



GenEpi-BioTrain

Molecular phylogenetics

May 2024

Intended Learning Objectives

1. Understand the basic principles of phylogeny, including substitutions models, optimality criteria and branch support.
2. Understand what is a phylogenetic tree and how to read it.
3. Learn the impacts of recombination on phylogenetic inference.

Outline

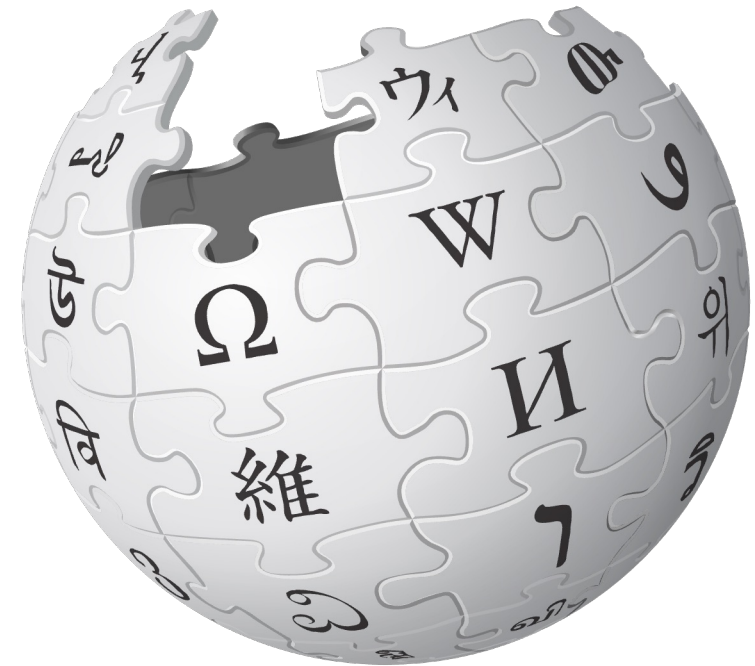
- 1. The tree:** tips, root, branches, etc.
- 2. The data:** distances, sequences, notion of homology
- 3. The methods:** clustering, MP, ML, Bayesian
- 4. Exploring the space of solutions:** tree rearrangements, hill climbing
- 5. Branch support:** Bootstrap, UF-boot, TBE, aLRT

Part 1: The tree

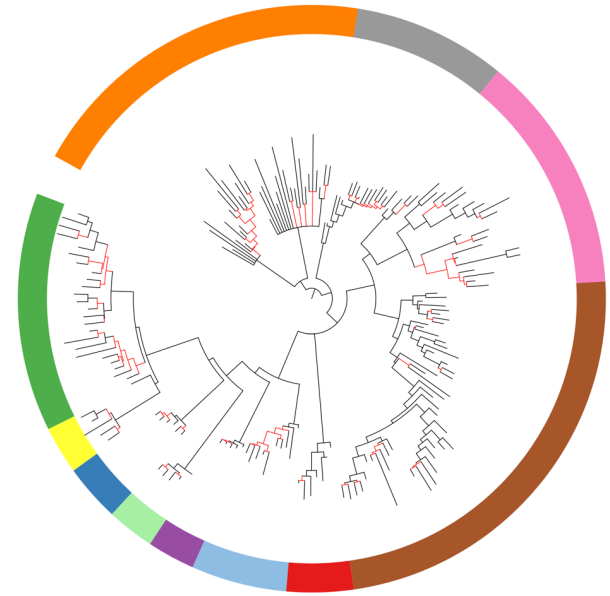
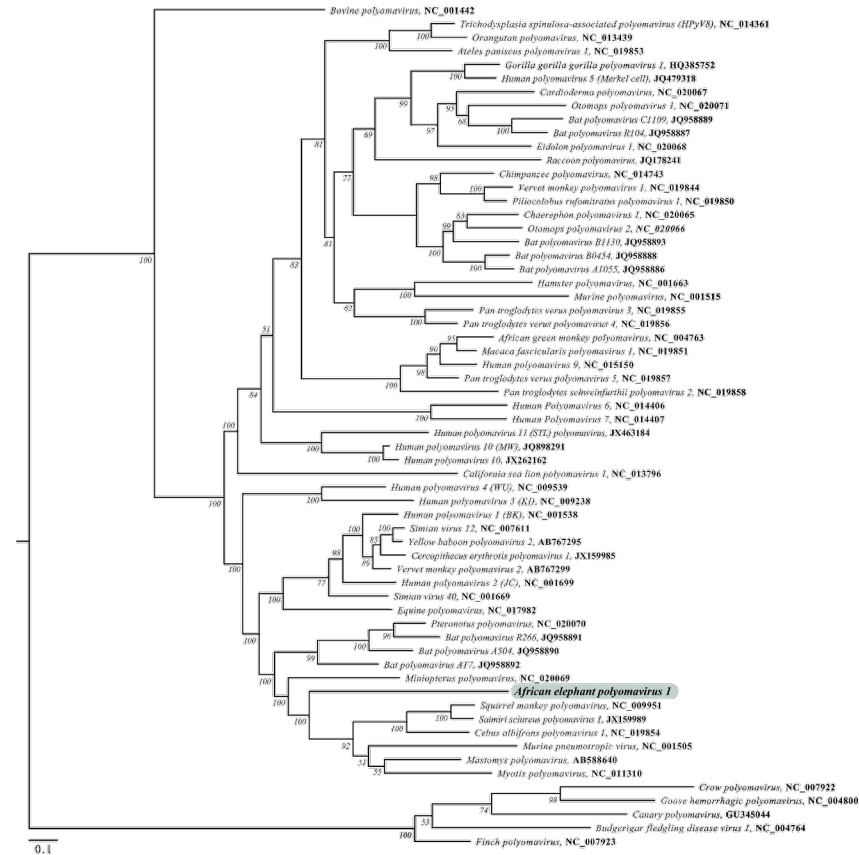
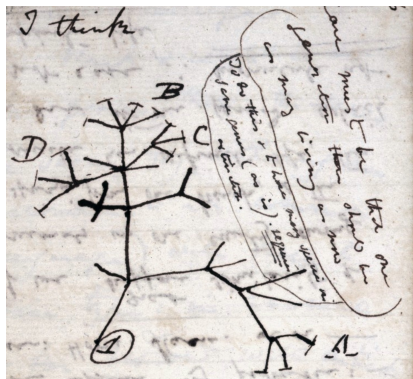
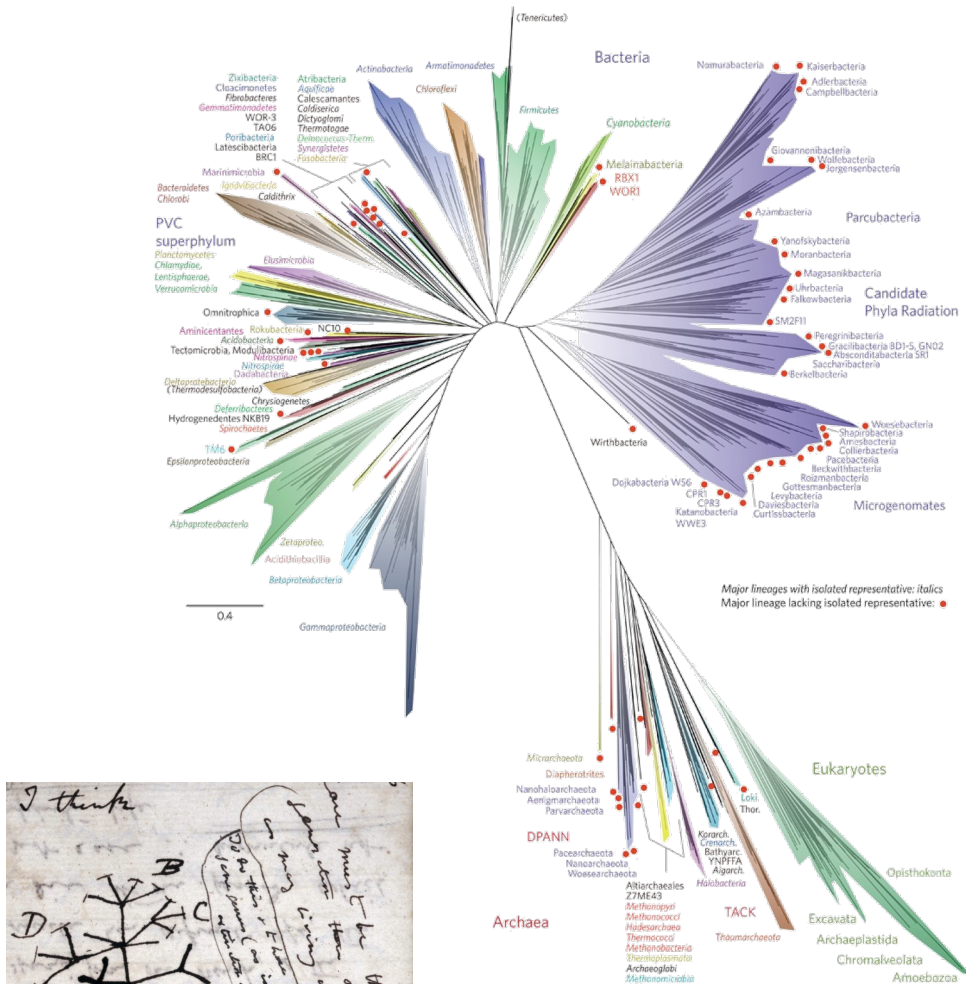
1. A phylogenetic tree

A phylogenetic tree, phylogeny or evolutionary tree is a graphical representation describing the **evolutionary history** between a set of species/taxa/sequences/etc.

In other words, it is a branching diagram or a tree showing the **evolutionary relationships** among various biological species or other entities based upon similarities and differences in their physical or genetic characteristics.

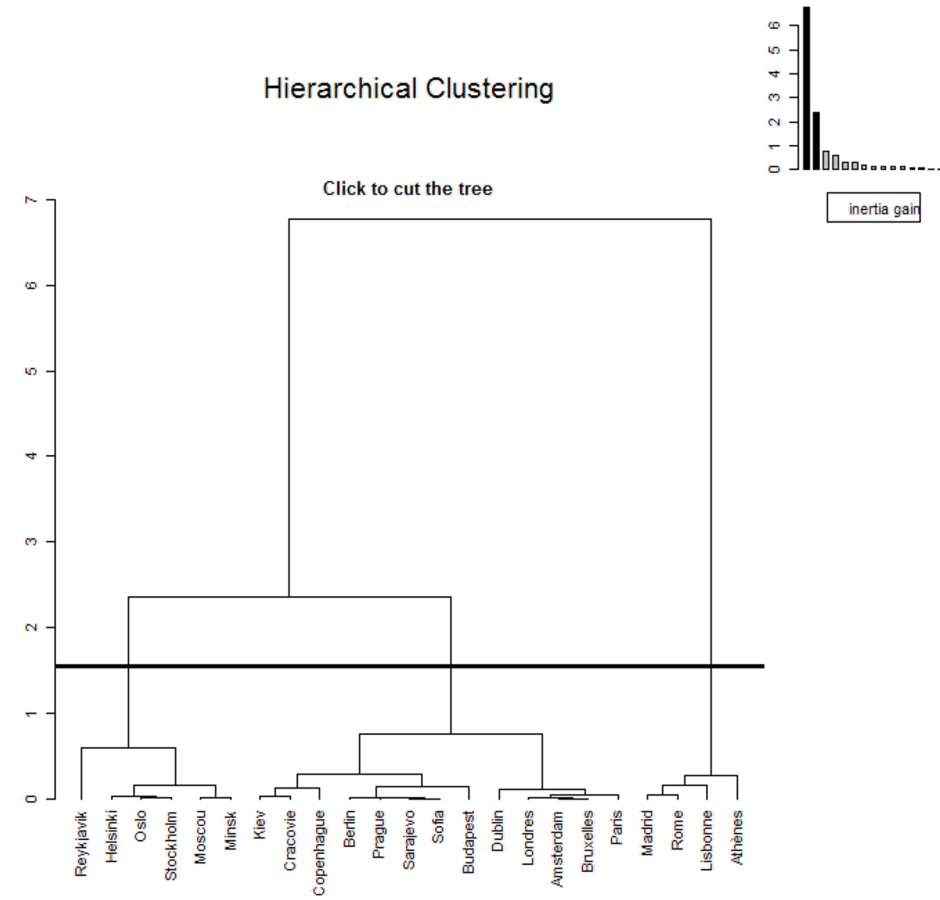
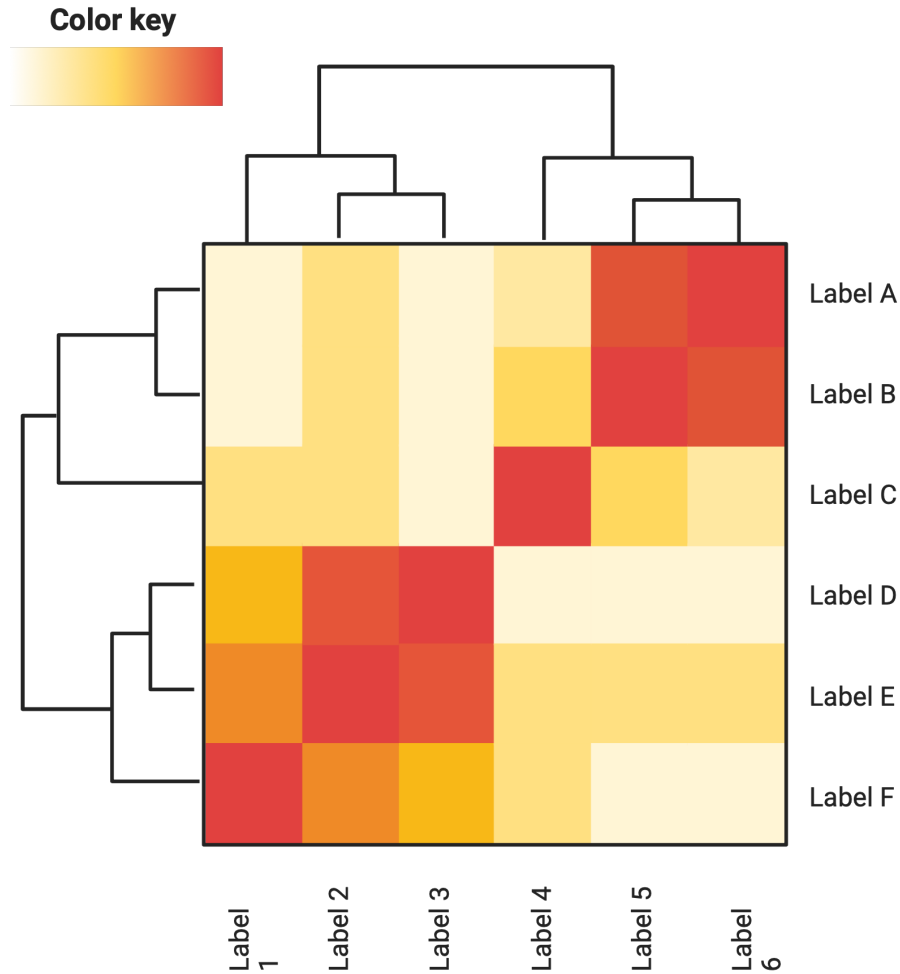


1. These are phylogenetic trees



From left to right:
I think by Charles Darwin / Public Domain
Figure 1 by Hug, L., Baker, B., Anantharaman, K. *et al.* / [CC BY 4.0](#)
Figure 3 by Stevenes, H.. *et al.* / [CC BY 2.0](#)
Modified from **Figure 4** by Colcombet-Cazenave, B. *et al.* / [CC0 1.0](#)

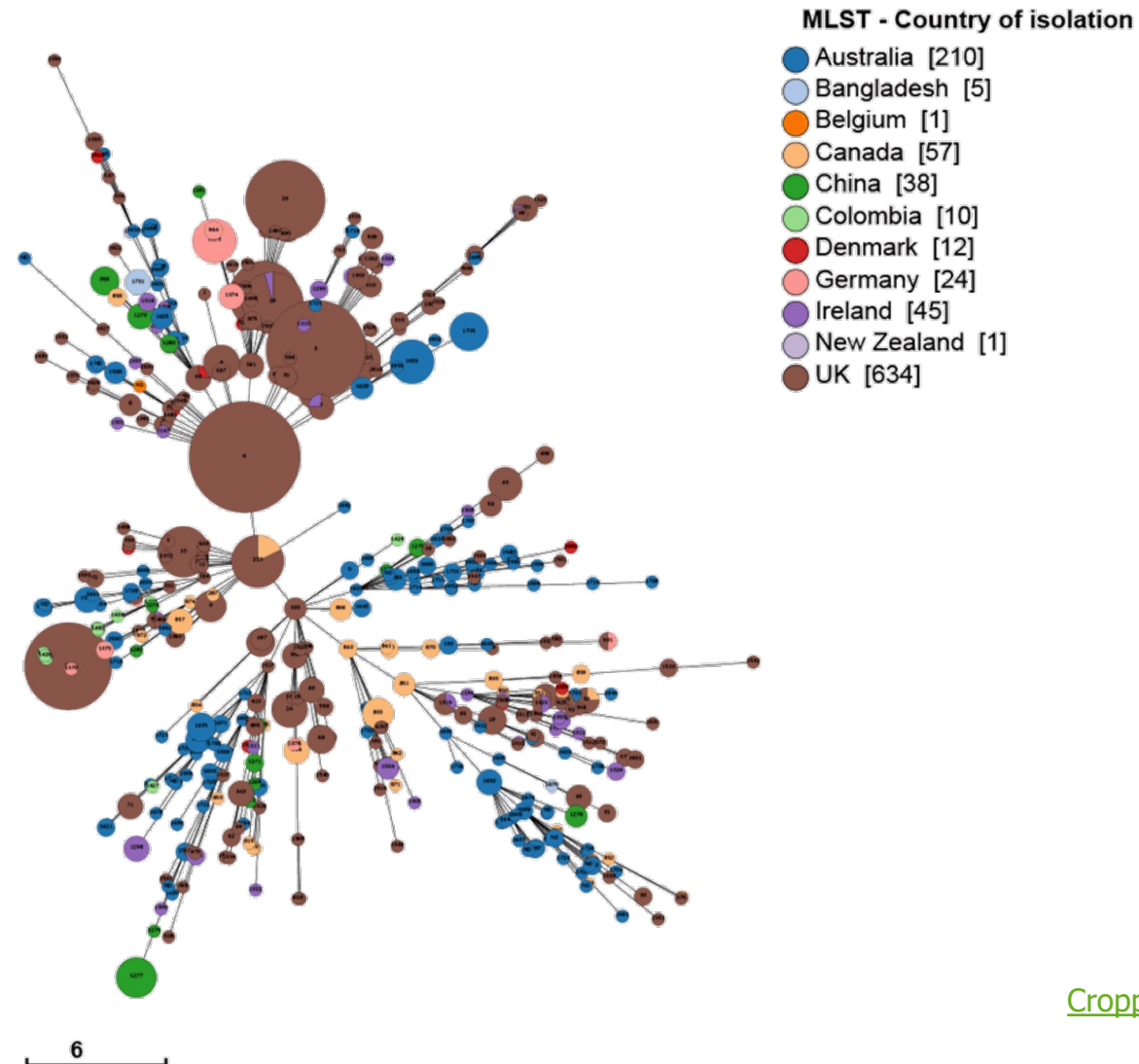
1. A hierarchical clustering dendrogram is NOT a phylogenetic tree



