Lesson 10. Content

1. Phylogenetic analyses

Phylogenetic analyses

With phylogenetic analyses we try to reconstruct the evolutionary history of life to show where different species or organisms diverged

These relationships are inferred from heritable traits, such as **DNA** or **protein** sequences (or morphology)

Result of such analyses are phylogenetic trees — diagrams containing hypotheses of relationships

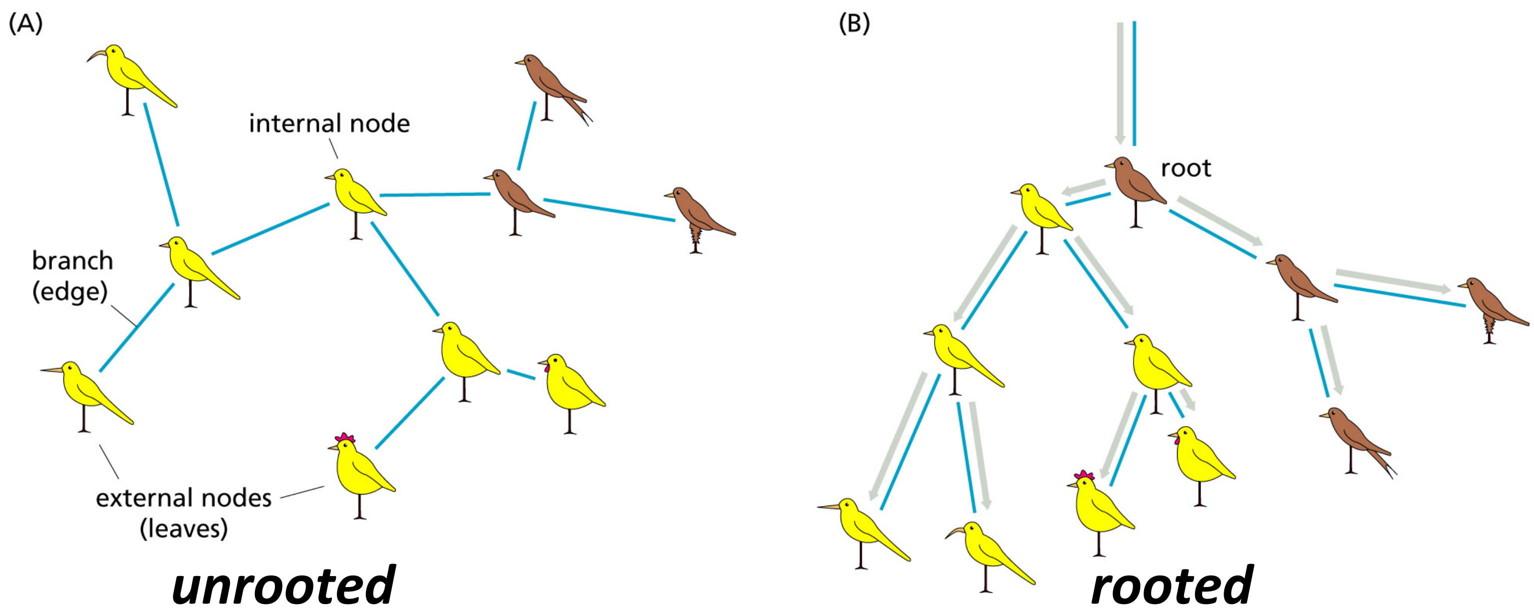
The key assumption when constructing a phylogenetic tree from a set of sequences is that they are all derived from a single ancestral sequence, i.e. they are homologous, specifically orthologous, that is pairs of genes whose last common ancestor occurred immediately before a speciation event

Phylogenetic trees

A phylogenetic tree is a diagram proposing an hypothesis for the evolutionary relationships between a set of objects (data), usually genes or proteins, used to derive it

These objects are referred to as: *taxa* or *operational taxonomic units (OTUs)*; in **species trees**, the taxa are labeled with species name

Example of species trees for imaginary bird species



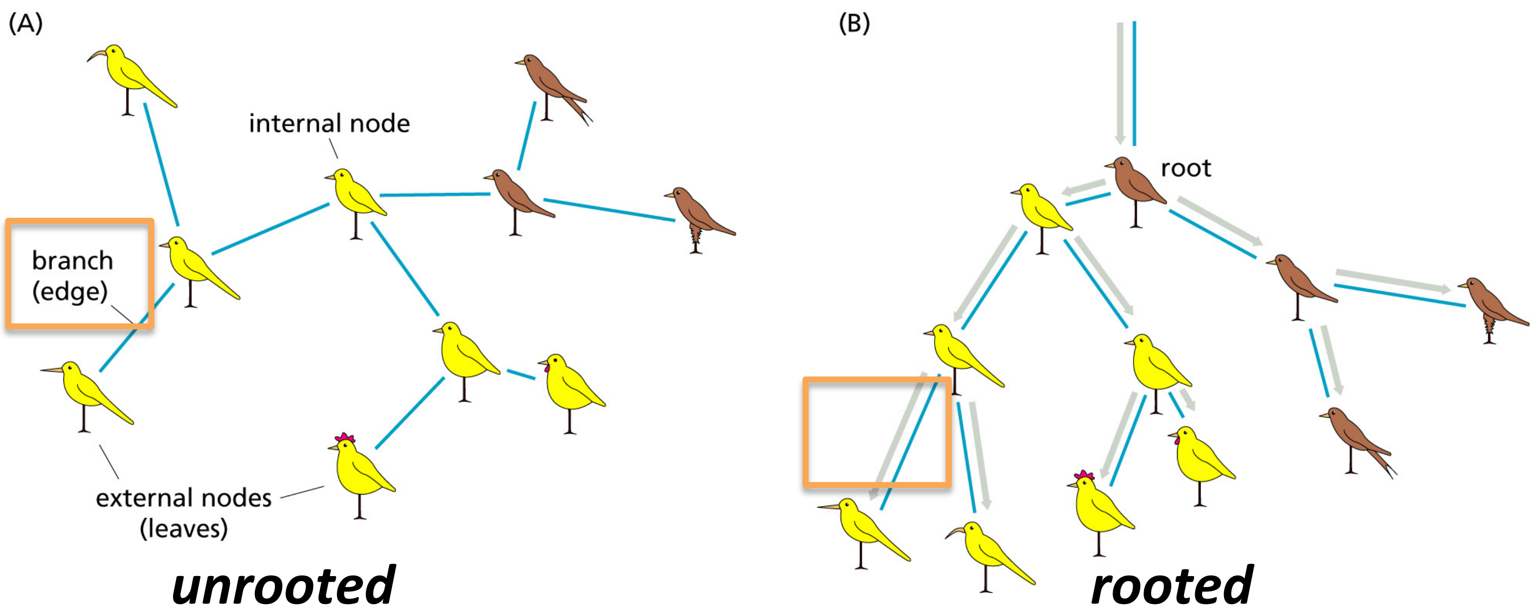
Phylogenetic trees

A phylogenetic tree is a diagram proposing an hypothesis for the evolutionary relationships between a set of objects (data), usually genes or proteins

These objects are referred to as: taxa or operational taxonomic units (OTUs); in **species trees**, the taxa are labeled with species name

Example of species trees for imaginary bird species

Represents the evolutionary relationship between species



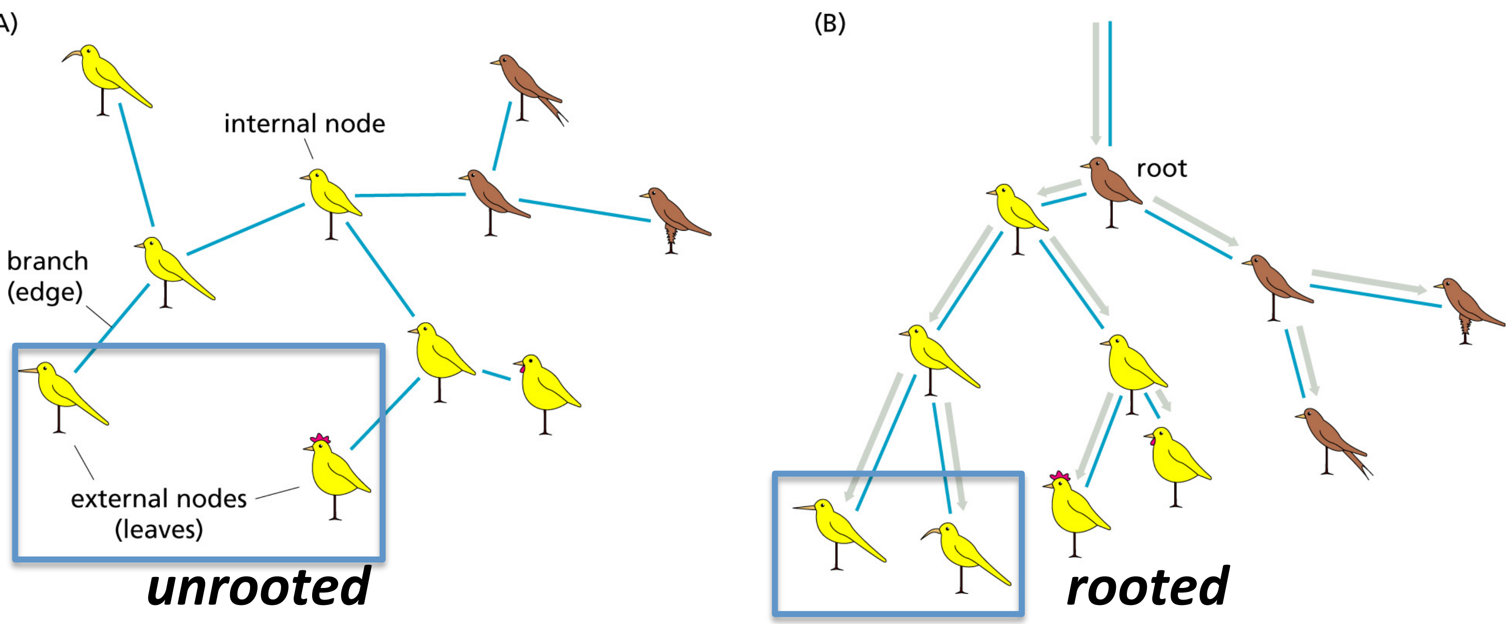
Phylogenetic trees

A phylogenetic tree is a diagram proposing an hypothesis for the evolutionary relationships between a set of objects (data), usually genes or proteins

