Phylogenetic Analysis

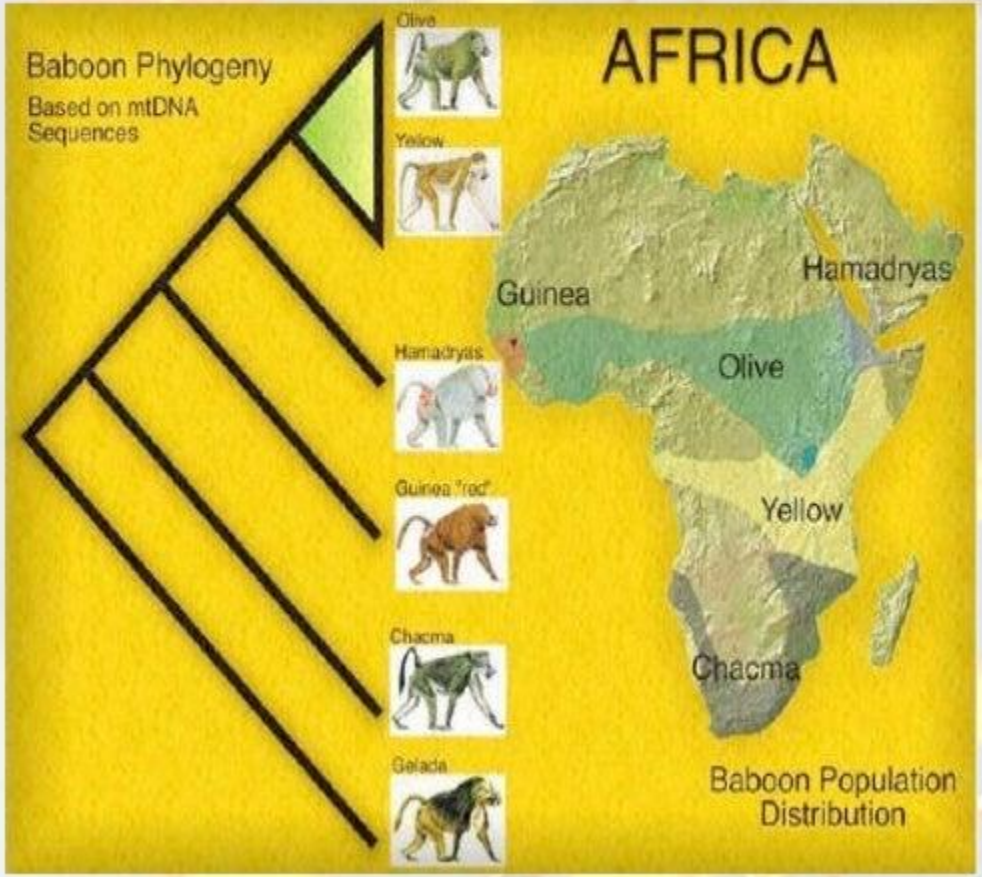
Definition

1. Greek origin➔ Phylon=tribe, Genetikos=relative to birth

2. The evolutionary development and history of a species or higher taxonomic grouping of organisms. (Speciation).

3. The evolutionary development of an organ or other part of an organism

4. The historical development of a tribe or racial group.



Macro- and Micro-Evolution

- Micro-evolution:

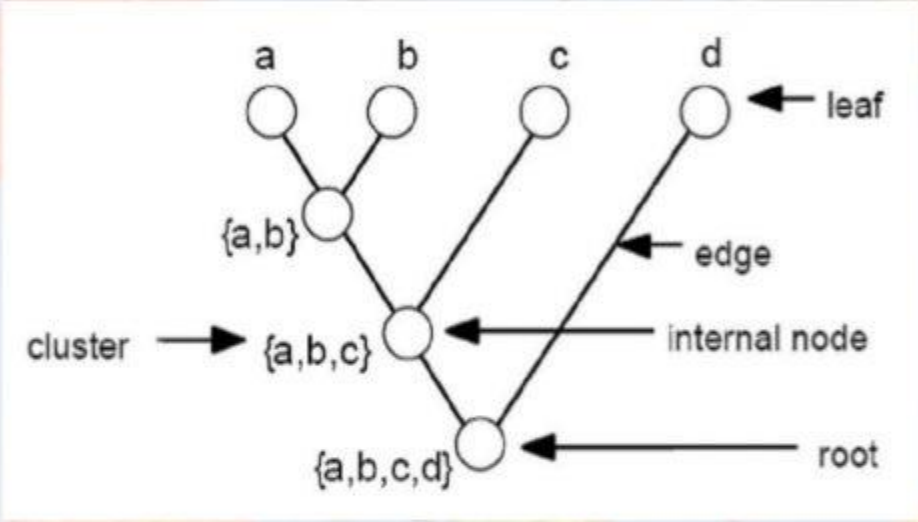
Change in gene frequency within a population- smaller time scale, smaller spatial scale.

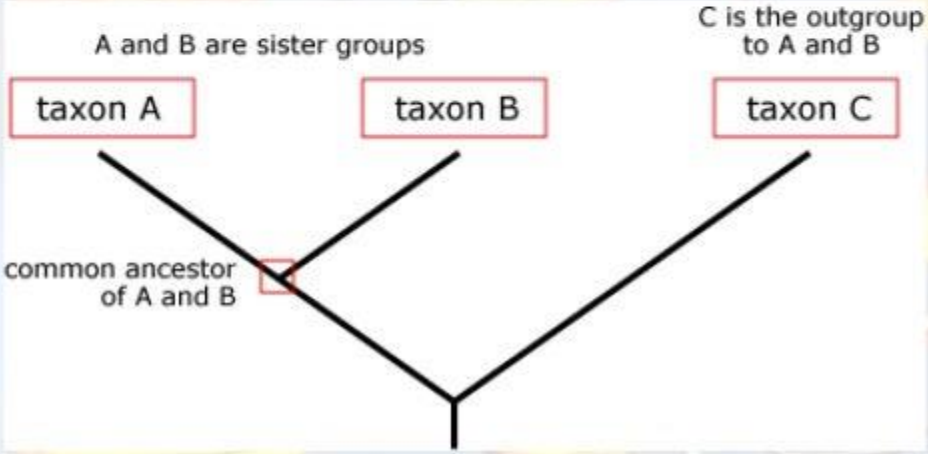
- Macro-evolution:

Origins of new taxonomic groups (taxa)-larger time scale, larger spatial time.

Parts of the tree

- Leaves (tips) that act the recent species,
  - The leaves that split from the same node are called sister groups.
- Nodes that represent the most recent species for the common ancestores
- Branches (edges) that represent the time from one speciation to another,
- Root that act the oldest common ancestor.

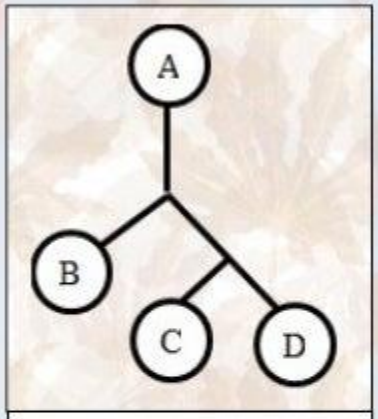




Types of Phylogeny

- There are tow types of the Phylogenetic tree: rooted tree and the un-rooted tree.

- The type depends on the existence of the common ancestor (i.e. there are sufficient information to define the common ancestor).



Rooted

Unrooted

Phylogenetic Analysis

1. Morphological features (Classic):
   Number of legs, lengths of legs, etc

2. Molecular features (Modern):
   Gene sequences
   Protein sequences

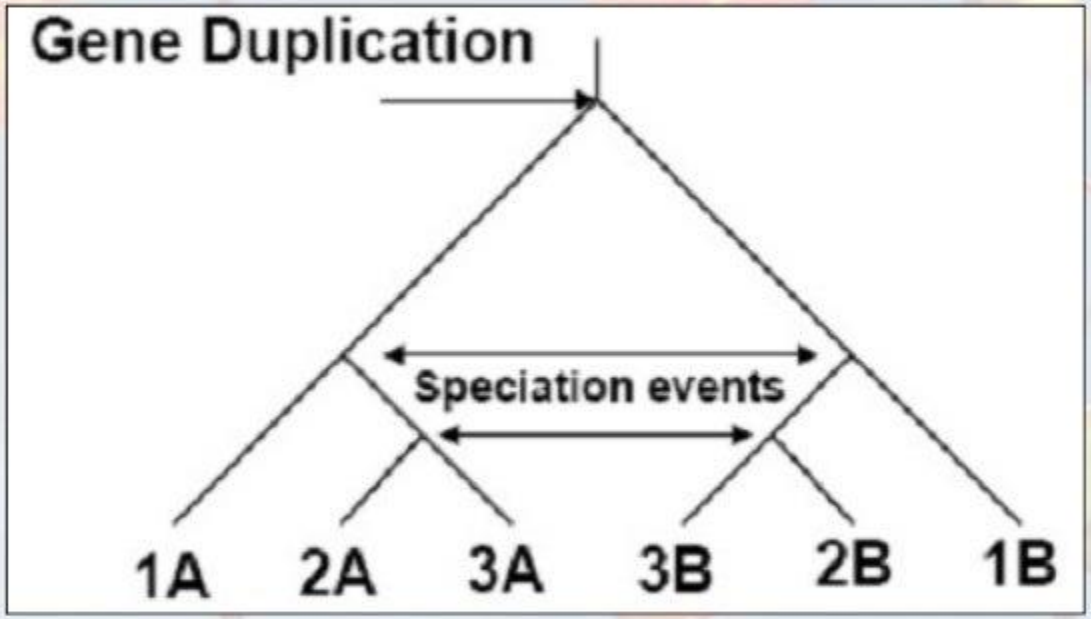
3 Ways to Use Your Tree

- Finding the closest relative of your organism
  - Usually done with a tree based on the ribosomal RNA
- Discovering the function of a gene
  - Finding the orthologues of your gene
- Finding the origin of your gene
  - Finding whether your gene comes from another species

