# Chapter 5: Phylogenetic Analysis

# LECTURE OBJECTIVES…

At the end of this chapter, Students should be able to:

- ✓ Describe what pylogeny is all about
- ✓ Recognise and interprete Pylogenitic trees
- ✓ Construct phylogentic trees using UPGMA and Neigbour joining methods
- ✓ molecular evolution

# What is Molecular Phylogenetics…

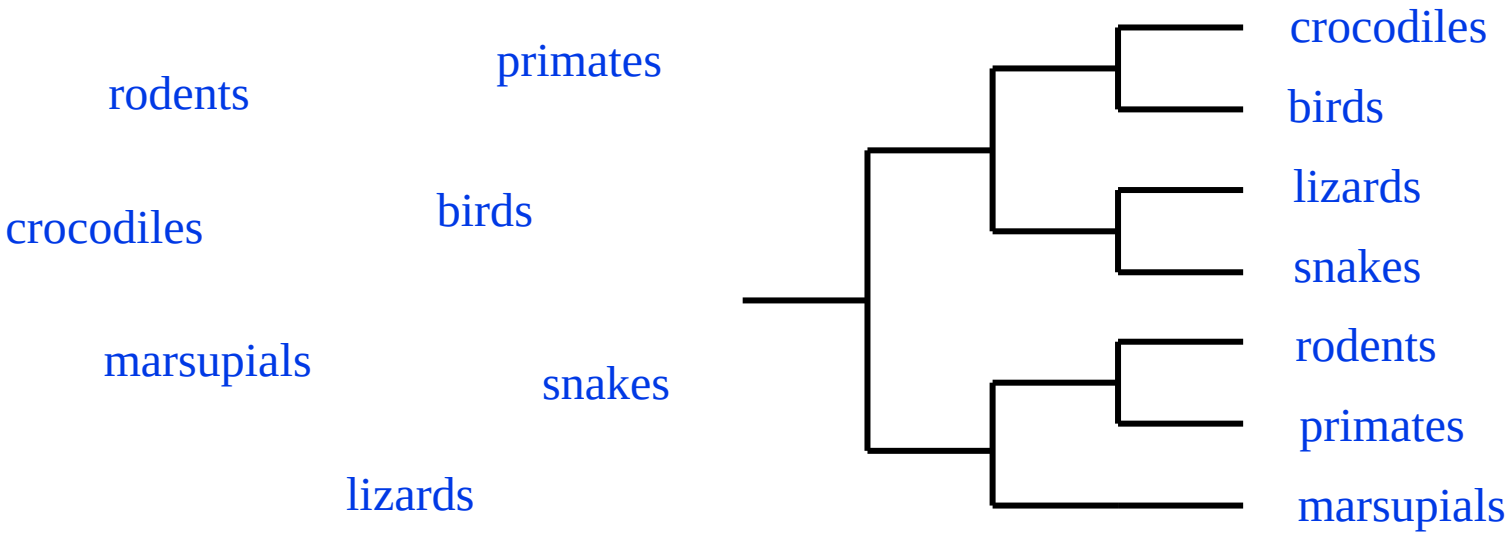
- Phylogenetics
  - the study of evolutionary relationships in organisms,
  - one part of the larger field of systematics, which also includes taxonomy.
    - The term taxonomy connotes the process and methodology for the naming and classification of organisms.
- The systematics
  - the branch of biology that deals with classification and nomenclature; taxonomy

# …What is Molecular Phylogenetics…

- The context of evolutionary biology is phylogeny,
  - the connections between all groups of organisms as understood by ancestor/descendant relationships.
- The molecular mechanisms of organisms studied strongly suggests that all organisms on earth have a common ancestor.
  - Thus, the species are related to each other by the virtue of having evolved from the same common ancestor.
- Such a relationship of species is called phylogeny and it's graphical representation is called a phylogenetic tree.

# …What is Molecular Phylogenetics

- Example:
  - relationship among species



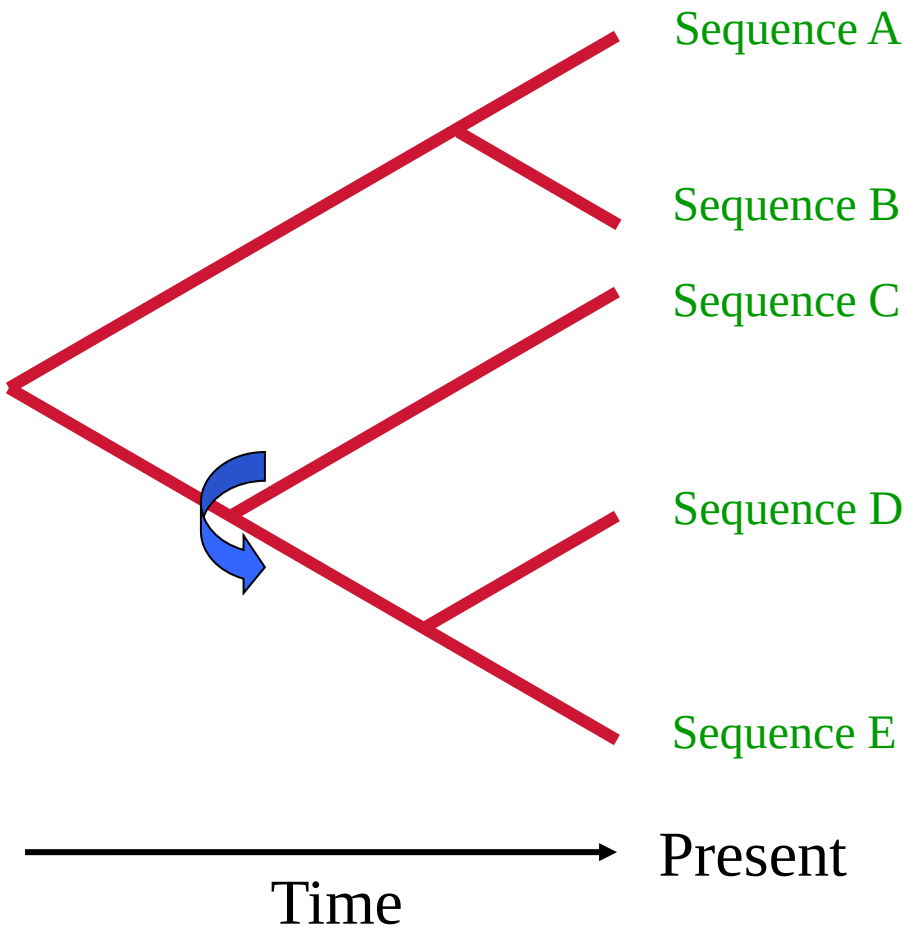
# Etymology

- phylogeny
  - the evolutionary history of a kind of organism
  - the evolution of a genetically related group of organisms
- phylogenetics
  - a branch of science that deals with phylogeny

# A Brief History of Molecular Phylogenetics

- 1900s
  - Immunochemical studies
    - cross-reactions stronger for closely related organisms
      - Nuttall (1902) - apes are closest relatives to humans!
- 1960s - 1970s
  - Protein sequencing methods, electrophoresis, DNA hybridization and Polymerase Chain Reaction (PCR) contributed to a boom in molecular phylogeny
- late 1970s to present
  - Discoveries using molecular phylogeny
    - Endosymbiosis - Margulis, 1978
    - Divergence of phyla and kingdom - Woese, 1987
    - Many Tree of Life projects completed or underway

# Phylogenetic concepts: Interpreting a Phylogeny




- Physical position in tree is not meaningful
- Swiveling can only be done at the nodes
- Only tree structure matters

# Tree Terminology

- Relationships are illustrated by a phylogenetic tree / dendrogram
  - Combination of Greek dendro/tree and gramma/drawing
  - A dendrogram is a tree diagram
    - frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.
  - Dendrograms are often used in computational biology
    - to illustrate the clustering of genes or samples, sometimes on top of heatmaps.