These objects are referred to as: taxa or operational taxonomic units (OTUs); in **species trees,** the taxa are labeled with species name

Example of species trees for imaginary bird species

*Also **leaves**, are:*
i. Existing species not yet evolved into new ones,
OR ii. Extinct species whose lineage died out



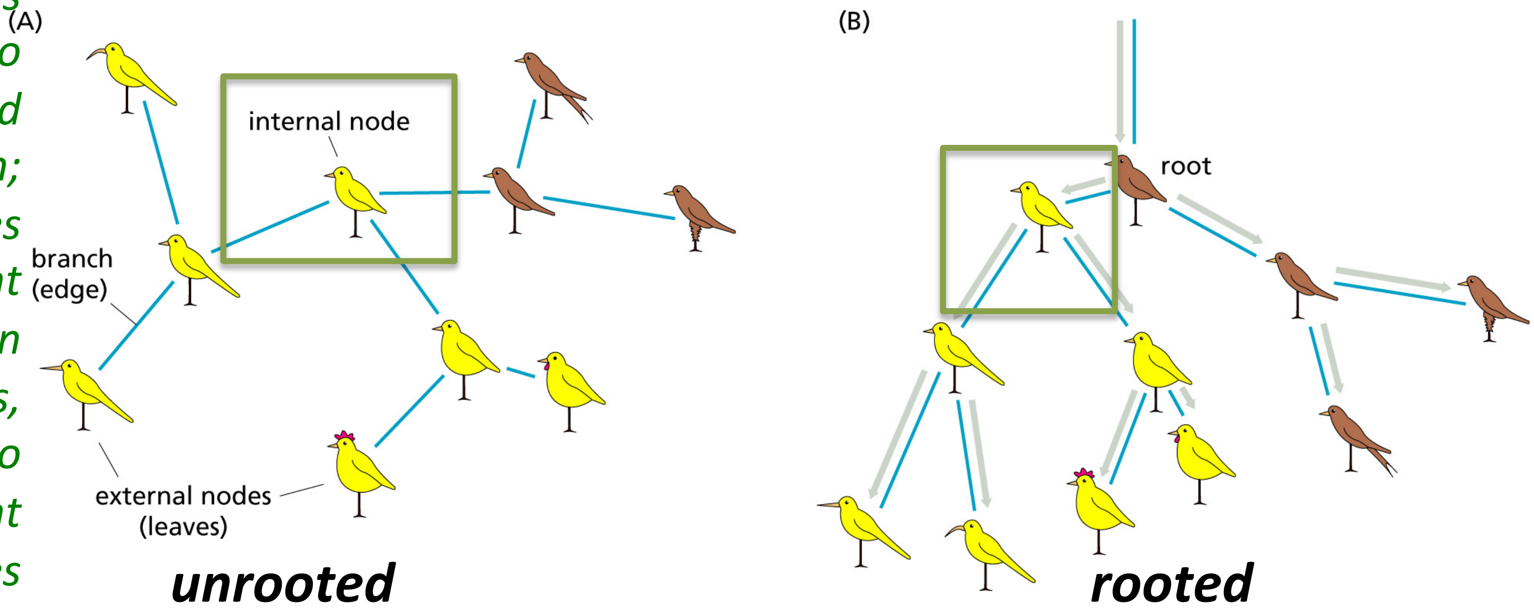
Phylogenetic trees

A phylogenetic tree is a diagram proposing an hypothesis for the evolutionary relationships between a set of objects (data), usually genes or proteins

These objects are referred to as: taxa or operational taxonomic units (OTUs); in **species trees,** the taxa are labeled with species name

Example of species trees for imaginary bird species

Ancestral states hypothesized to have occurred during evolution; Internal nodes represent speciation events, producing two descendant species

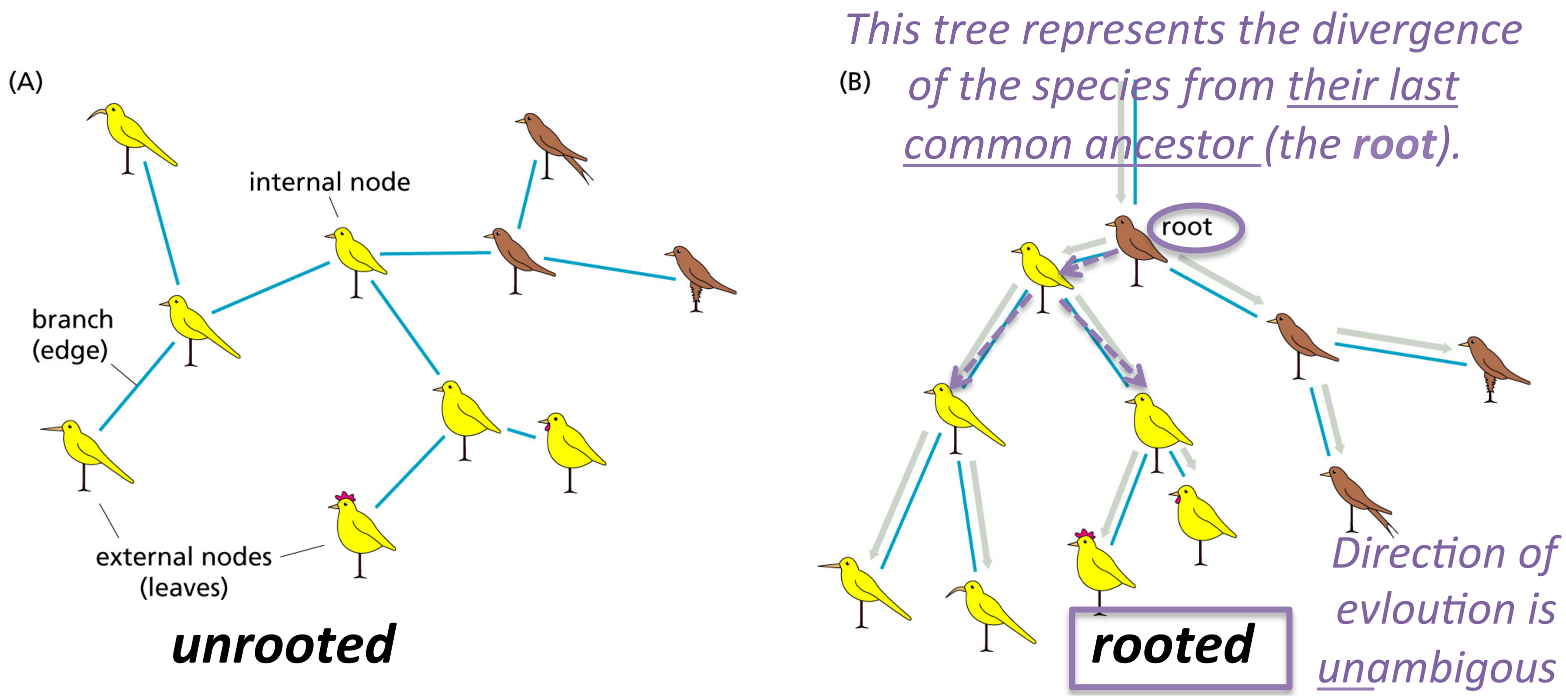


Phylogenetic trees

A phylogenetic tree is a diagram proposing an hypothesis for the evolutionary relationships between a set of objects (data), usually genes or proteins

These objects are referred to as: taxa or operational taxonomic units (OTUs); in **species trees,** the taxa are labeled with species name

