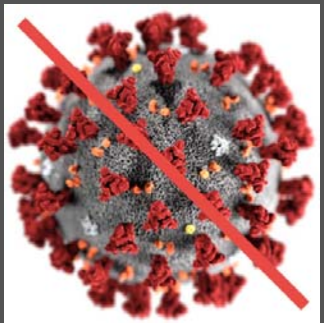


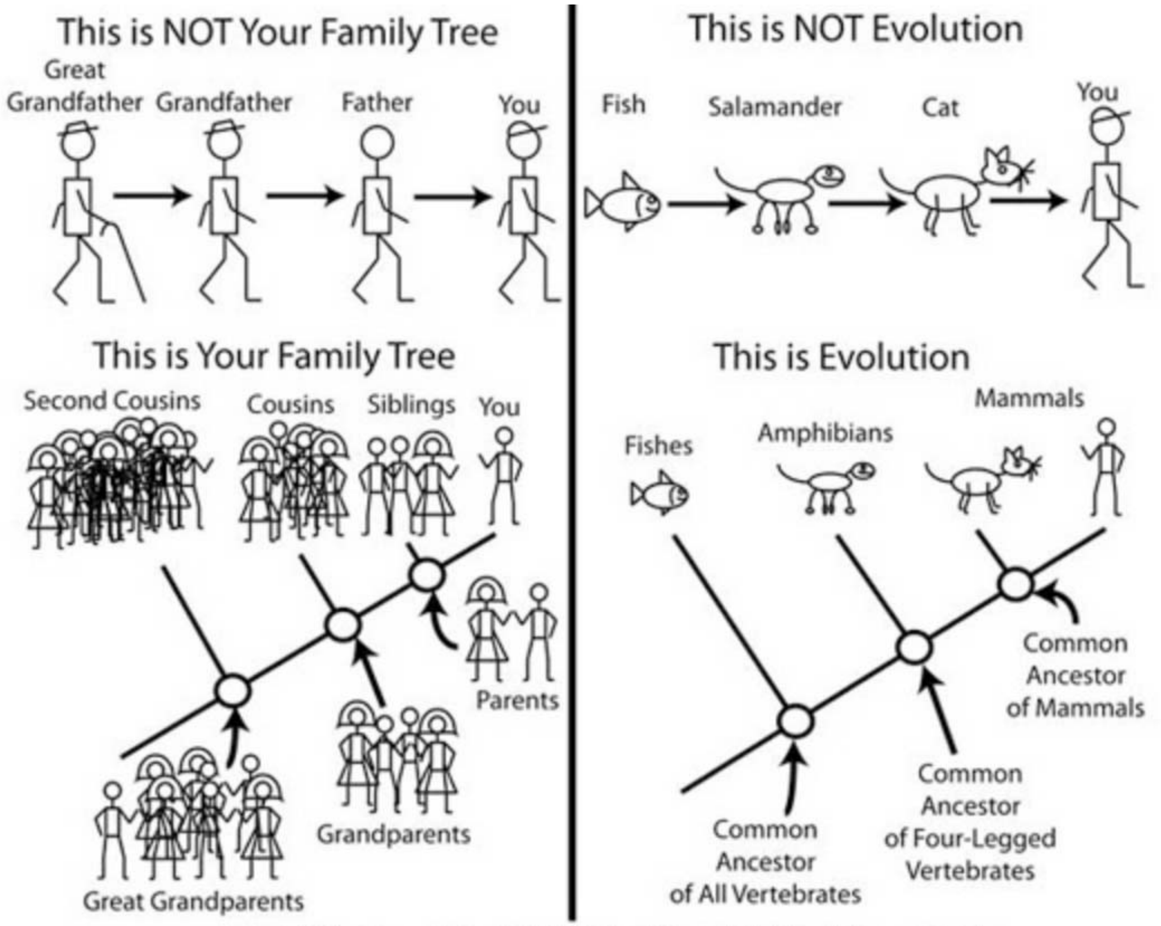
# CS123A Bioinformatics

# Module 3 – Week 9 – Presentation 1

Leonard Wesley

Computer Science Dept

San Jose State Univ



Cartoon by Matthew Bonnan of Macomb, IL, with kind permission of Florida Citizens for Science, Sept. 2010

# Agenda

- MSA Algorithm
- Phylogenetic Trees
  - Section **“Molecular Phylogeny: Properties of Trees”** starting on page 259 in textbook
- Definition & Terms of Phylogenetic Trees
- Hierarchical Clustering & UPGMA Tree Building Example

# Hierarchical MSA Is A Multiple Step Process.

- Given 3 or more sequences to align
- Sometimes random unrelated sequences are given to a MSA algorithm. Must determine significance by performing a randomization test.
- Two sequences are pair-wise aligned and the score (S) recorded.
- Then amino acids/nucleic acids in the sequences are shuffled so order is changed but length kept the same.