[A dendrogram plotted with R](#) by Jackverr / [CC BY-SA 3.0](#)

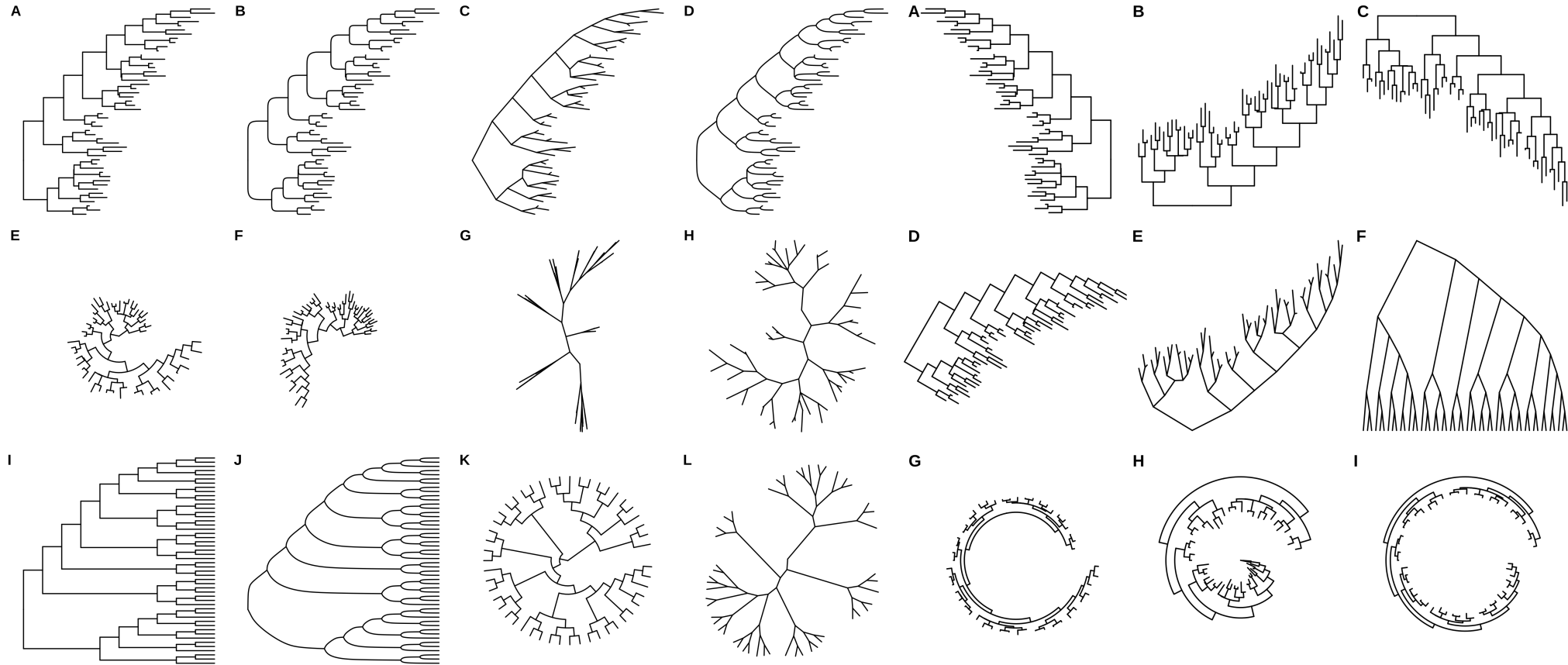


1. A Minimum Spanning Tree (MST) is NOT a phylogenetic tree



Cropped from Figure 3 by Whiley D *et al.* / CC BY 4.0

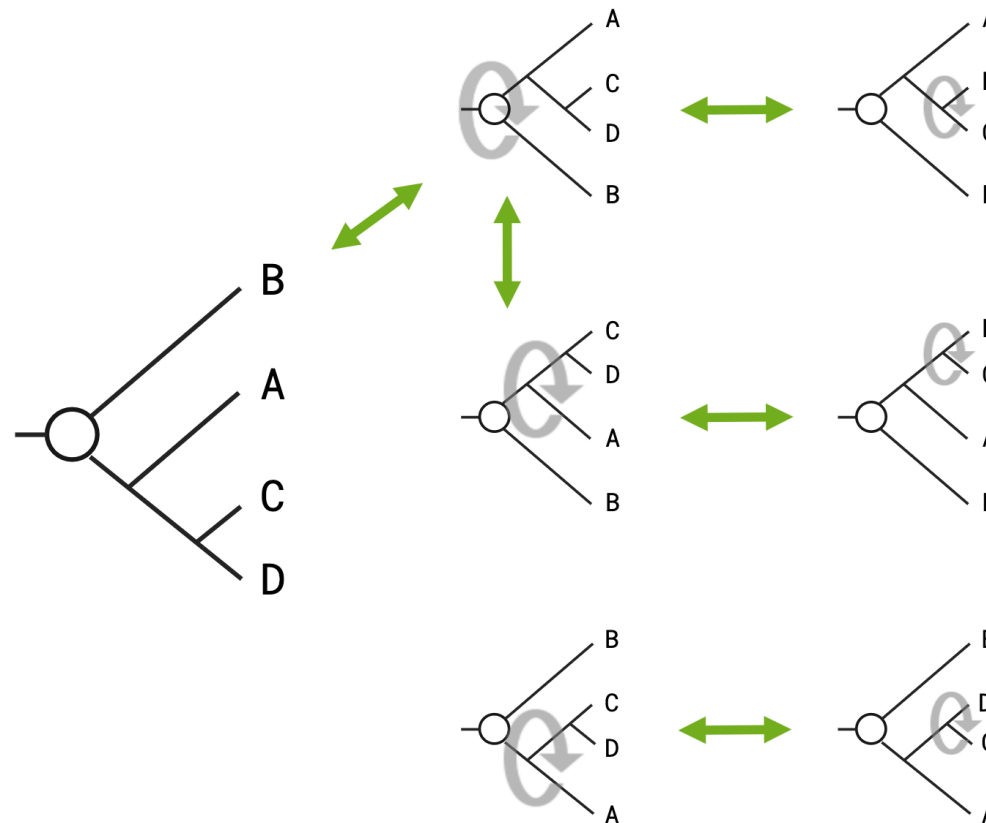
1. Phylogenetic trees come in all shapes and sizes



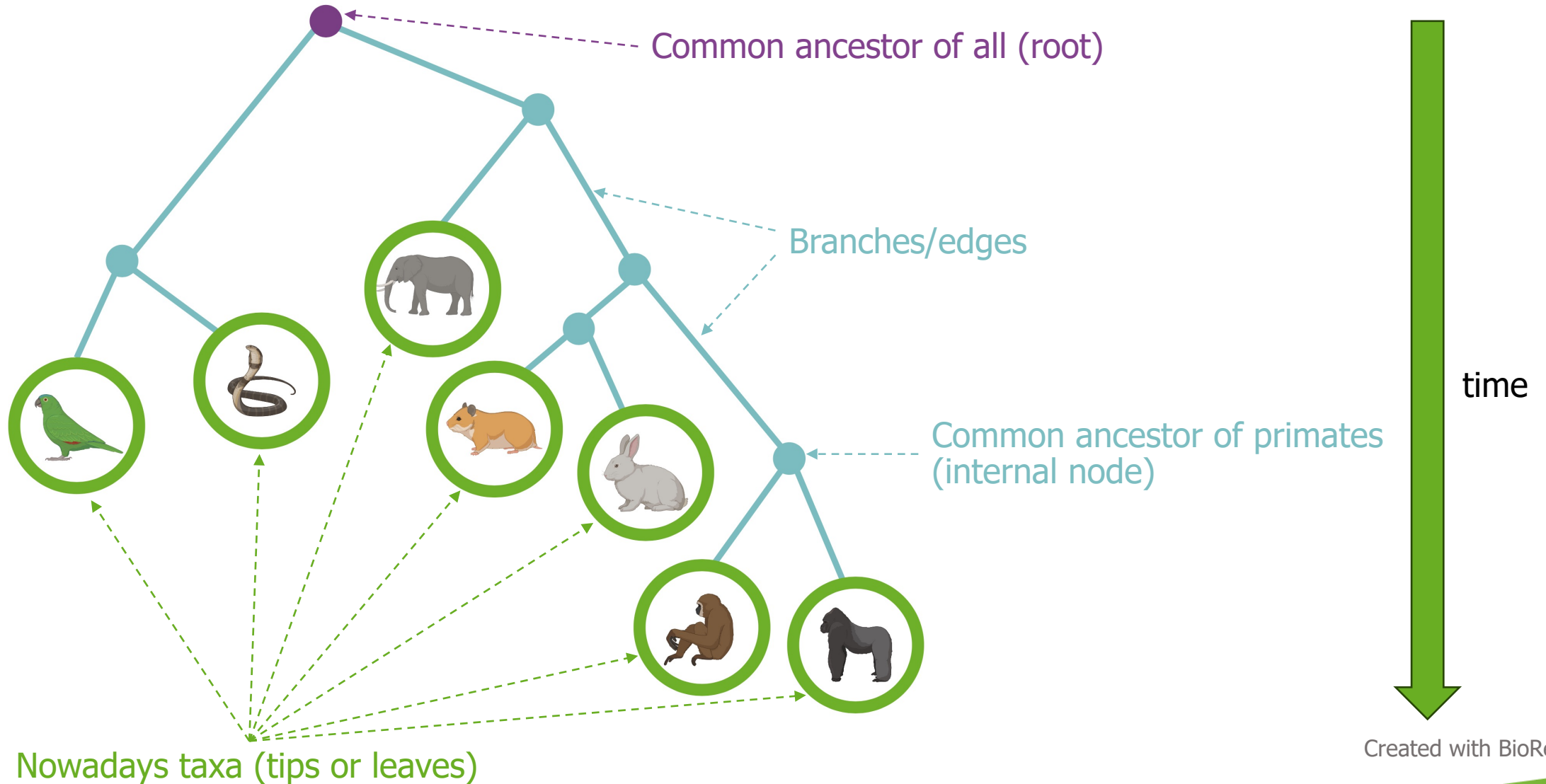
Figures 4.2 & 4.3 by Guangchuang Yu / [CC BY-NC-SA 4.0](#)

1. Tree rearrangements

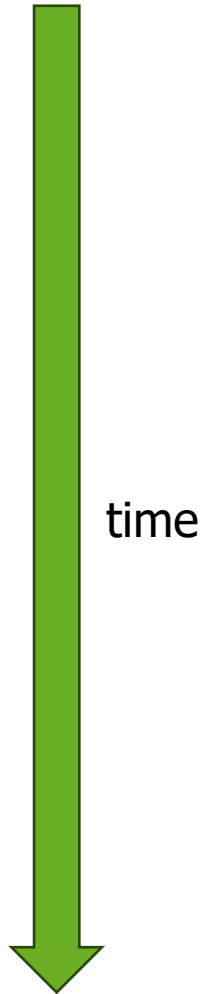
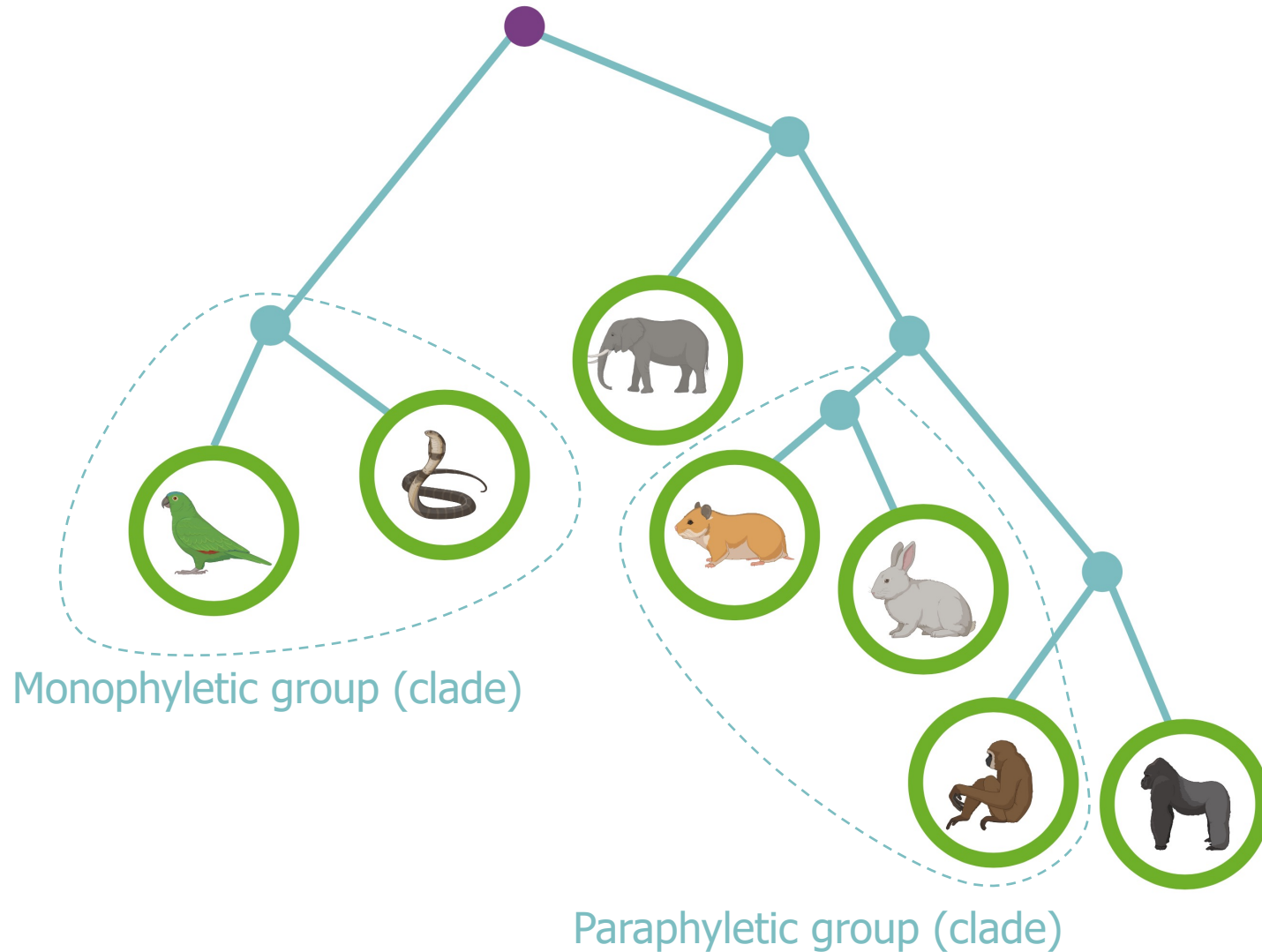
All of these rearrangements show the same evolutionary relationships between the taxa