Example of species trees for imaginary bird species



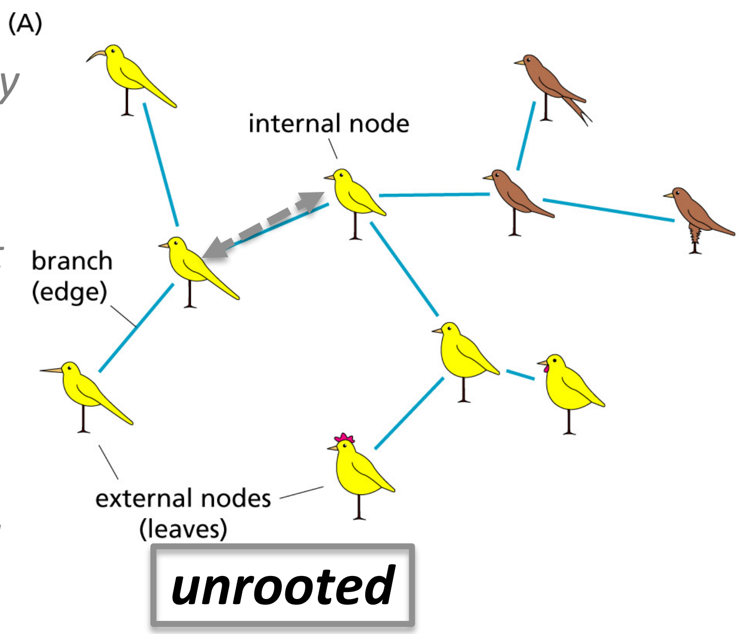
Phylogenetic trees

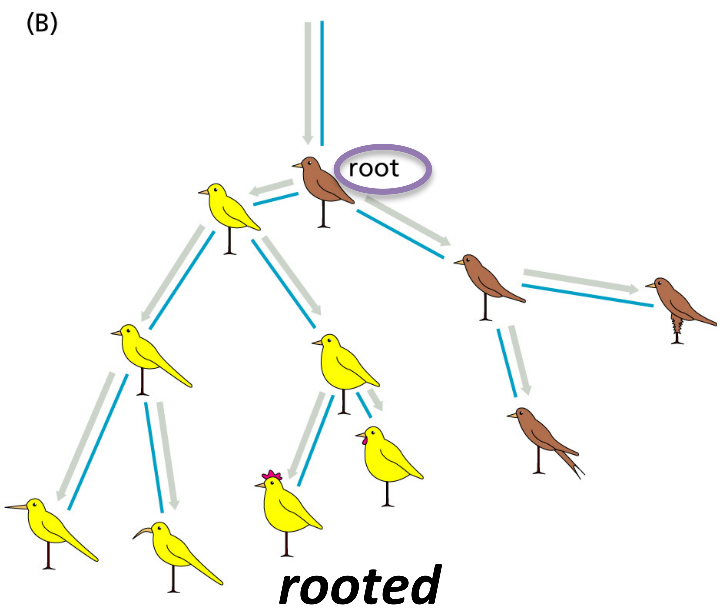
A phylogenetic tree is a diagram proposing an hypothesis for the evolutionary relationships between a set of objects (data), usually genes or proteins

These objects are referred to as: taxa or operational taxonomic units (OTUs); in **species trees**, the taxa are labeled with species name

Example of species trees for imaginary bird species

This tree shows the evolutionary relationships but does not identify the last common ancestor; unclear which ancestral species evolved from which



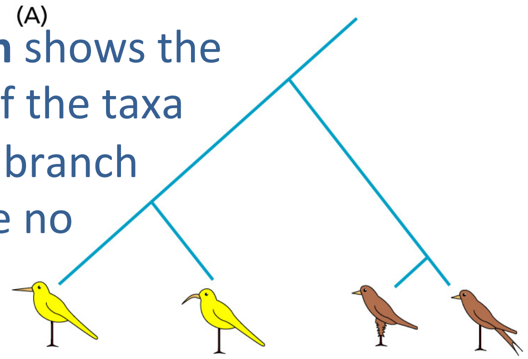


Phylogenetic trees

Phylogenetic trees may be shown in three basic types. Below examples from the same data are shown

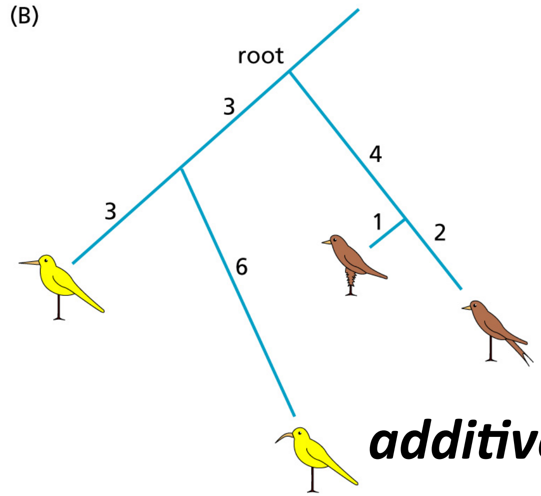
A **caldogram** shows the genealogy of the taxa but in it the branch lengths have no meaning

(no info on the timing and extent of divergence)

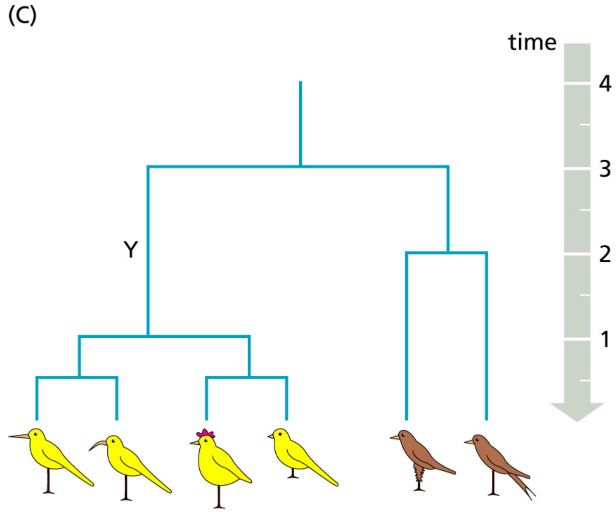


caldogram

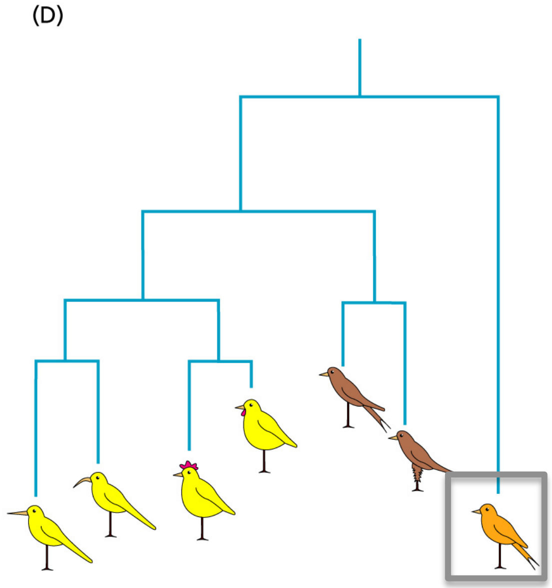
can be unrooted



additive tree



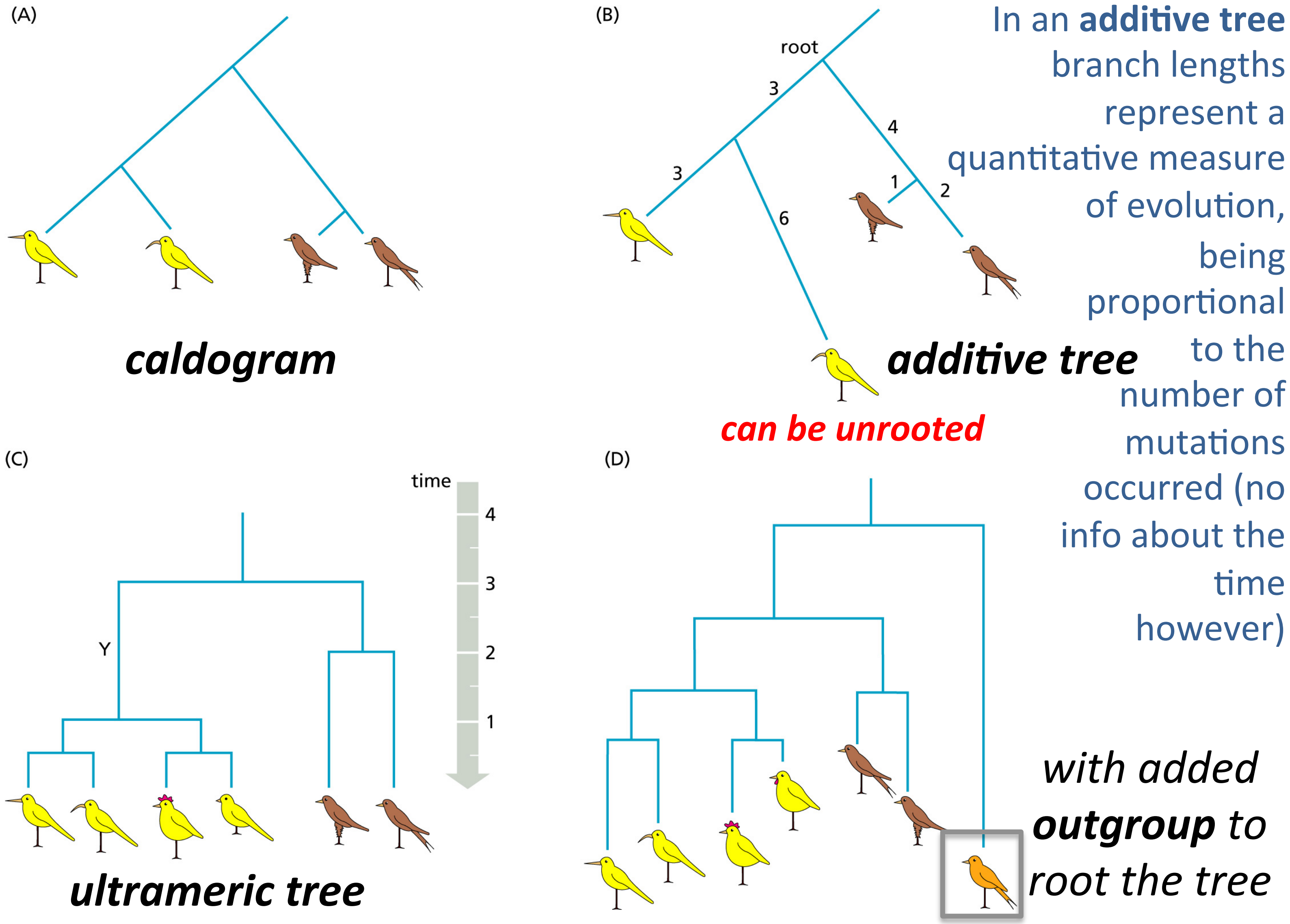
ultrameric tree



*with added **outgroup** to root the tree*

Phylogenetic trees

Phylogenetic trees may be shown in three basic types. Below examples form the same data are shown