- Different genes evolve at different speeds, and there is always the possibility of horizontal gene transfere
  - hibridization,
  - vector-mediated DNA movement,
  - or direct uptake of DNA

Convergence and Divergence

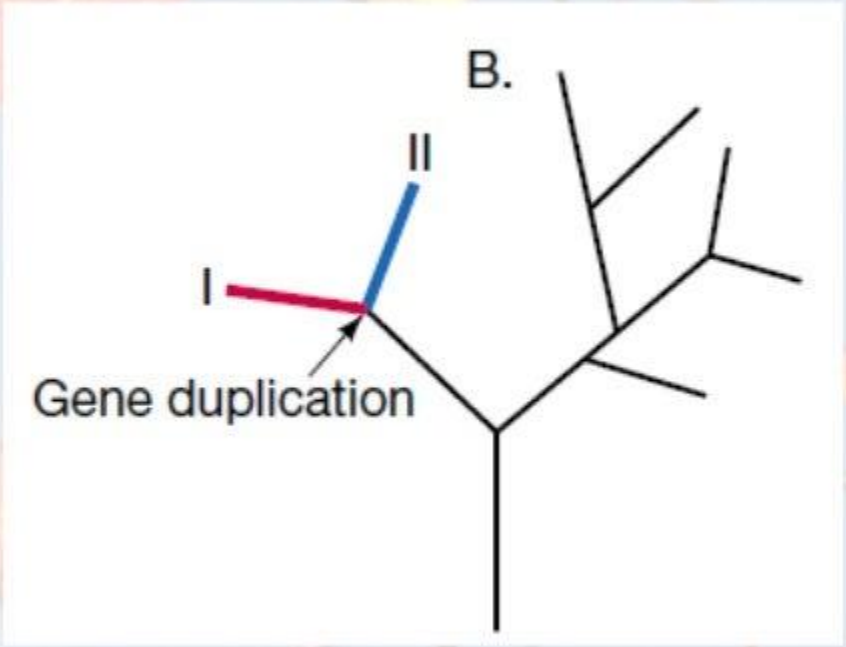
- **Homologs:** Sequences have different functions and with divergent evolution.
  - **Orthologs:** Sequences diverged after speciation event.
  - **Paralogs:** Sequences diverged after a duplication event.
  - **Xenologs:** Sequences diverged after a horizontal transfer (e.g. by virus)
- **Analogs:** Sequences have same function but with different history-convergent evolution.



There are two events cause the divergence.

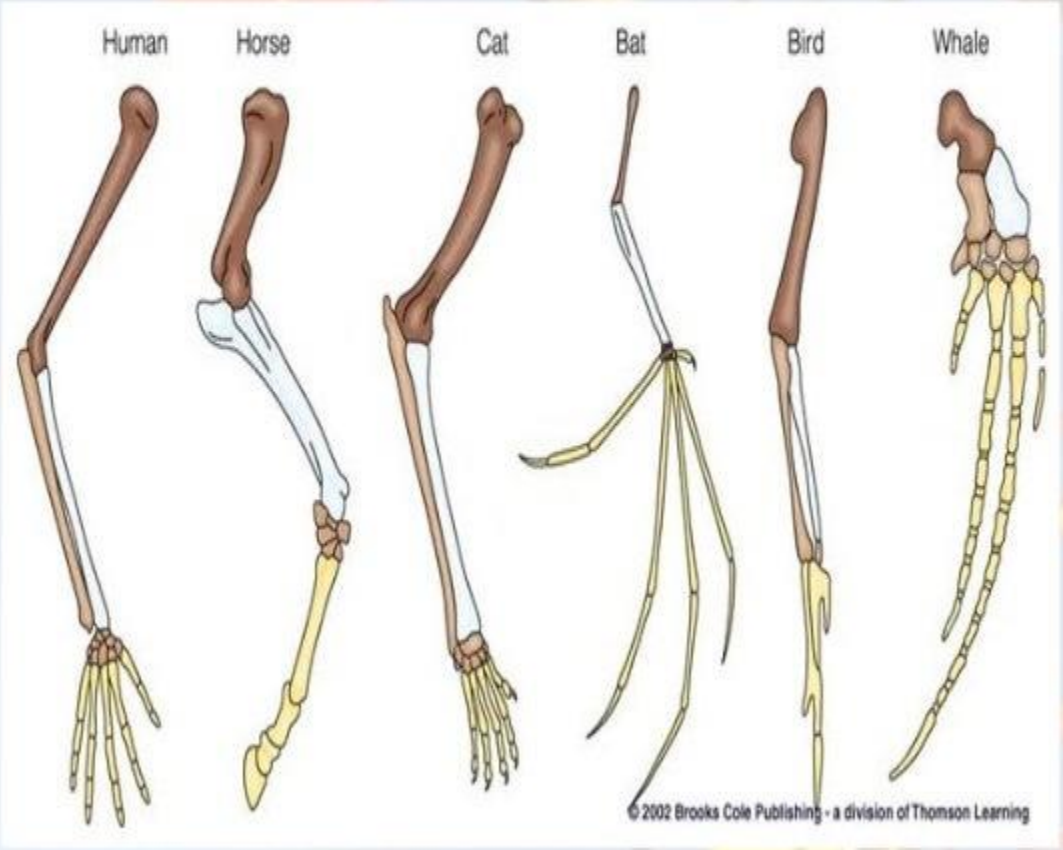
Gene duplication

- Separation of the duplicated region by speciation gives rise to two separate branches, shown as blue and red.



Homologs

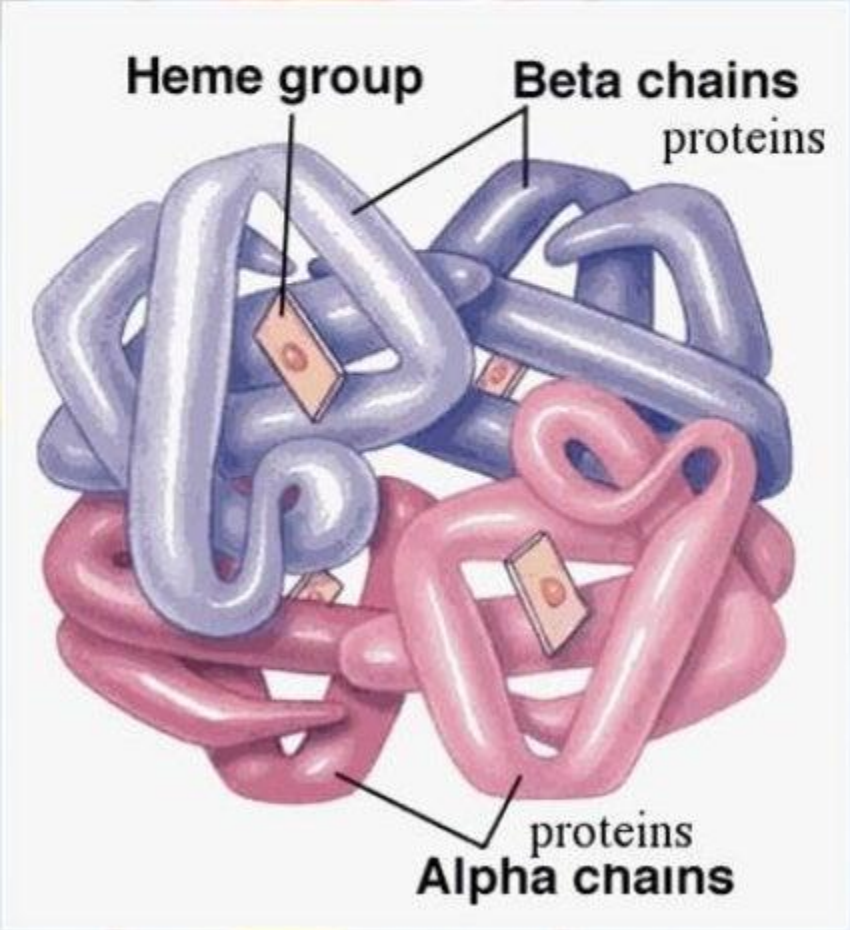
- Sequences have different functions and with divergent evolution.
- Ex. Vertibrate limbs; same bones, different purposes.

