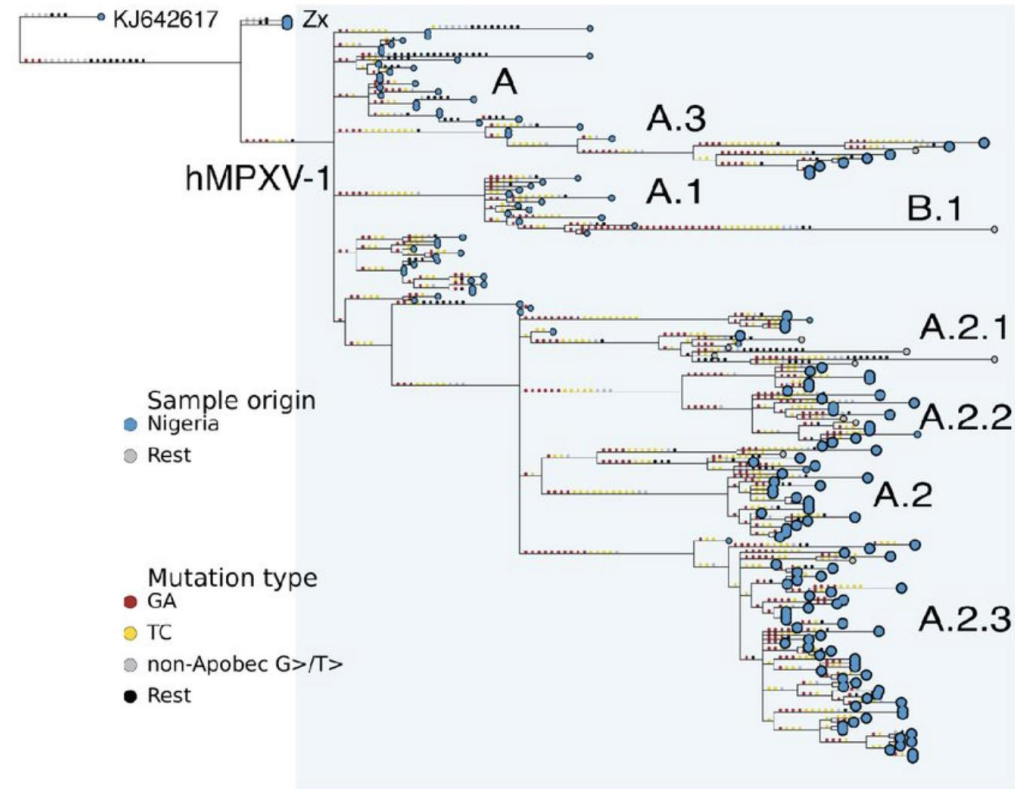


# 7 – Introduction to Phylogenetics

—  
Wytamma Wirth  
(@wytamma)