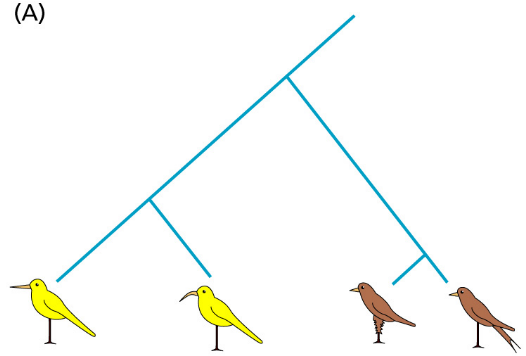


In an **additive tree** branch lengths represent a quantitative measure of evolution, being proportional to the number of mutations occurred (no info about the time however)

Phylogenetic trees

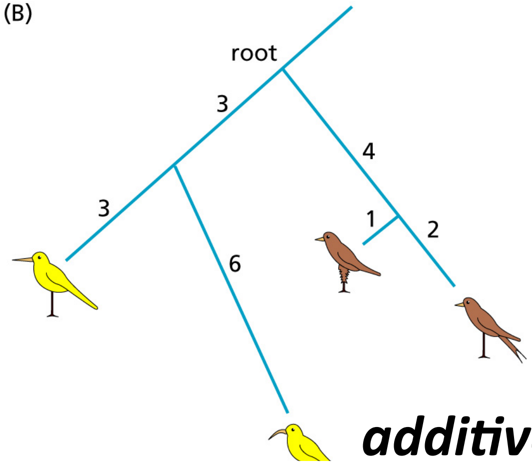
Phylogenetic trees may be shown in three basic types. Below examples form the same data are shown

(A)



caldogram

(B)

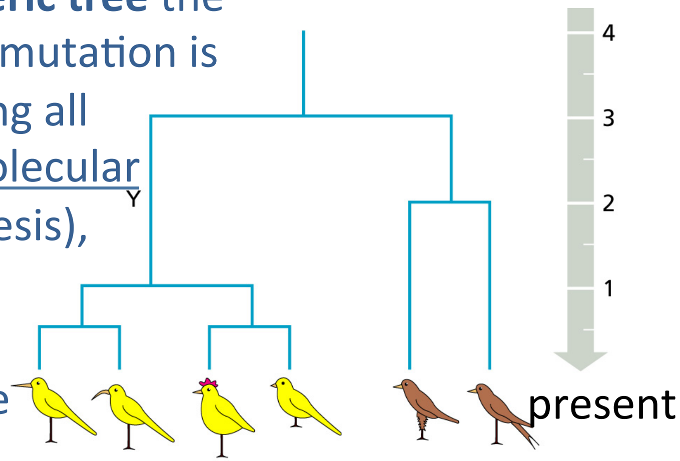


additive tree

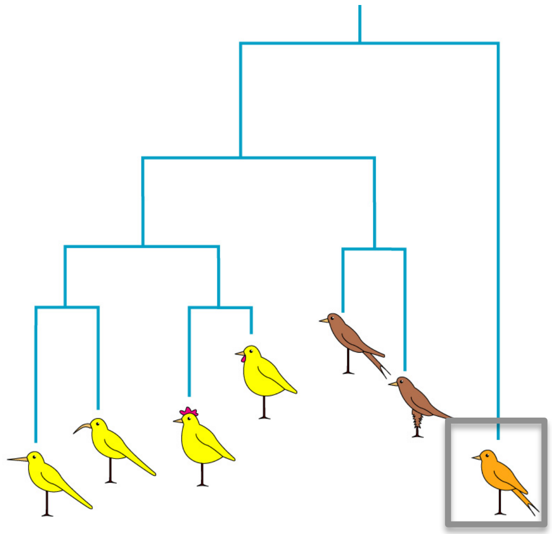
When the molecular clock hypothesis does not hold true, an unrooted additive tree can be more accurate

In an **ultrameric tree** the same rate of mutation is assumed along all branches (<u>molecular clock</u> hypothesis), the time of evolutionary events can be measured in principle

(C)



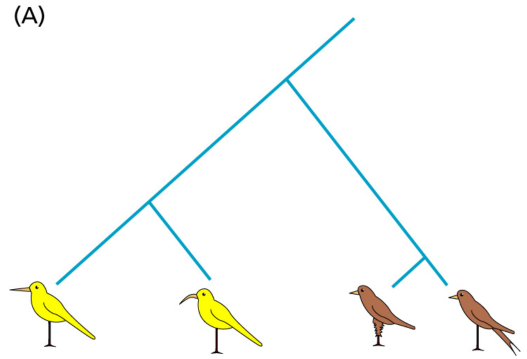
ultrameric tree



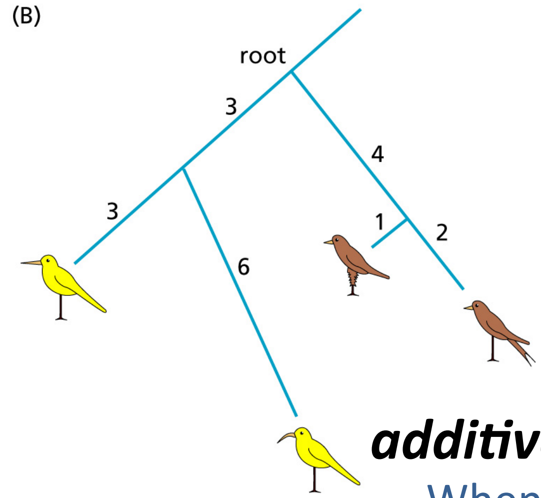
*with added **outgroup** to root the tree*

Phylogenetic trees

Phylogenetic trees may be shown in three basic types. Below examples form the same data are shown



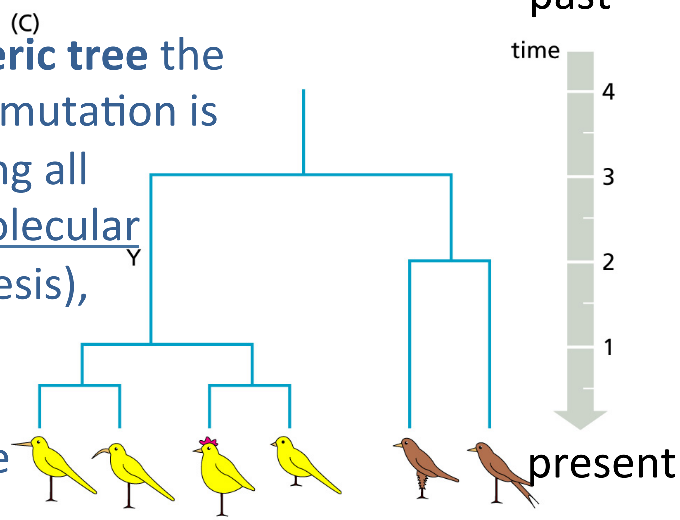
caldogram



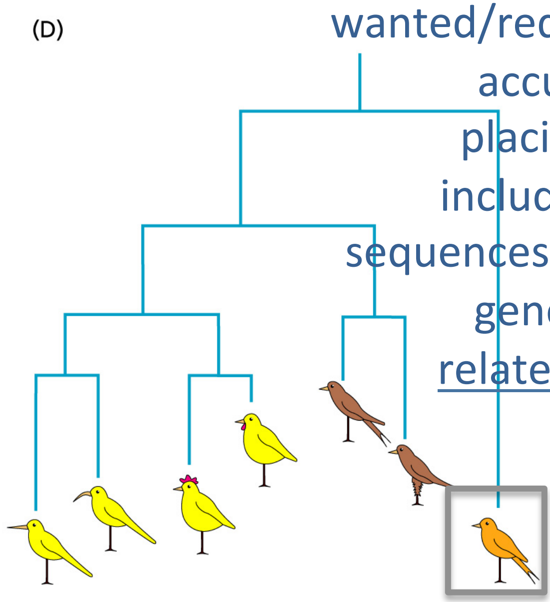
additive tree

When a rooted tree is wanted/required, the most accurate method of placing the root is to include in the dataset sequences from species or genes only distantly related (an **outgroup**)