1. Vocabulary

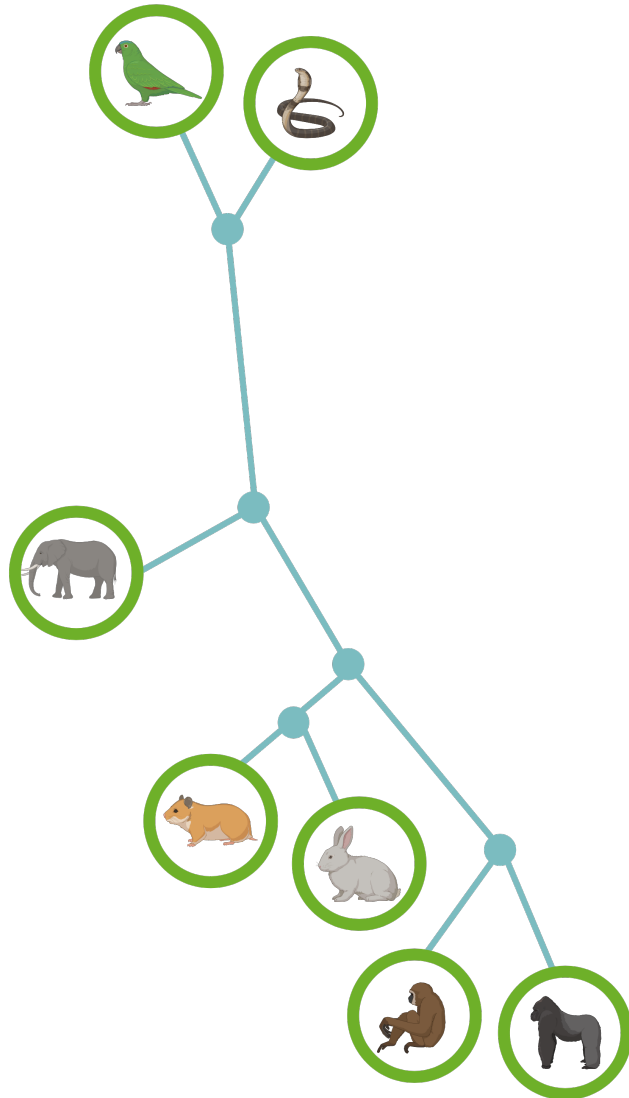


1. Vocabulary



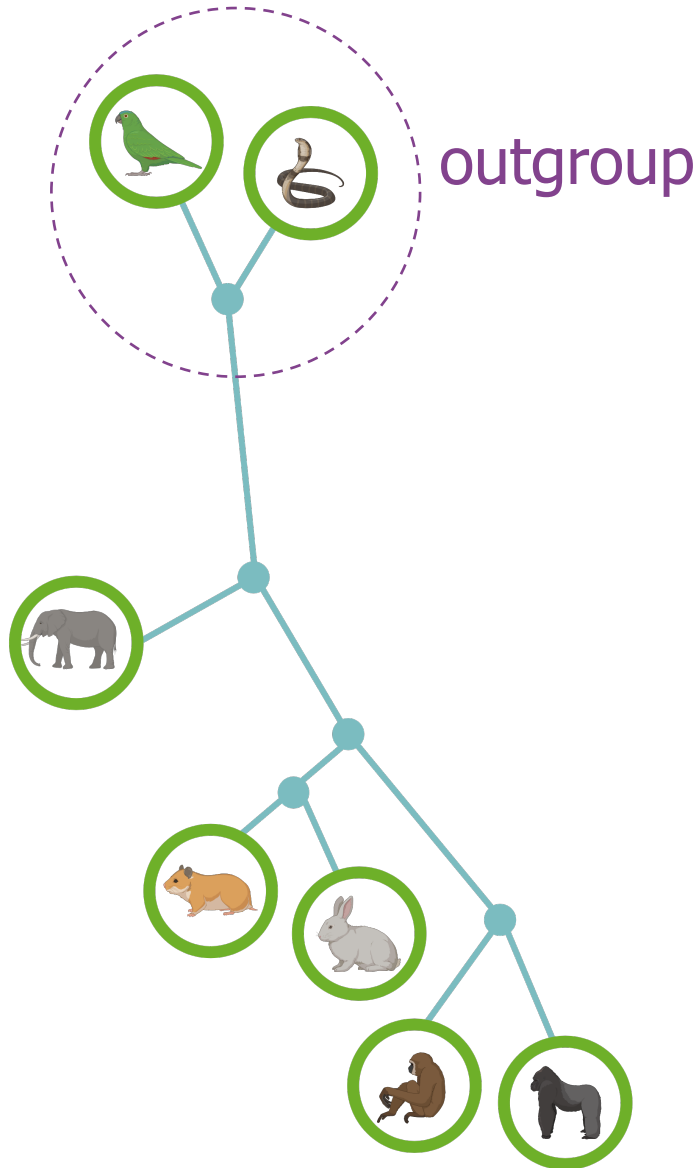
Created with BioRender.com

1. Rooting a tree



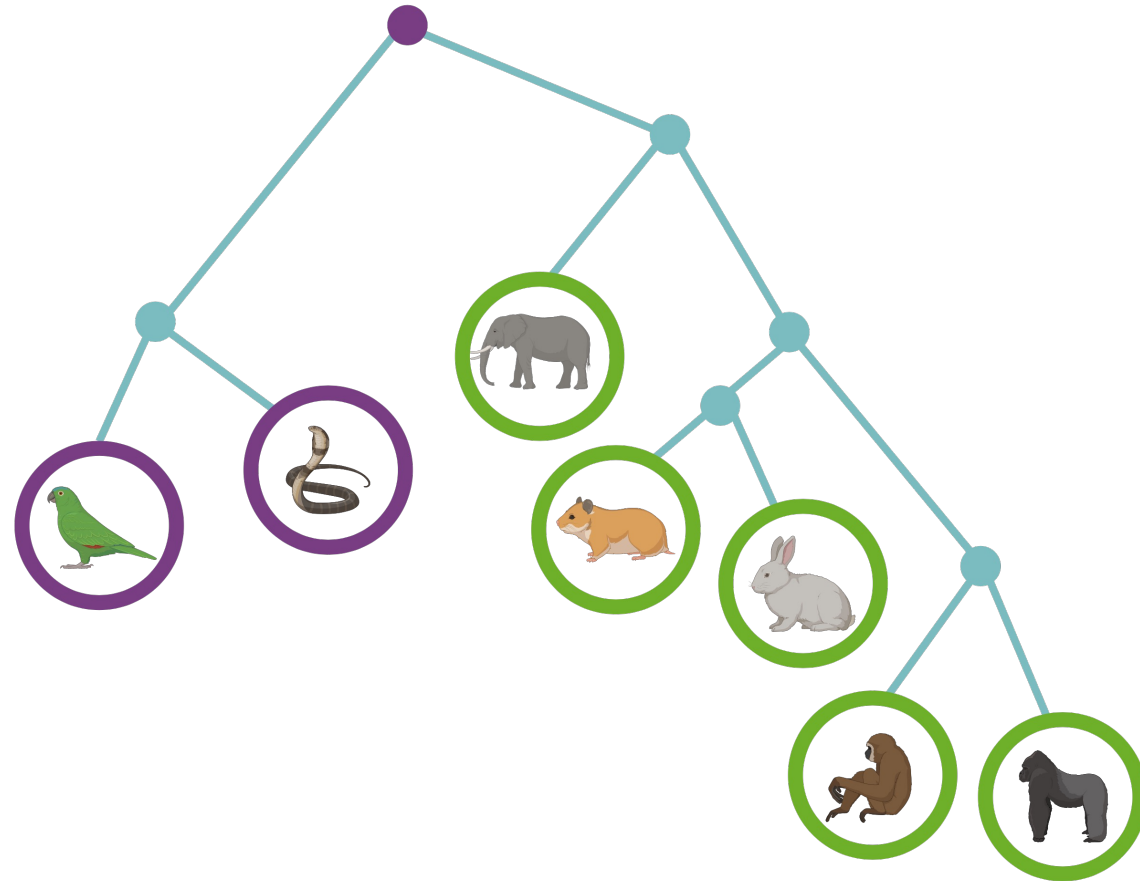
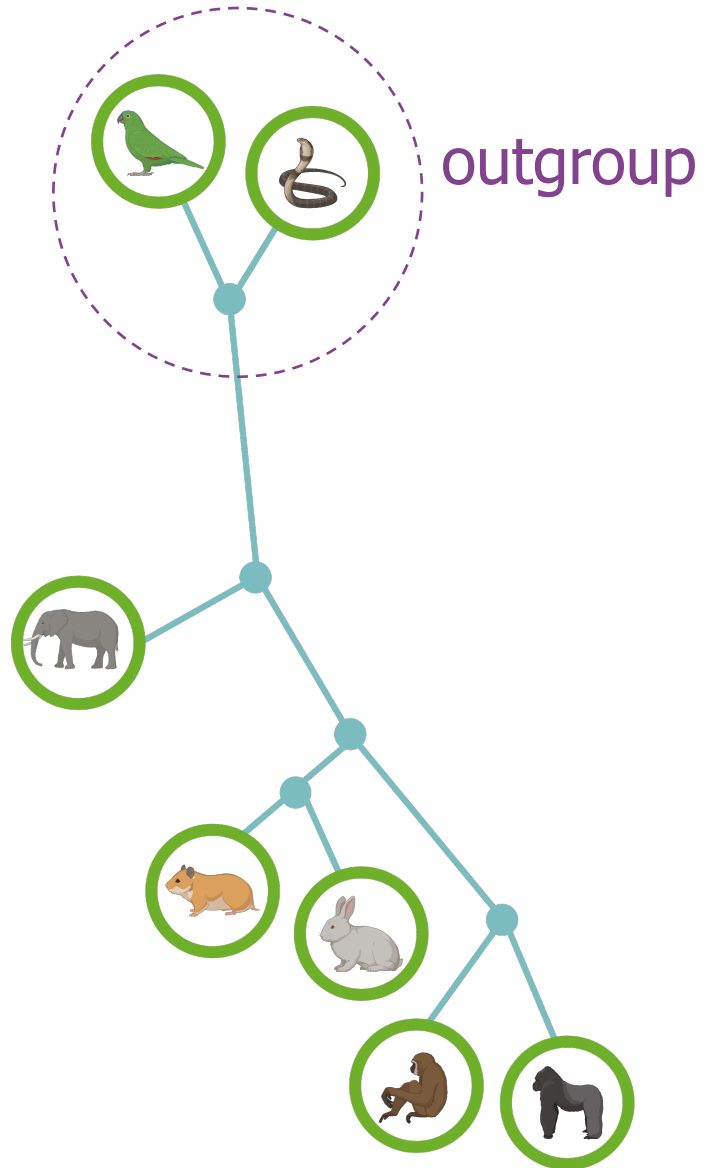
Additional data
needed to root

1. Rooting a tree

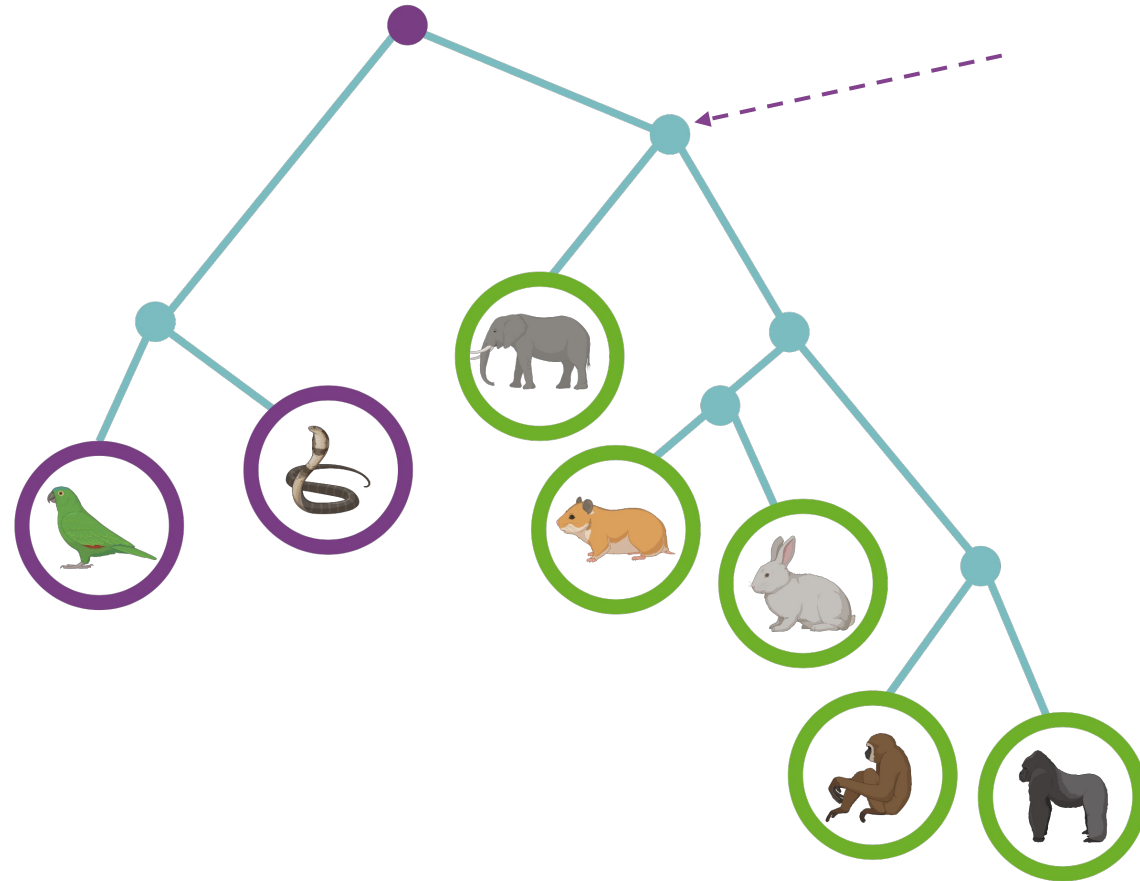
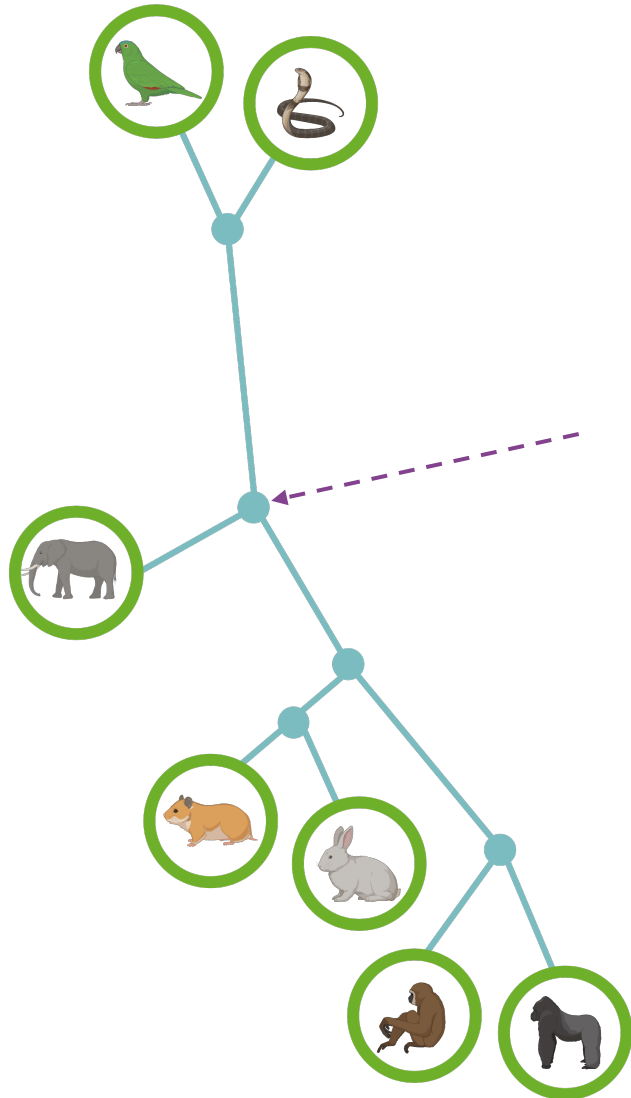


Additional data
needed to root

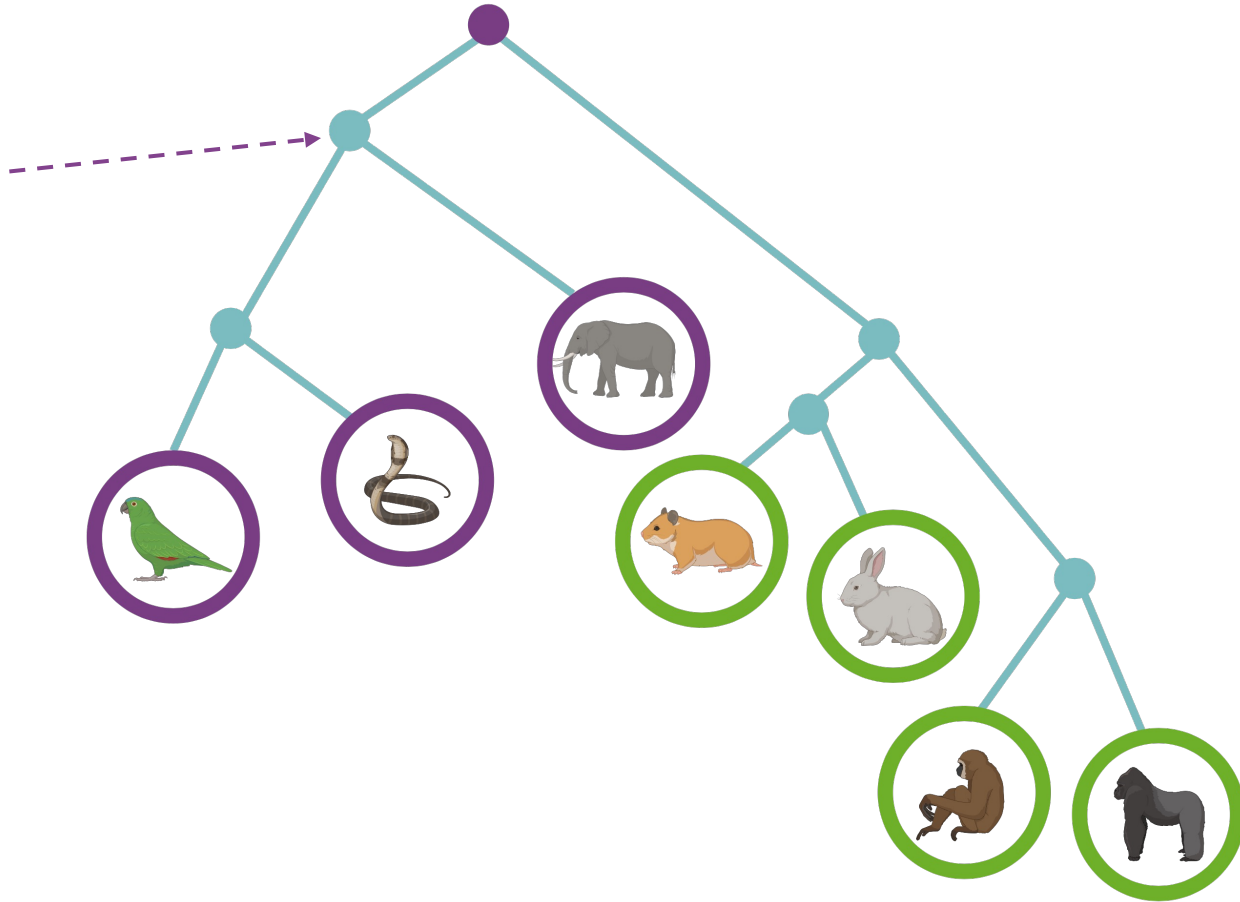
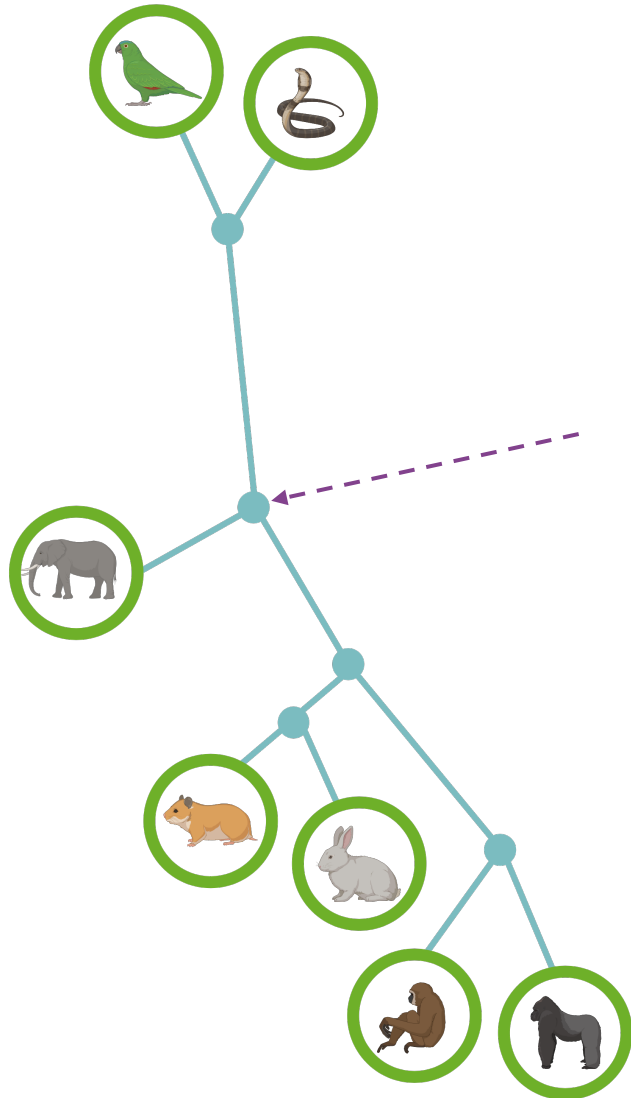
1. Rooting a tree



1. No root: no information about ancestry



1. No root: no information about ancestry

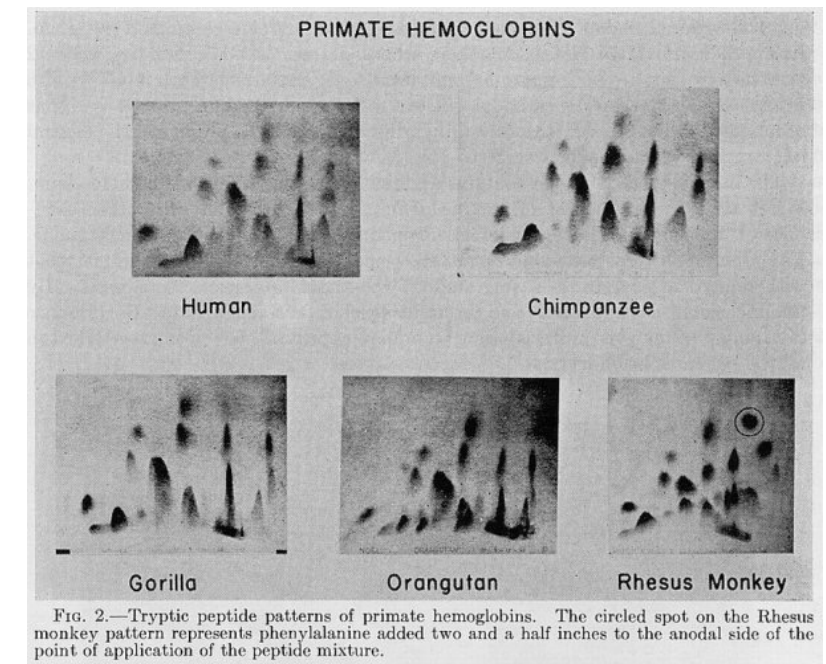


Part 2: The data

2. Molecular clock hypothesis

1960s: dissimilarity in protein fingerprints is approximately proportional to the distance between species.