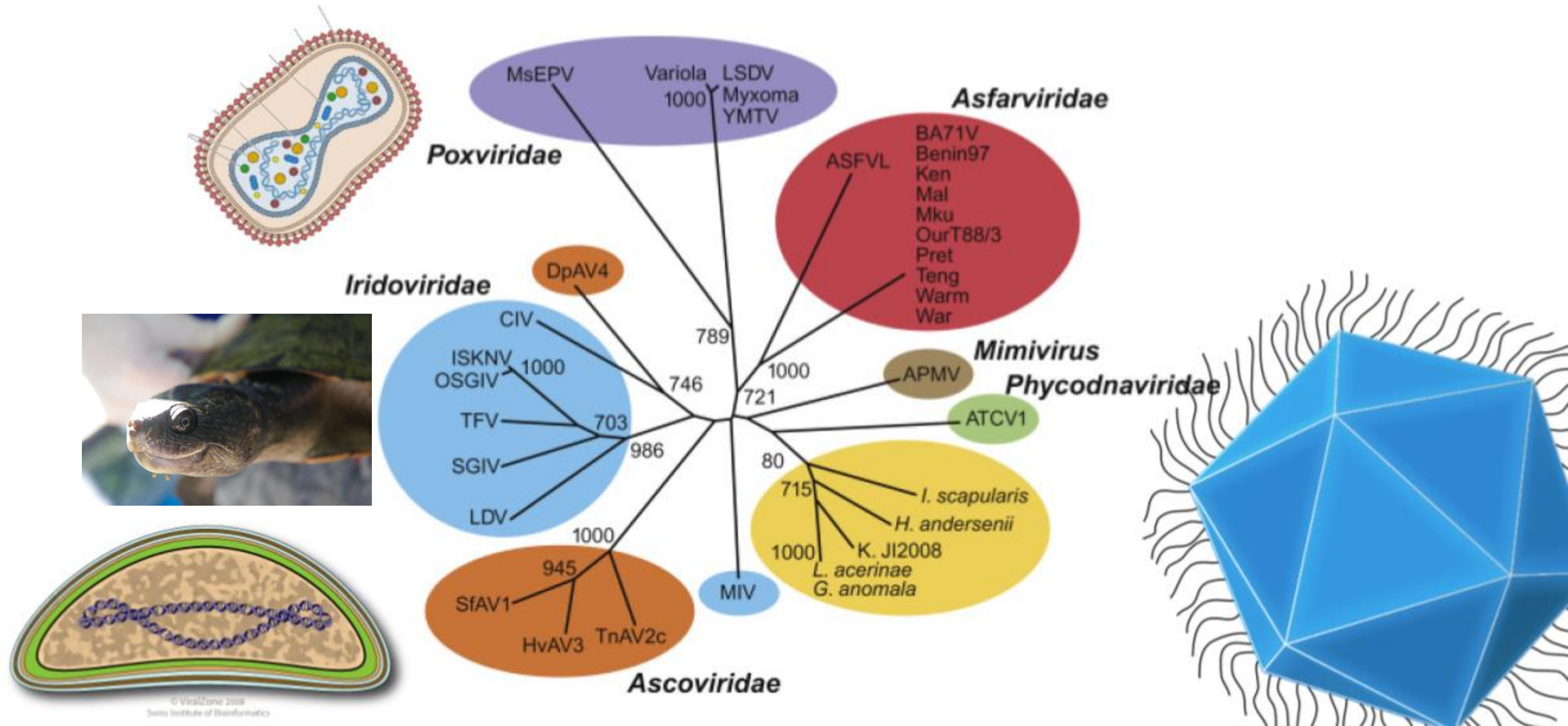
# Objectives

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





# Nucleocytoplasmic large DNA viruses (NCLDV)s



# What is Phylogenetics?

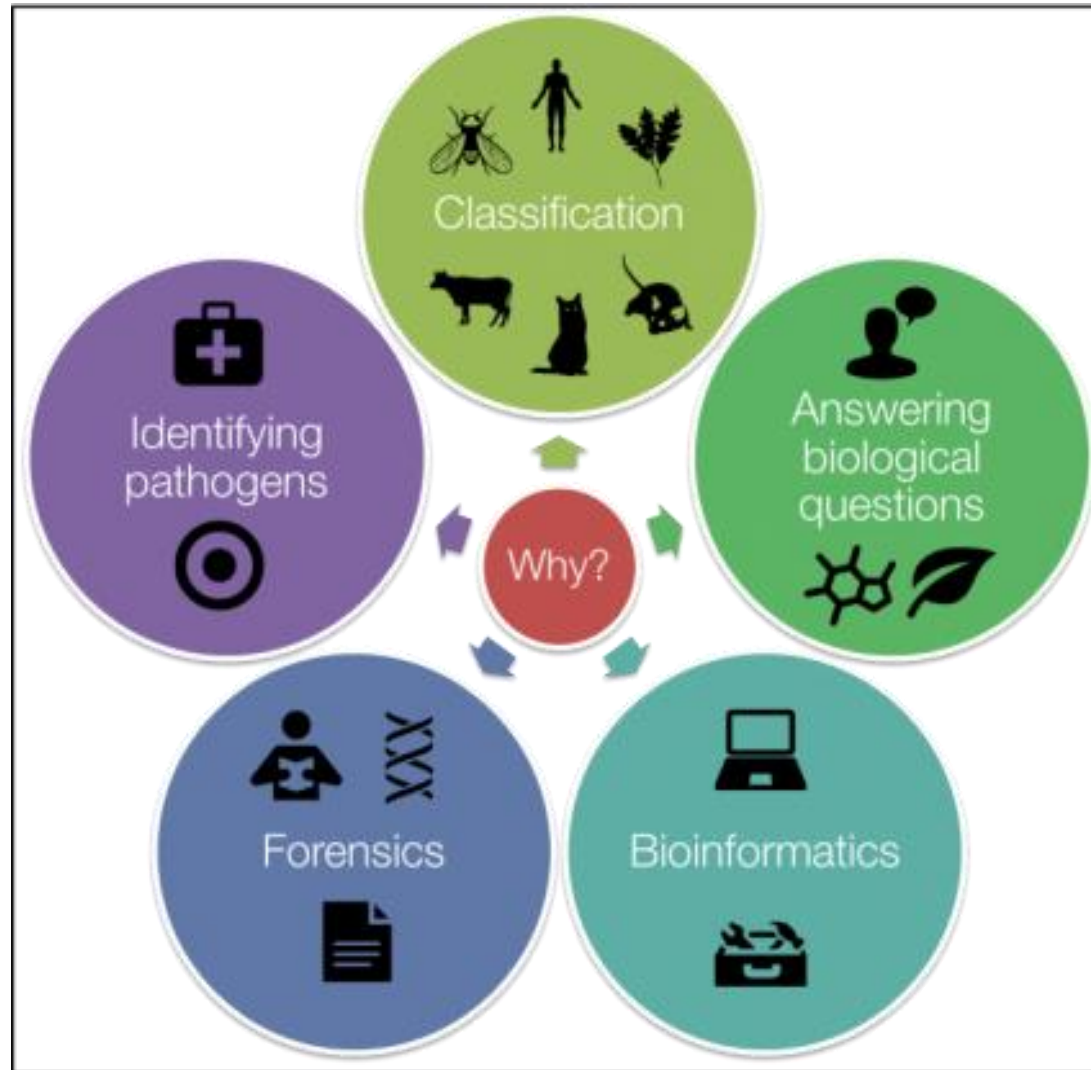
**Defintion:** Phylogenetics is the study of evolutionary relationships among biological entities – often species, individuals, genes, genomes. A phylogenetic tree is created (constructed) by looking at the nucleotide or protein sequences and combining with our understanding of sequence evolution.

## **Typical questions answered using phylogenetics:**

- 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.

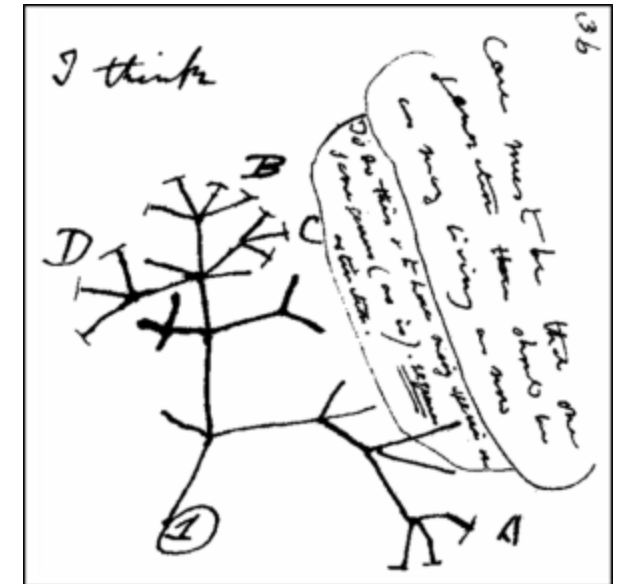
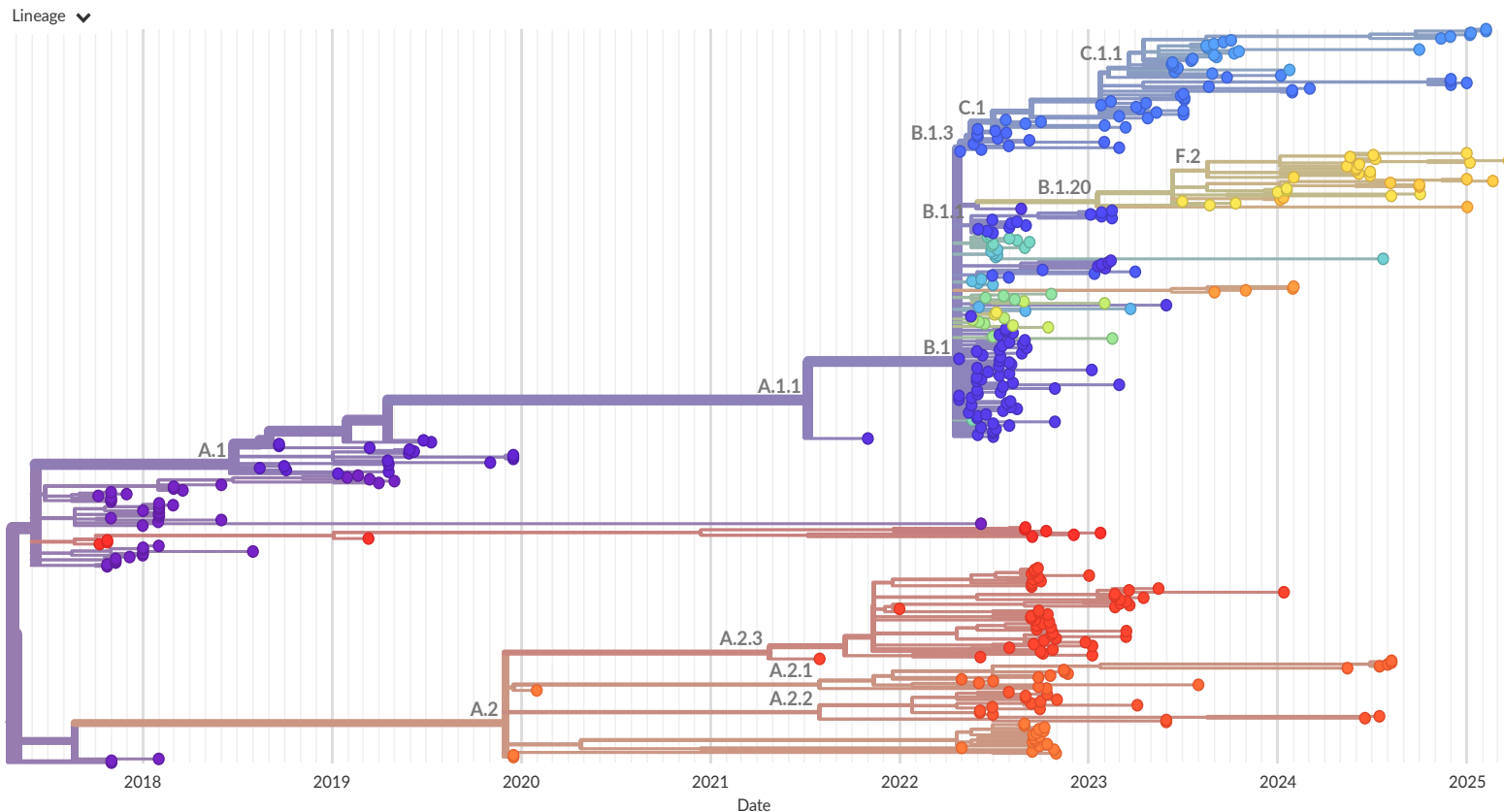
**Forensics:** Phylogenetics is used to assess DNA evidence presented in court cases to inform situations, e.g. where someone has committed a crime, when food is contaminated, or where the father of a child is unknown.