In an **ultrameric tree** the same rate of mutation is assumed along all branches (molecular clock hypothesis), the time of evolutionary events can be measured in principle



ultrameric tree



*with added **outgroup** to root the tree*

Phylogenetic analyses

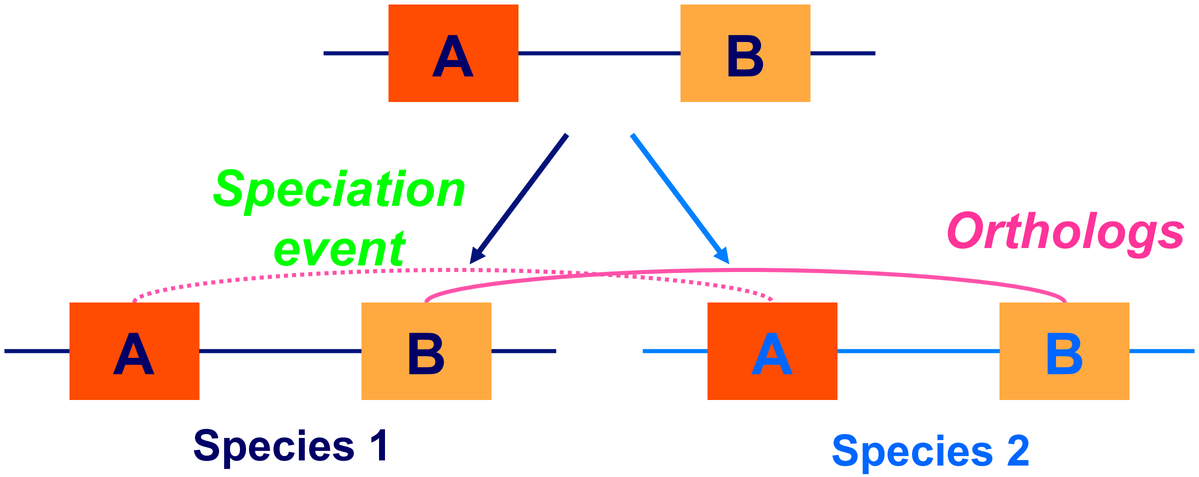
Before performing a phylogenetic analysis, four decisions must be made:

- which data to use
- which method
- which evolutionary model
- which (if any) tests to assess the robustness of prediction

The above decisions are usually inter-dependent

Phylogenetic analyses: which data

The ideal is a genomic region that occurs in every species but only once in the genome (to avoid misassignments of orthology)



It must have little (if any) horizontal gene transfer (HGT)

The rate of change in it must be fast enough to distinguish between closely related species but not so fast that regions from very distantly related species cannot be confidently aligned

Phylogenetic analyses: which data

For prokaryotes, the small ribosomal subunit rRNA (16S RNA) (although occurring in several copies in some genomes) has been found to be one of the best genomic segments for these analyses; the original proposal that prokaryotes comprised two distinct domains (bacteria and archaea) was in fact based on analysis of this region

A few protein-coding sequences have also been found to be suitable for determining the evolutionary relationships of species

For the animal kingdom a 658-bp segment of the gene for cytocrome c oxidase I, a component of the mitochondrial machinery involved in cellular aerobic respiration and present in all animals can be used

Phylogenetic analyses: which method

Methods for reconstructing phylogenetic trees can be divided in two broad groups:

- Distance-based: methods which derive a distance measure from each alignment pair sequences and uses these distances to obtain the tree

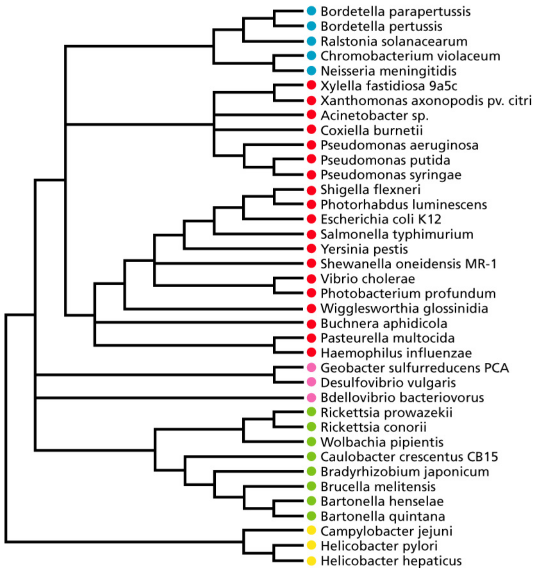
- Methods which use multiple alignments directly

Phylogenetic analyses: which method

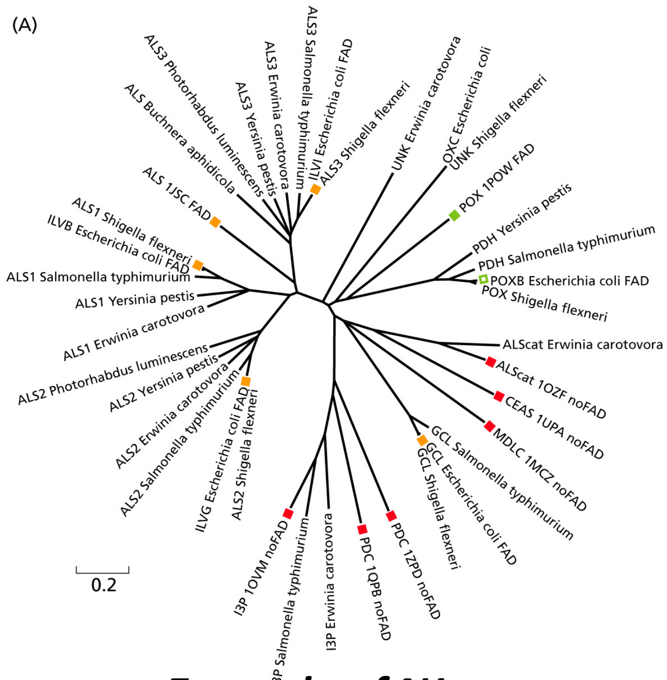
Commonly used <u>distance-based methods</u>, both producing a single tree with defined branch lengths, are the:

- **UPGMA** (Unweighted Pair-Group Method using Arithmetic Average)

- **Neighbor-joining** (NJ)



Example of UPGMA tree

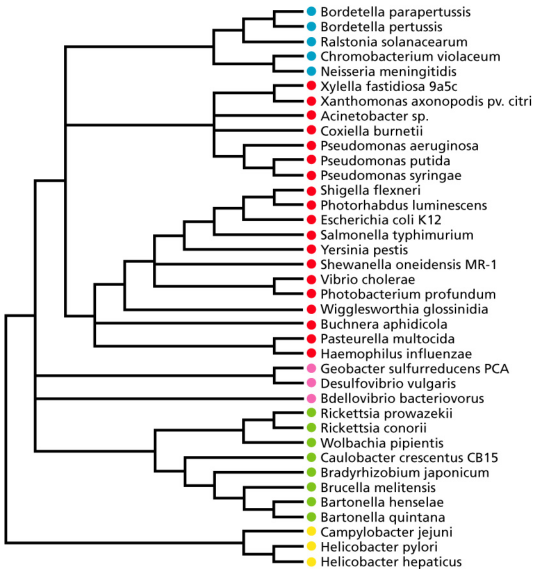


Example of NJ tree

Phylogenetic analyses: which method/model

The **UPGMA** (Unweighted Pair-Group Method using Arithmetic Average) makes the assumption that **the sequences evolved at a constant equal rate** over time (the molecular clock hypothesis)

It produces **rooted trees** with all sequences at the same distance from the last common ancestor



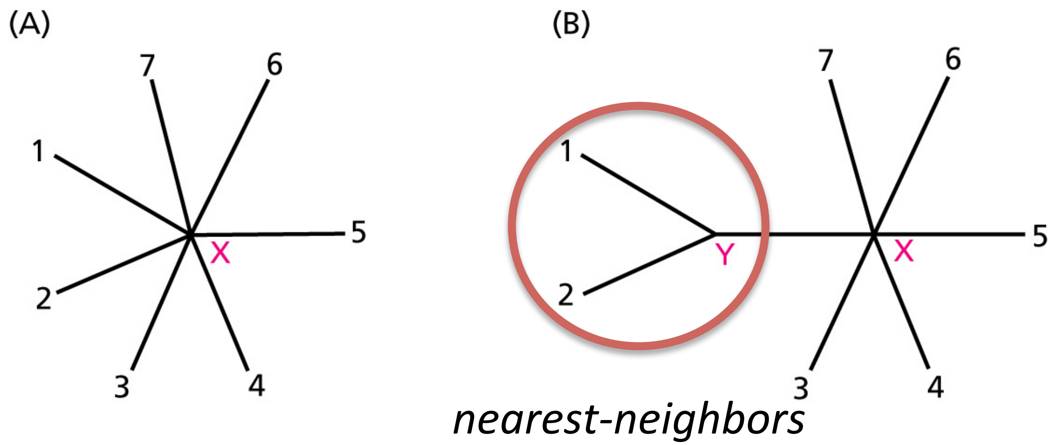
Example of UPGMA tree

Phylogenetic analyses: which method/model

The **Neighbor-Joining** (NJ) belongs to the group of **minimum evolution methods**, assuming that the most suitable tree will be the one proposing the least amount of evolution, i.e. that for which the total branch length, *S*, is shortest

It produces **unrooted trees** and is more generally applicable

Neighbors in this type of trees are defined as a pair of nodes that are separated by just one node, pairs of tree nodes are identified at each step of the method and used to gradually build up the tree