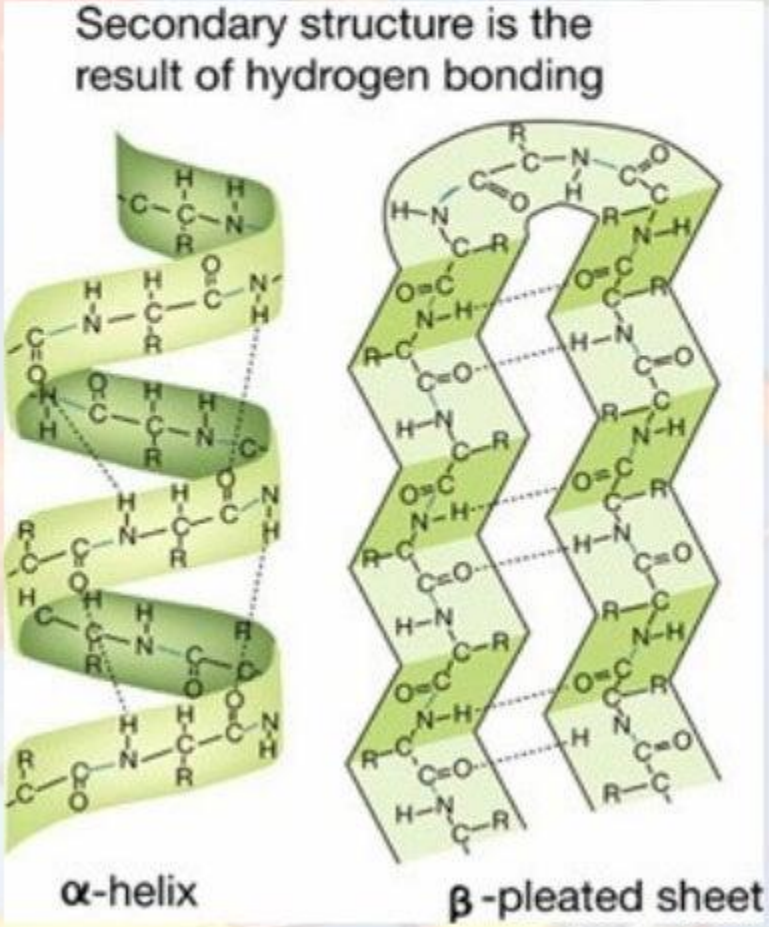


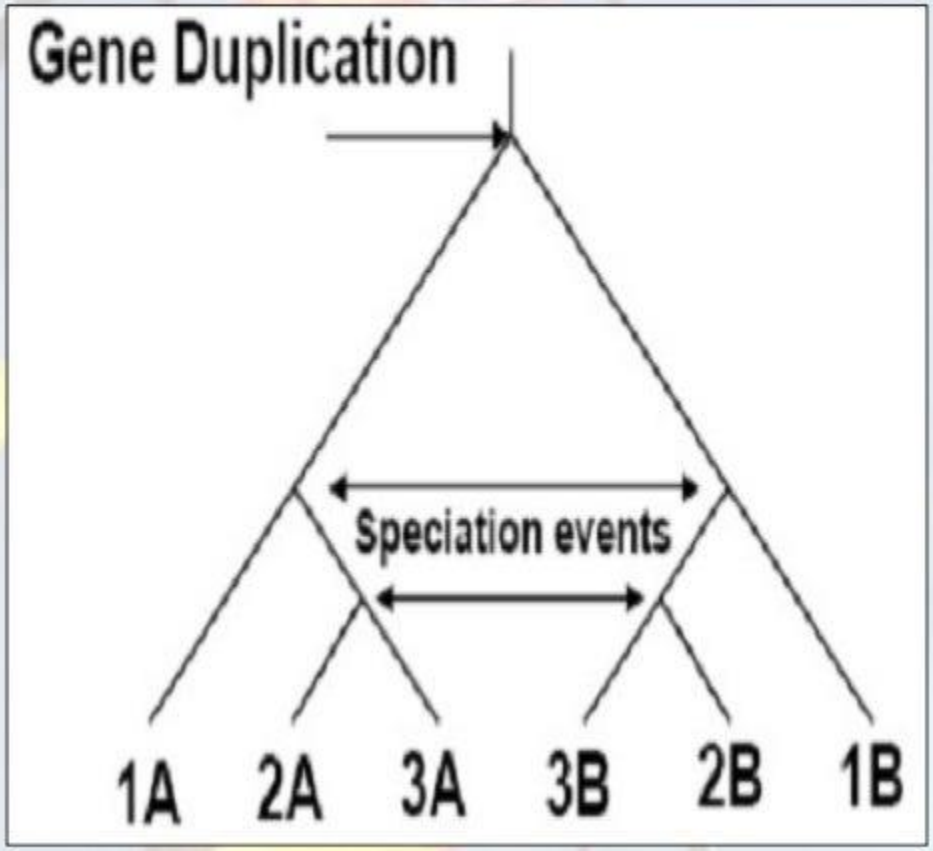
Orthologs & Paralogs

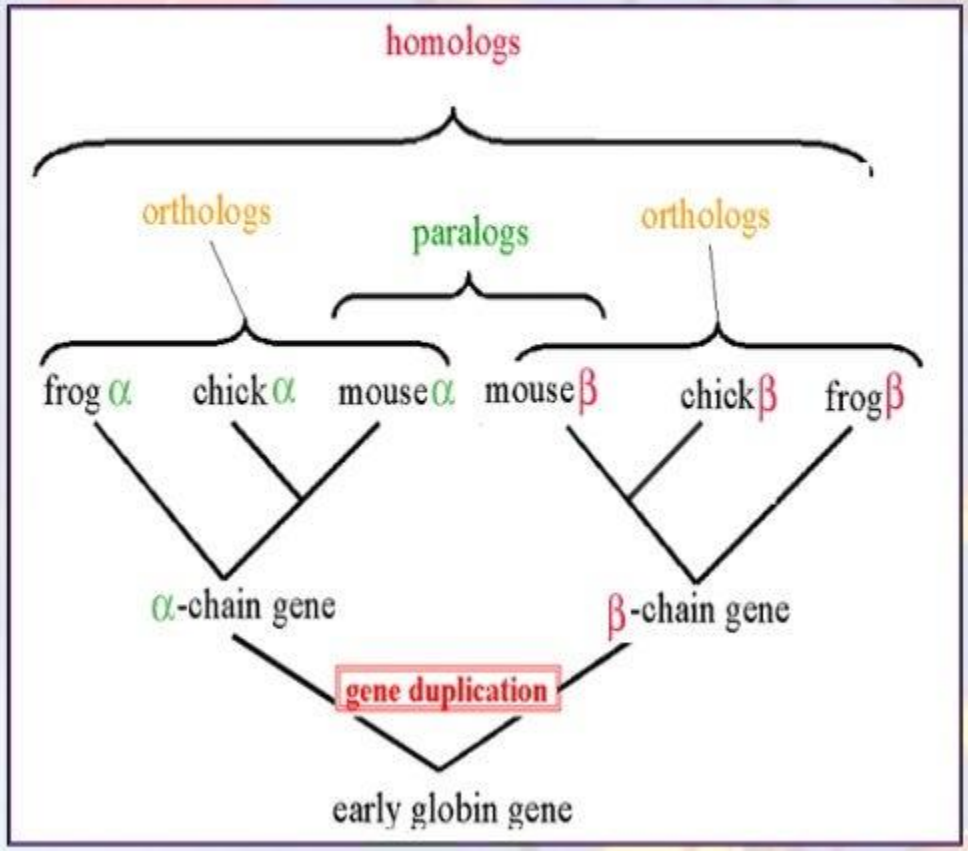
- **Orthologs:** Sequences diverged after speciation event.
- **Paralogs:** Sequences diverged after a duplication event.