**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.

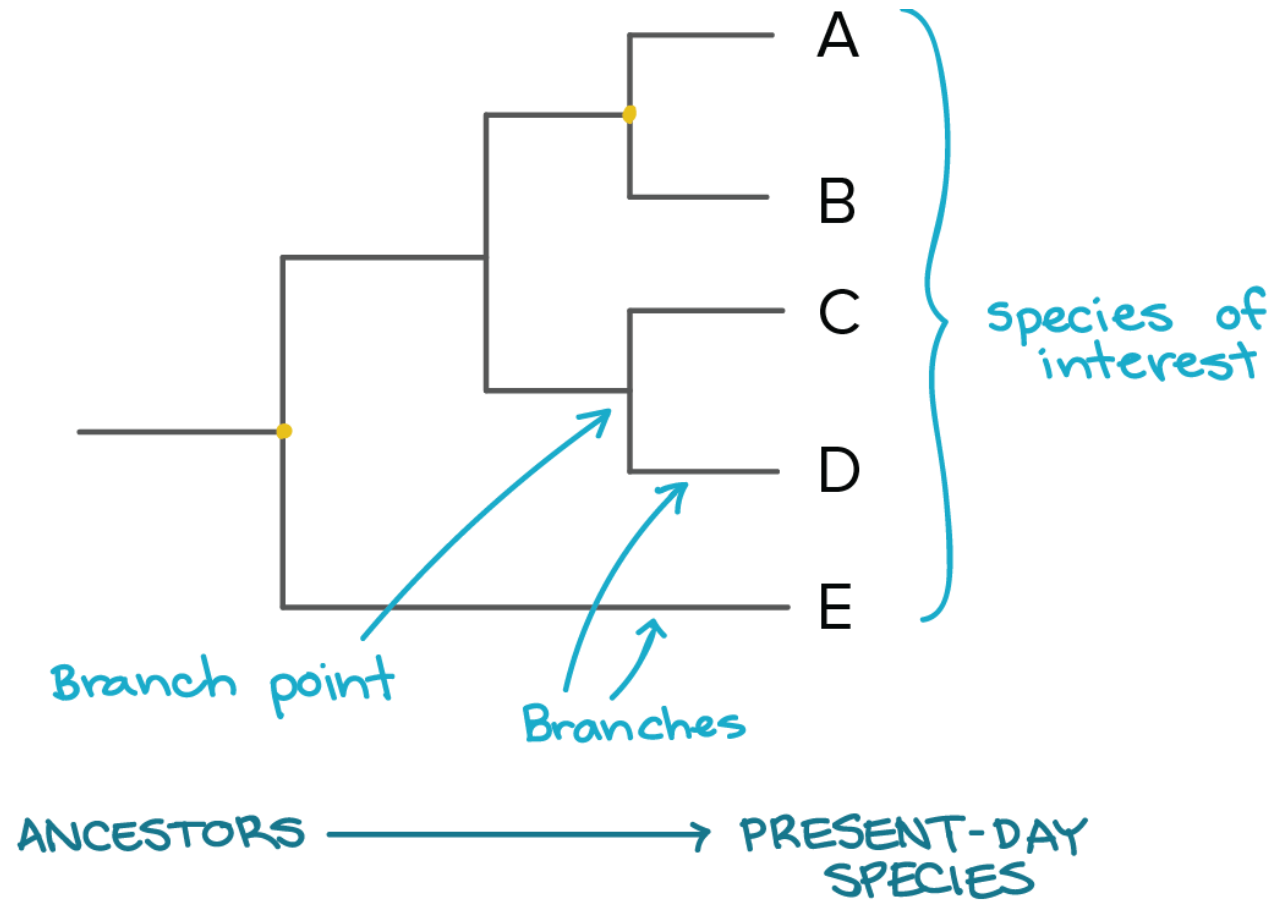
# What is a phylogeny?

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.



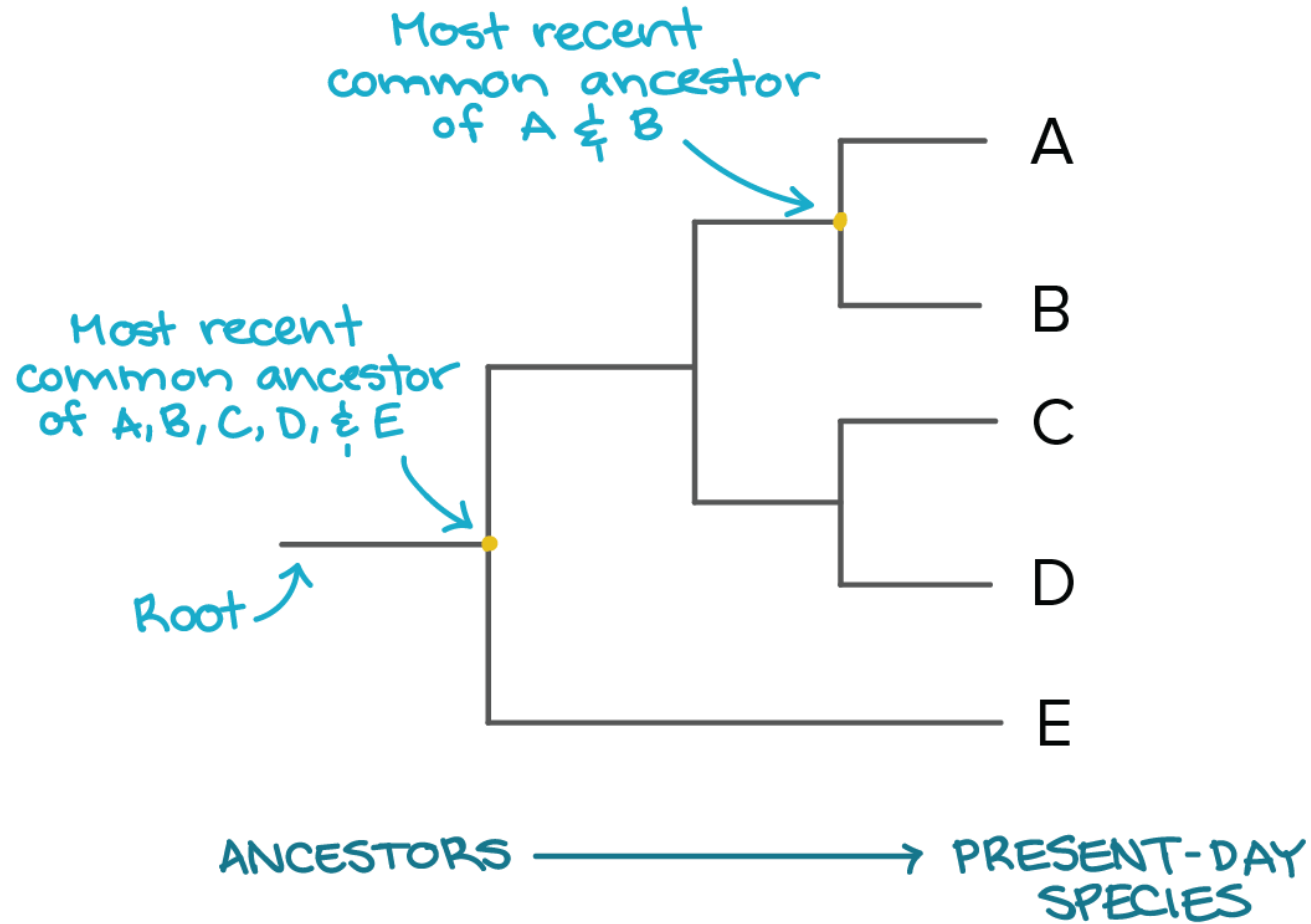
One of the first drawings of a phylogeny by Charles Darwin.