# SOME TERMINOLOGIES USED FOR PHYLOGENETIC TREES…

**Dendrogram** is a broad term used to represent a phylogenetic tree. More precisely, “dendrogram” is a generic term applied to any type of phylogenetic tree (scaled or unscaled).

**Cladogram** is a representation of the ancestor-to-descendant relationship through a branching tree. The length of a branch denotes nothing (neither genetic change nor time-scale). A clade simply means a common ancestor and the descendants (here, species) of that common ancestor.

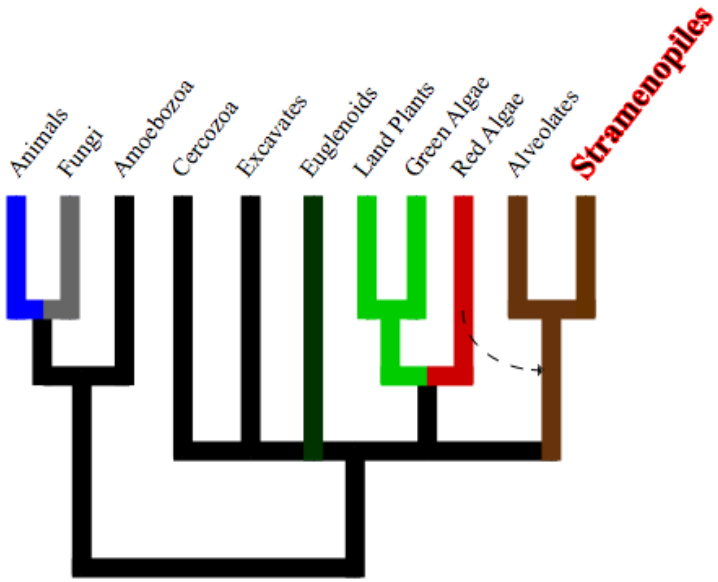
**Phylograms** are diagrammatic representations of molecular phylogeny, constructed by statistical analysis of molecular data. The tree shows the divergence of species from internal nodes, and the branch length signifies the degree of evolutionary change of the taxa. The common ancestor is shown only in rooted trees.

**Phenograms** are also statistically constructed trees which indicate only the degree of resemblance ...

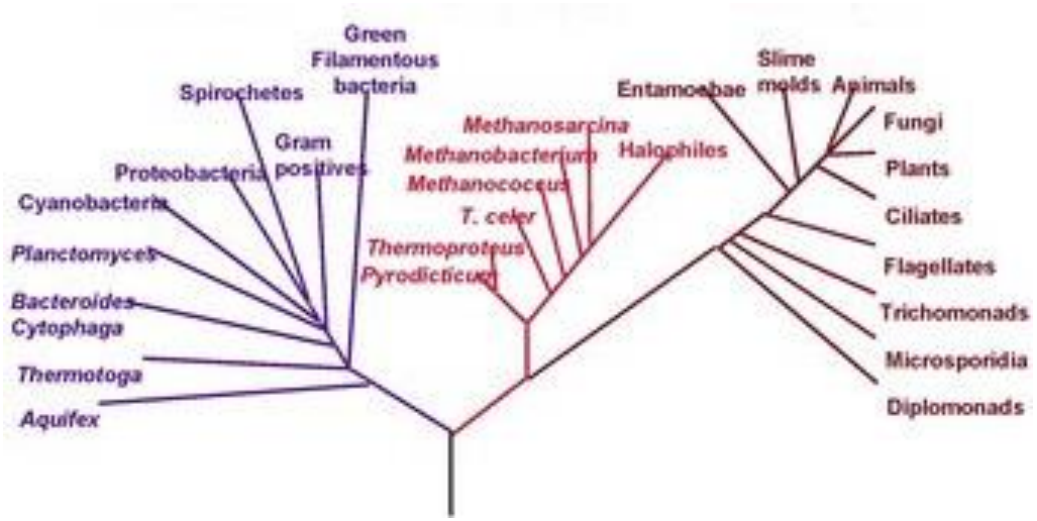
# Tree Terminology

- The branching pattern is called the tree's topology
- Trees can be represented in several forms:

Rectangular cladogram

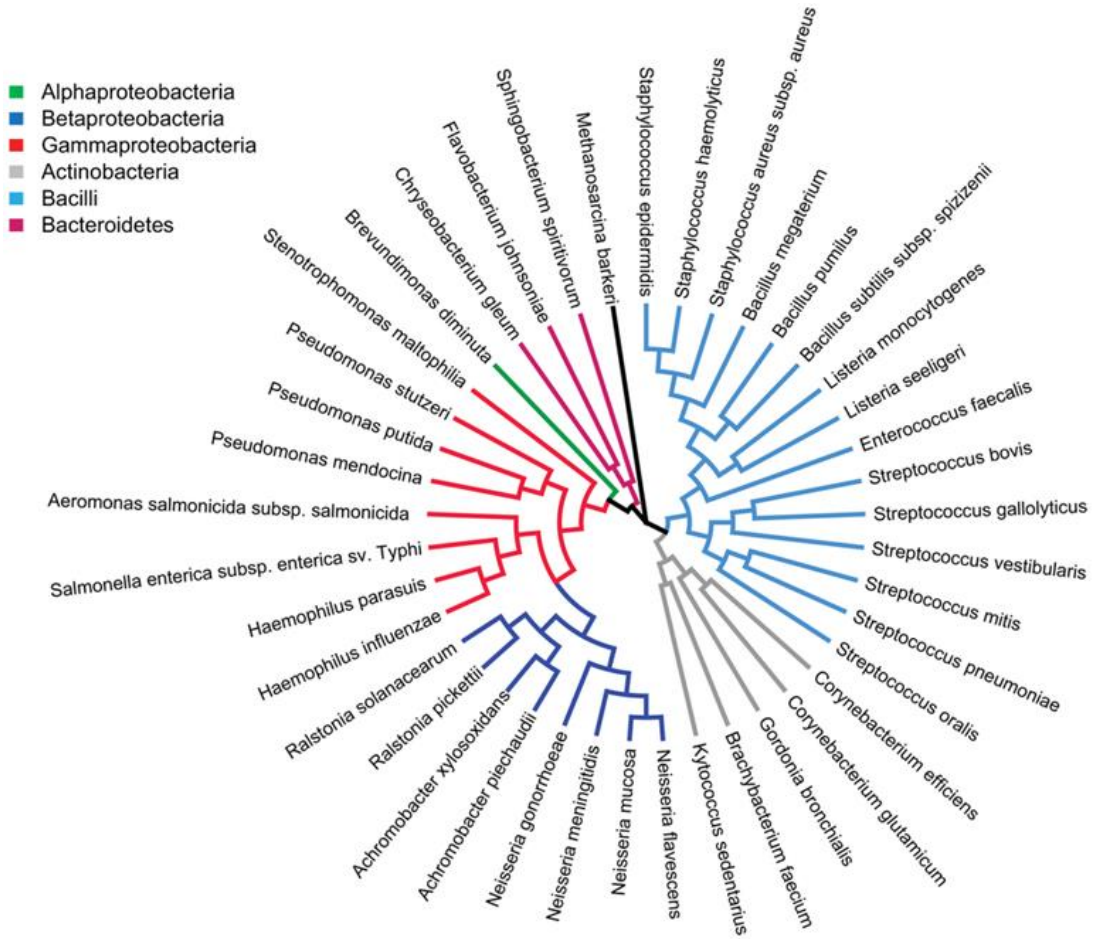



Slanted cladogram



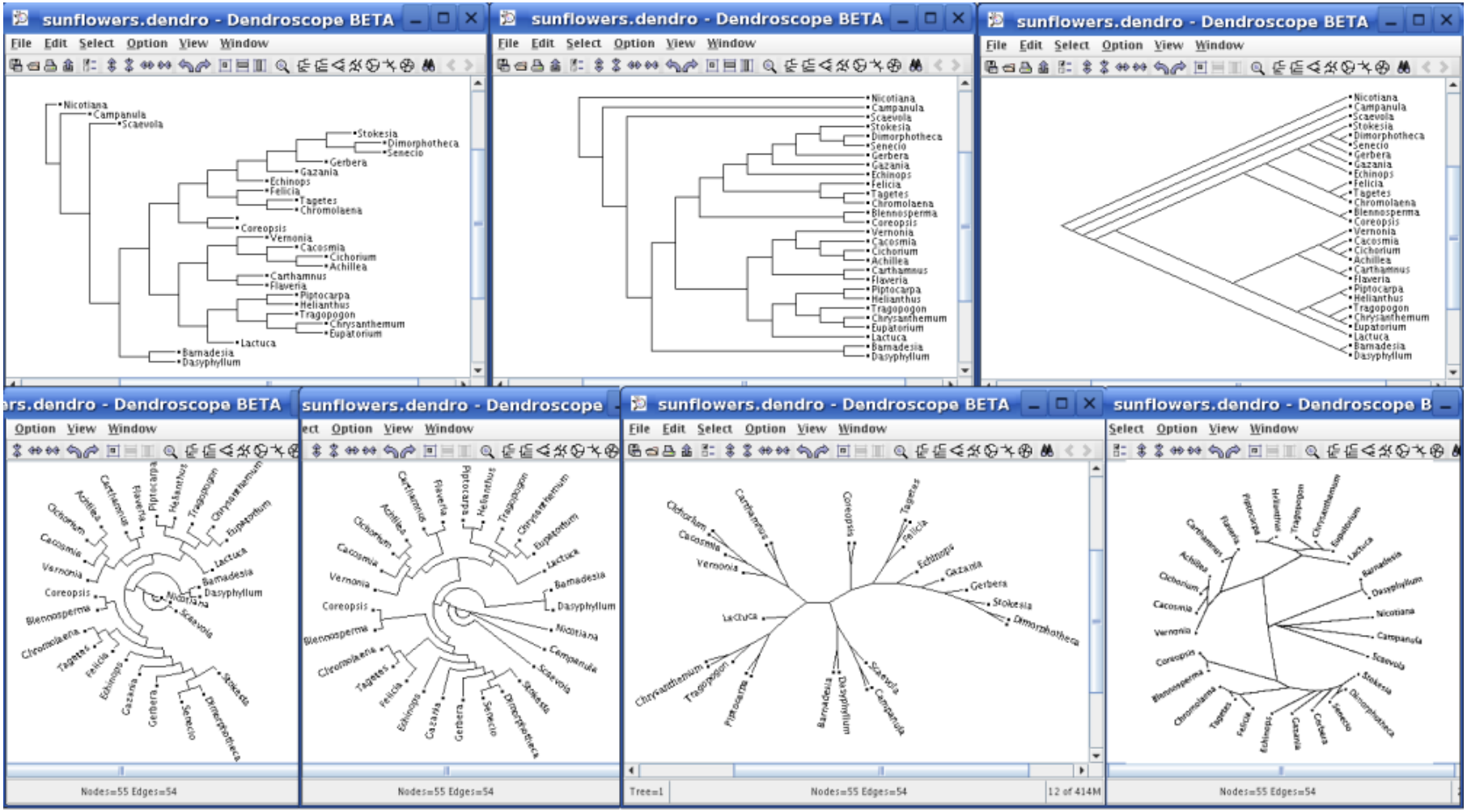
# Tree Terminology

- Circular cladogram

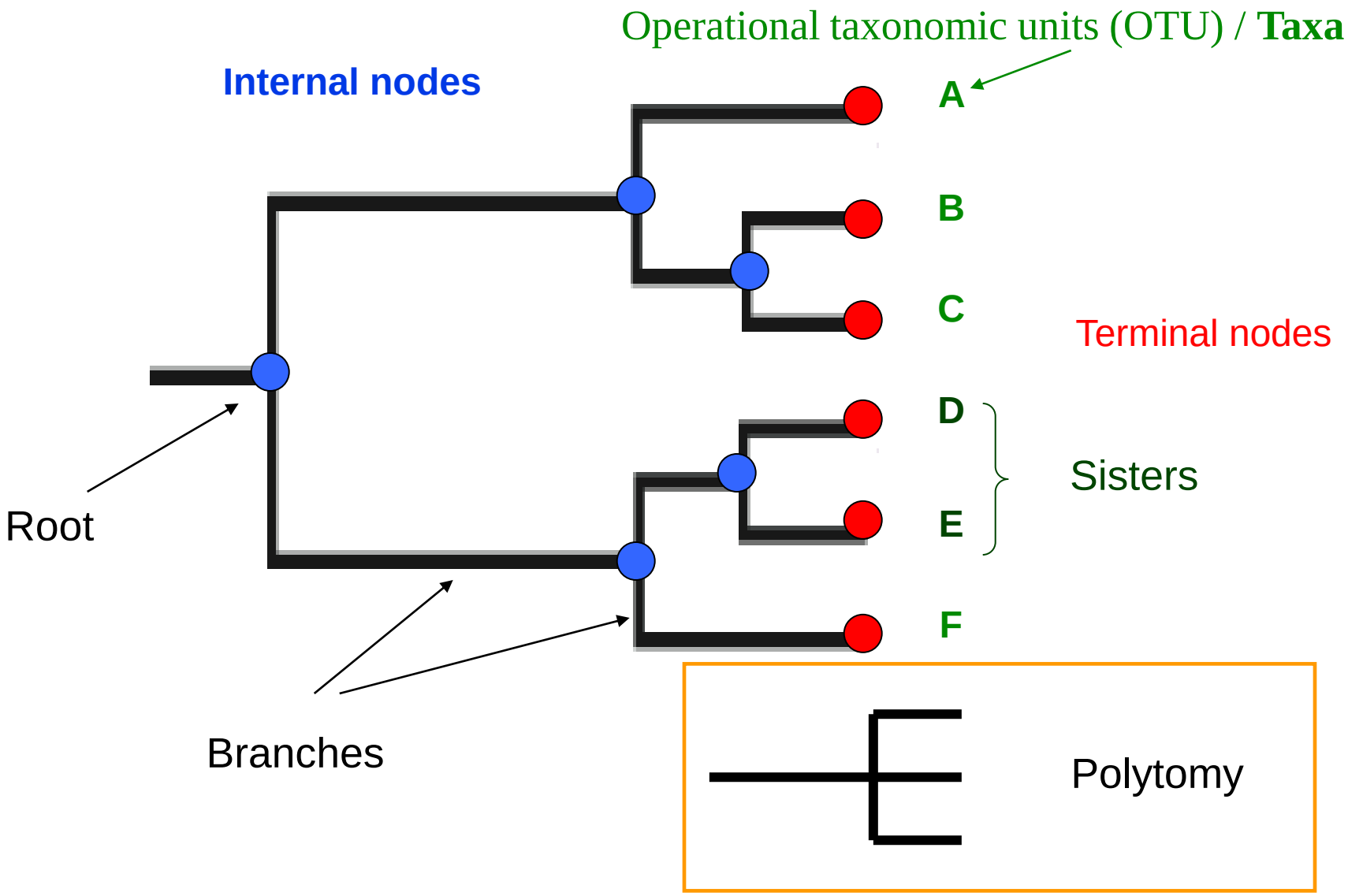


# Same tree - seven different views

Rectangular Phylogram, Rectangular Cladogram, Slanted Cladogram, Circular Phylogram, Circular Cladogram, Radial Phylogram and Radial Cladogram

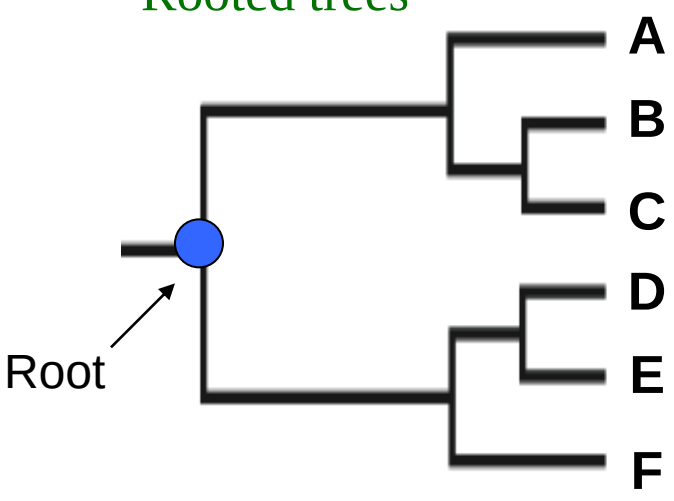