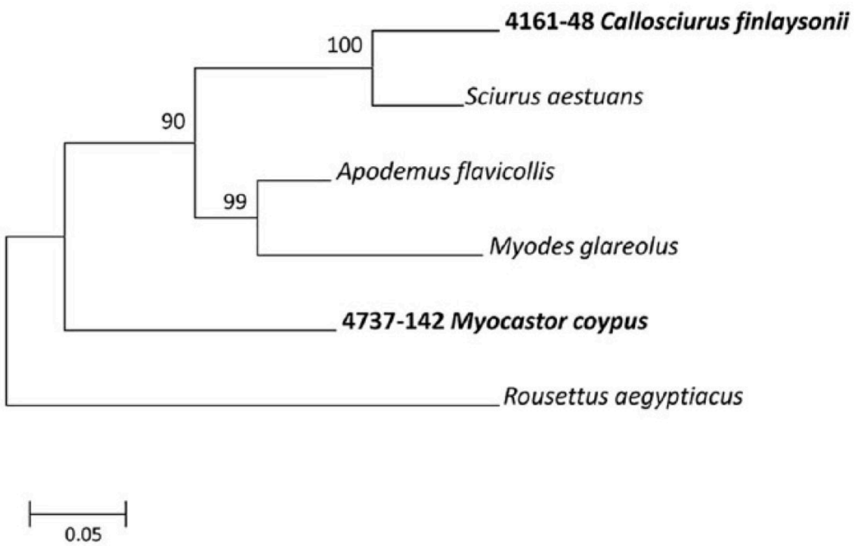


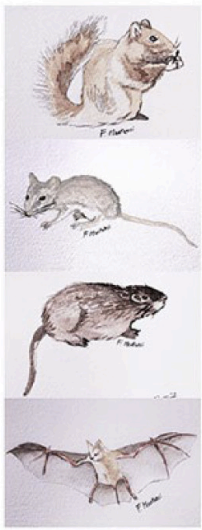
First step of the NJ method: sequences 1 and 2 are identified as the 1st pair of nearest-neighbors; they are separated from node X by an internal branch to internal node Y

Phylogenetic analyses: which method/model

The **Maximum Likelihood** (ML) method also belongs to maximum parsimony methods (based on the **minimum evolution** principle), but is directly based on a <u>multiple alignment</u>

It generates multiple tree topologies, then estimates the likelihood of a each tree topology to have produced the given data (alignment) assumed an evolutionary model and selects the topology that produces the greatest likelihood as the most appropriate hypothesis of the evolutionary history





The maximum-likelihood tree (the branch lengths represent the expected number of substitution per site)

Phylogenetic analyses: evolutionary model

Sequence data often do not conform to a **molecular clock** hypothesis (firstly hypothesized in 1962 for hemoglobin), i.e. that DNA and protein sequences evolve at a rate that is relatively constant over time and among different organisms

Not-clock-like sequence evolution results from a variety of causes, such as changes in evolutionary pressure and increasing biological constraints, i.e. factors which make populations resistant to evolutionary change in morphological structure and metabolism

For instance, in vertebrates the vertebral column is involved in the muscle, nerve, and vascular systems and provides support and flexibility, therefore it cannot be radically altered without causing severe functional disruption

Phylogenetic analyses: robustness tests

Trees obtained from the same data with different approaches or with the same approach from different datasets can differ in their **topology** and branch lengths

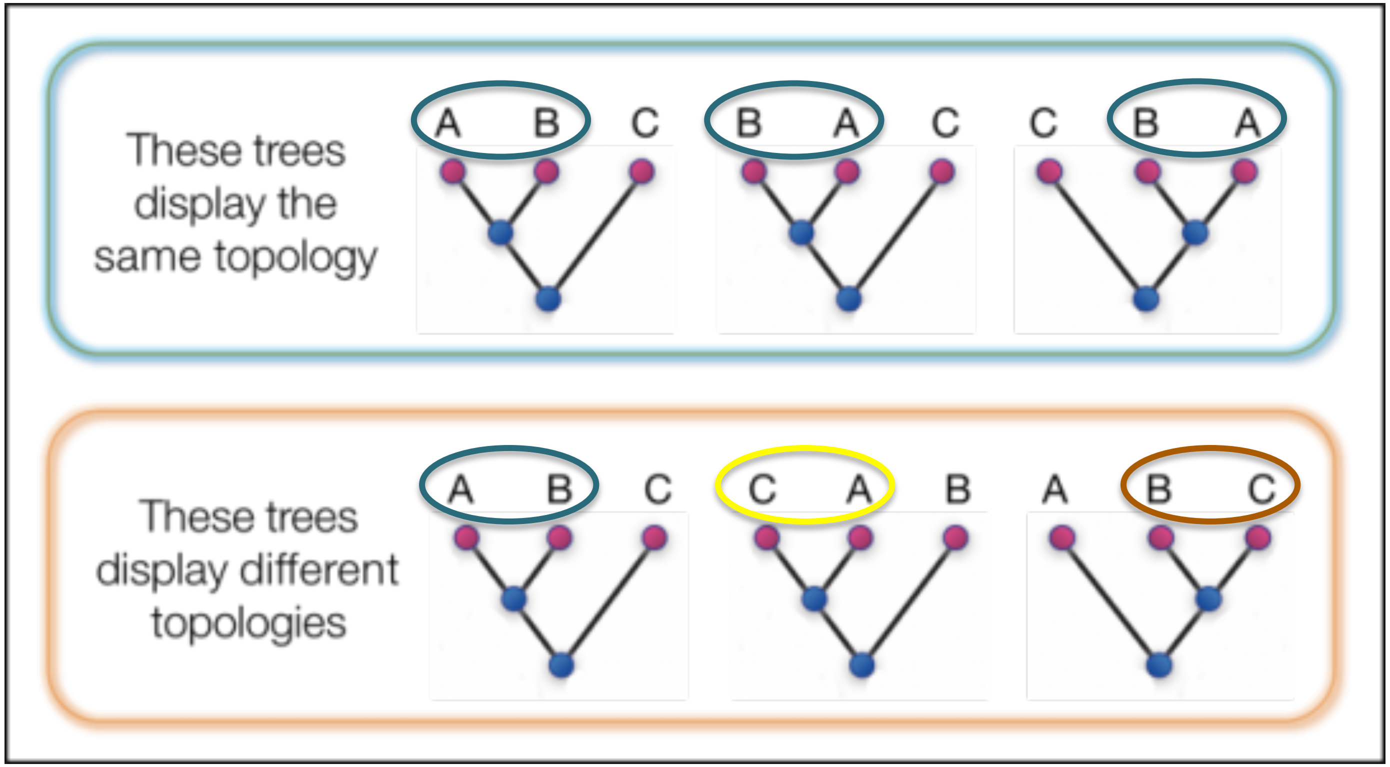
Differences in topology imply a disagreement about speciation and/or gene duplication events, therefore it is important to quantify these uncertainties

One may want to compare trees obtained from the same data with different methods, models or parameters or the reconstructed evolutionary history produced by two or more sets of data, e.g. different genes from the same set of species

A comparison among trees can identify support across a range of techniques or data

Phylogenetic analyses: robustness tests

Trees obtained from the same data with different approaches or with the same approach from different datasets can differ in their **topology** and branch lengths



Trees displaying the same topology have the same **subgroups**; trees displaying different topologies have different subgroups

Phylogenetic trees: comparison

We need therefore ways to describe a **tree topology** in a form that makes it comparable to other trees

Graphical views of trees are convenient for human visual interpretation, not for computers

A way of summarizing basic info about a tree in a computer-readable format is to subdivide (**split**) it in a collection of subgroups

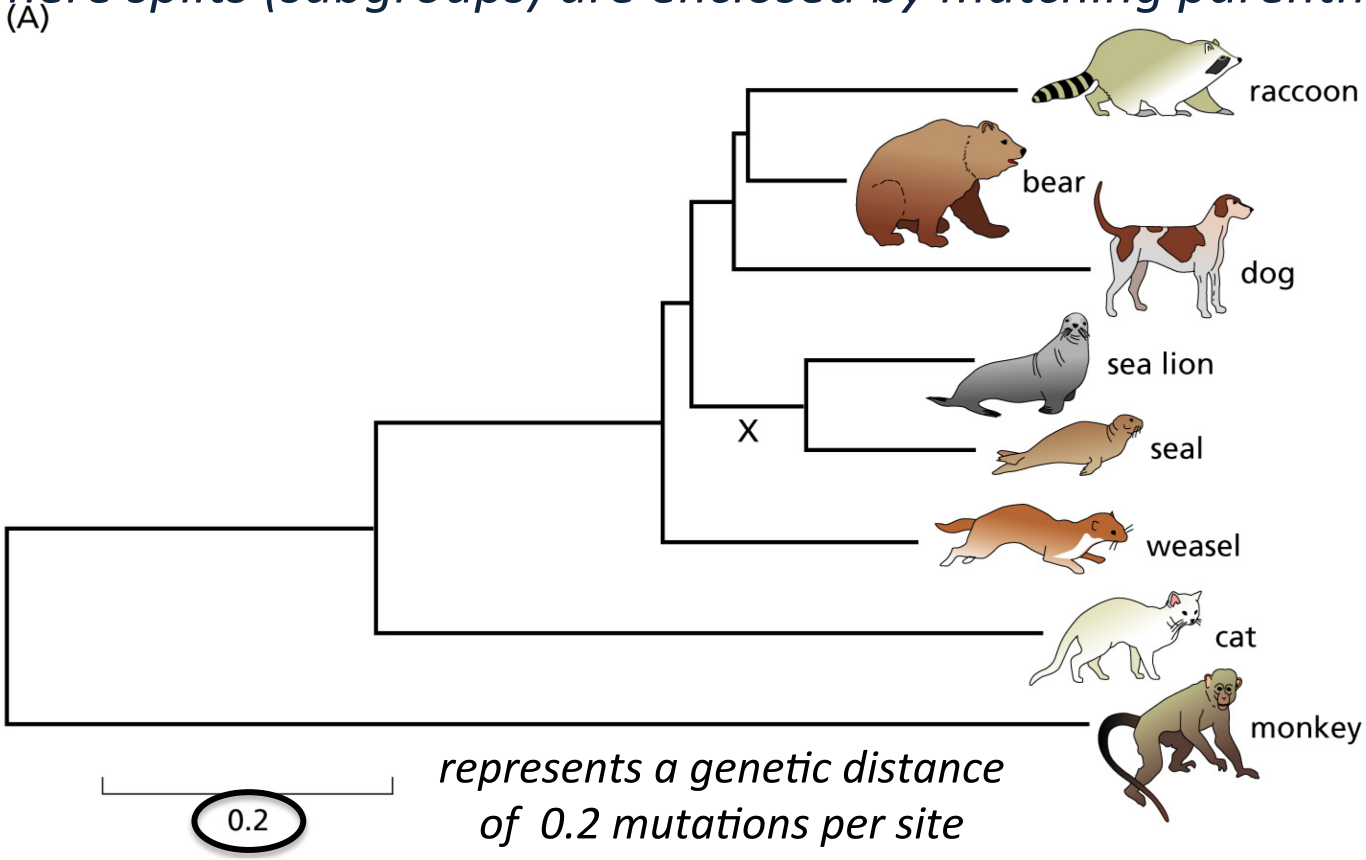
The topology comparison methods are then based on the concept of the <u>frequency of occurrence of particular splits</u> in the set of trees

