evolve independently
- Relationships amongst sequences can be represented by a bifurcating (division into two branches) tree



Common algorithms to construct a Phylotree

Table from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

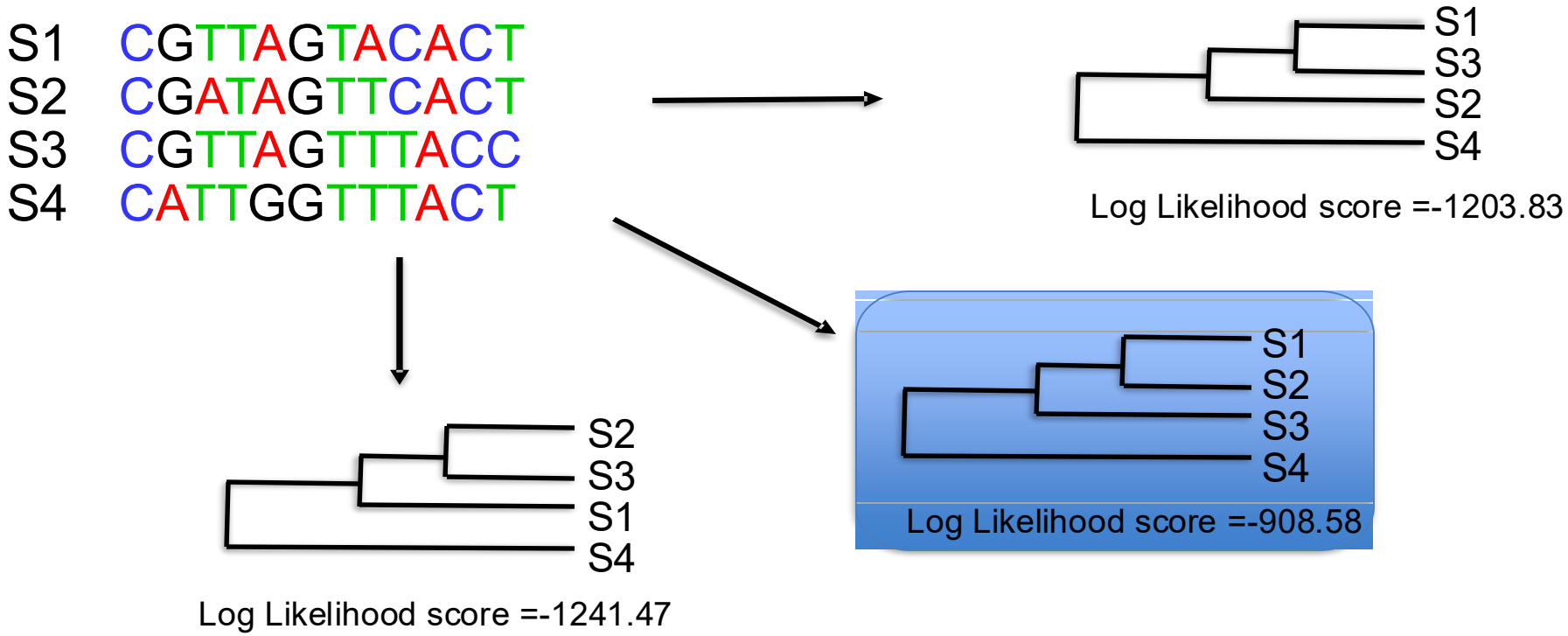
Table 1. The common algorithms used in phylogenetic tree construction.

Algorithm	Principle	Hypothesis	Criteria for Selecting the Final Tree	Scope of Application
NJ *	Minimal evolution: Minimizing the total branch length of the phylogenetic tree.	BME branch length estimation model: Ensuring general statistical consistency of minimum length phylogeny and non-negativity of its branch lengths [21].	In the end, only one tree was constructed.	Short sequences with small evolutionary distance and few informative sites.
MP	Maximum-parsimony criterion: Minimize the number of evolutionary steps required to explain the data set.	No model required.	The phylogenetic tree with the smallest number of base (or amino acid) substitutions during evolution.	Sequences with high sequence similarity, sequences for which it is difficult to design appropriate characteristic evolution models.
ML	Maximize likelihood value.	The sites in the alignment are independent; each branch is allowed to evolve at different rates.	Phylogenetic tree with maximum likelihood value.	Distantly related and small number of sequences.
BI	Bayes theorem.	Continuous-time Markov substitution model: Substitution probability is only related to the current nucleotide and has nothing to do with past nucleotides.	The most sampled phylogenetic tree in MCMC.	A small number of sequences.