# Tree Terminology

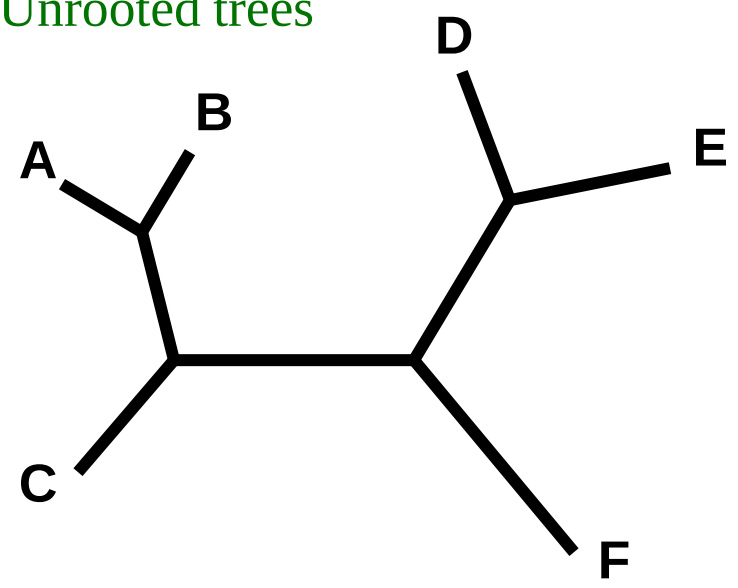


# Tree Terminology

Rooted trees




Unrooted trees

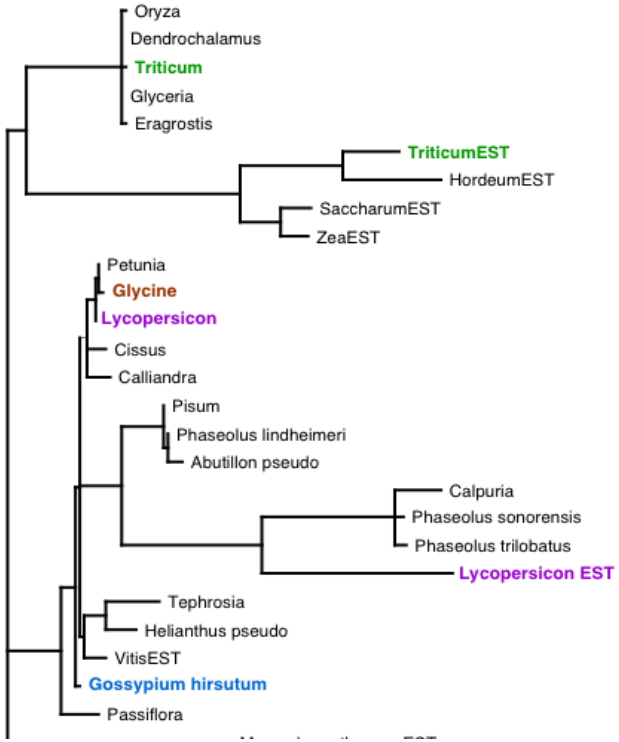



- Rooted trees:
  - has a root that denotes common ancestry
- Unrooted trees:
  - Only specifies the degree of kinship among taxa but not the evolutionary path

Taxon, plural taxa. (taxonomy): Any group or rank in a biological classification into which related organisms are classified.