Phylogenetic trees: comparison

Example of tree topology description in a computer-readable form (the Newick format):

((racoon, bear), ((sea_lion, seal), ((monkey, cat), weasel)), dog);

where splits (subgroups) are enclosed by matching parentheses



Phylogenetic analyses: bootstrap analysis

A **bootstrap analysis** is designed to estimate the degree of support (as opposed to variability or uncertainty) in a given dataset for particular topological features produced on applying a given tree construction method

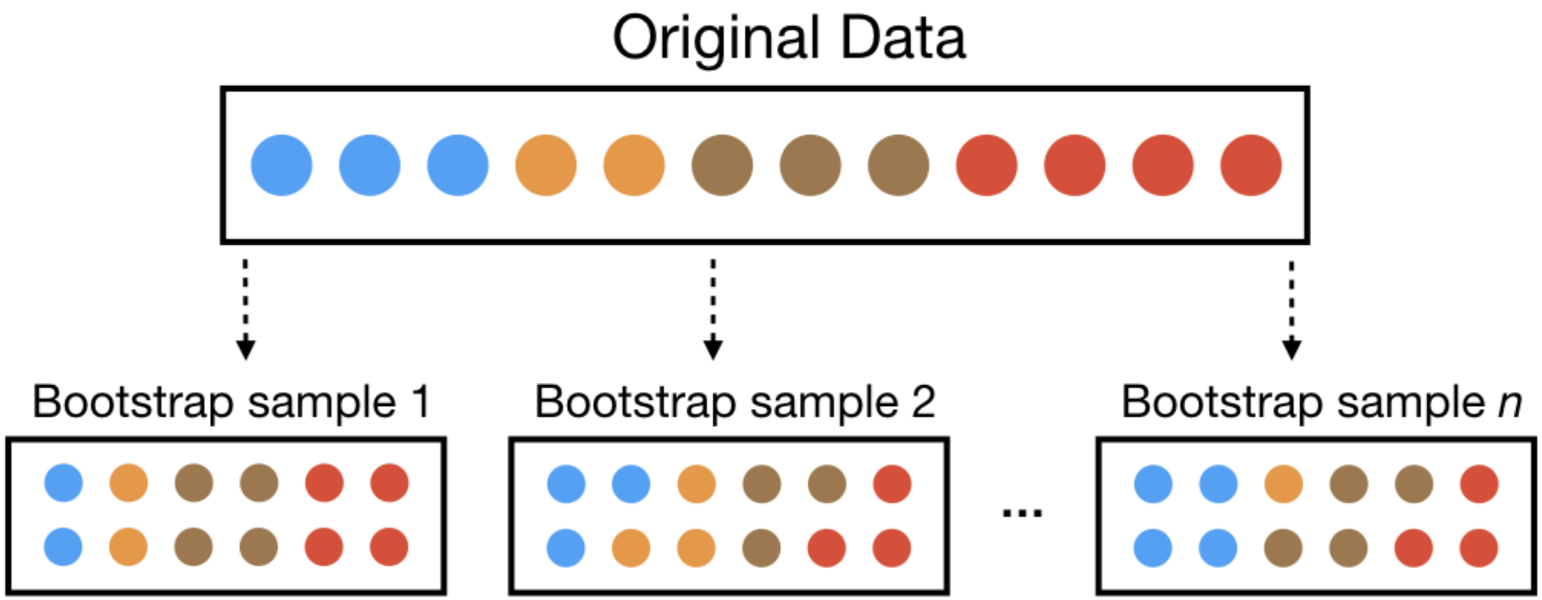
A **bootstrap analysis** is based on repeating the tree reconstruction for different samplings of the same dataset

The key assumption is that the original dataset is sampled in an unbiased manner, so a new dataset can be produced by sampling from the original one

Phylogenetic analyses: bootstrap analysis

To generate unbiased **replicate datasets**, data points are randomly selected from the original data; if there are *N* data points in the original dataset, bootstrapping normally implies selecting *N* replicate data points

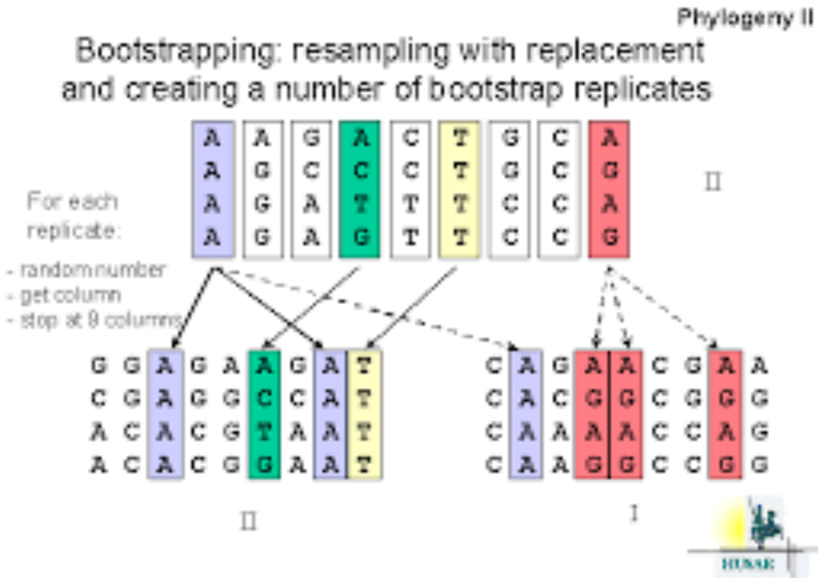
Every selection is from the complete set of the original data; some points may not be selected for the replicate while others may be selected more than once



Phylogenetic analyses: bootstrap analysis

To generate unbiased **replicate datasets**, data points are randomly selected from the original data; if there are *N* data points in the original dataset, bootstrapping normally implies selecting *N* replicate data points

Every selection is from the complete set of the original data; some points may not be selected for the replicate while others may be selected more than once



Phylogenetic analyses: bootstrap analysis

Once **bootstrap replicates** (normally several hundreds) have been obtained, a phylogenetic analysis identical to that performed on the original dataset is run on them

Then, obtained trees are compared: also in this case the frequency of each split is measured

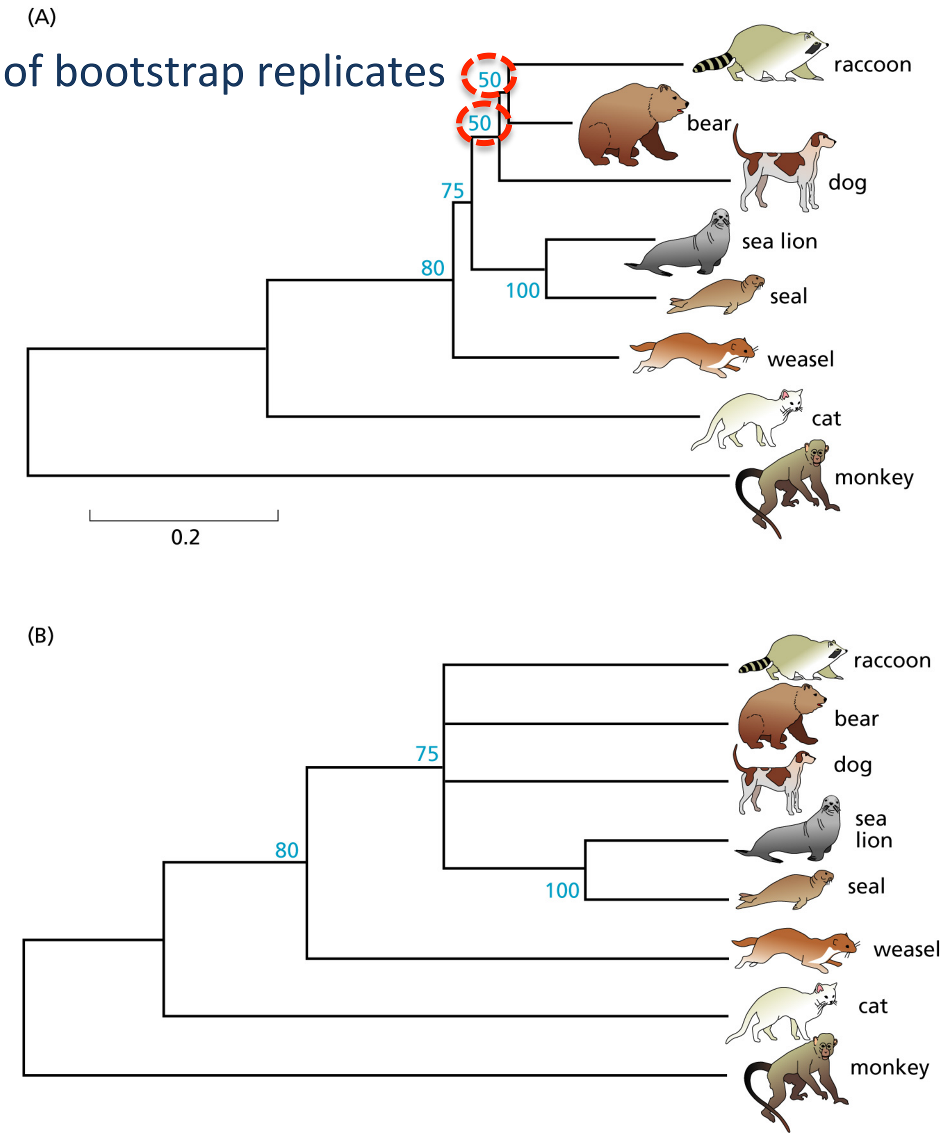
The percentage of bootstrap trees that contain each split is:

- reported in a splits list OR
- displayed on the tree itself as a number

Sometimes, as a visual aid, all <u>internal branches</u> not highly supported are <u>removed</u>, and a **condensed tree** is shown

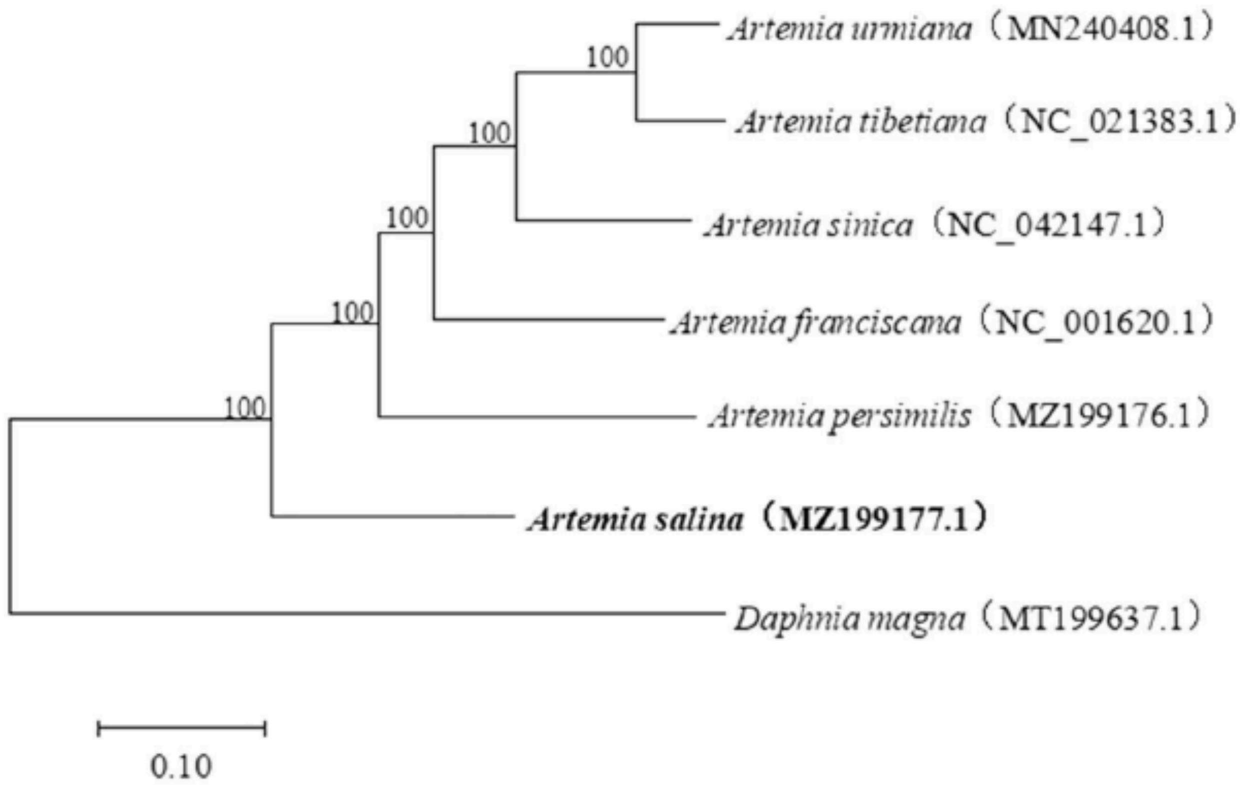
Phylogenetic analyses: bootstrap analysis

percentage of bootstrap replicates



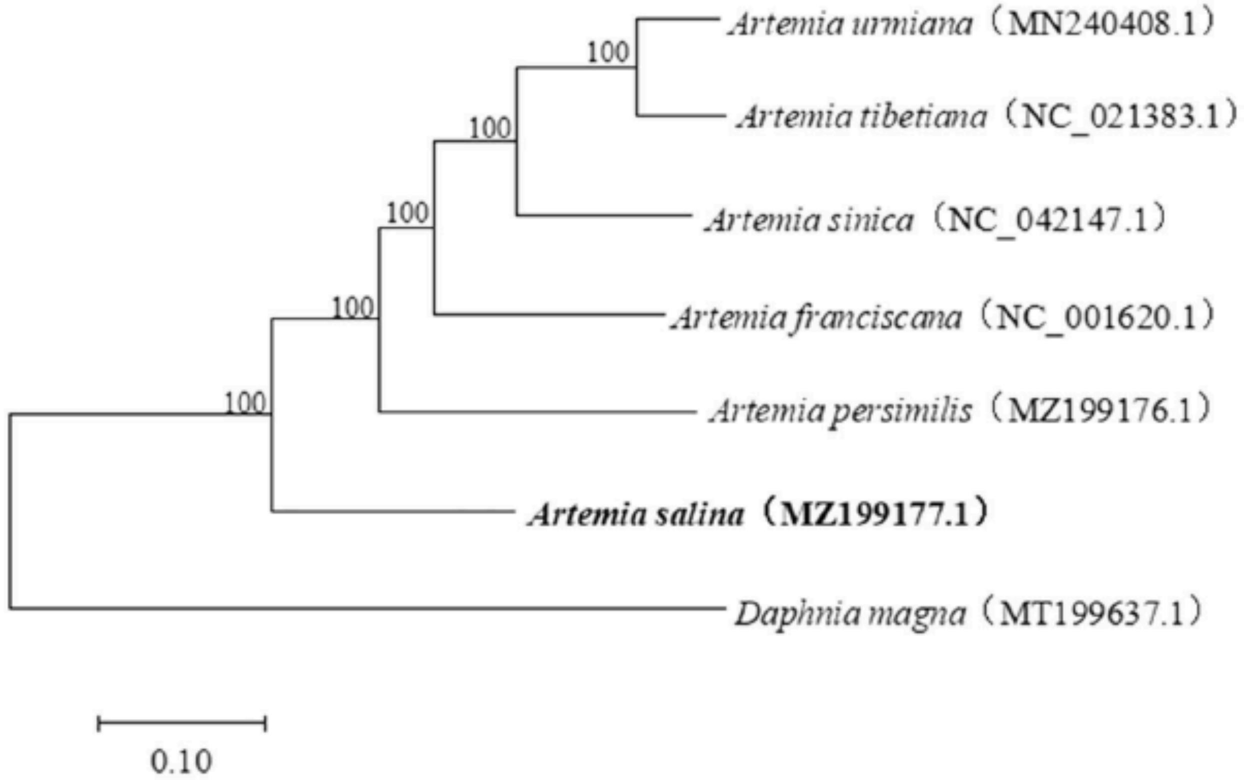
Condensed tree (threshold on splits frequency: 60%)

An example



Deji *et al*. (2021) *Mitochondrial DNA Part B* **6**:3255

An example



Data: Complete mt genomes
Method: Maximum-likelihood, ML (*Kimura 2-parameter*)
Model: Minimum evolution principle
Additive tree with *Daphnia magna* selected as an **outgroup**
Distance in terms of mutation events per site is shown & bootstrap support values are reported

Deji *et al.* (2021) *Mitochondrial DNA Part B* **6**:3255

IQ-TREE

Many software and servers are available for performing phylogenetic analyses

One of them is *IQ-TREE: fast and accurate phylogenetic trees under maximum likelihood*, available at

http://iqtree.cibiv.univie.ac.at/

It is a time-efficient and accurate implementation of the maximum likelihood (ML) method

It takes as an input a sequence alignment and supports any type of input data; by default it uses a bootstrap analysis

Nguyen *et al.*, *Mol Biol Evol* **32**:268

Lesson 10. Content

1. Phylogenetic analyses. Provide hypotheses on the evolutionary history and relationships between species or organisms. Can use different methods and evolutionary models. Are represented by different types of phylogenetic trees, which may be compared based on the number of common splits; their robustness can be tested by bootstrap analysis.