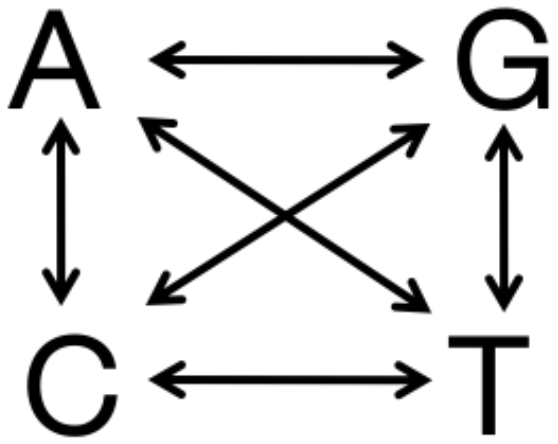
* NJ: a representative method and one of the most popular distance-based methods. The abbreviations in the table are as follows: NJ, neighbor-joining; MP, maximum parsimony; ML, maximum likelihood; BI, Bayesian inference; MCMC, Markov chain Monte Carlo.

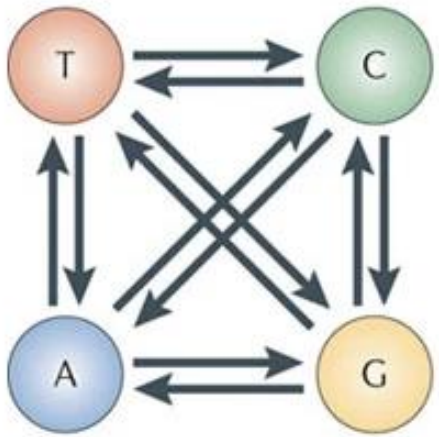
Focus: Maximum Likelihood (ML)

- Maximum Likelihood is a statistical method of estimating the probability distributions to assign probabilities to a particular possible phylogenetic tree
- It allows for varying rates of evolution across both lineages & nucleotide sites



Rate Matrix





Base Frequencies

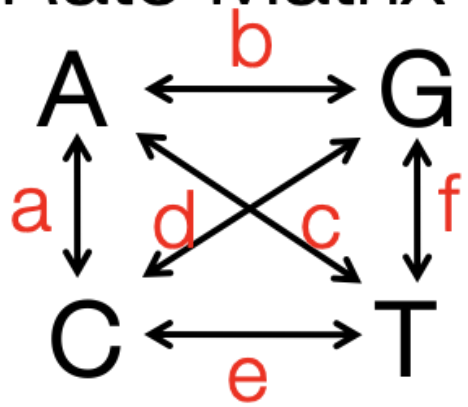
$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