# Parts of Phylogeny (or Tree)

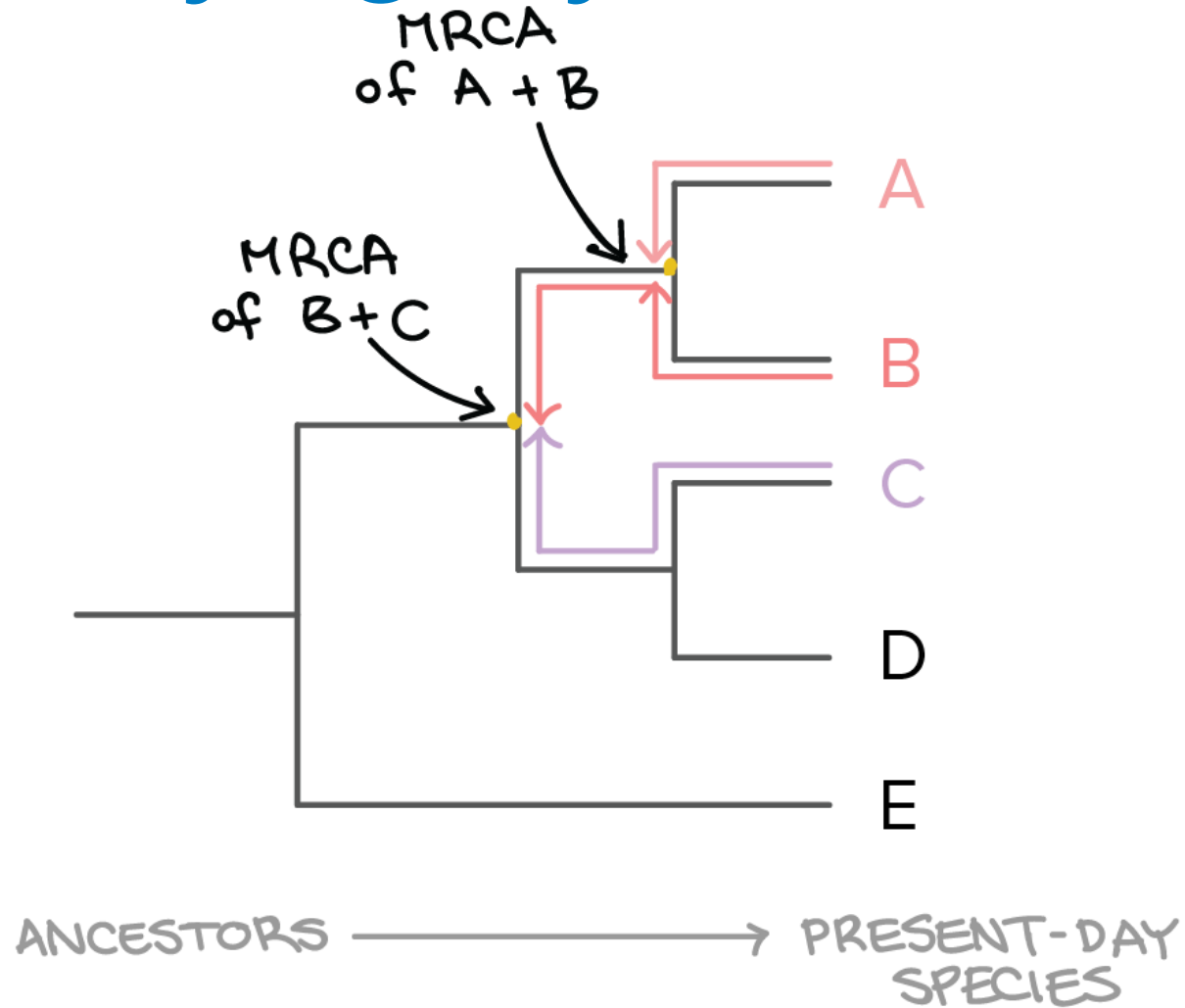




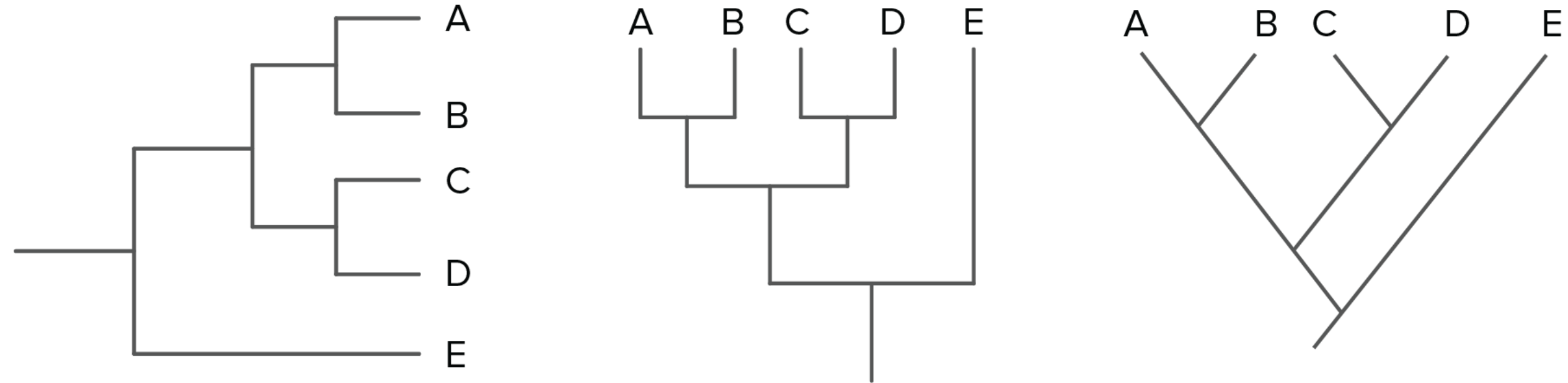
# Parts of Phylogeny (or Tree)