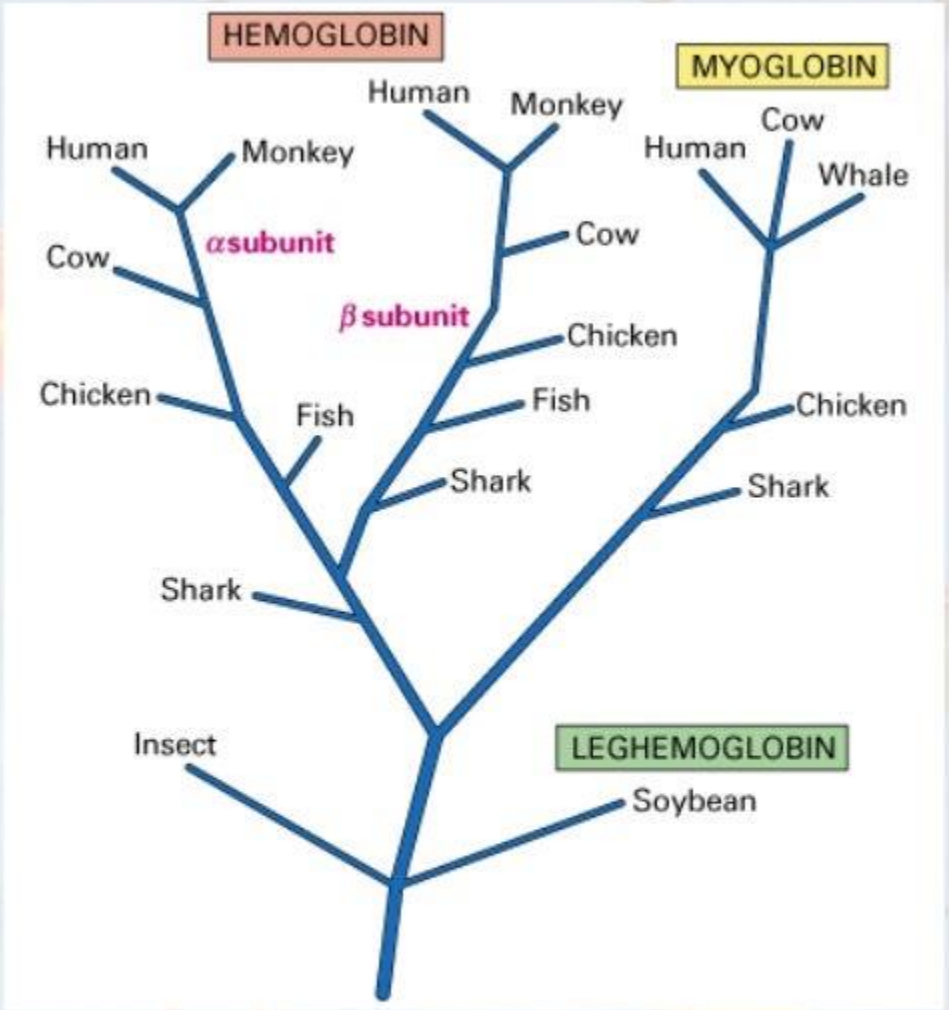
Globin protein



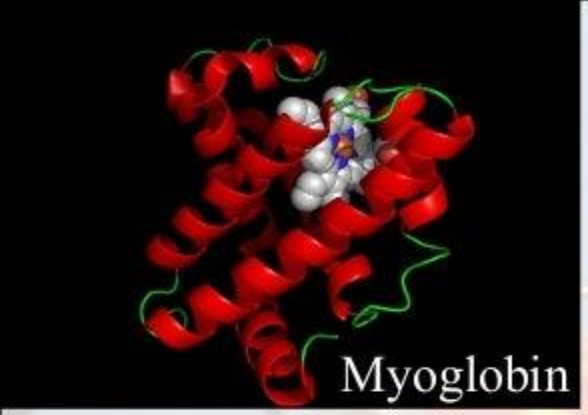



Gene Duplication
Speciation events
1A
2A
3A
3B
2B
1B

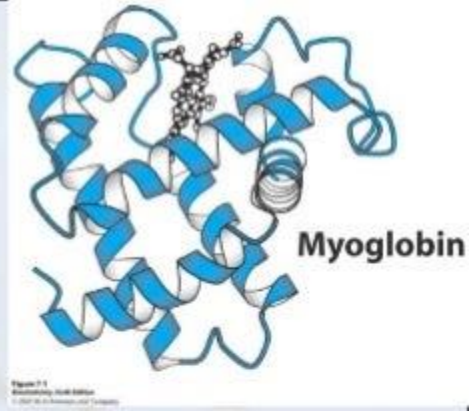

homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene

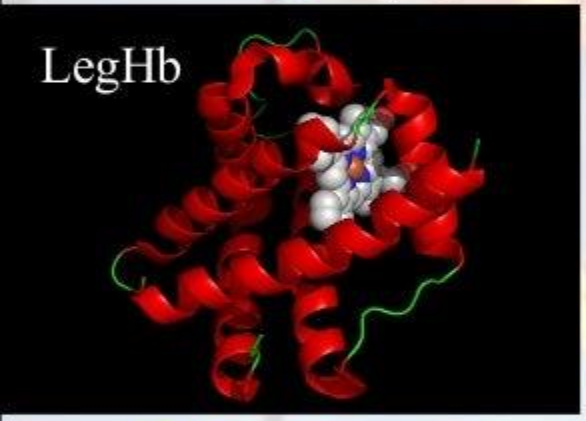


Globin protien evolution

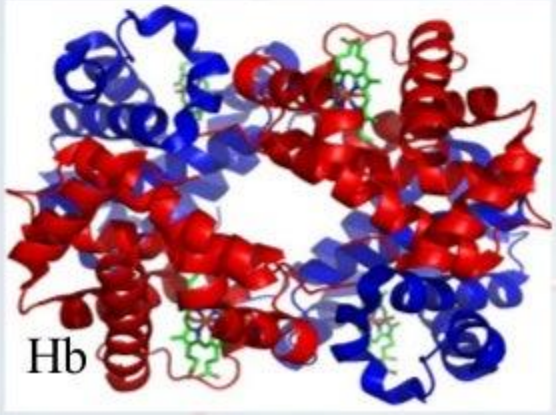


Myoglobin is monomeric protein, consists of a single protein chain with 153 amino acids and one heme group that stores oxygen in the muscle cells.

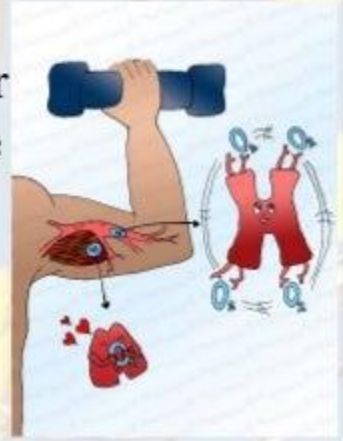




Leghemoglobin (also leghaemoglobin or legoglobin) is a monomeric, nitrogen or oxygen carrier, because naturally occurring oxygen and nitrogen interact similarly with this protein; and ahemoprotein found in the nitrogen-fixing root nodules of leguminous plants



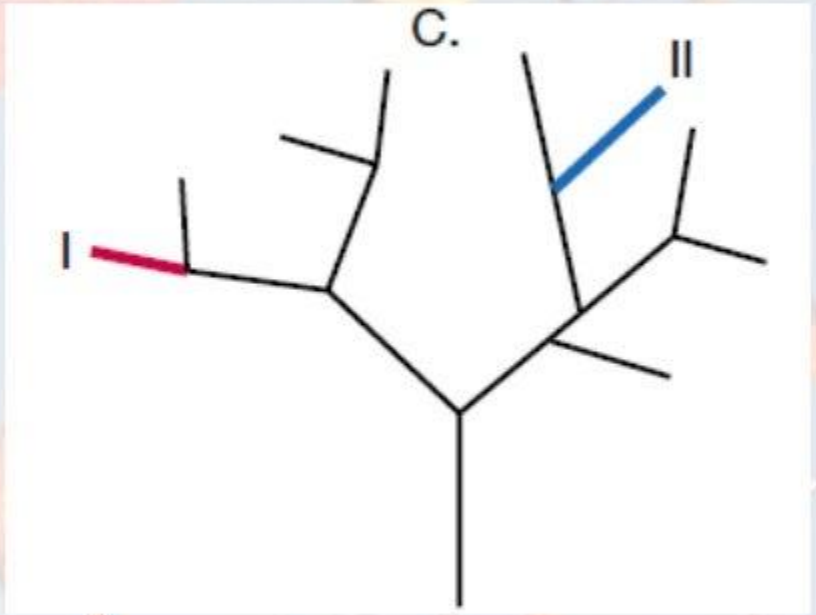
Hemoglobin consists of four protein chains and four heme groups that carry oxygen from the lungs to the tissue cells

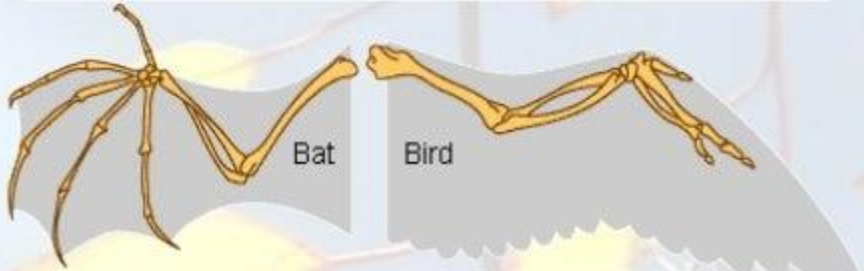


Analogs

- A gene in species I and a different gene in species II have converged on the same function by separate evolutionary paths.

- Such **analogous** genes, or genes that result from convergent evolution, include proteins that have a similar active site but within a different backbone sequence.





.An example of an analogous structure in two distantly related plants

Homoplasy

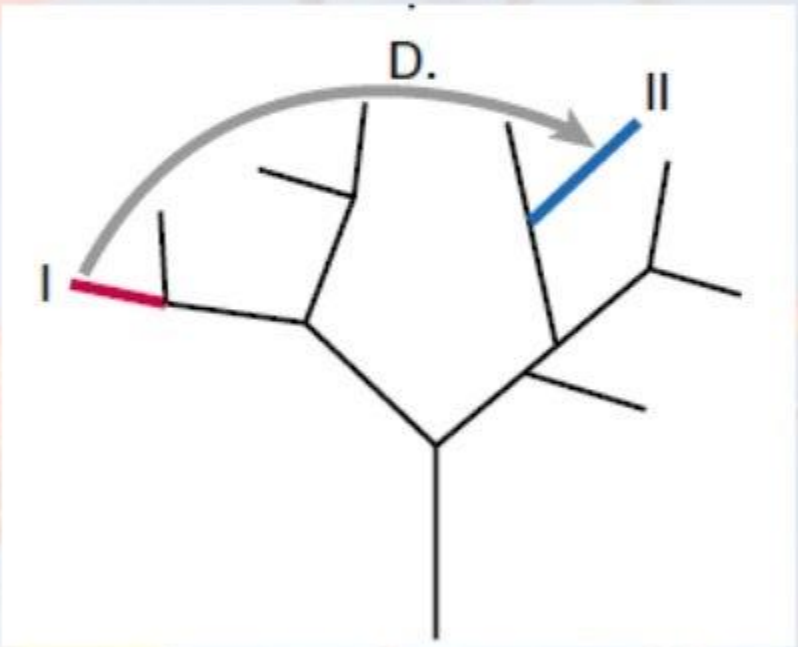
- It occurs when characters are similar, but are not derived from a common ancestor.
- often results from convergent evolution.

Xenologs

- Genes in species I and II are related through the transfer of genetic material between species, even though the two species are separated by a long evolutionary distance.

- Although the transfer is shown between outer branches of the evolutionary tree, it could also have occurred in lower-down branches, thus giving rise to a group of organisms with the transferred gene.

- Such genes are known as **xenologous** or horizontally transferred genes.



Which Sequences ?

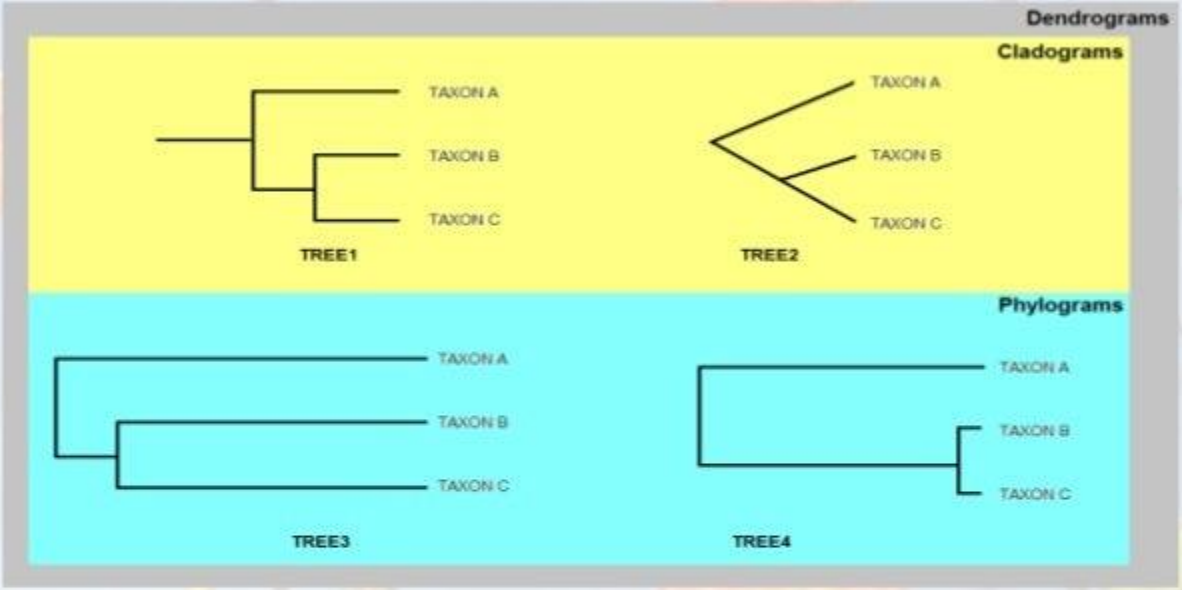
- Orthologous sequences
  - Produce a species tree
  - Show how the considered species have diverged

- Paralogous sequences
  - Produce a gene tree
  - Show the evolution of a protein family

Phylogeny, evolutionary tree, phylogenetic tree, dendogram and cladogram

- For general purpose they are the same
- But in specific analysis:
  - Cladogram emphasize the diagram represents a hypothesis about the actual evolutionary history of a group, or the length of the branches in the diagram are arbitrary.
  - Phylogenies represent true evolutionary history, or the branch lengths that indicate the amount of character change.
  - Phylogram interested in the changes of branches (edges) length.
  - Dendogram consider all the terminal nodes are equidistant from the root, and interested in molecular clock. It includes Cladogram and phylogram.