# Tree Terminology

- Scaled trees:



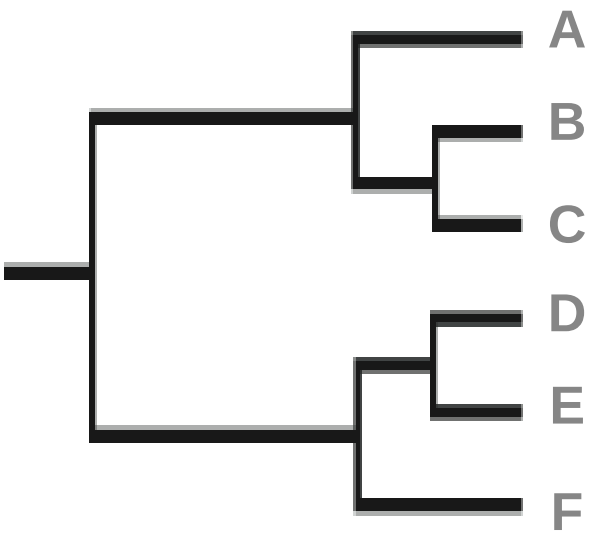

  - Branch lengths are proportional to the number of nucleotide/amino acid changes that occurred on that branch (usually a scale is included).

- In the best of cases, scaled trees are also additive, meaning that

  - the physical length of the branches connecting any two nodes is an accurate representation of their accumulated differences.

# Tree Terminology

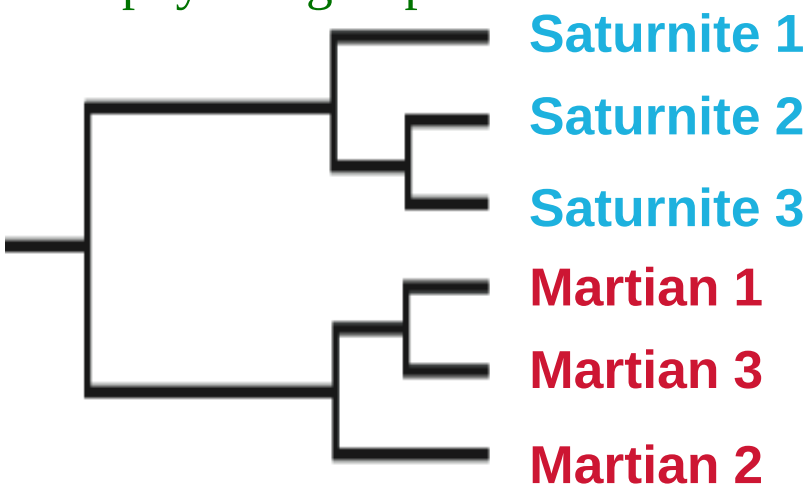
- Unscaled trees:



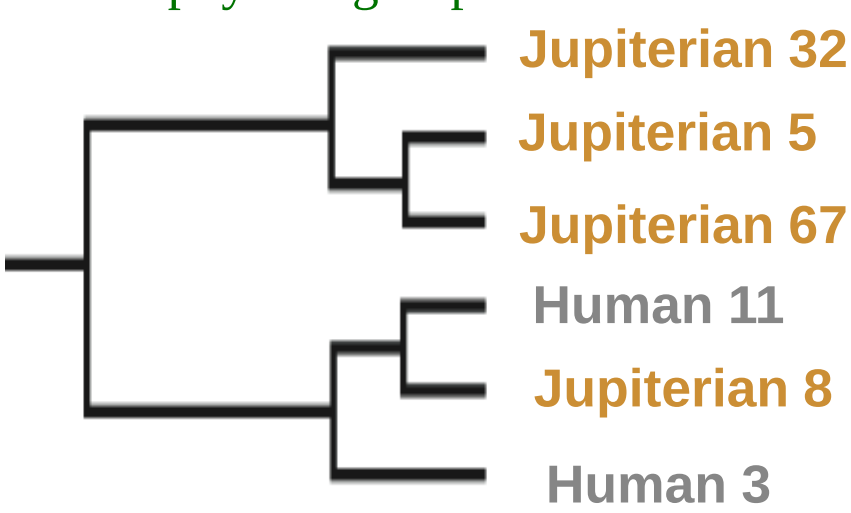

  - Branch lengths are not proportional to the number of nucleotide/amino acid changes
    - usually used to illustrate evolutionary relationships only.
  - line up all terminal nodes and convey only their relative kinship without making any representation regarding the number of changes that separate

# Tree Terminology

Monophyletic groups




Paraphyletic groups




- Monophyletic groups:
  - All taxa within the group are derived from a single common ancestor and members form a natural clade.
- Paraphyletic groups:
  - The common ancestor is shared by other taxon in the group and members do not form a natural clade.

# Gene vs. Species Trees

- Gene tree
  - a phylogenetic tree based on the divergence observed within a single homologous gene.
    - Such trees may represent the evolutionary history of a gene but not necessarily that of the species in which it is found.
- Species trees
  - best obtained from analyses that use data from multiple genes.
    - For example, a study on the evolution of plant species used more than 100 different genes to generate a species tree for plants.

# Character and Distance Data

- The molecular data used to generate phylogenetic trees fall into one of two categories:
  - Characters
    - a well-defined feature that can exist in a limited number of different states
  - Distances
    - a measure of the overall, pairwise difference between two data sets
- Both DNA and protein sequences are examples of data that describe a set of discrete character states.

# Methods in Phylogenetic Reconstruction

- Distance Based Methods
  - calculate pairwise distances between sequences, and group sequences that are most similar.
  - This approach has potential for computational simplicity and therefore speed
- Character Based Methods (Maximum parsimony)
  - assumes that shared characters in different entities result from common descent.
  - Groups are built on the basis of such shared characters, and the simplest explanation for the evolution of characters is taken to be the correct, or most parsimonious one.
- Probabilistic Methods (Maximum likelihood)
  - compute the probability that a data set fits a tree derived from that data set, given a specified model of sequence evolution.

# Comparison of Methods

**Distance**

- Uses only pairwise distances
- Minimizes distance between nearest neighbors
- Very fast
- Easily trapped in local optima
- Good for generating tentative tree, or choosing among multiple trees

**Maximum parsimony**

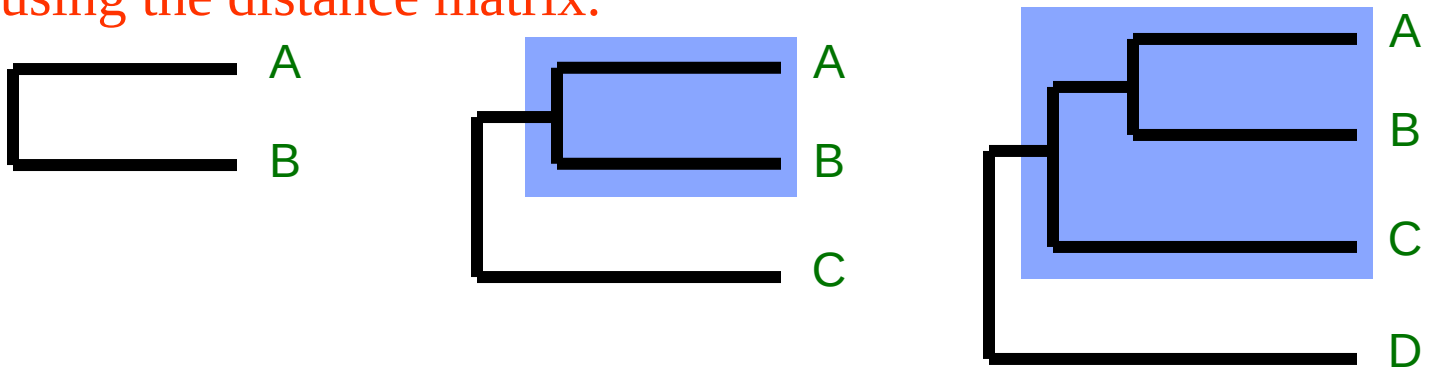
- Uses only shared derived characters
- Minimizes total distance
- Slow
- Assumptions fail when evolution is rapid
- Best option when tractable (<30 taxa, homoplasy rare)

**Maximum likelihood**

- Uses all data
- Maximizes tree likelihood given specific parameter values
- Very slow
- Highly dependent on assumed evolution model
- Good for very small data sets and for testing trees built using other methods

# Methods in Phylogenetic Reconstruction

- Distance Based Methods
  - Using a sequence alignment, pairwise distances/dissimilarities are calculated
  - Creates a distance/dissimilarity matrix
  - A phylogenetic tree is calculated with clustering algorithms, using the distance matrix.