Nucleotide frequencies: F

Site Rates

$$+ I + G$$

Rate Matrix



Base Frequencies

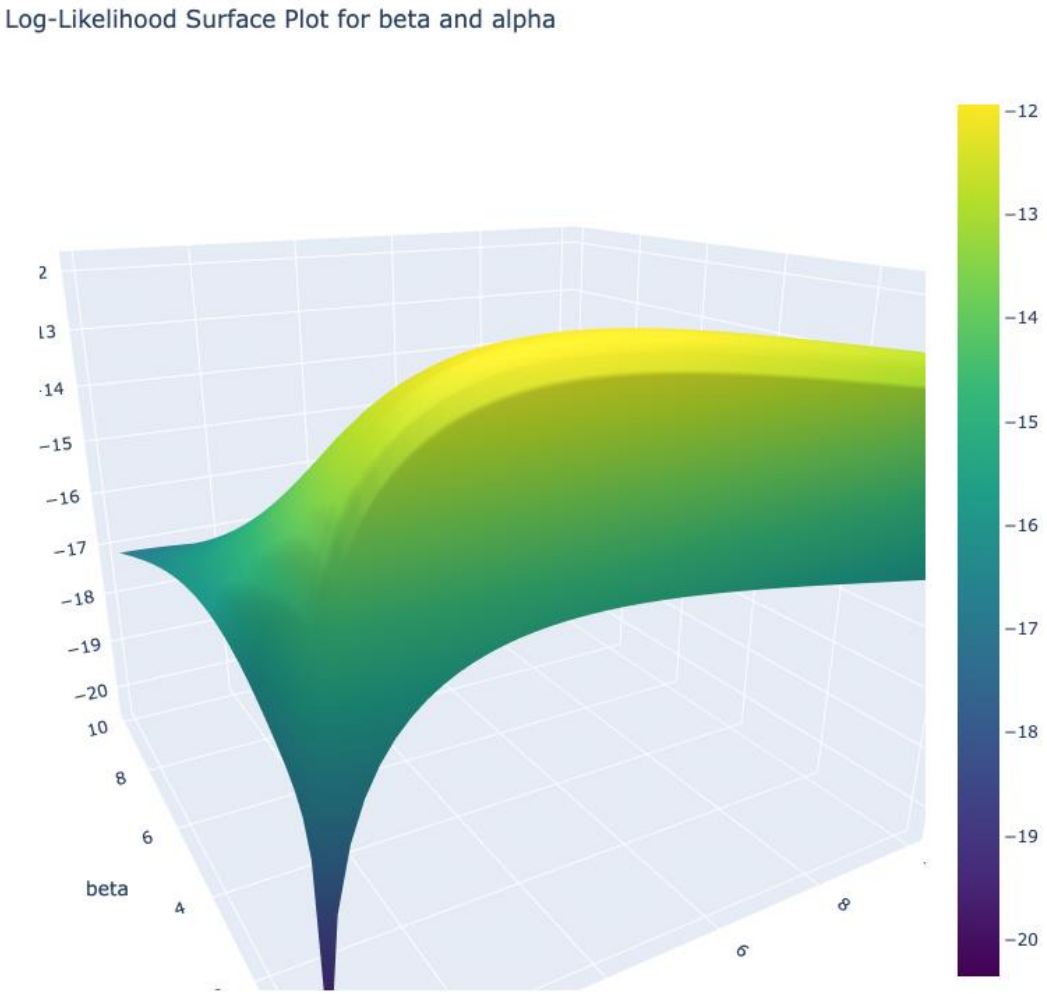
$\pi_A + \pi_C + \pi_G + \pi_T = 1$

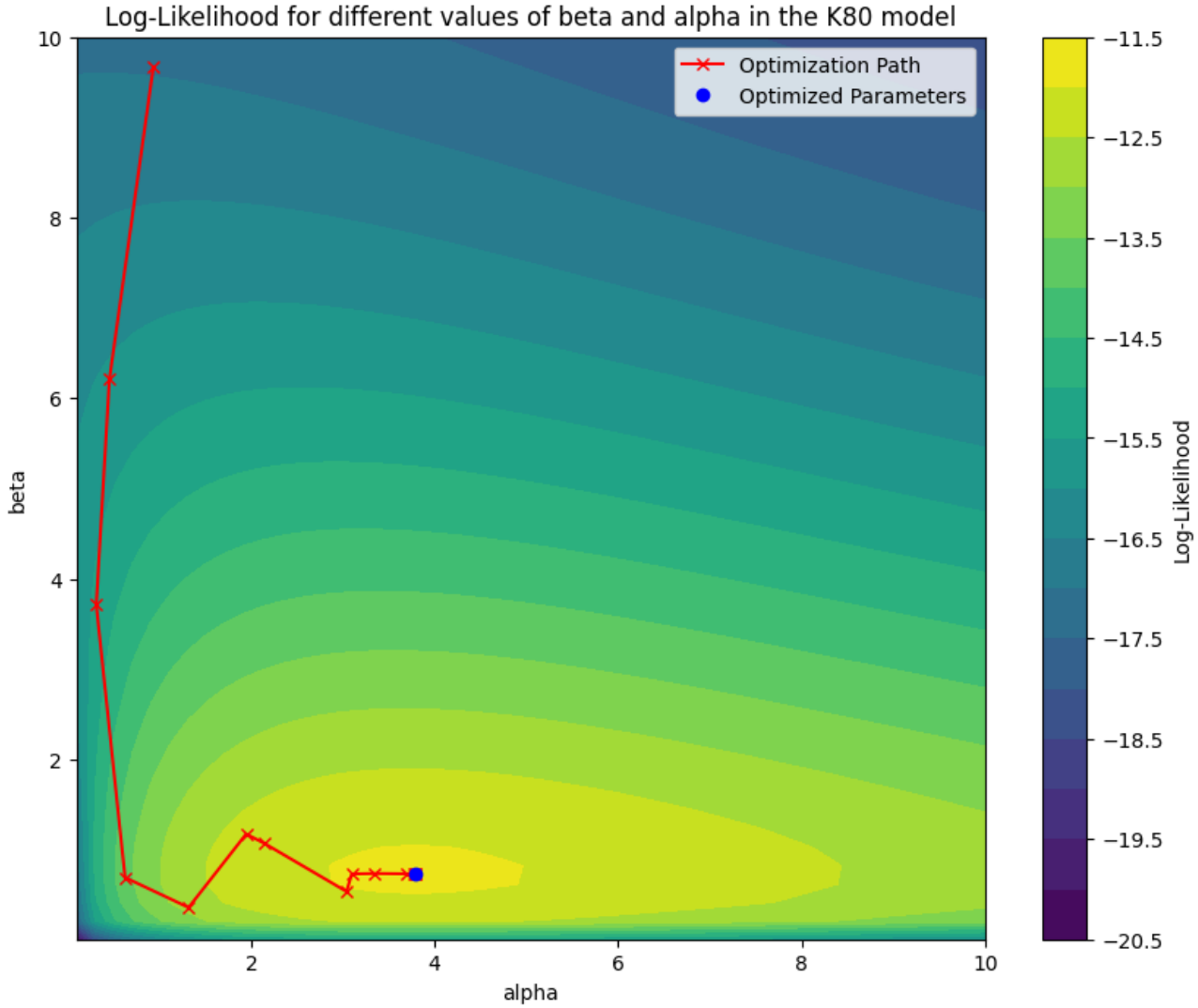
Site Rates

$+ I + G$

JC	HKY	GTR	GTR+I+G
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	π_A, π_C, π_G, π_T	π_A, π_C, π_G, π_T	π_A, π_C, π_G, π_T
No I or G	No I or G	No I or G	I, G
0 free parameters	4 free parameters	8 free parameters	10 free parameters

Advanced: Maximum Likelihood Internals



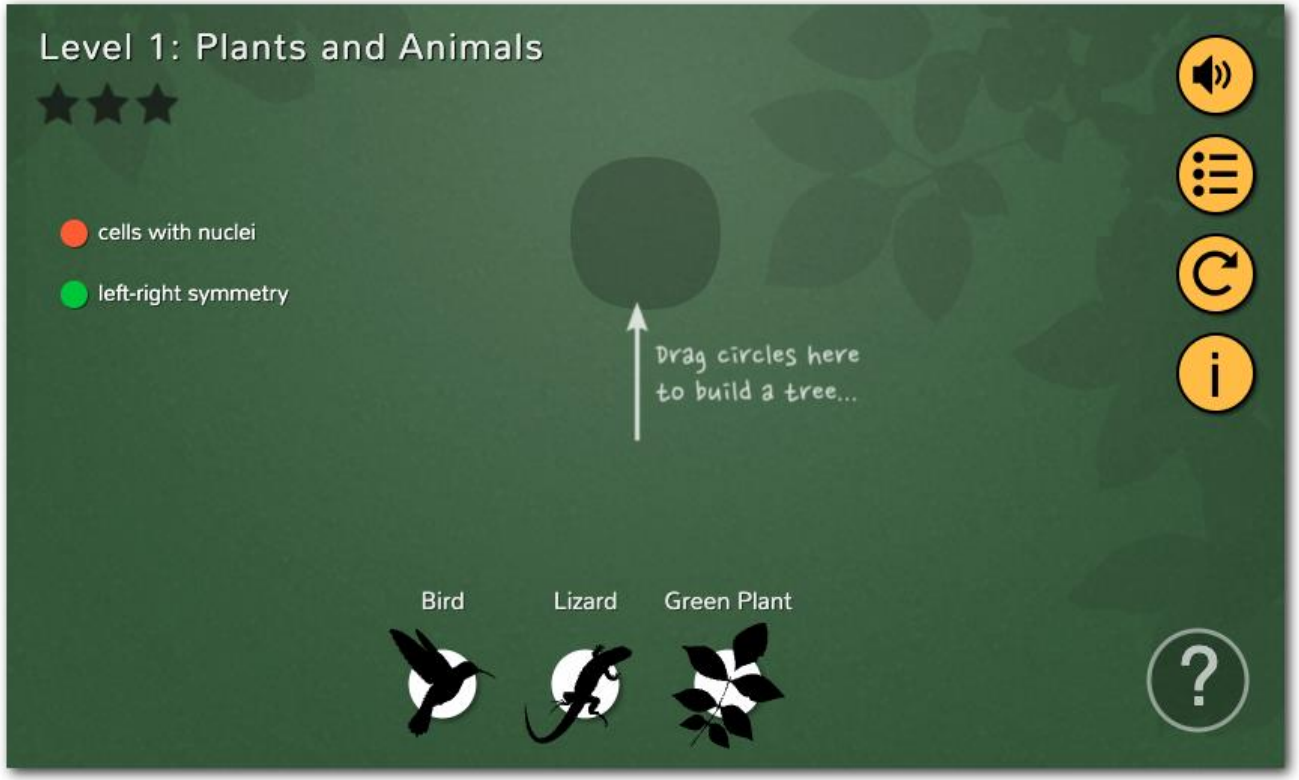


https://colab.research.google.com/gist/Wytamma/2304152006 39d8266dc2f4f34eef23d0/k80_likelihood.ipynb

Let's play a game!

Go to this webpage:
https://tidal.northwestern.edu/blog/bat/

Hard mode:
https://www.pbs.org/wgbh/nova/labs/lab/evolution/



Questions? + Resources

EBI Course on phylogenetics:

https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics

Khan Academy on Phylogenetic Trees
https://www.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/phylogenetic-trees

Good publication for deep dive into phylogenetic methods and processes:
https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635