Dendrogram, cladogram, phylogram



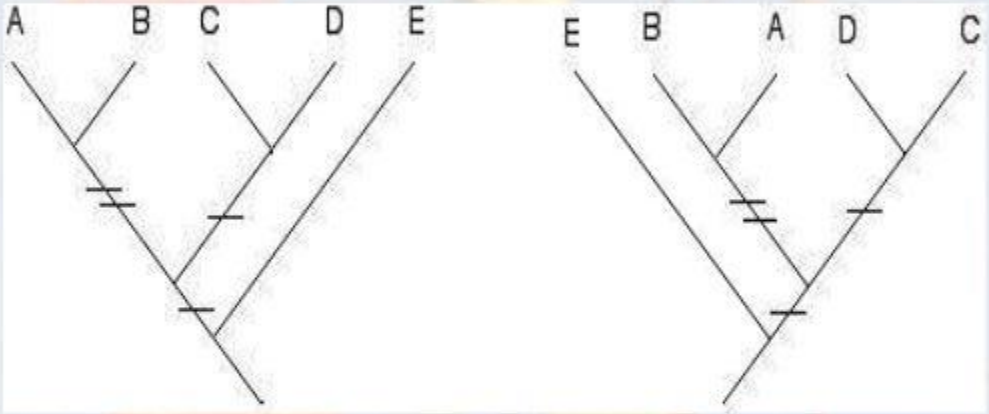
- **Dendrogram** is the 'generic' term applied to any type of diagrammatic representation of phylogenetic trees. All four trees depicted here are dendrograms.
- **Cladogram** (to some biologists) is a tree in which branch lengths DO NOT represent evolutionary time; clades just represent a hypothesis about actual evolutionary history

  TREE1 and TREE2 are cladograms and TREE1 = TREE2
- **Phylogram** (to some biologists) is a tree in which branch lengths DO represent evolutionary time; clades represent true evolutionary history (amount of character change) TREE3 and TREE4 are phylograms and TREE3 ≠ TREE4

Cladogram

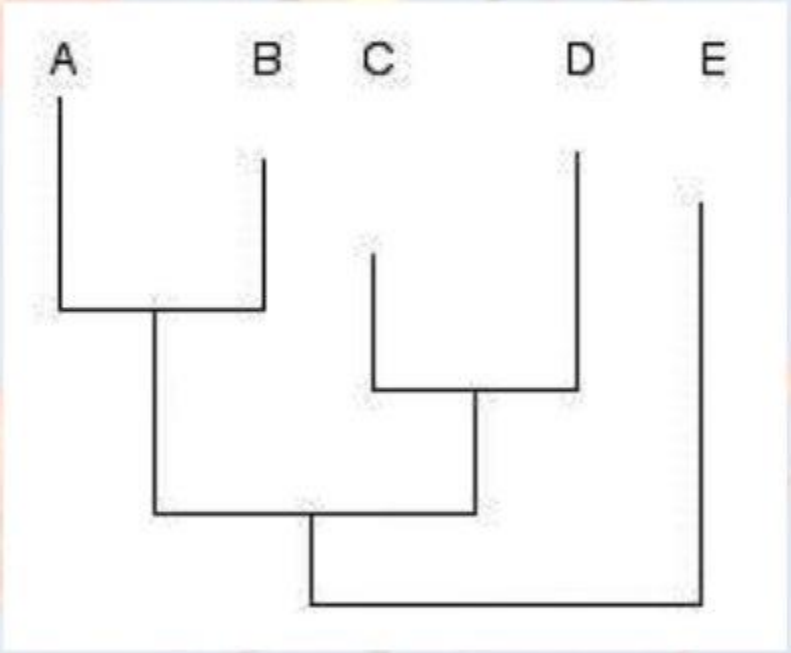
- A cladogram is simple tree depicting only relationships between terminal nodes.

- It can also show inferred character changes and is thus a diagram of *synapomorphies* (a character or trait that is shared by two or more taxonomic groups and is derived through evolution from a common ancestral form).



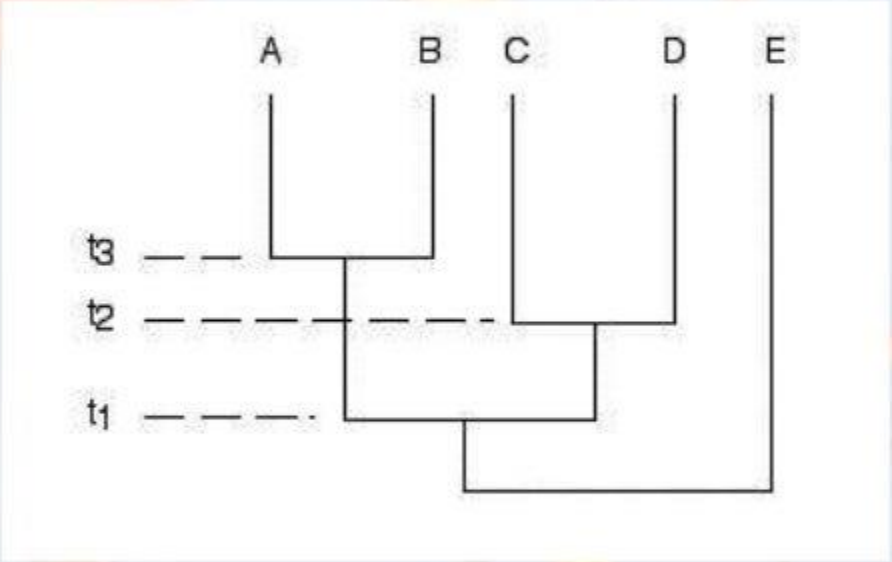
Phylogram

- An *additive tree* has additional information in that edge lengths are drawn proportional to some attribute such as amount of change.



Dendrogram

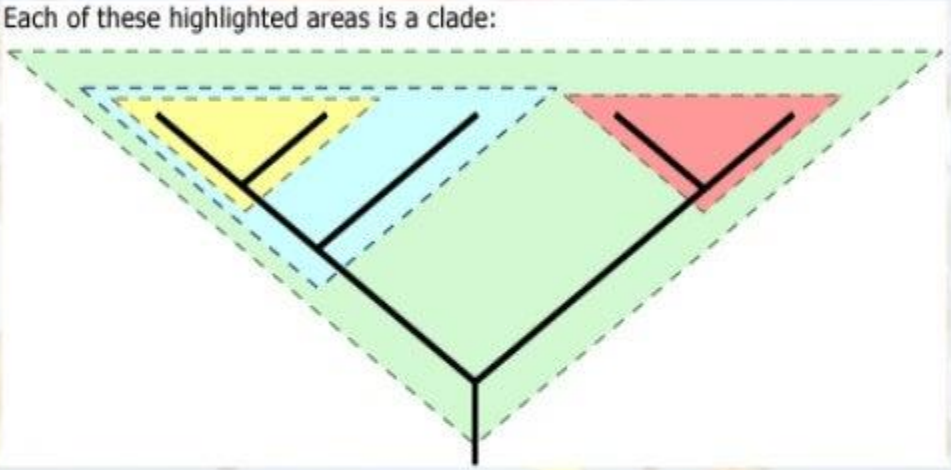
- An *ultrametric tree* is a special kind of additive tree where all pendant vertices (the "tips" or terminal nodes) are equidistant from the root.
- Ultrametric trees can thus depict evolutionary time (directly or as divergence with a molecular clock (rates of molecular change to speciation)).

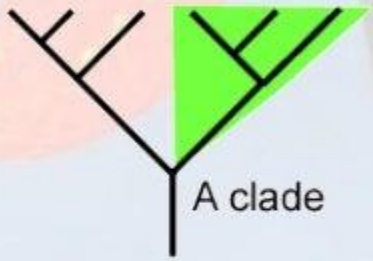


Phenetics versus Cladistics

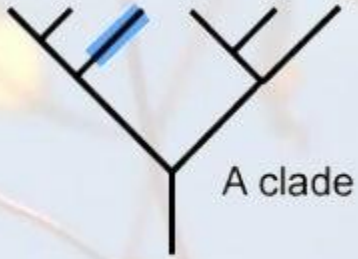
- **Cladistics:**
  - It is the study of the pathways of evolution
  - It interested in number of branches there are among a group of organisms; the connection between branches; and the branching sequence.
  - A tree-like network that expresses such ancestor-descendant relationships
  - Cladogram refers to the topology of a rooted phylogenetic tree.