Same principle applied today to molecular sequences (DNA, RNA, proteins).

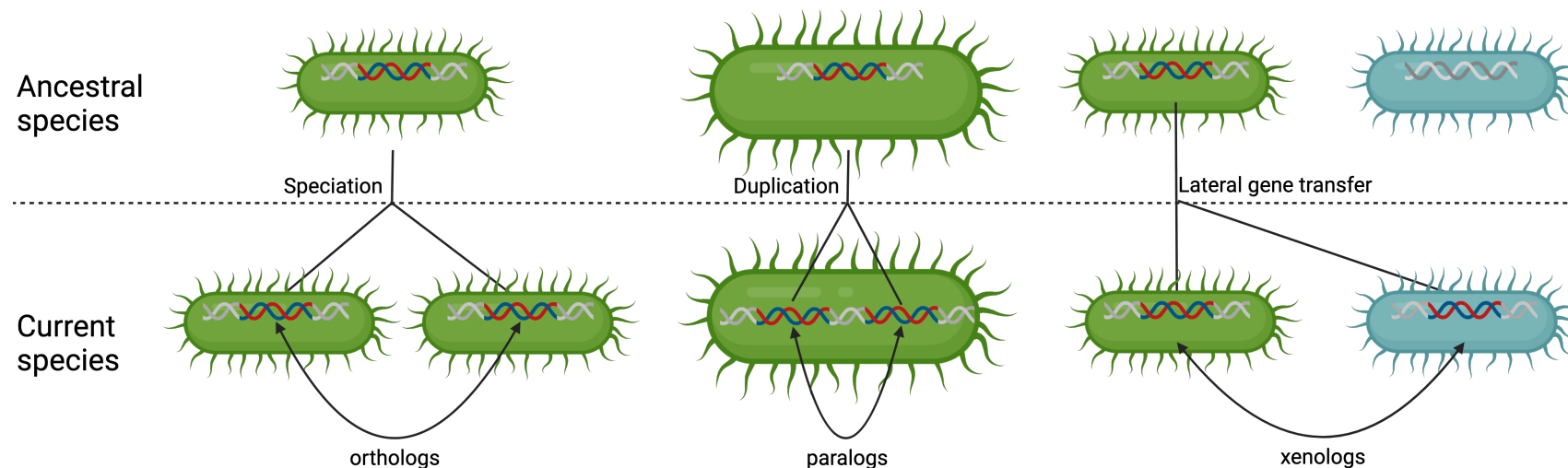


2. Sequence homology

Phylogenetic tree → evolution of a group of sequences since their **common ancestor**.

Sequences **need** to be homologs! Different types of homology:

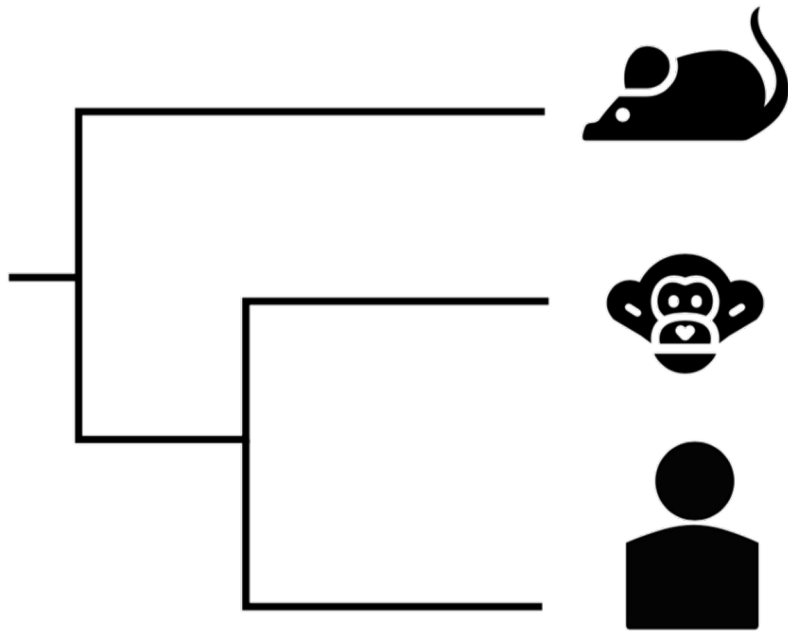
- Orthologs: originated from a speciation event.
- Paralogs: originated from a duplication event.
- Xenologs: originated from a lateral gene transfer.



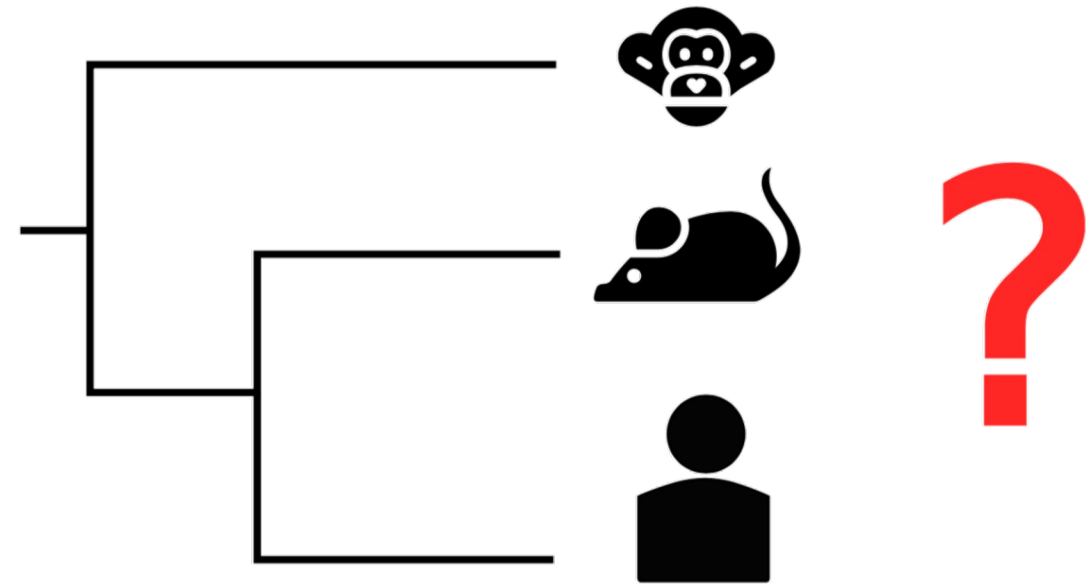
Created with BioRender.com

2. Species tree vs gene tree

Species tree

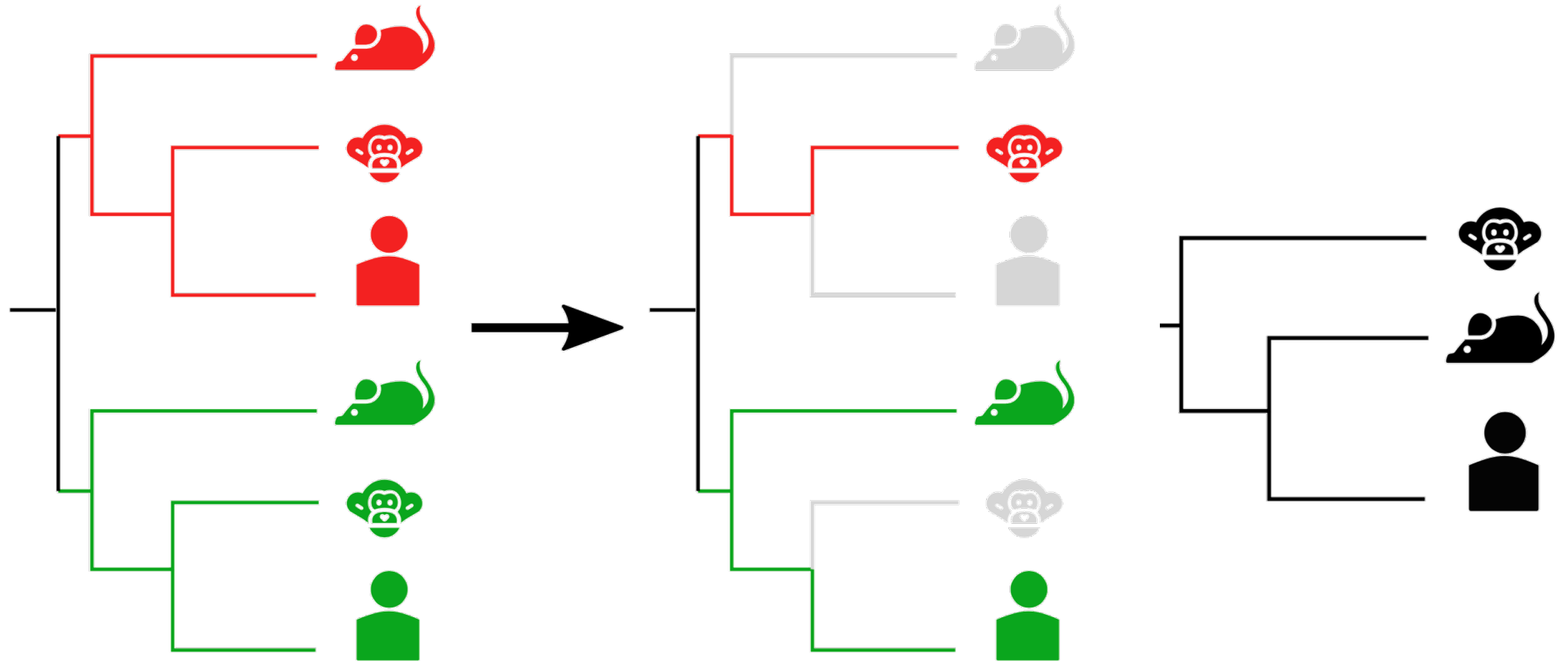


Gene tree

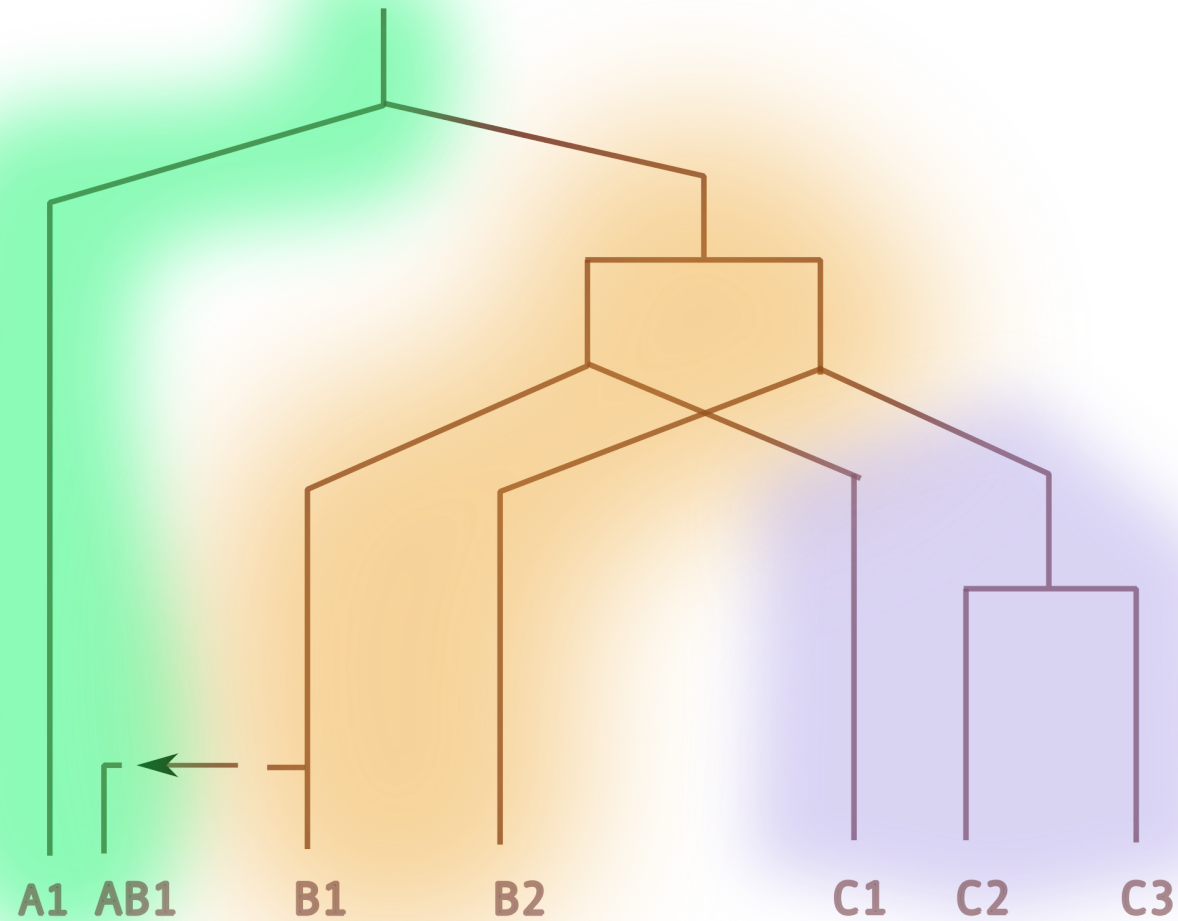


2. Orthology vs paralogy

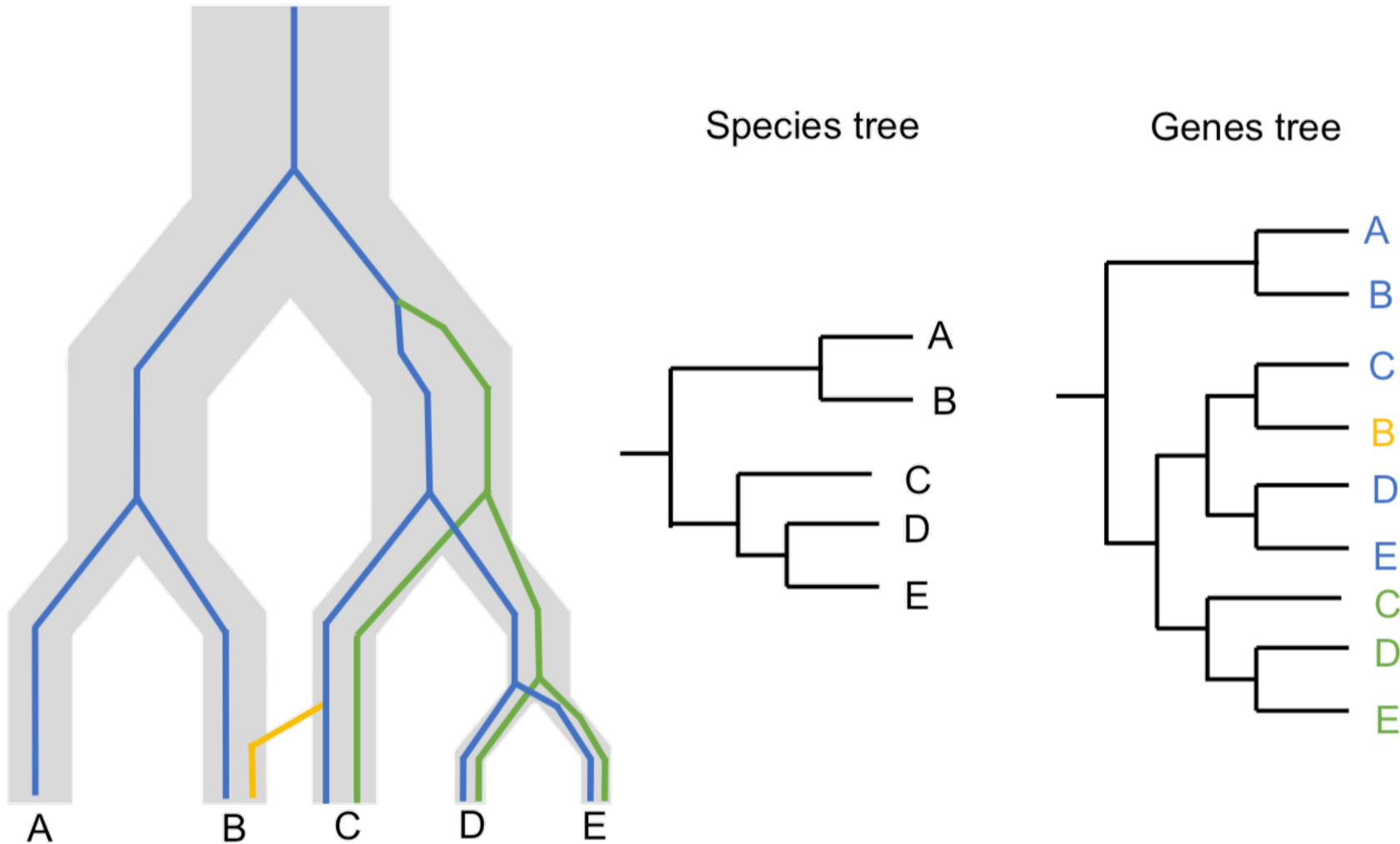
Gene history



2. Lateral gene transfer



2. Gene families can have complex histories



Part 3: The methods

3. Tree inference methods

		Data	
		Distances	Characters
Methods	Clustering	UPGMA Neighbor-joining (NJ)	
	Optimality criterion	Minimum Evolution (ME)	Maximum Parsimony (MP) Maximum Likelihood (ML) Bayesian

3. Inferring trees by clustering: UPGMA

Unweighted Pair Group Method with Arithmetic mean (**UPGMA**) (1950s).

Hierarchical clustering method.

Assumes strict molecular clock.

Gives an exact representation of a distance matrix, but exact tree/matrix correspondence never happens with real data.

UPGMA and more generally hierarchical clustering methods infer incorrect trees most of the time.