# Parts of Phylogeny (or Tree)

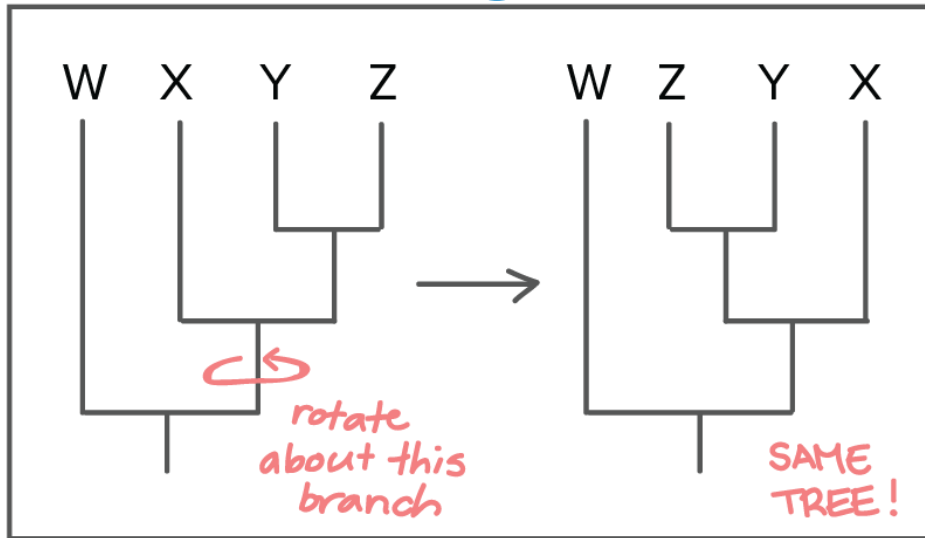


# Tips for reading trees

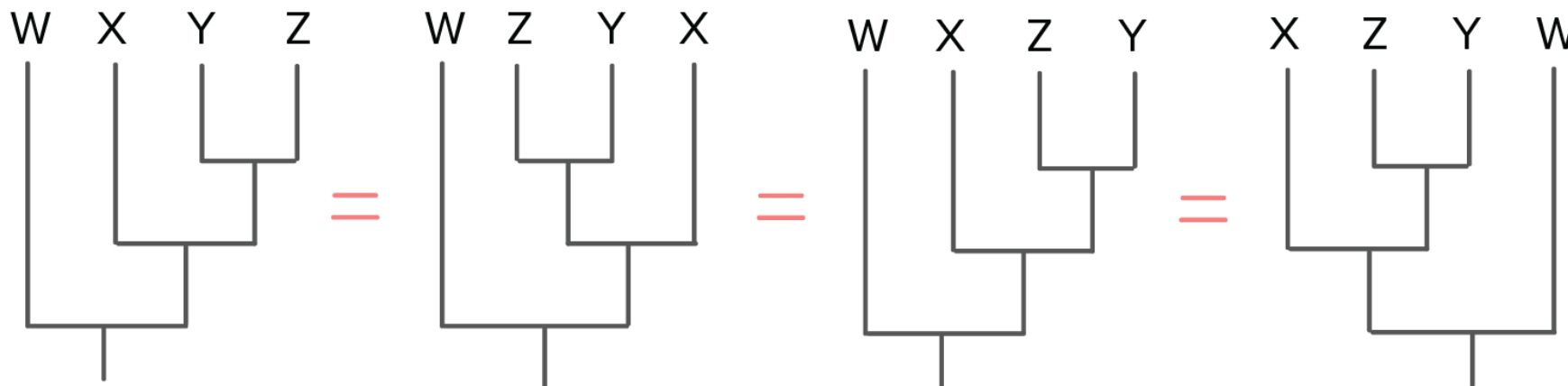


The three trees above represent identical relationships among species A, B, C, D, and E.

# Tips for reading trees

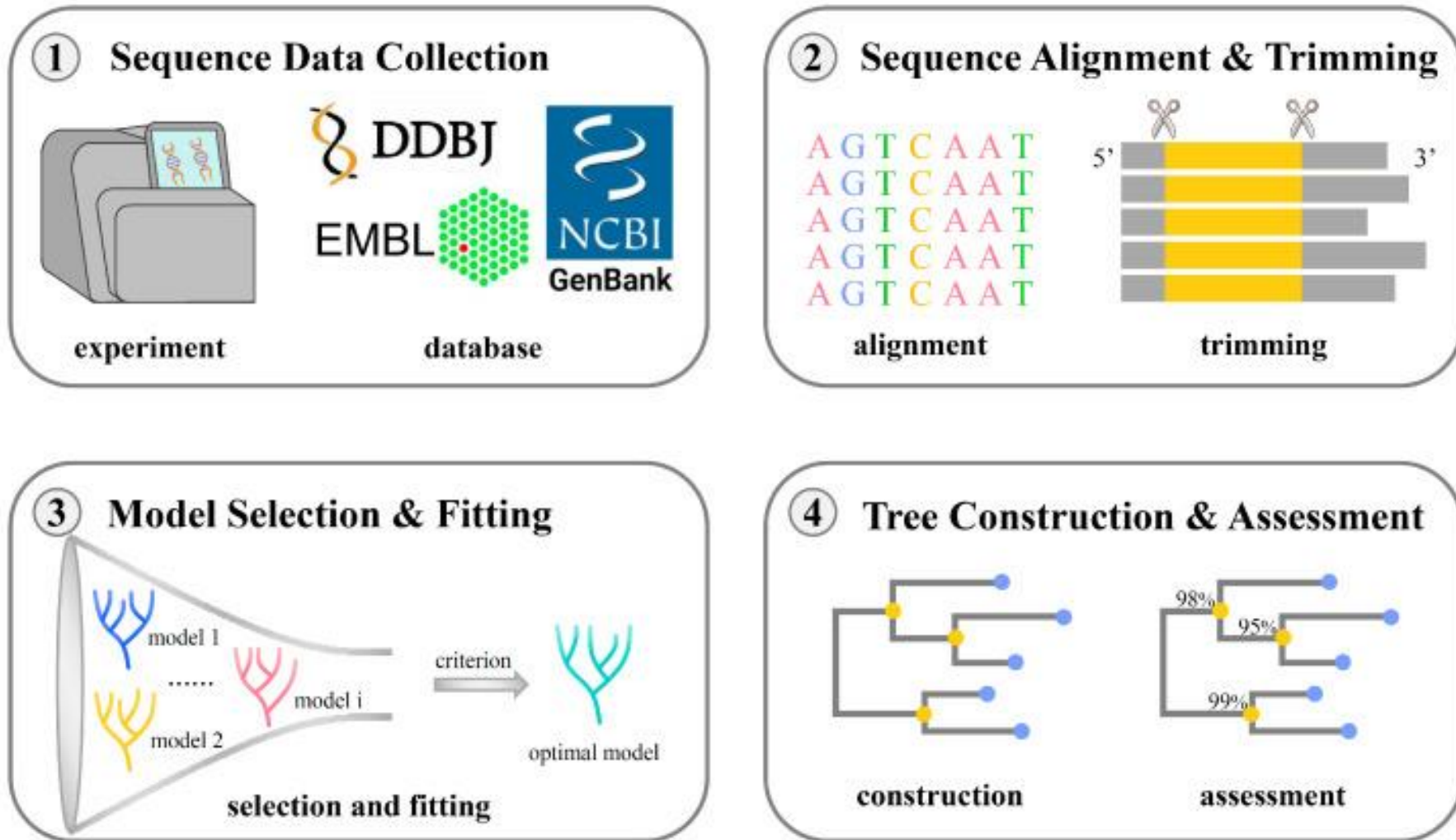


Another critical point about these trees is that if you rotate the structures, using one of the branch points as a pivot, you don't change the relationships.





# Major stages in phylogenetic tree analysis



# 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

126. A/New_Caledonia/99/2019_jul	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
127. A/South_Australia/39/2019e_feb	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
128. 131K_A/Victoria/23/2018_oct	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
129. A/Sri_Lanka/25/2019_jun	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
130. A/Sri_Lanka/27/2019_jul	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
131. A/Victoria/2/2019_jan	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
132. 135K_A/Fiji/71/2017_sep	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
133. 135K_A/VICTORIA/653/2017e	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
134. 135N_A/PHILIPPINES/13/2017_sep	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	A	T	C	
135. 135N_A/SYDNEY/22/2018e_mar	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
136. A/Malaysia/RP0049/2019	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
137. A/South_Africa/R03989/2019_apr	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
138. A/Hong_Kong/681/2018e	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
139. A/Timor_Leste/5/2019_mar	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
140. A/Macau/603415/2019_may	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
141. A/Victoria/213/2019_aug	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
142. cdcA/Hong_Kong/45/2019	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
143. A/South_Australia/2/2019e_jan	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C
144. A/South_Australia/4/2019e_jan	T	C	A	A	T	T	G	G	A	C	T	T	G	G	A	G	T	C	A	A	A	C	G	G	A	A	C	A	A	G	T	T	C	T	G	C	T	T	G	C	A	T	A	A	G	G	G	A	T	C