Each of these highlighted areas is a clade:





A clade



A clade

Not a clade

Not a clade

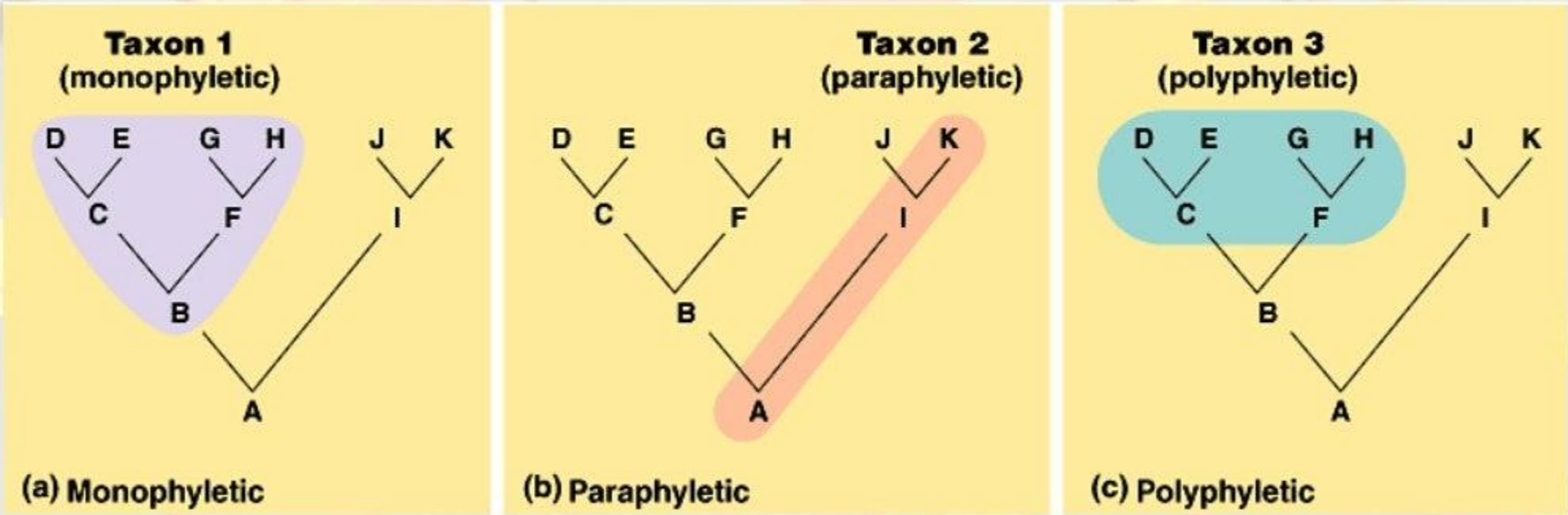
What is the clade? Which of them represents a clade?

Types of clades

- Monophyletic pertains to a taxon that is derived from a single ancestral species.

- Polyphyletic pertains to a taxon whose members were derived from two or more ancestors not common to all members.

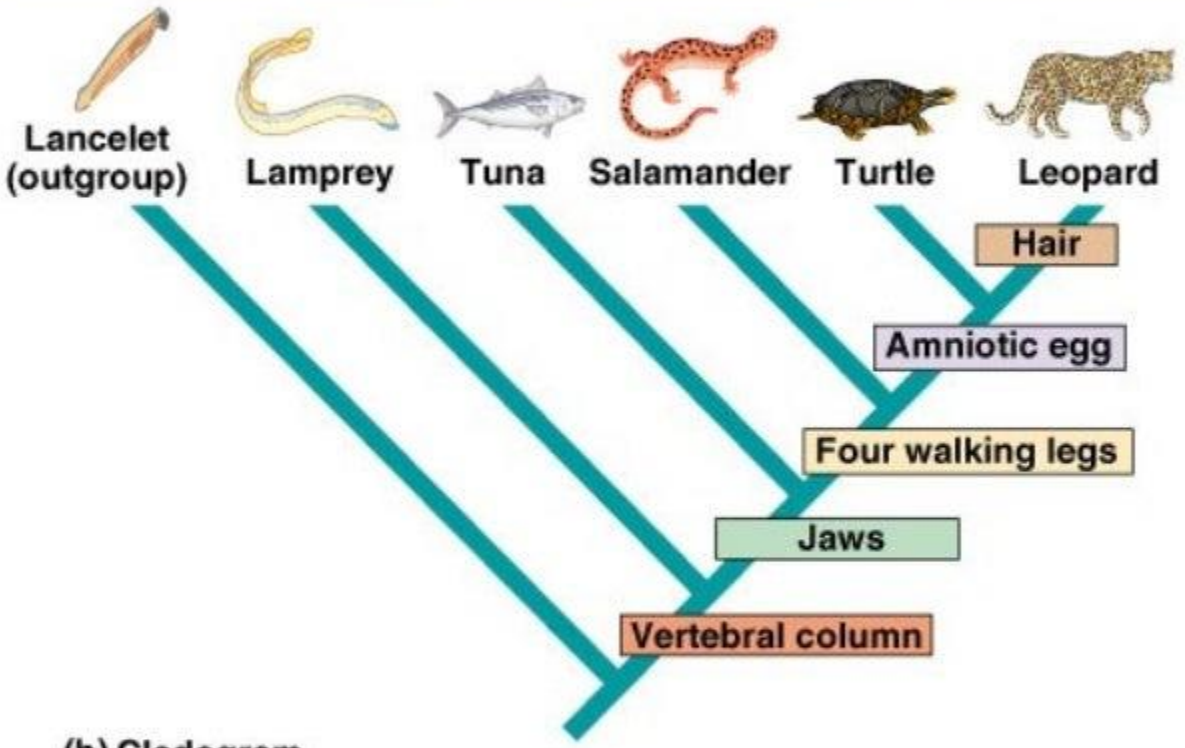
- Paraphyletic pertains to a taxon that excludes some members that share a common ancestor with members included in the taxon. "Foxes" are paraphyletic with respect to dogs, wolves, jackals, coyotes, etc.


Taxon 1
(monophyletic)
D E G H J K
C F I
B
A
(a) Monophyletic
Taxon 2
(paraphyletic)
D E G H J K
C F I
B
A
(b) Paraphyletic
Taxon 3
(polyphyletic)
D E G H J K
C F I
B
A
(c) Polyphyletic

TAXA

CHARACTERS	Lancelet (outgroup)	Lamprey	Tuna	Salamander	Turtle	Leopard
Hair	0	0	0	0	0	1
Amniotic (shelled) egg	0	0	0	0	1	1
Four walking legs	0	0	0	1	1	1
Jaws	0	0	1	1	1	1
Vertebral column (backbone)	0	1	1	1	1	1

(a) Character table



(b) Cladogram

Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

- **Phenetics:**
  - It is the study of relationships among a group of organisms on the basis of the degree of similarity between them (molecular, phenotypic, or anatomical).
  - Phenogram refers to a tree-like network expressing phenetic relationships.

- Branch lengths are only meaningful when the tree is scaled
  - **Cladograms** are often scaled
  - **Phenograms** are usualy unscaled

- Maximum Parsimony considered as typical representative of Cladogram approach.

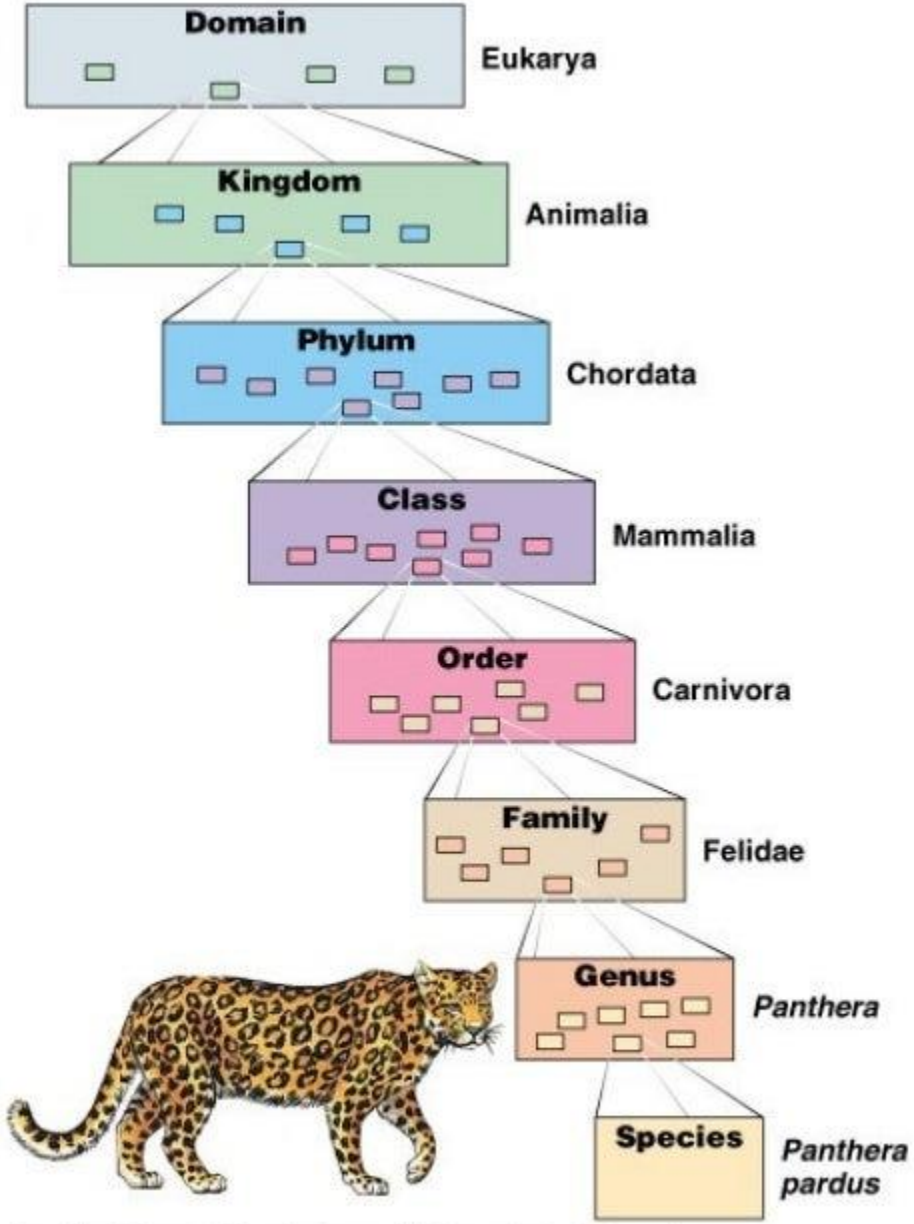
- UPGMA and Neighbor joining considered as typical representative of Phenogram approach.