3. Inferring trees by clustering: NJ

Neighbor-joining (NJ) and its descendants (e.g., BioNJ, 1990s) are widely used now.

They infer **unrooted**, non-molecular-clock trees using an algorithm comparable to UPGMA.

Branch lengths are interpreted in number of substitutions per site (**not time**).

Evolutionary distances between sequences are estimated using probabilistic models accounting for hidden substitutions.

AA**T**GCTT

AA**G**GCTT

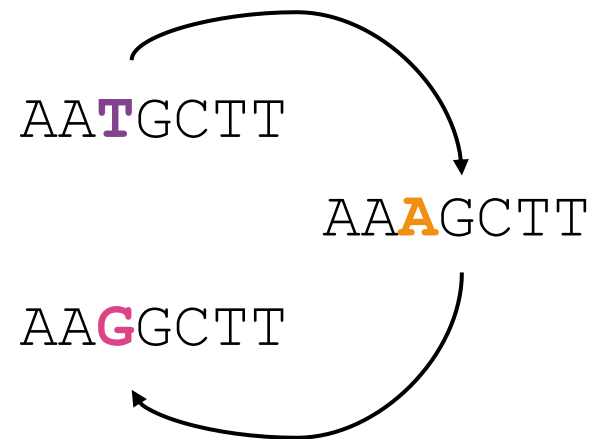
3. Inferring trees by clustering: NJ

Neighbor-joining (NJ) and its descendants (e.g., BioNJ, 1990s) are widely used now.

They infer **unrooted**, non-molecular-clock trees using an algorithm comparable to UPGMA.

Branch lengths are interpreted in number of substitutions per site (**not time**).

Evolutionary distances between sequences are estimated using probabilistic models accounting for hidden substitutions.



3. Inferring trees by clustering

NJ uses a **criterion** to evaluate at each step the leaf pair to agglomerate.

A criterion is an objective value allowing to compare a set of phylogenetic trees.

3. Optimality criterion: Minimum Evolution (ME)

For **distance-based** methods (using a distance matrix).

Consider a tree with branch lengths. The length of the tree is simply the sum of its branch lengths.

According to the ME criterion, the best tree is the shortest.

Fast and accurate distance-based tree reconstruction methods have been implemented using the ME criterion (FastME).

Also used in NJ.

		Data	
		Distances	Characters
Methods	Clustering	UPGMA Neighbor-joining (NJ)	
	Optimality criterion	Minimum Evolution (ME)	Maximum Parsimony (MP) Maximum Likelihood (ML) Bayesian

3. Summary: Distance-based methods

Distance-based methods are fast.

We need a way to estimate distances.

Correction method (multiple substitutions).

Pairwise distance estimation
is not reliable for large
divergence times.

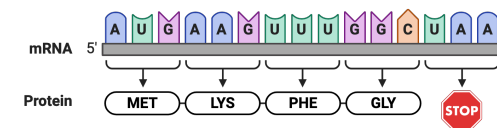
		Data	
		Distances	Characters
Methods	Clustering	UPGMA Neighbor-joining (NJ)	
	Optimality criterion	Minimum Evolution (ME)	Maximum Parsimony (MP) Maximum Likelihood (ML) Bayesian

3. Multiple sequences alignments

ATGTTTGACCCGTTCTAC
ATGTTGGCGTTCTAC
ATGTATAACCCGTTTAC

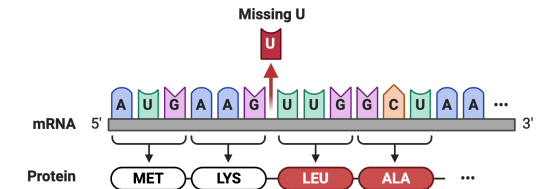
Wild type

mRNA sequence without any mutation



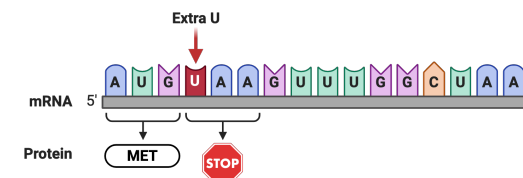
Base-pair deletion

Frameshift causing extensive missense



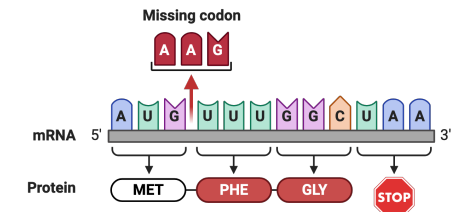
Base-pair insertion

Frameshift causing immediate nonsense



Three-nucleotide insertion/deletion

Extra/missing amino acids



3. Multiple sequences alignments

ATGTTTGACCCGTTCTAC
ATGTTGGCGTTCTAC
ATGTATAACCCGTTCTAC



ATGTTTGACCCGTTCTAC
ATGTTG---GCGTTCTAC
ATGTAT-AACCCGTTCTAC

3. Tree inference procedure

Input: multiple alignment (or distance matrix).

Output: tree.

Goal: find a tree that explains the input.

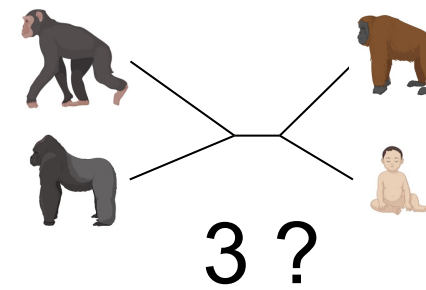
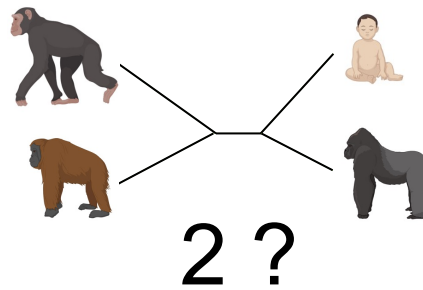
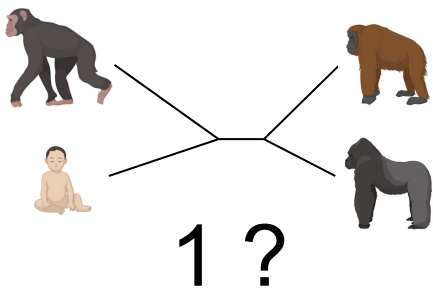
3. Optimality criteria: Maximum Parsimony

Select the tree that **minimizes the number of mutations (or steps)** needed to explain the data.

Character-based approach.

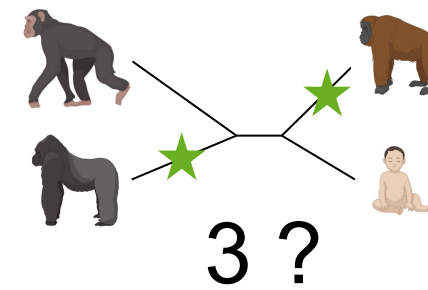
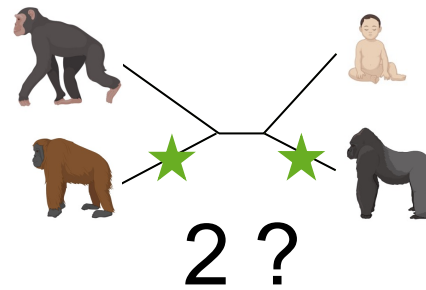
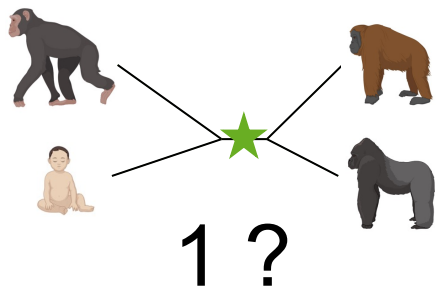
		Data	
		Distances	Characters
Methods	Clustering	UPGMA Neighbor-joining (NJ)	
	Optimality criterion	Minimum Evolution (ME)	Maximum Parsimony (MP) Maximum Likelihood (ML) Bayesian

3. Optimality criteria: Maximum Parsimony







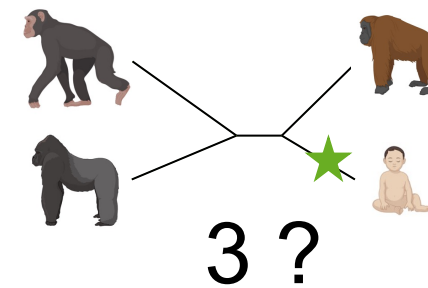
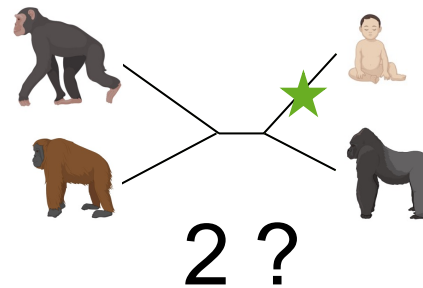
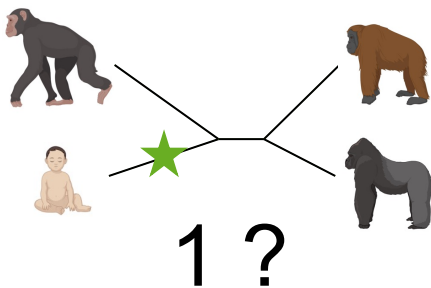
3. Optimality criteria: Maximum Parsimony

A
A
T
T
1







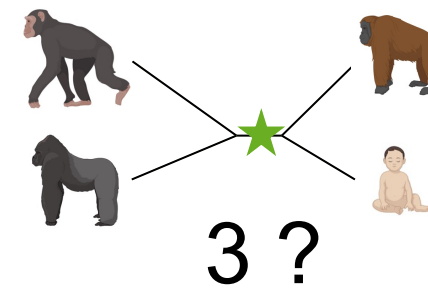
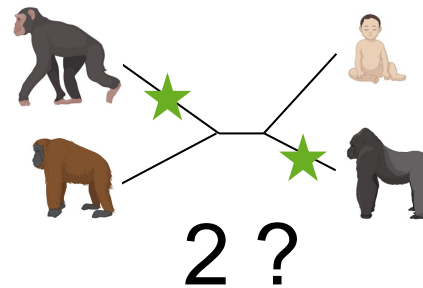
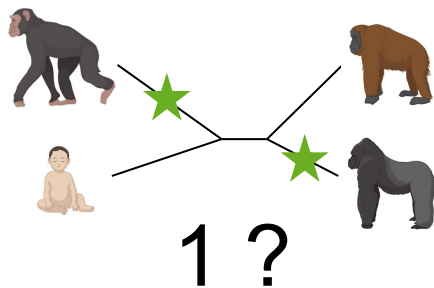
3. Optimality criteria: Maximum Parsimony

	A	A
	A	T
	T	T
	T	T
	1	







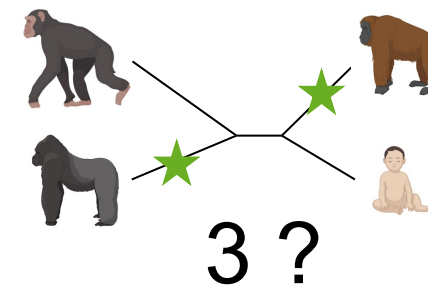
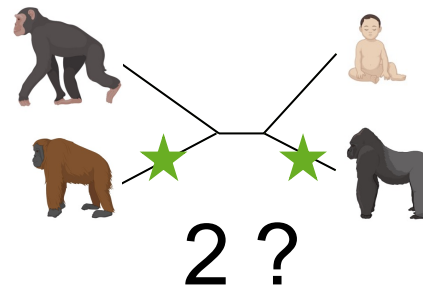
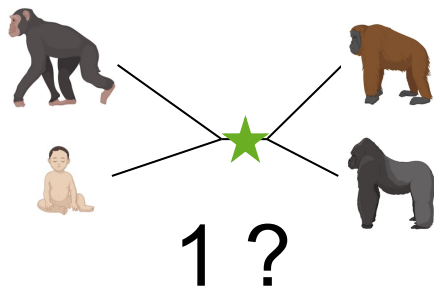
3. Optimality criteria: Maximum Parsimony

	A	A	G
	A	T	T
	T	T	T
	T	T	G
	1		3



3. Optimality criteria: Maximum Parsimony

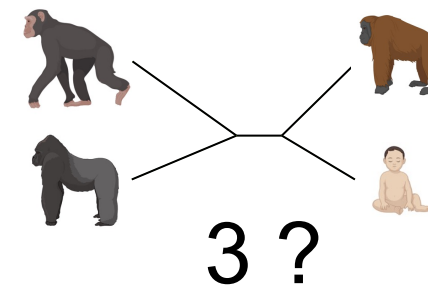
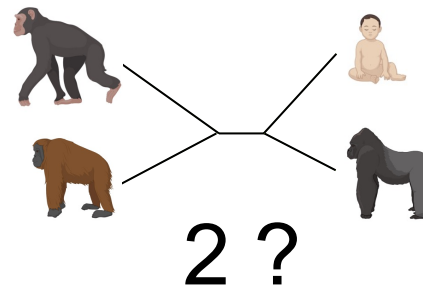
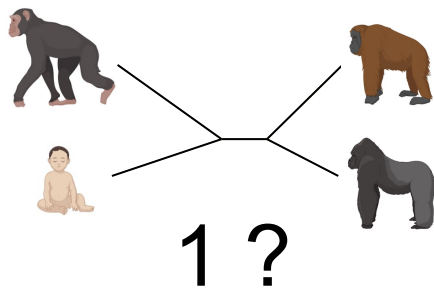
	A	A	G	T
	A	T	T	T
	T	T	T	C
	T	T	G	C
	1	3	1	



3. Optimality criteria: Maximum Parsimony

	A	A	G	T	T	G	...
	A	T	T	T	T	G	...
	T	T	T	C	C	A	...
	T	T	G	C	G	A	...
	1	3	1		1		

Select the majority (*i.e.* 1)



3. Optimality criteria: Maximum Parsimony

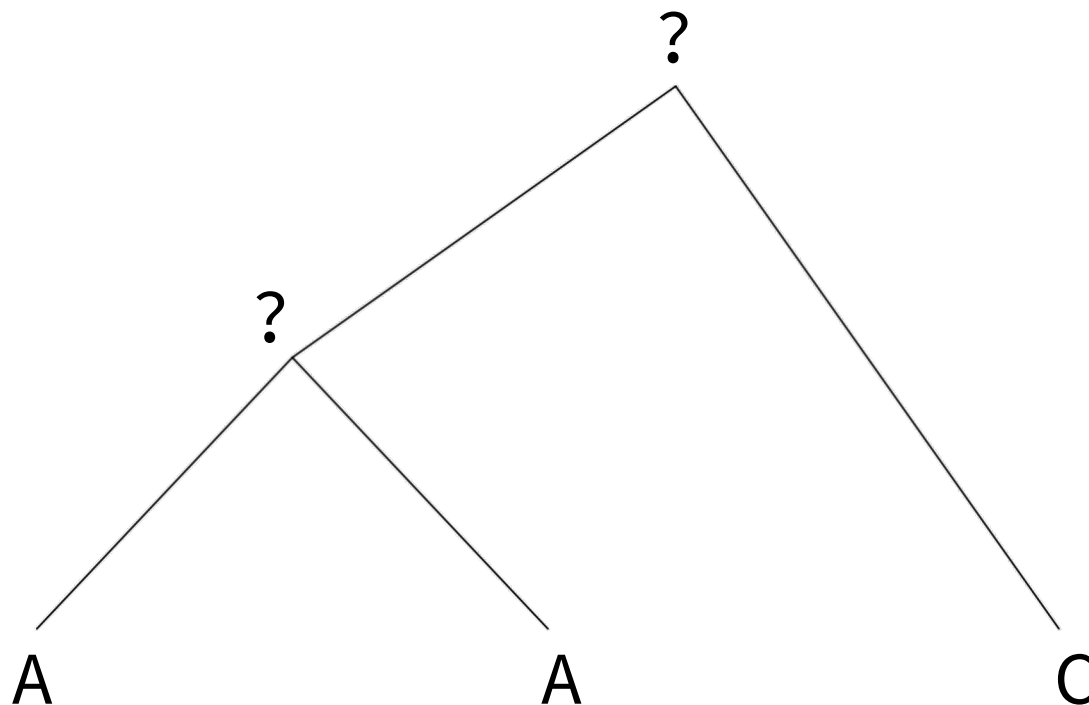
Select the tree that **minimizes the number of mutations (or steps)** needed to explain the data.

Character-based approach.

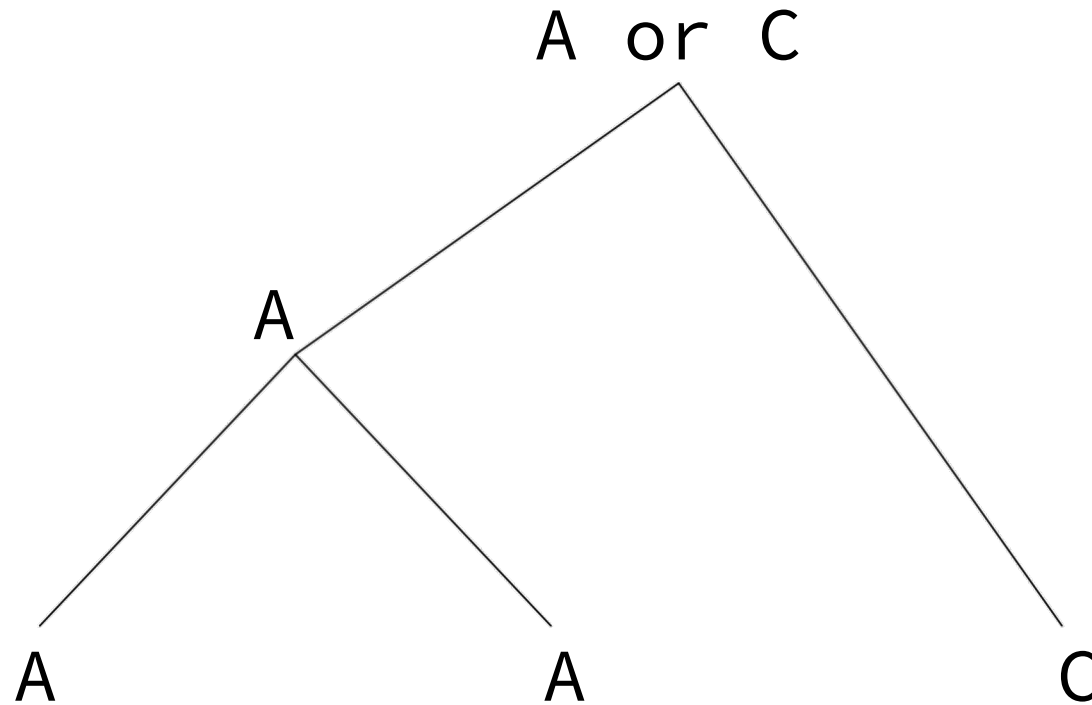
No root, no branch lengths.