## Homologous sites:

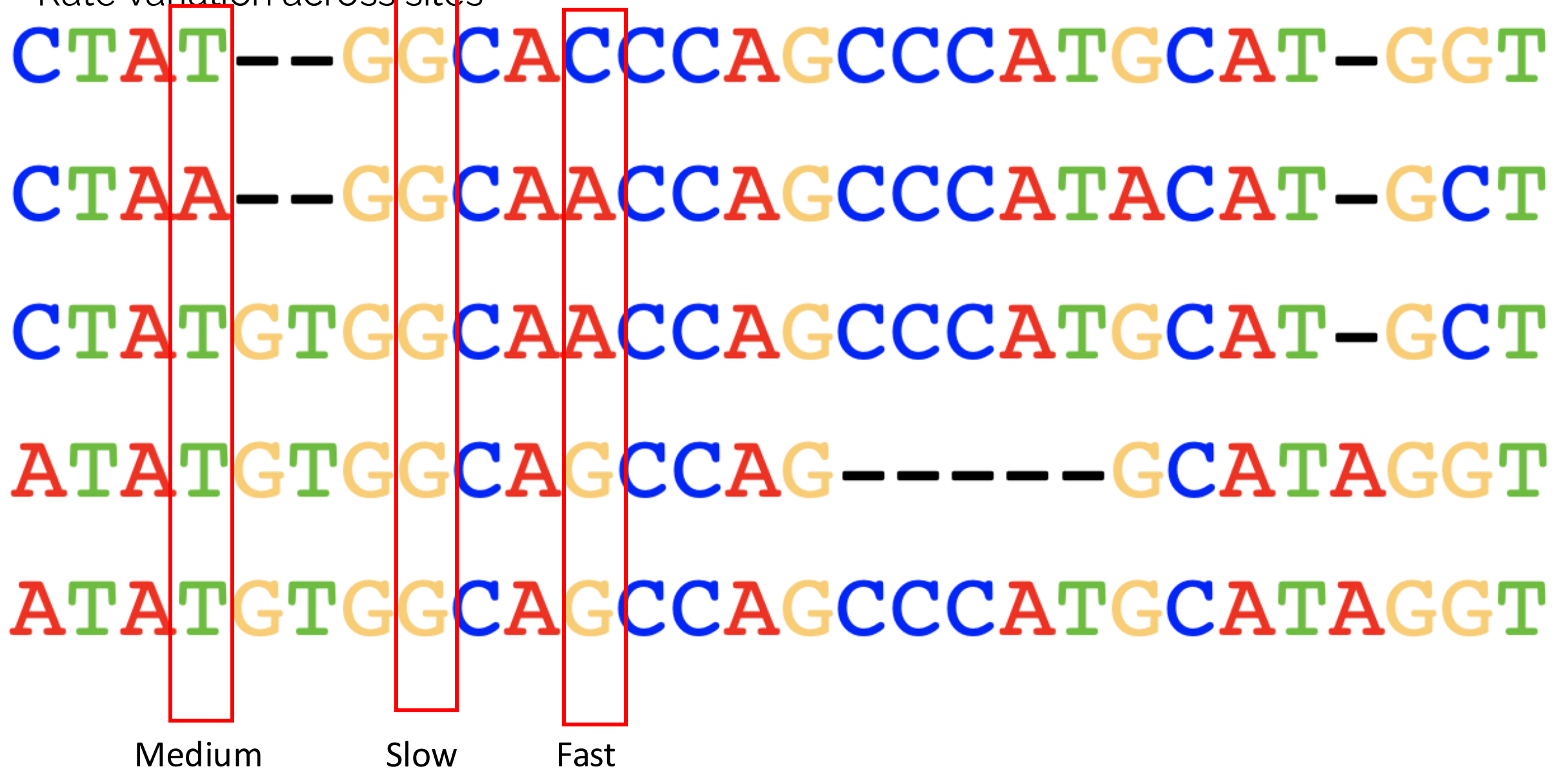
- Inherited from the common ancestor of all the species in the alignment
- Not informative for inferring information on the amount of evolution  
i.e. branch lengths, times, evolutionary rates

## 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

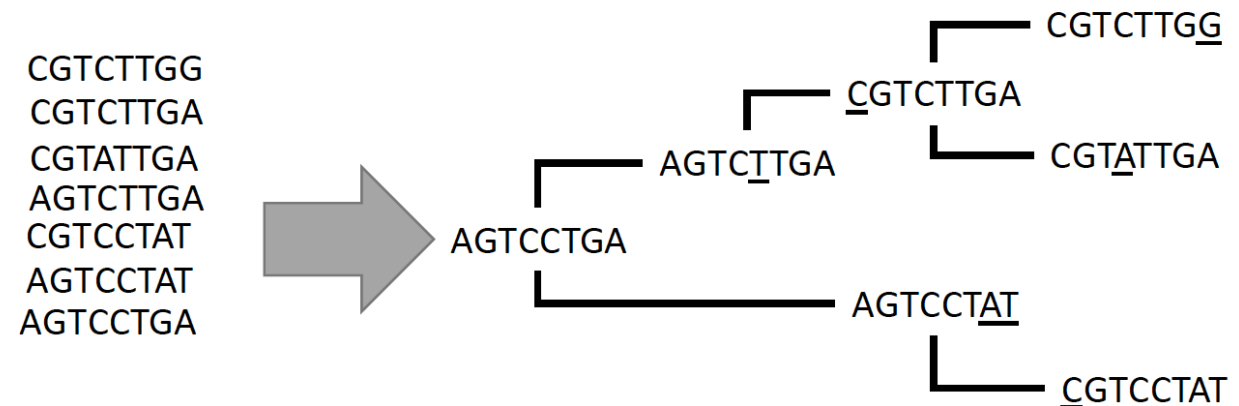




# Phylogenetic Methods

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

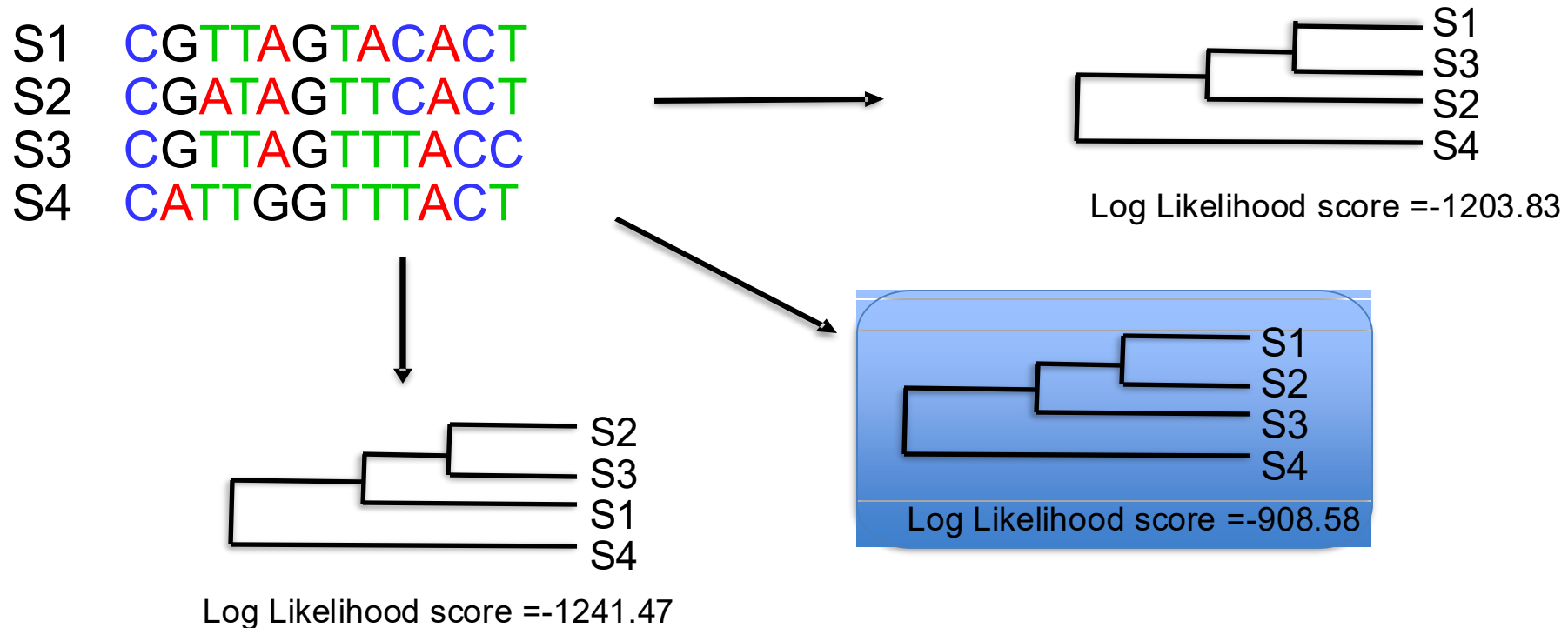
**Table 1.** The common algorithms used in phylogenetic tree construction.