Systematics: Connecting classification to phylogeny

- **Systematics**: the study of biological diversity in an evolutionary context, including taxonomy and phylogenetics.

- Taxonomy uses a hierarchical classification system
- The hierarchical classfication: Kingdom, Phylum, Class, Order, Family, Genus, Specie
- A named taxonomic unit at any level is called a taxon.

- Phylogenetic trees are used to place different taxonomic schemes together, and to show connection between classification and phylogeny.
- Modern phylogenetic systematics are based on cladistic analysis



Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Methodologies for Tree Construction

1) Selection of sequences/traits used for comparison

The importance of this step is highlighted by the earlier discussion on homology and homoplasy.

2) Multiple Sequence Alignment

3) Tree Building

The tree is constructed using calculated distances using; maximum parsimony, maximum likelihood, and distance methods.

4) Tree Evaluation

The optimal tree is selected using three different approaches will be evaluate using bootstrap or Jacknife.

DNA or Proteins

- Most phylogenetic methods work on **Proteins** and **DNA** sequences
- If your DNA sequences are coding and have more than 70% identity . . .
  - Compute the **tree** on the **DNA** multiple-sequence alignment
- If your DNA sequences are coding and have less than 70% identity . . .
  - Compute the **tree** on the **protein** multiple-sequence alignment

Creating the Perfect Dataset

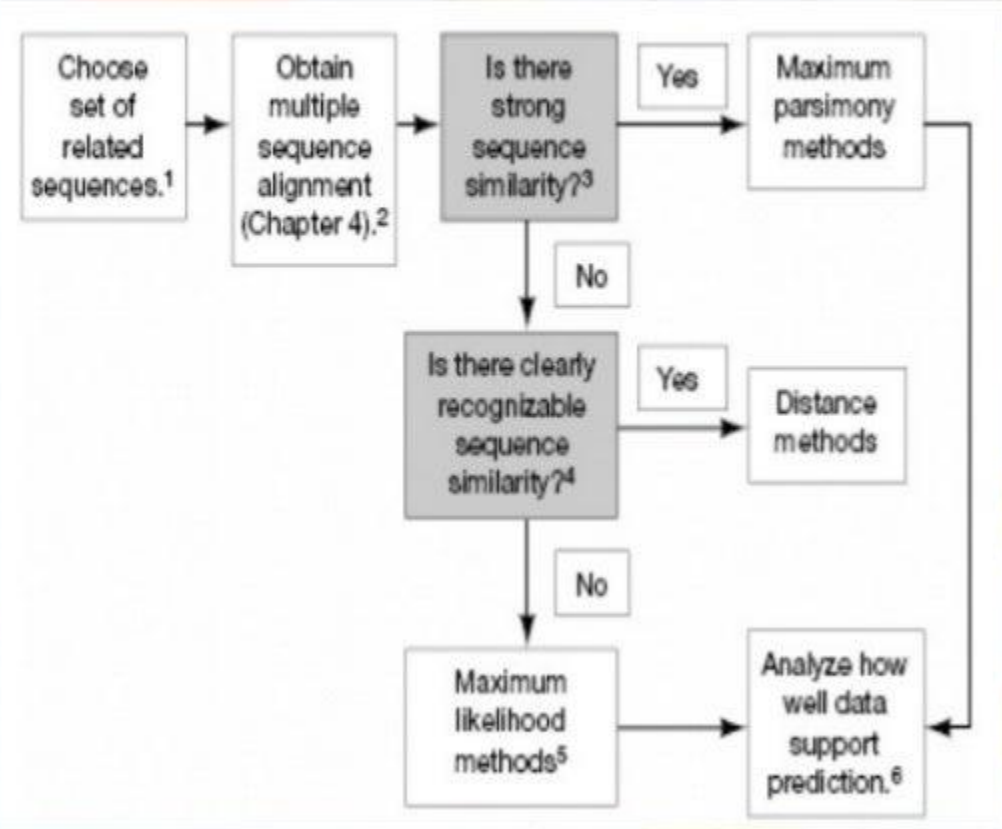
Problem	*Reason and Solution*
Avoid sequence fragments	Sequence fragments produce low-quality MSAs. Use the same fragments for all the sequences.
Avoid xenologs	Avoid genes resulting from a horizontal gene transfer (HGT).
Avoid recombinant sequences	Recombinant Sequences have two ancestors. They confuse the tree reconstruction.
Avoid large complex families	Avoid proteins that have many paralogous in each genome. It is hard to find orthologous in large families. Avoid multi-domain proteins. Work on one domain at the same time.
Try to make a small set	Large datasets are difficult to align.
Add an outgroup to your dataset	An outgroup is an organism whose last common ancestor with the dataset is older than the common ancestor of this dataset. For instance, chicken is an outgroup for man, dog and horse.

Building the Right MSA

- Your MSA should have as few gaps as possible.
- Some variability but not too much!
- Some conservation but not too much!

Phylogenetic tree approaches

- Three general types of methods:
- Distance: find tree that accounts for estimated evolutionary distances
- Maximum parsimony: find the tree that requires minimum number of changes to explain the data
- Maximum likelihood: find the tree that maximizes the likelihood of the data



Building the Right Tree

- There are two types of tree-reconstruction methods
  - Distance-based methods
  - Statistical methods

- Statistical methods are the most accurate
  - Maximum likelihood of success
  - Parsimony

- Statistical methods take more time
  - Limited to small datasets

Distance-based methods for tree reconstruction

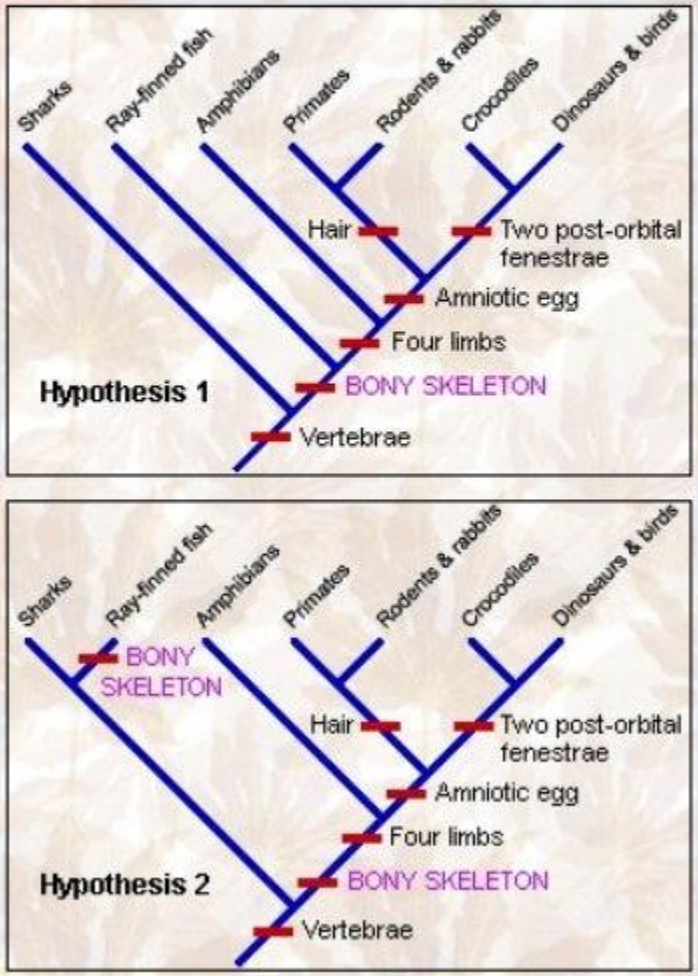
- Distance-based methods are the most popular
  - Neighbor Joining (NJ)
  - UPGMA

- Distance-based methods involve 2 steps:
  - Measure the distances between pairs of sequences in the MSA
  - Transform the distance matrix into a tree

Maximum Parsimony Method

- It is a nonparametric method.
- It searches for the tree with the least number of mutations along its branches needed to explain the data.

- Hypothesis 1 requires six evolutionary changes and Hypothesis 2 requires seven evolutionary changes, with a bony skeleton evolving independently, twice.



The Distance Method

- It is a semiparametric method.
- It uses the evolutionary model to estimate distances between sequences and then use methodology akin to hierarchical clustering to build the tree.
- It employs the number of changes between each pair in a group of sequences.
- The sequence pairs that have the smallest number of sequence changes between them are termed "**neighbors**."

A. Sequences

sequence A	ACGCGTTGGGCGATGGCAAC
sequence B	ACGCGTTGGGCGACGGTAAT
sequence C	ACGCATTGAATGATGATAAT
sequence D	ACACATTGAGTGATAATAAT

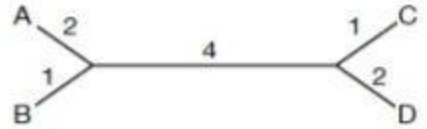
B. Distances between sequences, the number of steps required to change one sequence into the other.

n_{AB}	3
n_{AC}	7
n_{AD}	8
n_{BC}	6
n_{BD}	7
n_{CD}	3

C. Distance table

	A	B	C	D
A	–	3	7	8
B	–	–	6	7
C	–	–	–	3
D	–	–	–	–

D. The assumed phylogenetic tree for the sequences A-D showing branch lengths. The sum of the branch lengths between any two sequences on the trees has the same value as the distance between the sequences.



Distance analysis programs

1. FITCH estimates a phylogenetic tree assuming additivity of branch lengths using the Fitch-Margoliash method and does not assume a molecular clock (allows rates of evolution along branches to vary).

2. KITSCH estimates a phylogenetic tree using the Fitch-Margoliash method but under the assumption of a molecular clock.

3. Neighbor-joining method its principle is to find pairs of neighbors that minimize the total branch length at each stage of clustering.

it starting with a starlike tree.

It does not assume a molecular clock

It produces an unrooted tree.