		Data	
		Distances	Characters
Methods	Clustering	UPGMA Neighbor-joining (NJ)	
	Optimality criterion	Minimum Evolution (ME)	Maximum Parsimony (MP) Maximum Likelihood (ML) Bayesian

3. The probabilistic framework

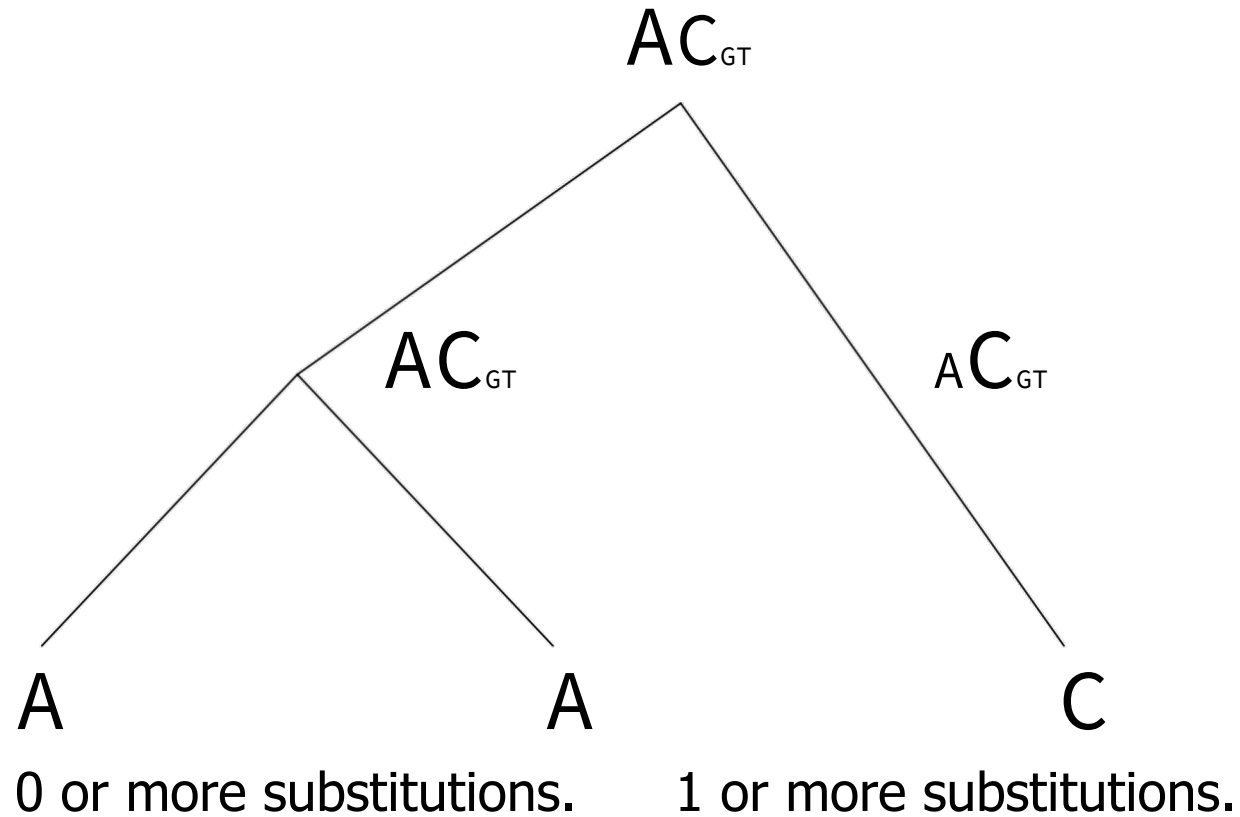


3. The probabilistic framework



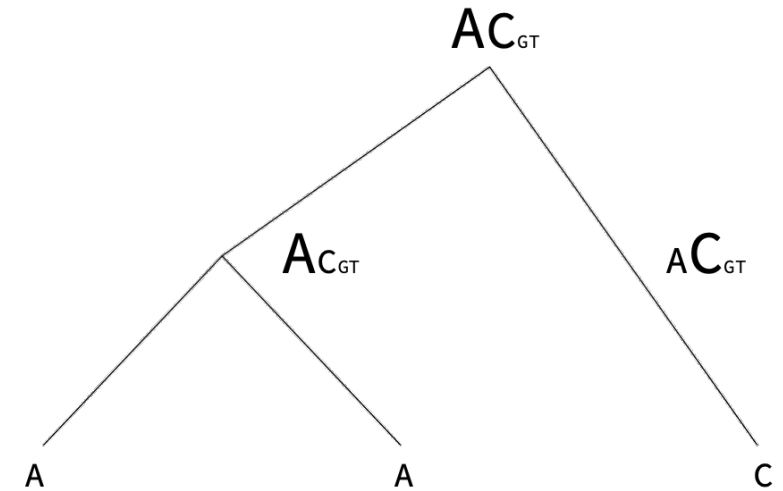
This is parsimony.

3. The probabilistic framework



3. The probabilistic framework

- Hidden substitutions may occur.
- Nucleotides have **probabilities**.
- We need a **model** that gives the probabilities of substitution between all possible different characters during a given amount of time.



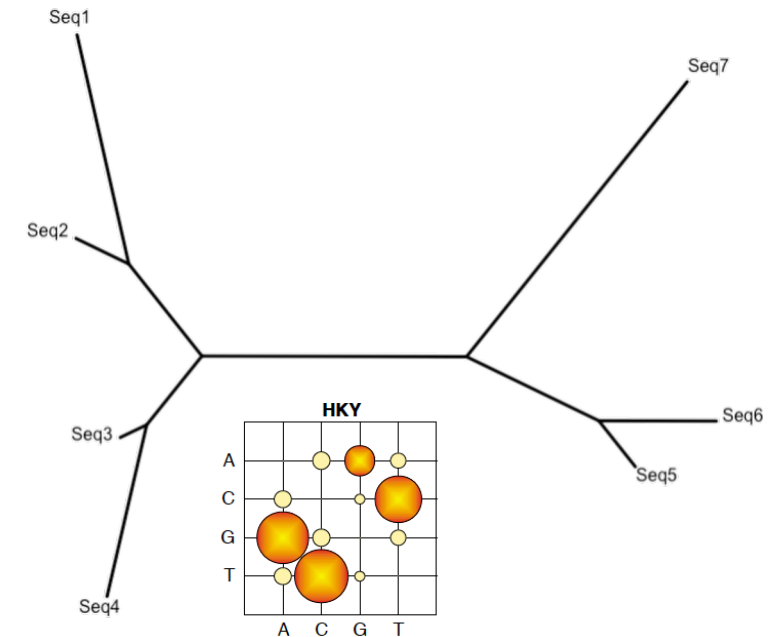
3. Optimality criterion: Maximum Likelihood

First proposed by Felsenstein (1981, ML) and Yang and Rannala (1996, Bayesian).

Character-based approaches, using multiple alignment.

Standard model:

- Tree with branch lengths
- Substitution model



3. Optimality criterion: Maximum Likelihood

In phylogeny:

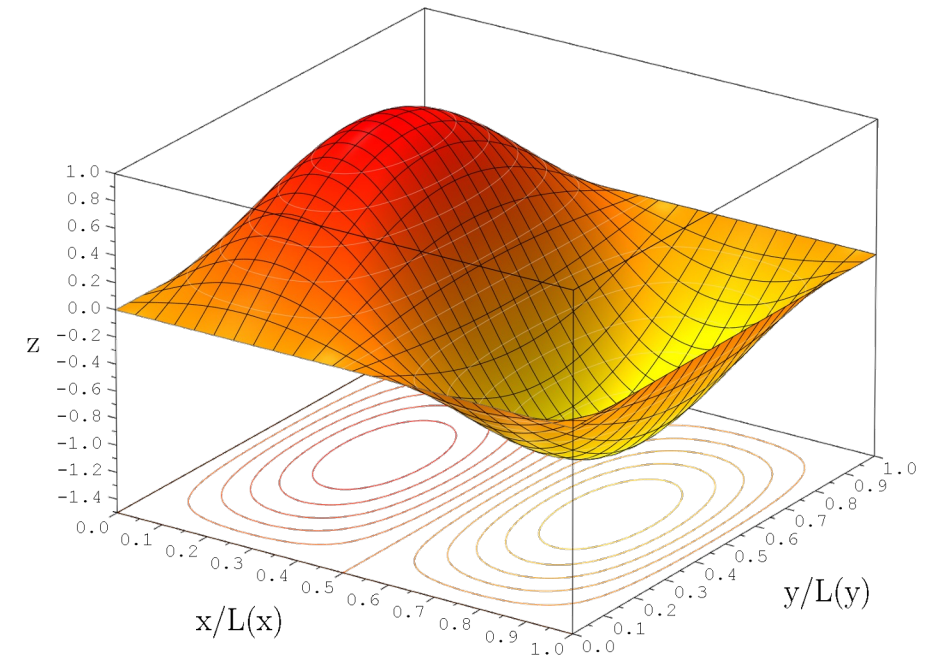
- The **data** is the sequence alignment.
- The set of parameters representing the **model** are the parameters of the evolutionary scenario (at the very least, a tree topology, branch lengths, and the parameters of the substitution model).
- The inferred **criterion** is the **likelihood** of the model (the **probability** of observing this model given our data).
- Maximum likelihood allows estimating the parameters of a model that describe the data the best.

3. Optimality criteria: Maximum Likelihood

ML and Bayesian methods aim at maximizing the probability of a model given the data (likelihood).

ML methods maximize the likelihood and provide a **unique tree**.

Bayesian methods incorporate prior knowledge, maximize the “posteriors” using dedicated algorithms, and provide a **collection of alternative trees** (forest).



3. Models of evolution: DNA

GTR + G + I

GTR model

Gamma distribution (evolutionary rates of the sites may vary)

Proportion of invariant sites (some sites do not vary at all)

3. Models of evolution: protein

LG + G

LG model

Gamma distribution

No Proportion of invariant sites

3. A note on recombination

Vertical or clonal evolution occurs via different mechanisms. Point mutations are the simplest ones.

Mobile genetic elements could influence genome-wide similarity measures but are not shared by all members of a species and thus easily ignored.

Yet homologous recombination events occur, commonly in naturally transformable species but sometimes outside of these species. This leads to genome regions with dramatic levels of sequence divergence which does not reflect the real evolutionary signal, leading to incorrect trees inference.

3. A note on recombination

Recombination can affect only branch lengths, but also tree topologies for extreme cases.

Note that recombination is not likely to affect statistical support of the branches.

Detection of regions affected by recombination might prove useful (*e.g.*, Gubbins).

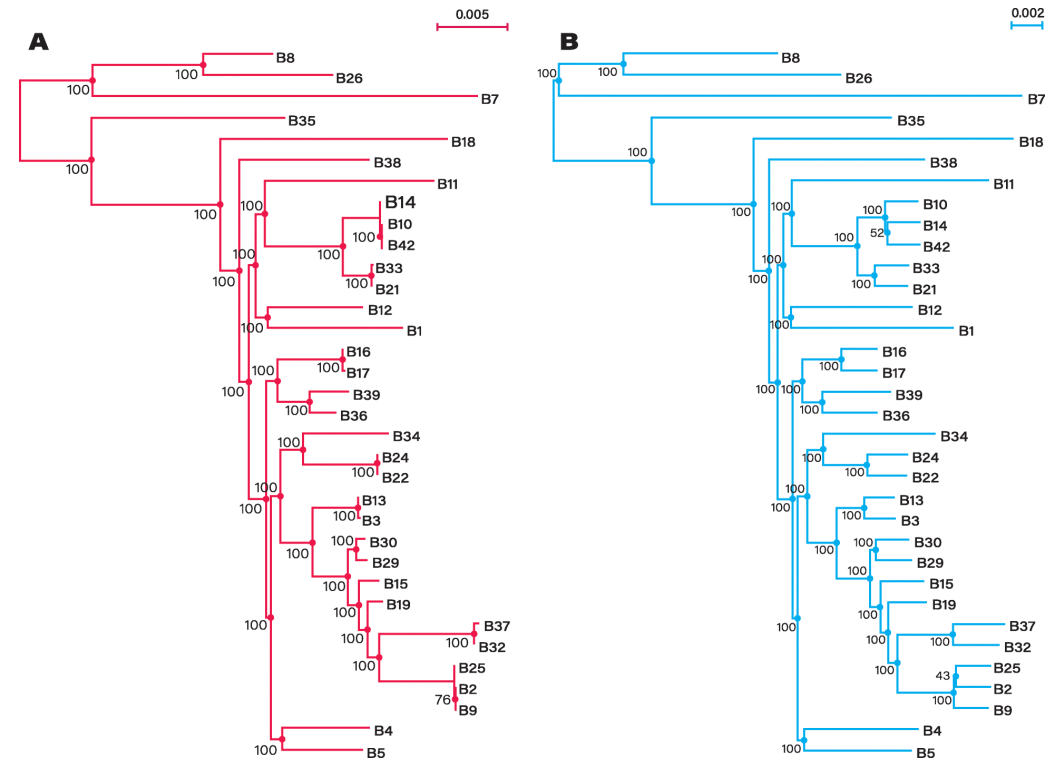


Figure 1 by Stott & Bobay / [CC BY 4.0](#)

3. Summary

Recommended methods:

Distance-based (NJ, BioNJ, ME) especially when the level of divergence is low (<10%)

Maximum Likelihood (software: IQ-TREE, RAxML-NG)

Bayesian (software: PhyloBayes, MrBayes, BEAST)

Part 4: Exploring the space of solutions

4. Search the tree space

We cannot estimate a criterion for all possible trees.

Heuristic search: does not guarantee to find the optimal tree.

Taxa	Number of rooted trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
15	213,458,046,676,875
20	8,200,794,532,637,891,559,375
30	$4,9518 \cdot 10^{38}$
40	$1.00985 \cdot 10^{57}$
50	$2.75292 \cdot 10^{76}$

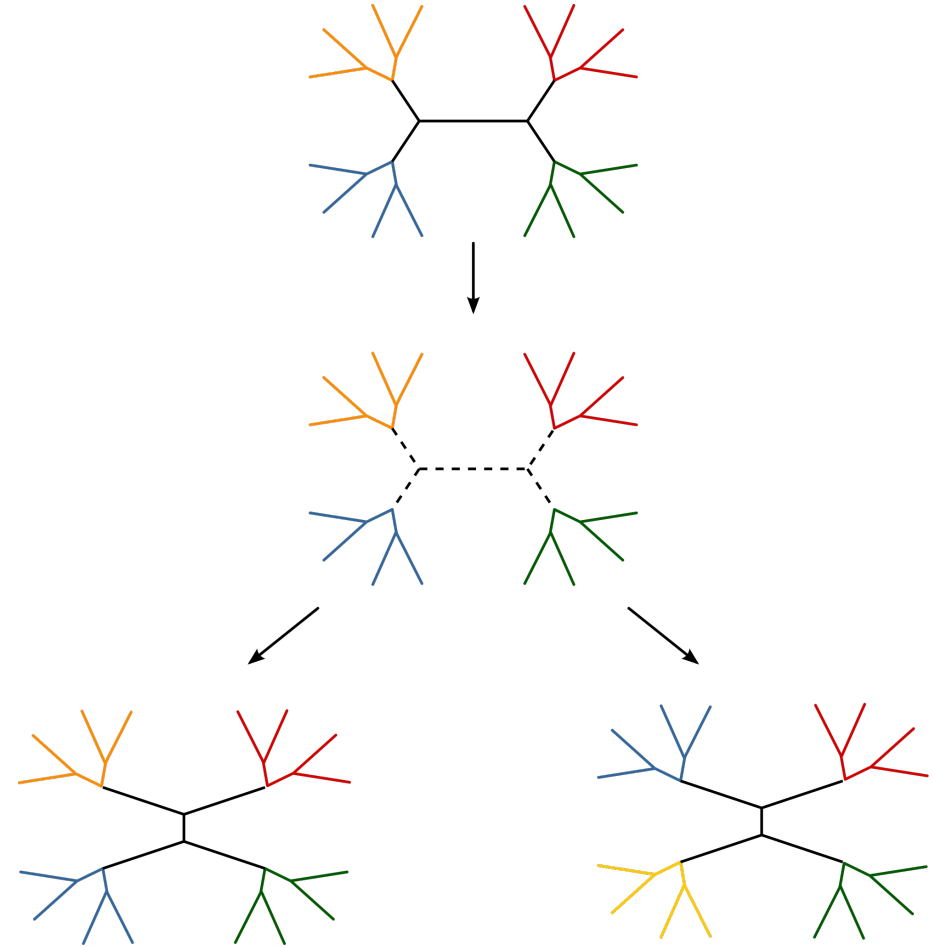
4. Tree rearrangement: Nearest-Neighbor Interchange (NNI)

If we chose a single **internal** branch of a tree, 2 new trees can be obtained by swapping two subtrees.

→ fast because the total number of new trees that can be obtained is small

Number of NNI for a tree with n taxa:

$$2(n - 3) = 2n - 6$$



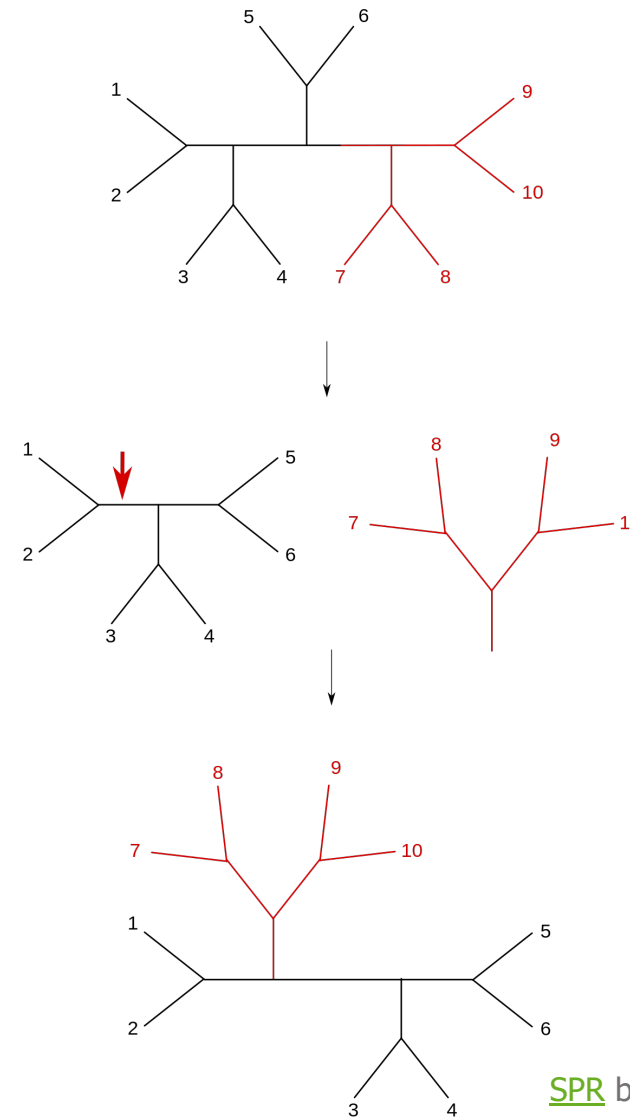
4. Tree rearrangement : Subtree Prune and Regraft (SPR)

Using SPR we can obtain many new trees.