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.

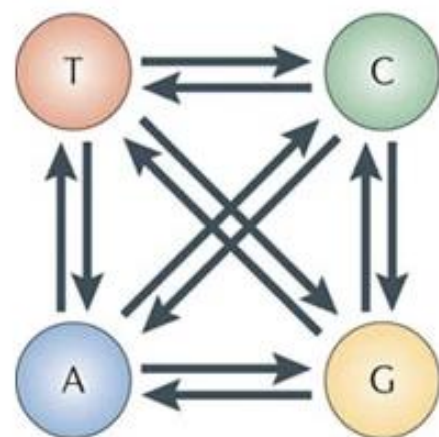
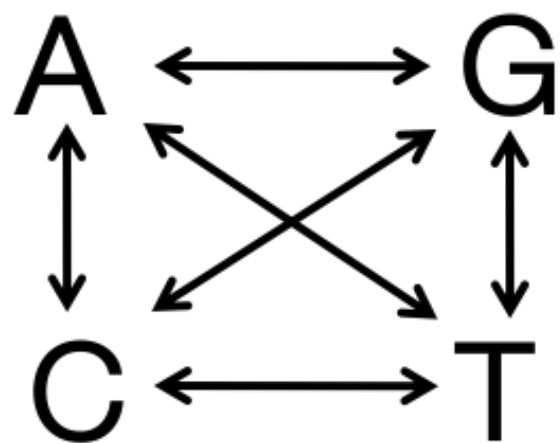
# Focus: Maximum Likelihood (ML)