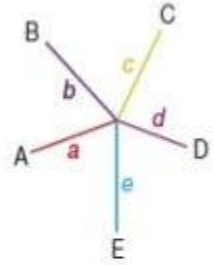


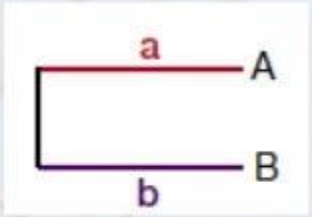
Figure 6.14. Tree for five sequences with no pairing of sequences. In the neighbor-joining method, the sum of the branch lengths $S_0 = a + b + c + d + e$ is calculated. The known distances from (1) A to B, $D_{AB} = a + b$; (2) A to C = $D_{AC} = a + c$; (3) B to C = $D_{BC} = b + c$; and finally (4) D to E, $D_{DE} = d + e$ for a total of 4 + 3 + 2 + 1 = 10 combinations. In summing the 10 distances = 22 + 39 + . . . + 10 = 314, each branch a, b, c, etc., is counted four times. Hence, the sum of branch lengths is 314/4 = 78.5. In general, for N sequences, $S_0 = \Sigma D_{ij}/(N - 1)$, where D_{ij} represents the distances between sequences i and j, $i < j$.

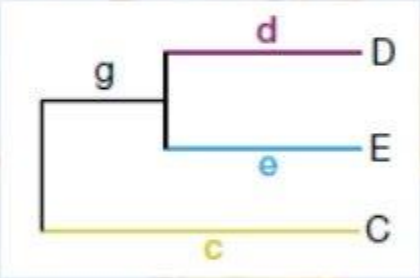
4. UPGMA a simple method assumes the rate of change along the branches of the tree is a constant and the distances are approximately ultrametric.

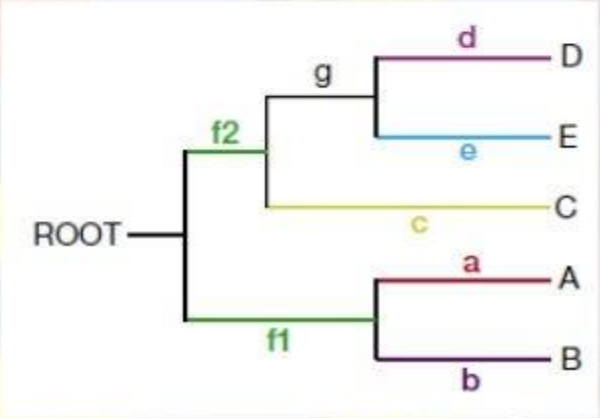
   It assumes a molecular clock and produces a rooted tree.

The Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

- It assumes the rate of change along the branches of the tree is a constant and the distances are approximately ultrametric.
- It starts by calculating branch lengths between the most closely related sequences, then averages the distance between this pair or sequence cluster and the next sequence or sequence cluster, and continues until all the sequences are included in the tree.
- Finally, the method predicts a position for the root of the tree.







Maximum Likelihood

- It is computationally intense.
- It uses probability calculations to find a tree that best accounts for the variation in a set of sequences.
- It estimate the tree by choosing the tree with the highest probability at a position in the scoring matrix by a combination of mutations and gaps.
- This path provides an indication of the evolutionary distance between the sequences.

- It considered the number and type of sequence changes (mutations; substitutions, insertions, and deletions) that may have occurred to give the sequence variation for each tree.

- The rate of appearance of new mutations is very small, so more mutations needed to fit a tree to the data.

- It has been used for an analysis of mutations in overlapping reading frames in viruses.

Bayesian phylogenetics

- It is similar to ML, but ML shows what the data is telling about the parameters.
- It produces both a tree estimate and measures of uncertainty for the groups on the tree
- Optimal hypothesis is the one that maximizes the posterior probability, = ML x PRIOR PROBABILITY of that hypothesis.
- It uses parametric evolutionary models

- It based on Markov chain Monte Carlo computations

- It uses the method known as MCMCMC ("Metropolis-coupled Markov chain Monte Carlo") to empirically determine the posterior probability distribution of trees, branch lengths and substitution parameters.

- Prior probabilities of different hypotheses shows the scientist's beliefs about the parameters before having seen the data. i.e; accumulate scientific knowledge.

Methods of evaluating trees (Reliability)

- Bootstrap: resample initial data set with one datum removed and replaced with another member (more preferable)

- Jackknife: resample initial distribution with one datum missing and not replaced. The purpose of this is to see if excluding certain characters has a big effect on the shape of the tree

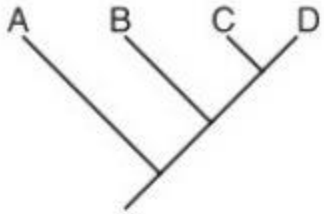
Bootstrapping

- It is used to verify the solidity and reliability of each node
- It involves these steps:
  - Select a subset of your MSA
  - Redo the tree
  - Repeat this operation N times (100 or 1000 times if you can)
  - Compute a consensus tree of the N trees
  - Measure how many of the N trees agree with the consensus tree on each node
- Each node gets a bootstrap figure between 0 and N

- High bootstrap ⇔ good node
- For branches in the predicted tree topology to be significant, the value should be ≥50
- Bootstrap analysis is supported by most of the commonly used phylogenetic inference software packages

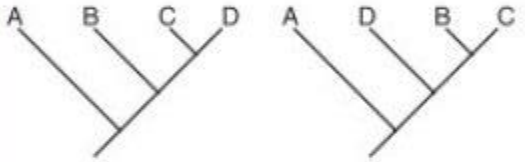
Bootstrapping

Tree from Original Data Set

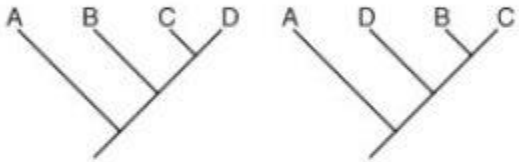


Trees from Bootstrapped Data Sets

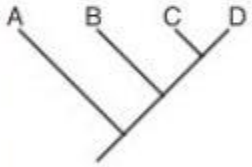
Bootstrap Pseudoreplicate 1:



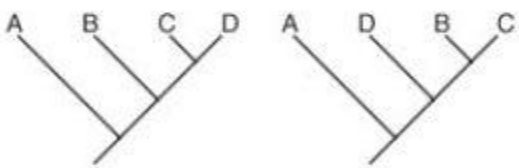
Bootstrap Pseudoreplicate 2:



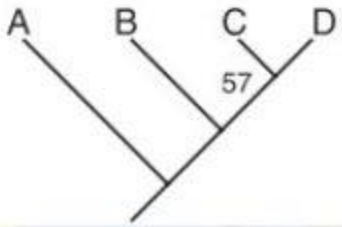
Bootstrap Pseudoreplicate 3:



Bootstrap Pseudoreplicate 4:



Bootstrap Consensus Tree:



A Bootstrapped Tree

- 93 means that the species were siblings in 93% of the bootstrap replications;
- 49 means that the sequences CONS B34, CONS N2, CONS-CPZ and CONS O4 were grouped together in what is called a monophyletic

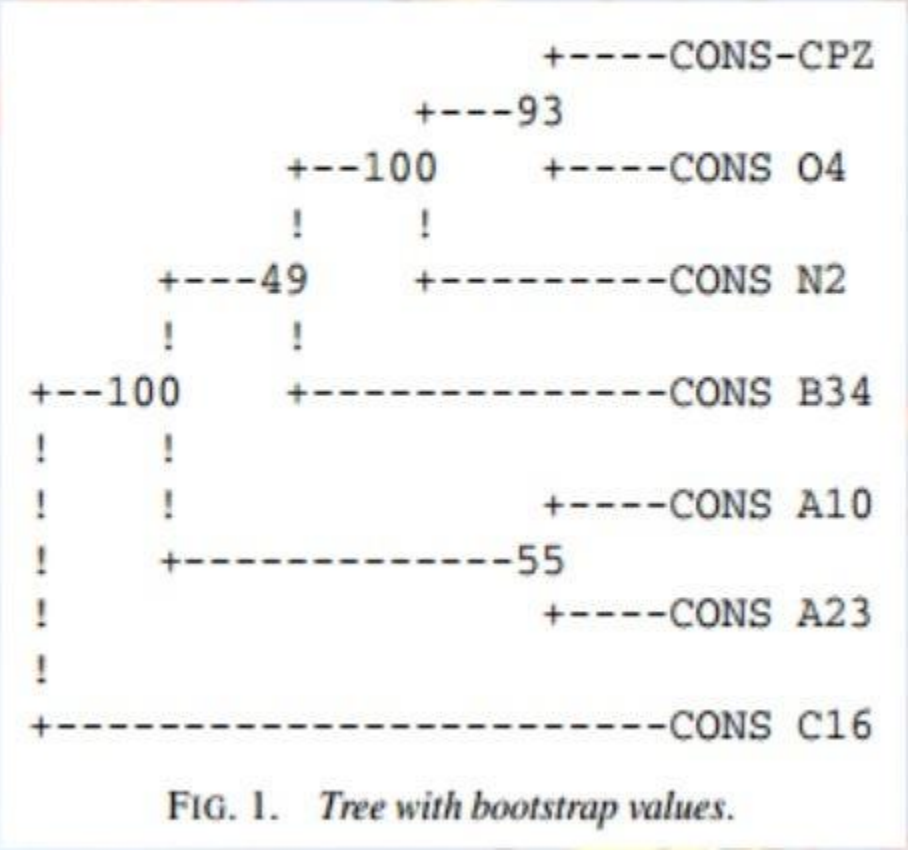


FIG. 1. *Tree with bootstrap values.*

Phylogenetic software

Software packages

- Freely available
  - Phylip
  - BioNJ
  - PhyML
  - Tree Puzzle
  - MrBayes
  - MEGA
- Commercial
  - PAUP

Output format

- Newick format.

  Can be written as ((A, B), (C, D)), E), or if one wishes to add information on branch lengths ((A:0.1, B:0.2):0.2, (C:0.15, D:0.3):0.4):0.3, E:0.4).

- Graphic format.