  - Examples of clustering algorithms include
    - Unweighted Pair Group Method using Arithmetic averages (UPGMA)
    - Neighbor Joining clustering.

# UPGMA

- Unweighted-Pair-Group Method with Arithmetic mean (UPGMA)
  - Oldest and simplest distance matrix method
  - Originally proposed in the early 1960s to help with the evolutionary analysis of morphological characters,
  - requires data that can be condensed to a measure of genetic distance between all pairs of taxa being considered.
  - requires a distance matrix such as one that might be created for a group of 4 taxa called A, B, C, and D.

# UPGMA

- Assume that the pairwise distances between each of the taxa are given in the following matrix:

| Species | A | B | C |
|---|---|---|---|
| **B** | $d_{AB}$ | - | - |
| **C** | $d_{AC}$ | $d_{BC}$ | - |
| **D** | $d_{AD}$ | $d_{BD}$ | $d_{CD}$ |

  - $d_{AB}$ : distance between species A and B
  - $d_{AC}$ : distance between species A and C
  - ...

# UPGMA

- UPGMA begins by clustering the two species with the smallest distance separating them into a single, composite group.
  - Assume that the smallest value in the distance matrix corresponds to $d_{AB}$ in which case species A and B are the first to be grouped (AB).
    - After the first clustering, a new distance matrix is computed with the distance between the new group (AB) and species C and D being calculated as

$$d_{(AB)C} = \frac{d_{AC}+d_{BC}}{2} \text{ and } d_{(AB)D} = \frac{d_{AD}+d_{BD}}{2}$$

# UPGMA

- The species separated by the smallest distance in the new matrix are then clustered to make another new composite species.
- The process is repeated until all species have been grouped.
  - If scaled branch lengths are to be used on the tree to represent the evolutionary distance between species, branch points are positioned at a distance halfway between each of the species being grouped
    - i.e., at $d_{AB}/2$ for the first clustering

# UPGMA - example

- Consider the following alignment between five different DNA sequences

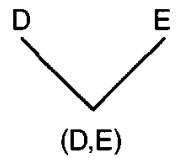
```
                10         20         30         40         50
A:      GTGCTGCACGG CTCAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA
B:      ACGCTGCACGG CTCAGTGCG GTGCTTACCC TCCCATCTTC AGATCCTGAA
C:      GTGCTGCACGG CTCGGCGCA GCATTTACCC TCCCATCTTC AGATCCTATC
D:      GTATCACACGA CTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCCTAAA
E:      GTATCACATAG CTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCTAAAA
```

- Pairwise distance matrix

| Species | A | B | C | D |
|---|---|---|---|---|
| **B** | 9 | - | - | - |
| **C** | 8 | 11 | - | - |
| **D** | 12 | 15 | 10 | - |
| **E** | 15 | 18 | 13 | 5 |

– Smallest distance:

- $d_{DE}$ , so
  – Species D and species E are grouped



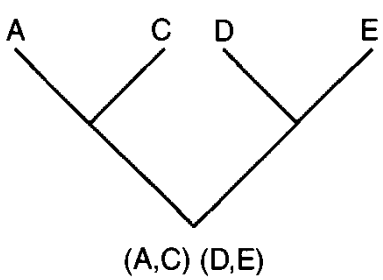
# UPGMA - example

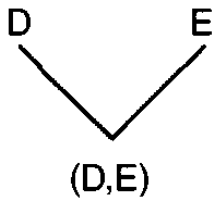
| Species | A | B | C | D |
|---|---|---|---|---|
| B | 9 | - | - | - |
| C | 8 | 11 | - | - |
| D | 12 | 15 | 10 | - |
| E | 15 | 18 | 13 | 5 |

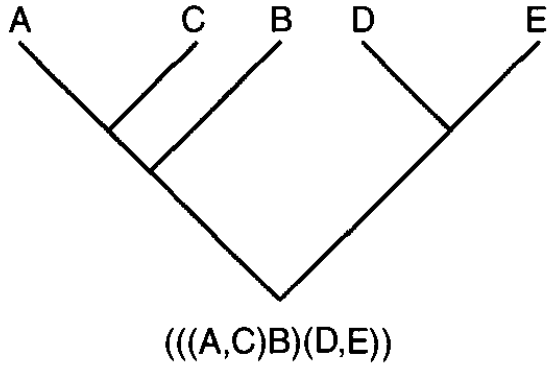
| Species | A | B | C |
|---|---|---|---|
| B | 9 | - | - |
| C | 8 | 11 | - |
| DE | $\frac{12+15}{2} = 13.5$ | $\frac{15+18}{2} = 16.5$ | $\frac{10+13}{2} = 11.5$ |

| Species | B | AC |
|---|---|---|
| AC | $\frac{9+11}{2} = 10$ | - |
| DE | $\frac{15+18}{2} = 16.5$ | $\frac{13.5+11.5}{2} = 12.5$ |

| Species | (AC)B |
|---|---|
| (AC)B | – |
| DE | $\frac{16.5+12.5}{2} = 19.5$ |



# Estimation of Branch Lengths

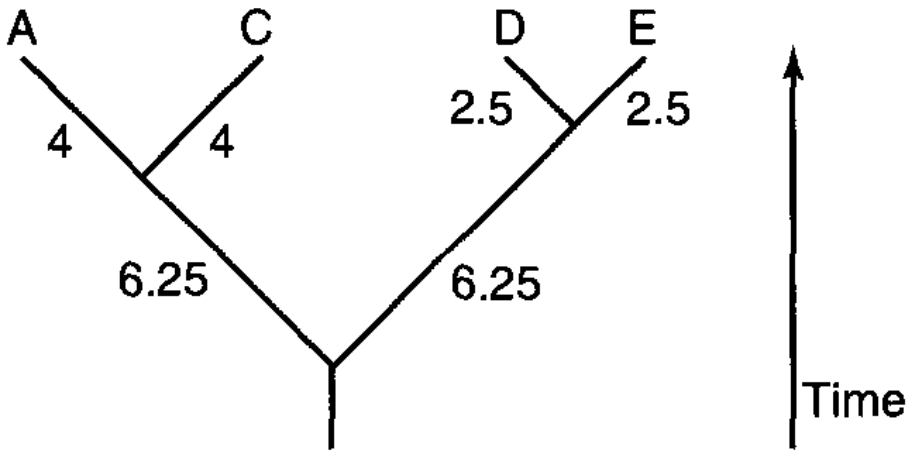
- Tree describes the relatedness of sequences
- It is possible for the topology of phylogenetic trees to convey information about
  - the relative degree to which sequences have diverged.
  - Scaled trees that convey that information, often referred to as cladograms,
    - the length of branches correspond to the inferred amount of time that the sequences have been accumulating substitutions independently.