→ better exploration, but costly.

$2(n - 3)(2n - 7)$ new topologies.

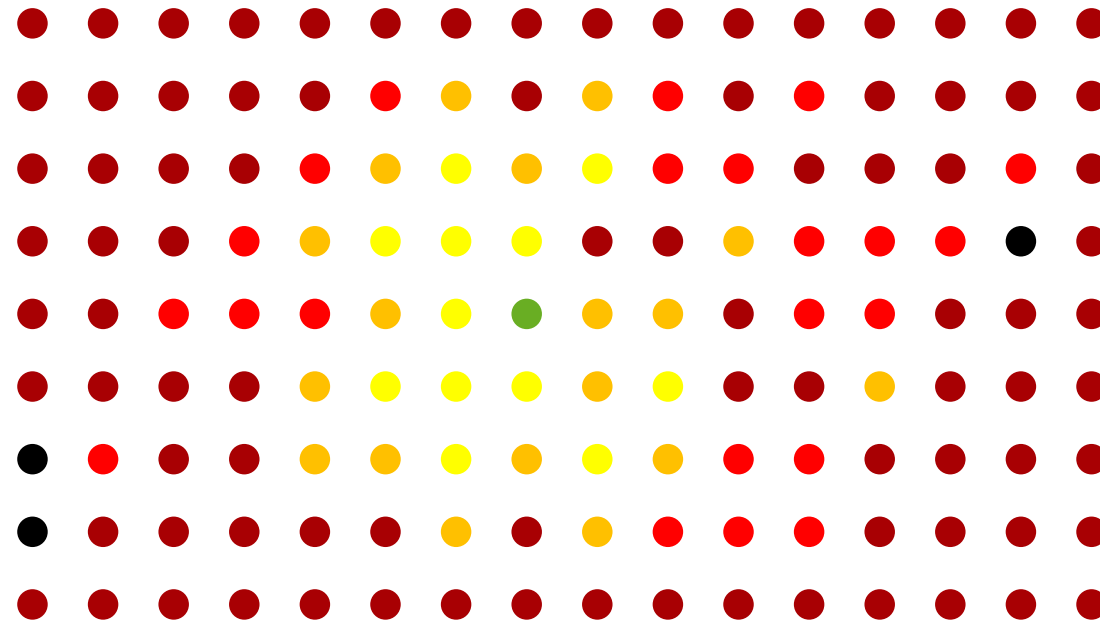
Of note, among all possible SPR moves some are NNIs.



SPR by François M / Public domain

4. Exploring the tree space

one dot = one tree
two neighbor dots are separated by exactly
one tree rearrangement

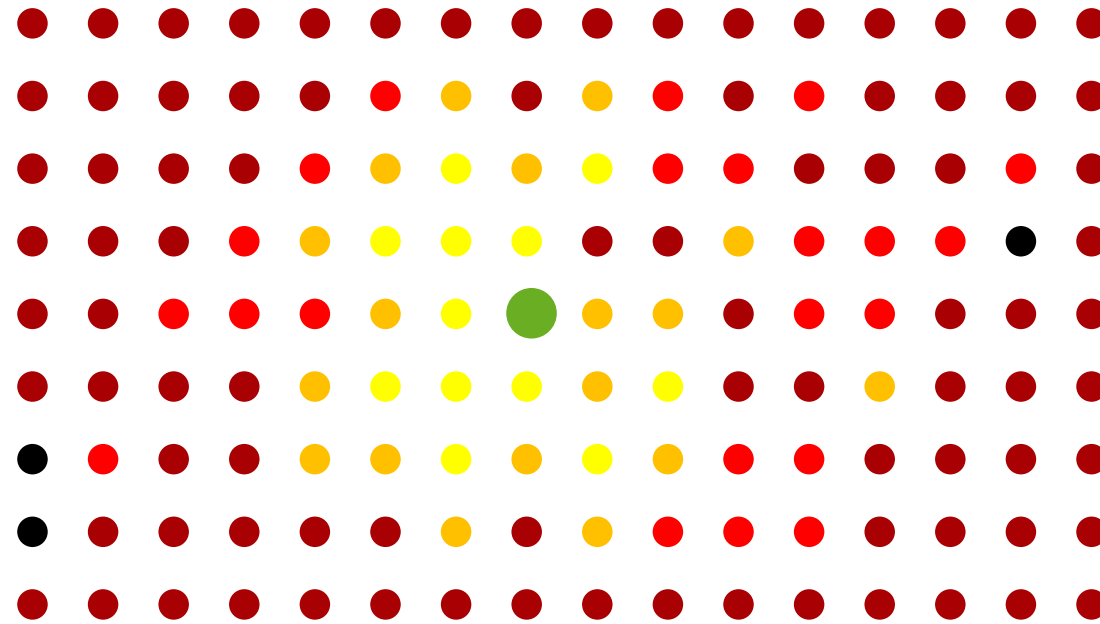


criterion

worst ● ● ● ● ● ● *best*

4. Exploring the tree space

If we obtain this NJ tree, then tree rearrangements cannot lead to an improvement...

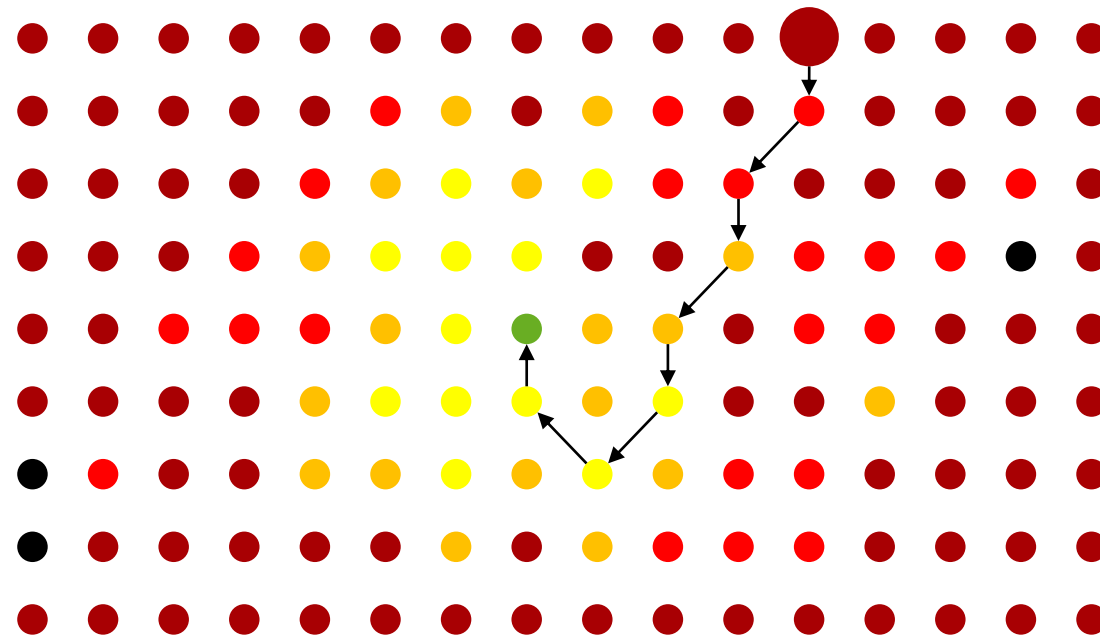


criterion

worst ● ● ● ● ● *best*

4. Exploring the tree space: hill climbing

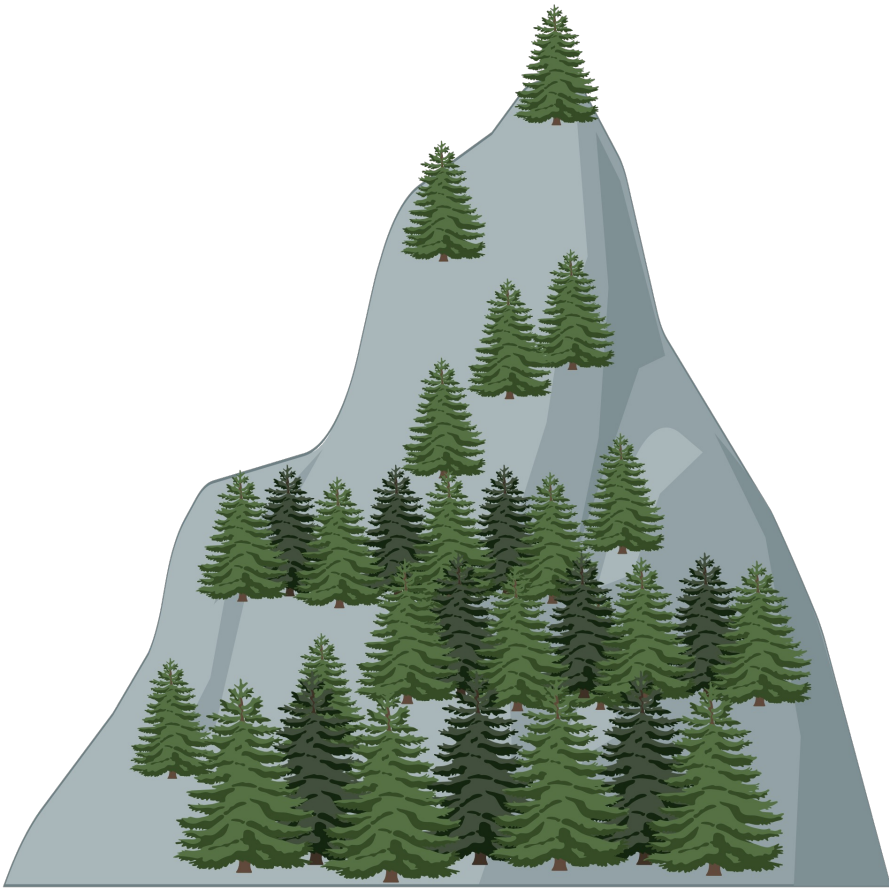
... but generally, we do not obtain the optimal tree and so rearrangements are useful



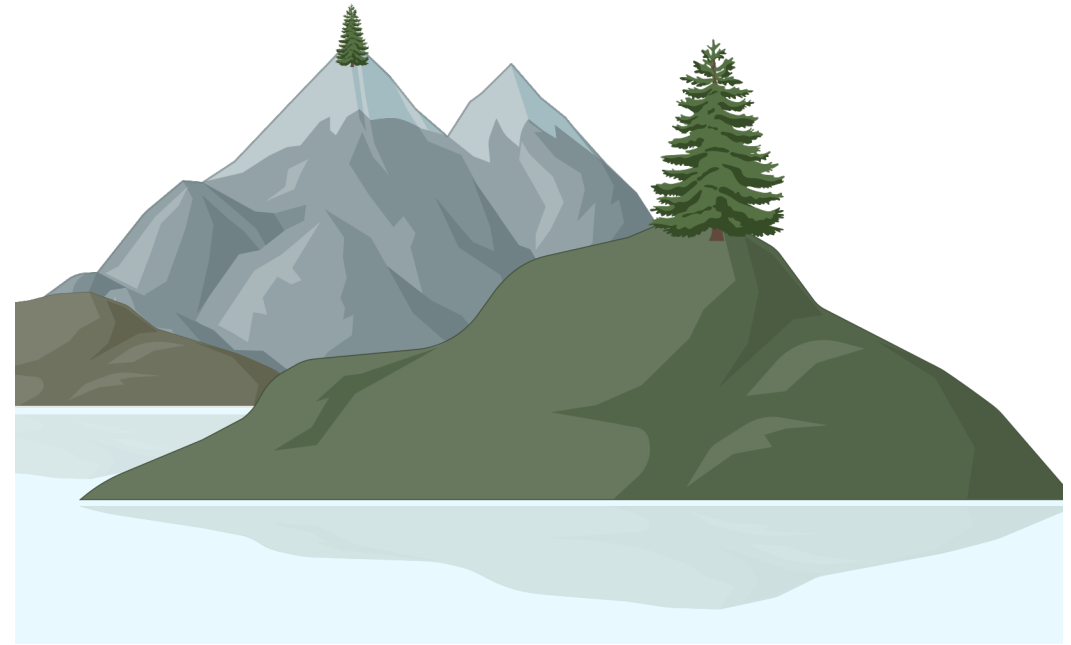
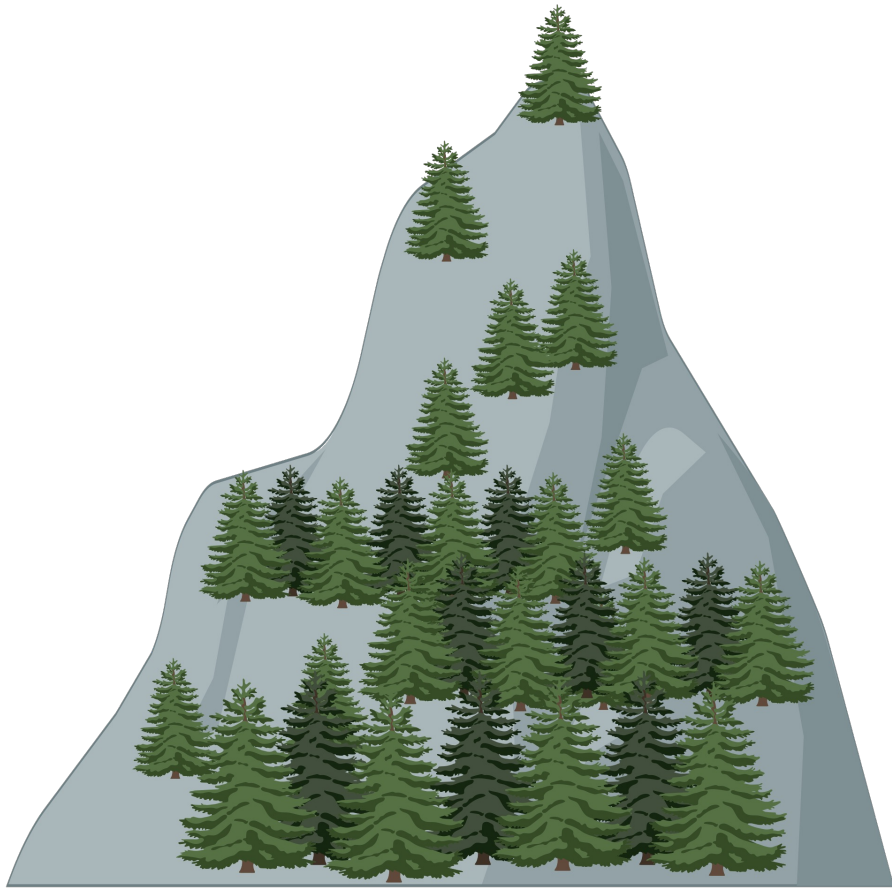
criterion

worst ● ● ● ● ● *best*

4. Exploring the tree space: hill climbing



4. Exploring the tree space: hill climbing

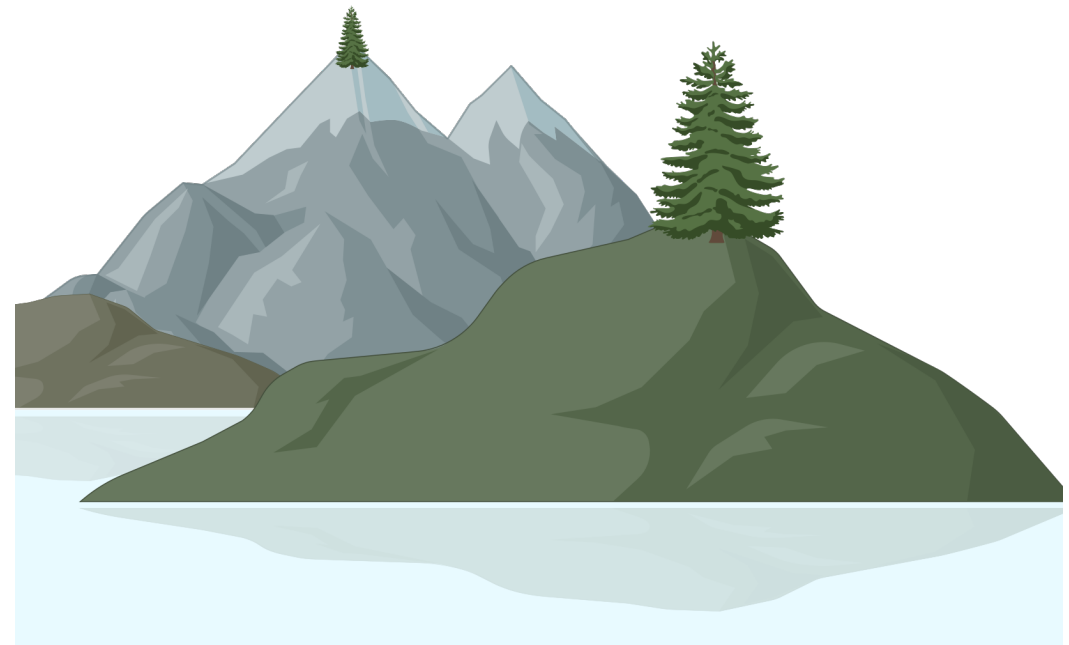


4. Exploring the tree space: hill climbing

Using only NNI might result in getting stuck in a **local optimum**.

SPR might avoid these local optima.

Another strategy is to start from a different tree.