- 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

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

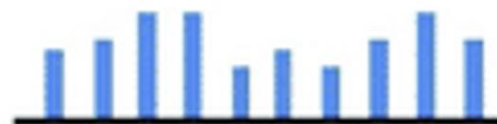


Nucleotide  
frequencies: **F**

## Site Rates

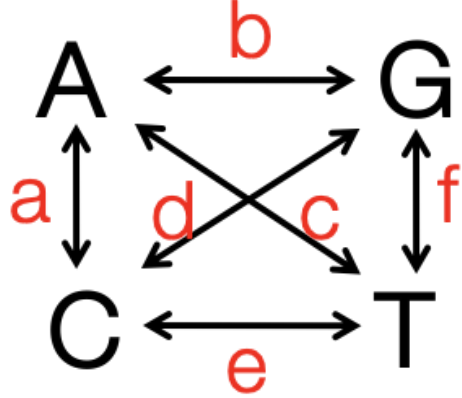
+ I + G

- C **G** A A A C A **T** G  
 A A **G** A G - - A **T** A  
 T A **G** A G C - T **T** G  
 A A **G** A A A A A **T** G



- C G A A A C A T G  
 A A G A G - - A T A  
 T A G A G C - T T G  
 A A G A A A A A T G

Rate Matrix



Base Frequencies

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

Site Rates

$$+ I + G$$

JC

$$a=b=c=d=e=f$$

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

No I or G

0 free  
parameters

HKY

$$a=c=d=f, b=e$$

$$\pi_A, \pi_C, \pi_G, \pi_T$$

No I or G

4 free  
parameters

GTR

$$a, b, c, d, e, f$$

$$\pi_A, \pi_C, \pi_G, \pi_T$$

No I or G

8 free  
parameters

GTR+I+G

$$a, b, c, d, e, f$$

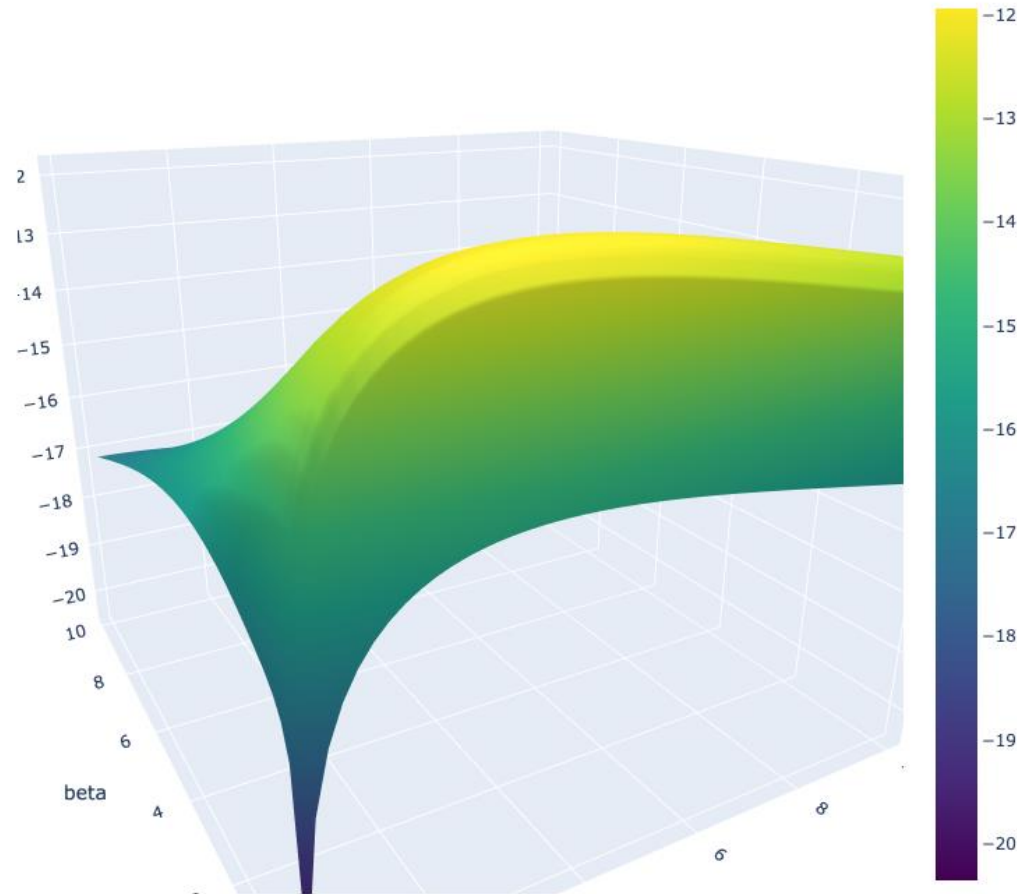
$$\pi_A, \pi_C, \pi_G, \pi_T$$

I, G

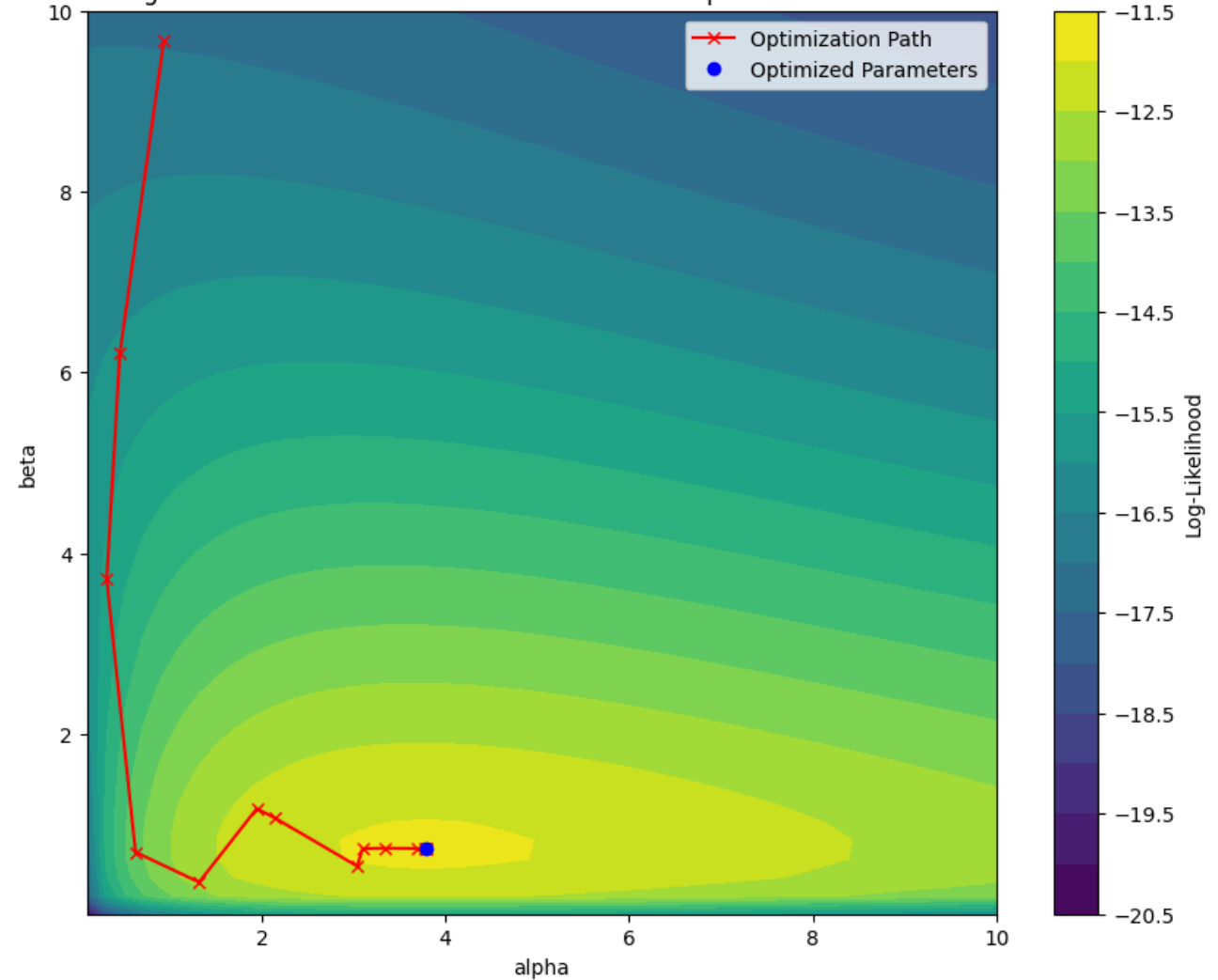
10 free  
parameters

# Advanced: Maximum Likelihood Internals

Log-Likelihood Surface Plot for beta and alpha



Log-Likelihood for different values of beta and alpha in the K80 model



[https://colab.research.google.com/gist/Wytamma/230415200639d8266dc2f4f34eef23d0/k80\\_likelihood.ipynb](https://colab.research.google.com/gist/Wytamma/230415200639d8266dc2f4f34eef23d0/k80_likelihood.ipynb)

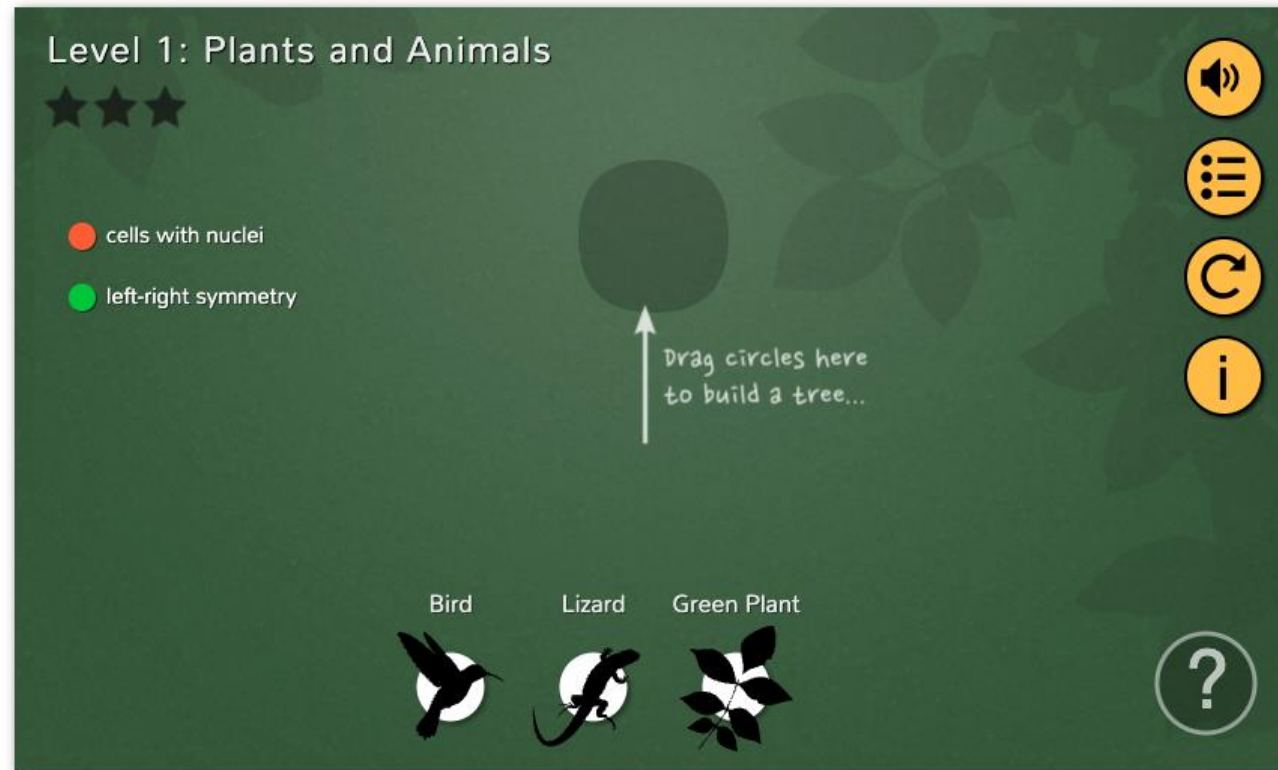
# Let's play a game!

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>