# Estimation of Branch Lengths

- The relative length of each branch in a cladogram can be calculated using the information in a distance matrix.
  - In the example, the $d_{DE}$ is 5,
    - the pair of branches connecting each of those species to their common ancestor should each be $d_{DE}/2$ or 2.5 units long on a tree with scaled branch lengths.
  - A and C should be connected to their common ancestor by branches that are $d_{AC}/2$ or 4 units long,
  - The branch point between (AC) and (DE) should be connected to (AC) and (DE) by branches that are both $d_{(AC)(DE)}/2$ or 6.25 units long,

# Estimation of Branch Lengths

- A scaled tree showing the branch lengths separating four of the species depicted in slide 30.

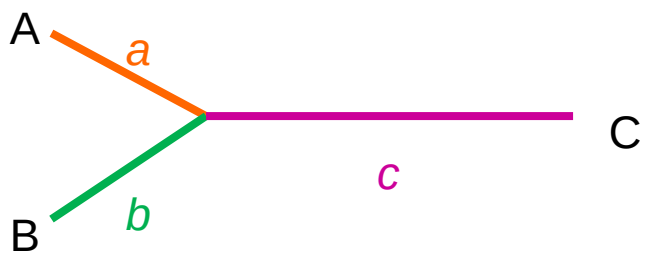



- Branch lengths are shown next to each branch.
- Branches are also drawn to scale to reflect the amount of differences between all species.

- This very simple approach to estimating branch lengths actually allows UPGMA to intrinsically generate rooted phylogenetic trees.

# Estimation of Branch Lengths

- Determining branch lengths for a scaled tree is only slightly more complicated
  - when it cannot be assumed that evolutionary rates are the same for all lineages.




- The simplest tree whose branch lengths might have some meaningful information is one with just three species (A, B, C) and one branch point, such as the one shown.
- On such a tree, the length of each of the three branches can be represented by a single letter (*a*, *b*, and *c*) for which we know the following must be true:

$d_{AB} = a + b$ ; $d_{BC} = b + c$ ; $d_{AC} = a + c$

# Estimation of Branch Lengths

- Phylogeny reconstruction for 3 sequences
  - There is a single tree topology
  - The branch lengths ($a$, $b$, $c$) :

$a + b = d_{AB}$

$b + c = d_{BC}$

$a + c = d_{AC}$

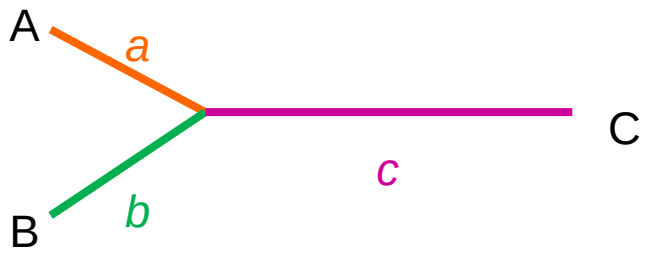
| | A | B | C |
|---|---|---|---|
| A | - | $a+b$ | $a+c$ |
| B | - | - | $b+c$ |
| C | - | - | - |

- Input:
  - $d_{AB}$, $d_{BC}$ and $d_{AC}$ (pairwise distances)
- Output:

$a = (d_{AB} + d_{AC} - d_{BC}) / 2$

$b = (d_{AB} + d_{BC} - d_{AC}) / 2$

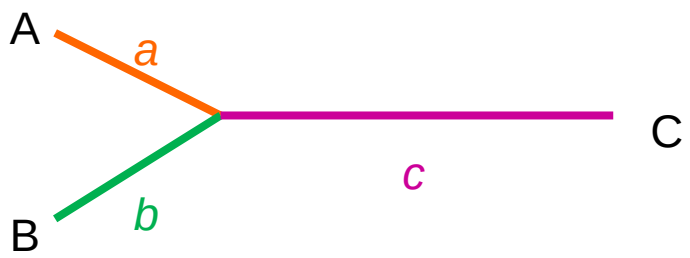
$c = (d_{AC} + d_{BC} - d_{AB}) / 2$



# Estimation of Branch Lengths - example

- Distance matrix of 3 sequences and unrooted tree

| | A | B | C |
|---|---|---|---|
| A | -- | 22 | 39 |
| B | -- | -- | 41 |
| C | -- | -- | -- |




  - distance from A to B = $a + b = 22$ (1)
  - distance from A to C = $a + c = 39$ (2)
  - distance from B to C = $b + c = 41$ (3)

- subtracting (2) from (3) yields:

$$b + c - (a + c) = b - a = 41 - 39 = 2 \quad (4)$$

# Estimation of Branch Lengths - example

- adding (1) and (4) yields

  $a + b + b - a = 2b = 22 + 2 = 24$

  $2b = 24$

  $b = 24 / 2 = 12$

- so

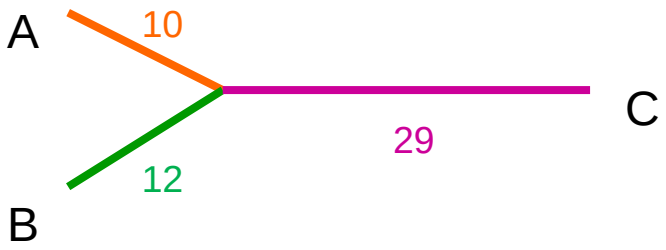
  $a + b = a + 12 = 22;$

  $a = 22 - 12 = 10$

- finally

  $a + c = 10 + c = 39;$

  $c = 39 - 10 = 29$



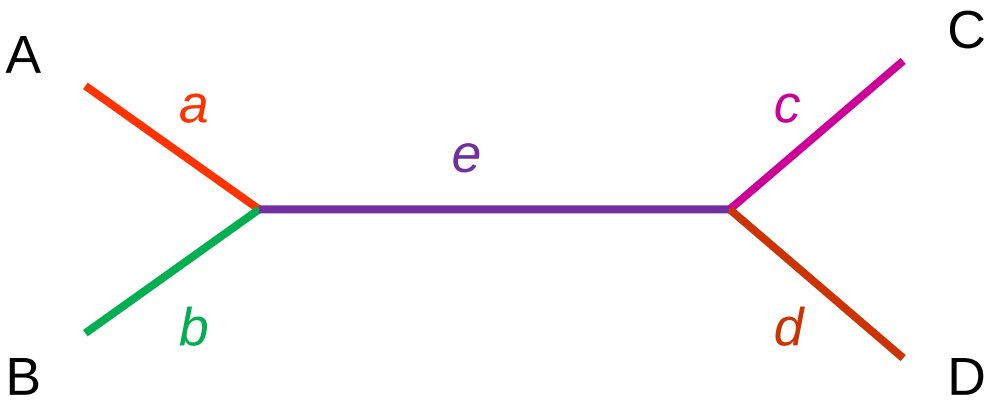
# Neighbor's Relation Method

- Popular variant of the UPGMA method
- emphasizes pairing species in such a way that
  - a tree is created with the smallest possible branch lengths overall.