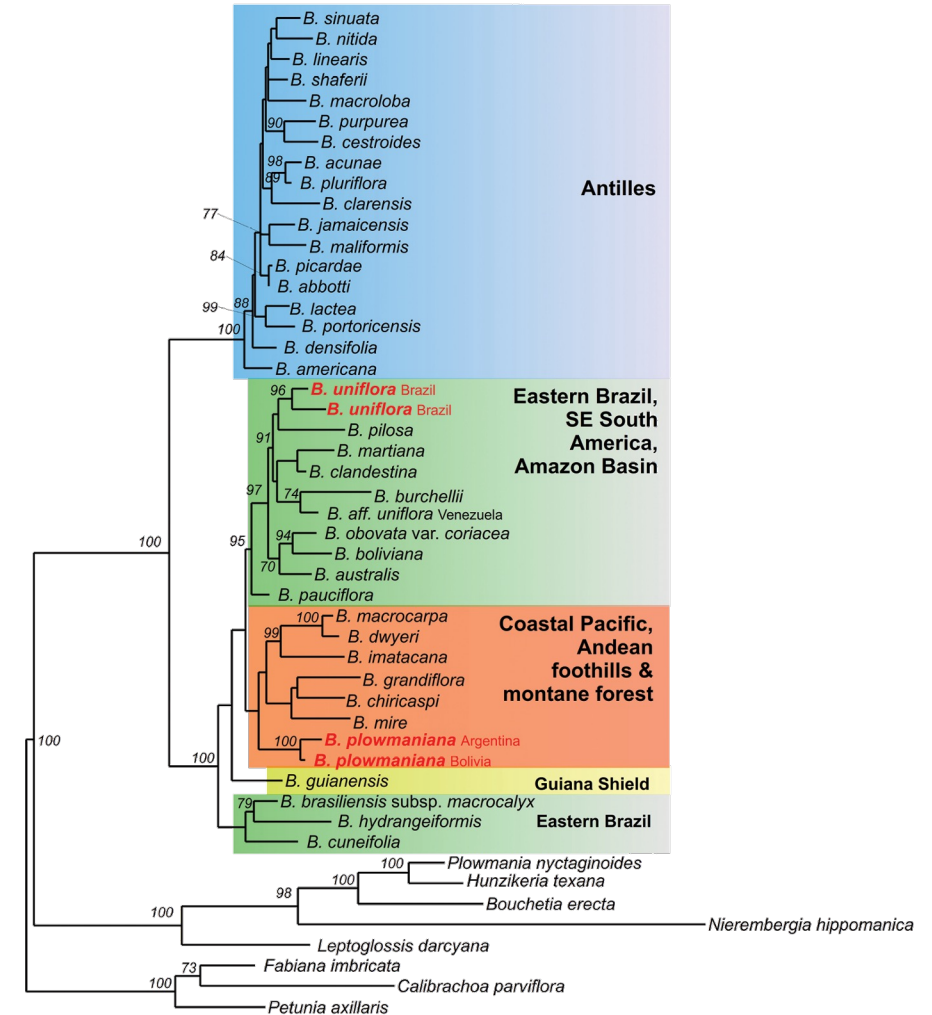
Part 5: Branch support

5. Bootstrap

The bootstrap is a computer-based technique for assessing the accuracy of a almost any statistical estimate.

A phylogenetic tree can be considered as a statistical estimate.

Joseph Felsenstein introduced the use of bootstrap in phylogenetic analyses to assess the confidence of each clade of the tree.



Phylogenetic tree by Filipowicz, N et al. / CC BY 4.0

5. Phylogenetic bootstrap

From the alignment, sites are **sampled randomly with replacement**.

This is equivalent to assigning **random weights** to each sites.

For each new alignment we infer a phylogeny (same parameters).

A branch found in $x\%$ of the bootstrap trees have a support of $x\%$.

Seq1	A	T	A	A	T	C	T
Seq2	C	T	A	G	T	C	A
Seq3	G	G	C	T	C	G	C
Seq4	C	A	C	A	A	C	C
Seq5	C	A	C	A	A	C	G

Seq1							
Seq2							
Seq3							
Seq4							
Seq5							

Seq1							
Seq2							
Seq3							
Seq4							
Seq5							

Seq1							
Seq2							
Seq3							
Seq4							
Seq5							

5. Limitations

A great number of replicates should be performed (1,000 is recommended) → **slow**.

High bootstrap value ($>90\%$) does not imply a true branch. A tree made of fully supported branches can be entirely wrong.

→ bootstrap is as relevant as the initial analysis. If the initial analysis is wrong (e.g., from a set of completely unrelated genes), the bootstrap confidence values will be meaningless.

5. Limitations

Bootstrap is highly sensitive.

If a single taxa can be placed elsewhere in the tree without affecting the optimality criterion (e.g., randomly resolved polytomies) the bootstrap value will drop.

5. Alternatives: Ultra-fast bootstrap

Fast approximation of the bootstrap. Need at least 1,000 replicates to produce meaningful scores.

Pros: almost costless.

Cons: not as reliable as bootstrap. Still a nice alternative when dealing with very large trees.

Implemented only in the software IQ-TREE.

Bui *et al.*, MBE 2013

Hoang *et al.*, MBE 2018

5. Alternatives: Transfer Bootstrap Expectation (TBE)



Bootstrap trees must be produced just like for standard bootstrap.

The difference is how support values are computed.

Pros: responds to the sensitivity issue. If an almost identical branch is found in a bootstrap tree, this will increase the score. As a result, deep branches can get better scores.

Cons: Still slow because it requires at least 100 bootstrap trees.

Implemented all modern ML softwares.

5. Alternatives: approximate likelihood ratio test (aLRT)



For each branch of the tree, the likelihood is compared to the likelihood of the tree obtained via the best possible NNI. From the difference of these two likelihood scores, a p-value can be derived.

By repeating this for each branch, we obtain a p-value for each branch of the tree.

Pros: performing a single NNI and updating the likelihood score is very fast. Suitable for very large trees.

Cons: most reviewers demand bootstraps. aLRT is considered as less reliable.

Summary

In summary

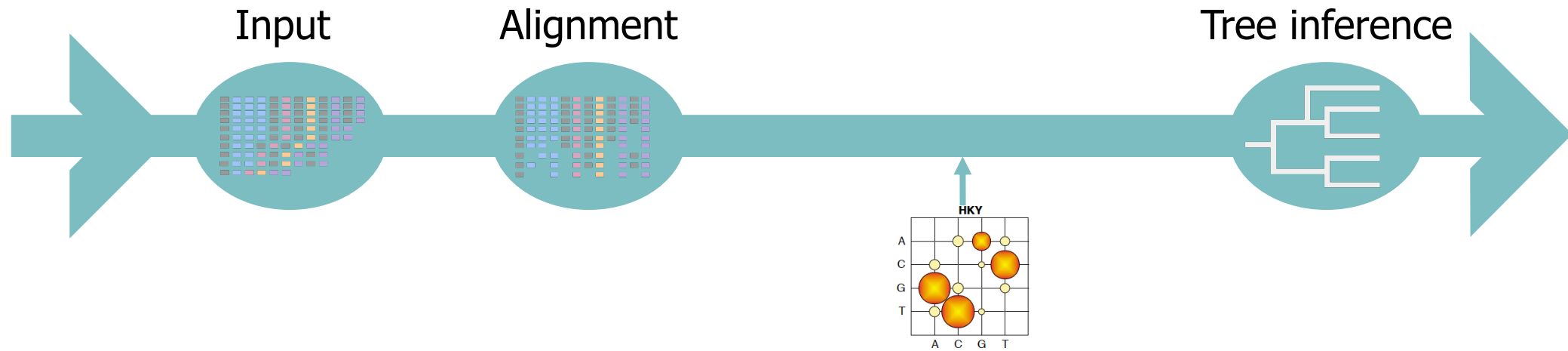
The data can be both distances or sequences.

The methods can be clustering methods (UPGMA, NJ) or they can rely on an optimality criterion (ME, Parsimony, ML, Bayesian).

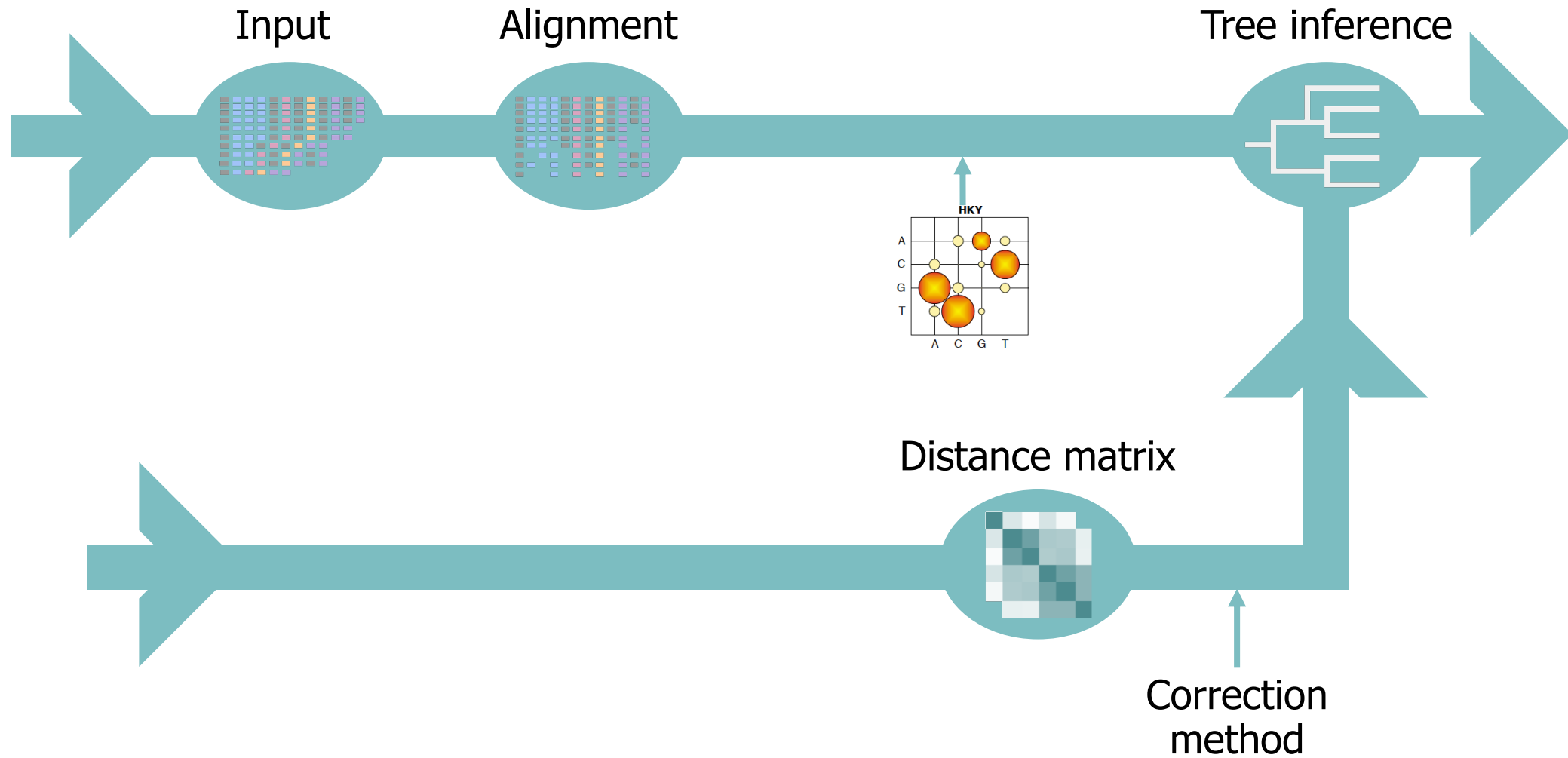
Distance-based analyses are fast but have limitations.

Character-based analyses are more robust but can be resource- and time-consuming.

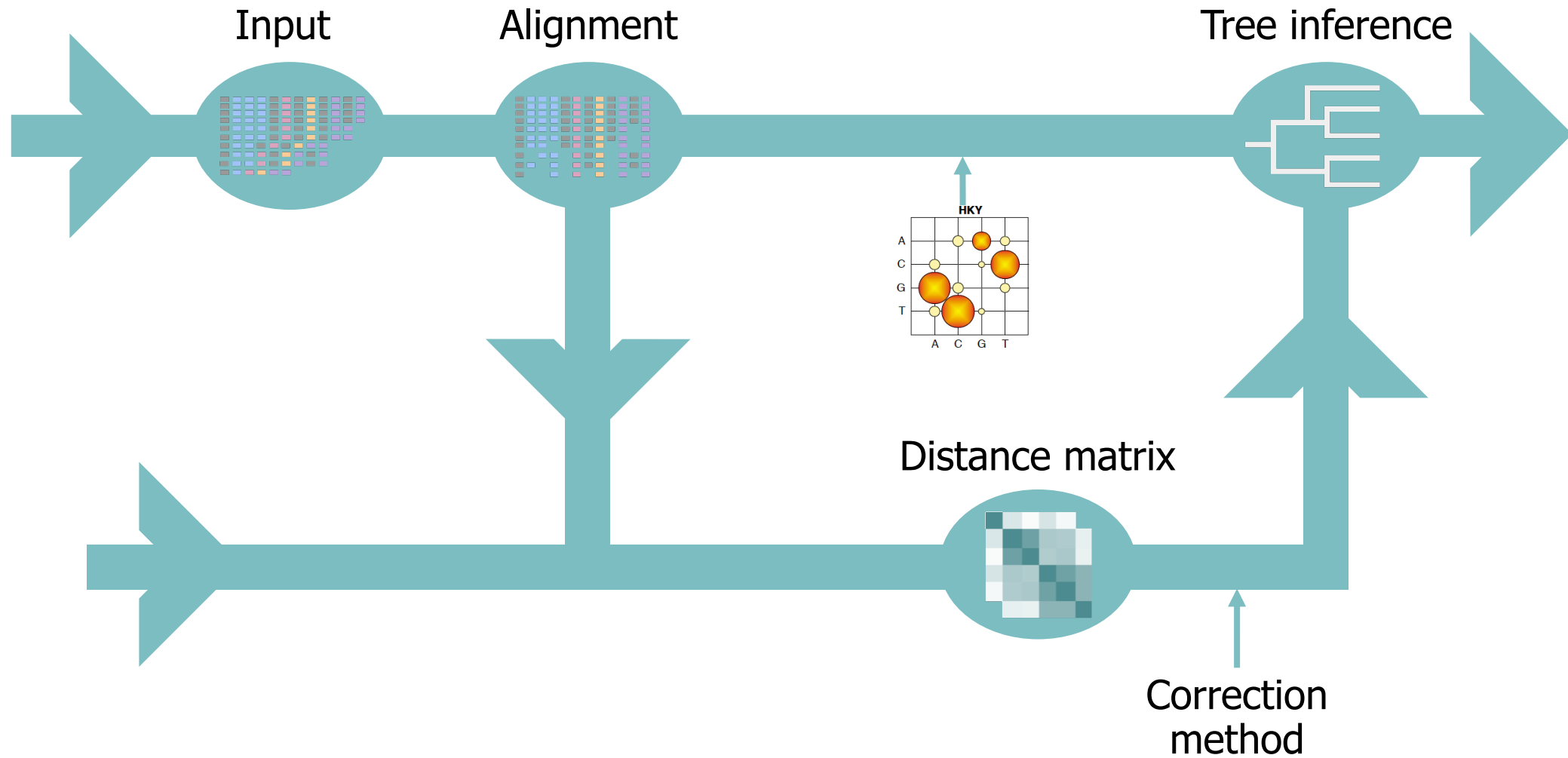
In summary: general workflow



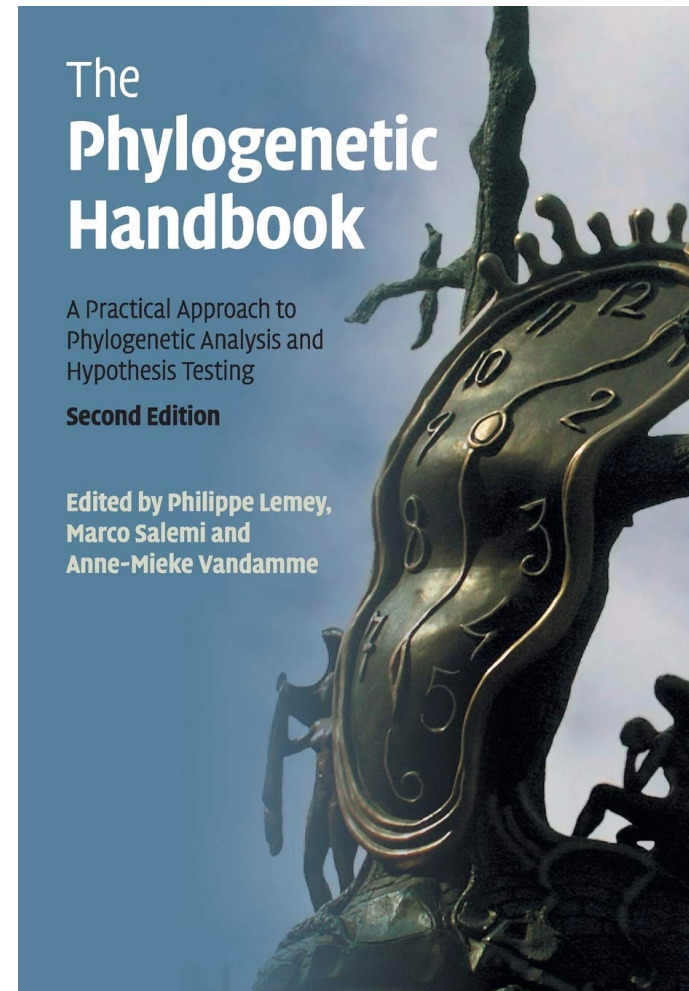
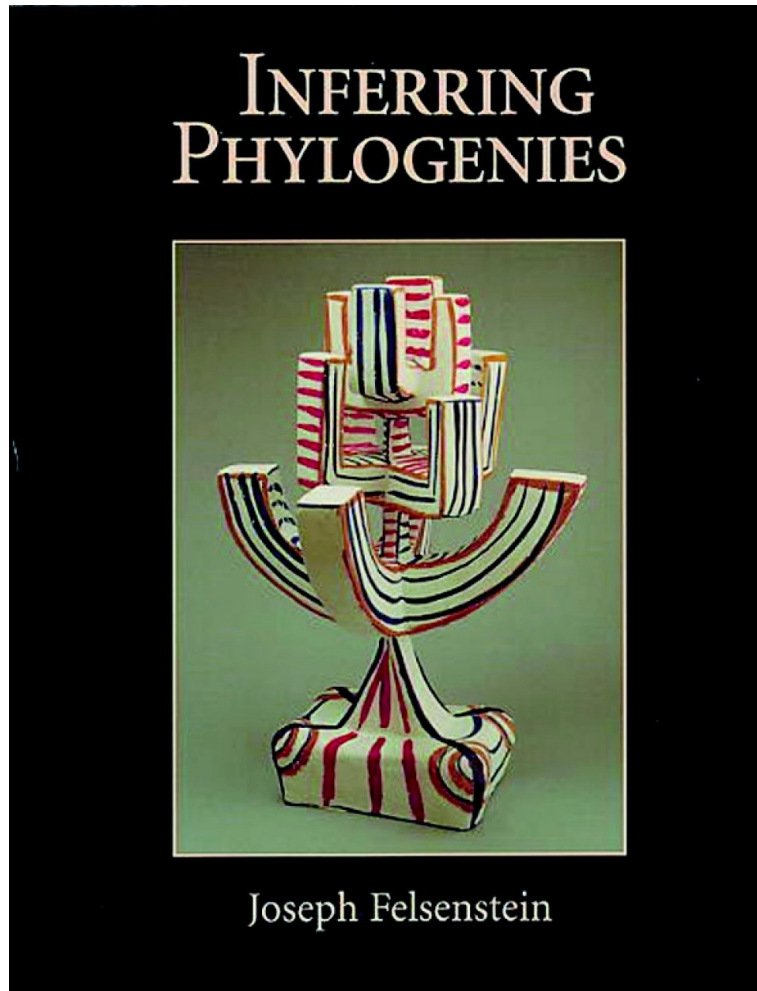
In summary: general workflow



In summary: general workflow



Further reading



Acknowledgements

The creation of this training material was commissioned by ECDC to Institut Pasteur with the direct involvement of Julien Guglielmini