- On any unrooted tree, pairs of species that are separated from each other by just one internal node are said to be neighbors.

# Neighbor's Relation Method

- The topology of a phylogenetic tree such as the one shown below gives rise to some useful algebraic relationships between neighbors.




| Species | A | B | C |
|---|---|---|---|
| B | $d_{AB}$ | - | - |
| C | $d_{AC}$ | $d_{BC}$ | - |
| D | $d_{AD}$ | $d_{BD}$ | $d_{CD}$ |

| Species | A | B | C |
|---|---|---|---|
| B | $a+b$ | - | - |
| C | $a+e+c$ | $b+e+c$ | - |
| D | $a+e+d$ | $b+e+d$ | $c+d$ |

- If the tree above is a true tree for which additivity holds, then the following should be true:

$$d_{AC} + d_{BD} = d_{AD} + d_{BC} = a + b + c + d + 2e = d_{AB} + d_{CD} + 2e$$

  - where $a$, $b$, $c$, and $d$ are the lengths of the terminal branches and $e$ is the length of the single central branch.

# Neighbor's Relation Method

- The following conditions, known together as the four-point condition, will also be true:

$$d_{AB} + d_{CD} < d_{AC} + d_{BD}\ ;\quad d_{AB} + d_{CD} < d_{AD} + d_{BC}$$

- It is in this way that a neighborliness approach considers all possible pairwise arrangements of four species and determines which arrangement satisfies the four-point condition.
  - An important assumption of the four-point condition is that branch lengths on a phylogenetic tree should be additive and, while it is not especially sensitive to departures from that assumption, data sets that are not additive can cause this method to generate a tree with an incorrect topology.

# Neighbor's Relation Method - example

- Consider the alignment:

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

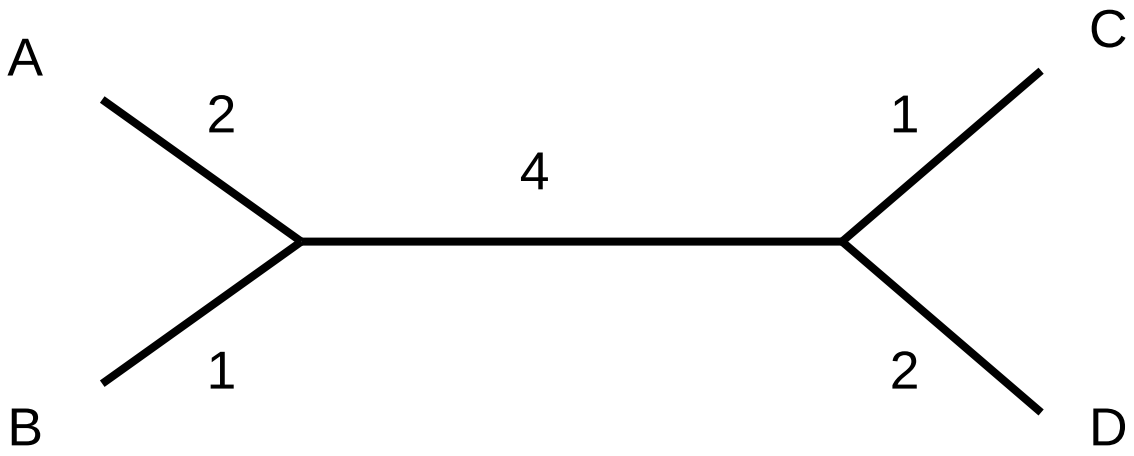
- The distances between these sequences can be shown as a table:

| | A | B | C | D |
|---|---|---|---|---|
| A | - | 3 | 7 | 8 |
| B | - | - | 6 | 7 |
| C | - | - | - | 3 |
| D | - | - | - | - |

# Estimation of Branch Lengths - example

- Using this information, an unrooted tree showing the relationship between these sequences can be drawn:

| | A | B | C | D |
|---|---|---|---|---|
| A | - | 3 | 7 | 8 |
| B | - | - | 6 | 7 |
| C | - | - | - | 3 |
| D | - | - | - | - |



# Neighbor-Joining Methods

- A variant of neighborliness
- an agglomerative technique, and so operates using iteration,
  - building the tree from the bottom-up
- Start with a star-like tree with all species coming off a single central node regardless of their number.
- Neighbors are then sequentially found that minimize the total length of the branches on the tree.

# Neighbor-Joining Methods

- The input is an $n \times n$ dissimilarity/distance matrix $d$.
- In the first iteration,
  - the $n$ leaves are all in their own clusters;
- In subsequent iterations,
  - each cluster is a set of leaves,
    - but the clusters are disjoint.
- At the beginning of each iteration, the taxa are partitioned into clusters, and for each cluster we have a rooted tree that is leaf-labelled by the elements in the cluster.

# Neighbor-Joining Methods

- During the iteration, a pair of clusters is selected to be made siblings;
  - this results in the trees for the two clusters being merged into a larger rooted tree by making their roots siblings.
- When there are only three subtrees, then the three subtrees are merged into a tree on all the taxa by adding a new node, *r*, and making the roots of the three subtrees adjacent to *r*.
- Note that this description suggests that neighbor joining produces a rooted tree.
  - However, after the tree is produced, the root is ignored, so that neighbor joining actually returns an unrooted tree.

# Neighbor-Joining Methods

*The neighbor joining algorithm.* *Input:* $n \times n$ dissimilarity matrix $\mathbf{d}$ with $n \geq 4$
*Output:* Unrooted tree with $n$ leaves labelled $1 \ldots n$

*Initialization:* Compute the $n \times n$ matrix $\mathbf{Q}$, defined by

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n}(d_{ik} + d_{jk}).$$

While $n > 3$, DO:

Find the pair $i, j$ minimizing $Q_{i,j}$. Without loss of generality, we will call that pair $a, b$.
Make the rooted trees associated with taxa $a$ and $b$ siblings, and call the root of the tree you form $u$.
Update the distance matrix by deleting the rows and columns for $a$ and $b$, and including a new row and column for $u$, and set $d_{u,k} = \frac{d_{ak} + d_{bk} - d_{ab}}{2}$ for all $k \neq u$. Decrement $n$ by 1.

Now $n = 3$; return the star tree with a single internal node $v$ where the roots of the three rooted trees are all adjacent to $v$.

- The neighbor-joining method: a new method for reconstructing phylogenetic trees. (1987). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a040454

# PRACTICAL SESSION ON UPGMA AND NEIGHBOUR JOINING METHOD FOR PHYLOGENETIC TREE CONSTRUCTION.