# Hierarchical MSA Is A Multiple Step Process. *(cont. #1)*

- Shuffled sequences are compared again and scores (S) recorded again. This is repeated ~100 times.

- The mean $\bar{S}$ and the standard deviation $\sigma$ for the scores is calculated.

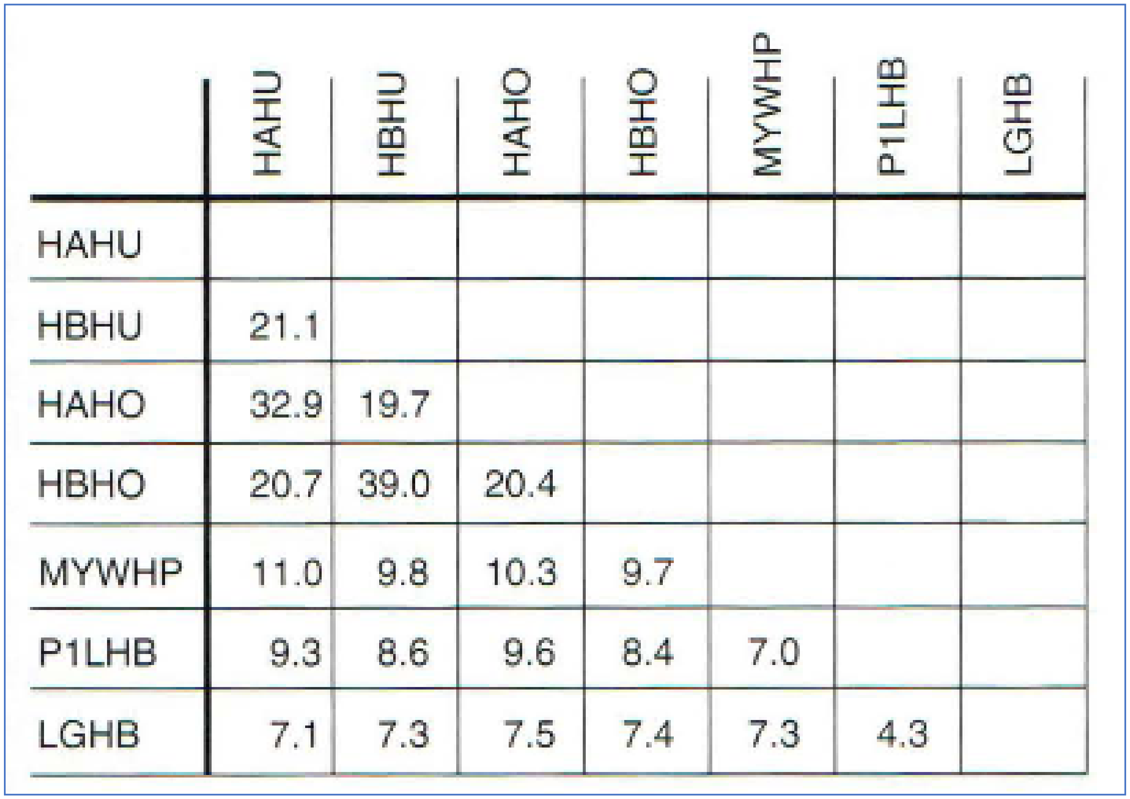
- A Z score = $(S - \bar{S}) / \sigma$ provides an indication of the significance of the two sequences.

# Hierarchical MSA Is A Multiple Step Process. *(cont. #2)*

- A  Z score > 6 means high likelihood the two sequences can be aligned and aligned correctly in a way that can give insight into function, structure, …and so forth.

- However, some alignments with Z score < 6 can be correct. If and when this happens, one needs to consider the possibility that sequence similarity might have diverged faster than structural or functional similarity.

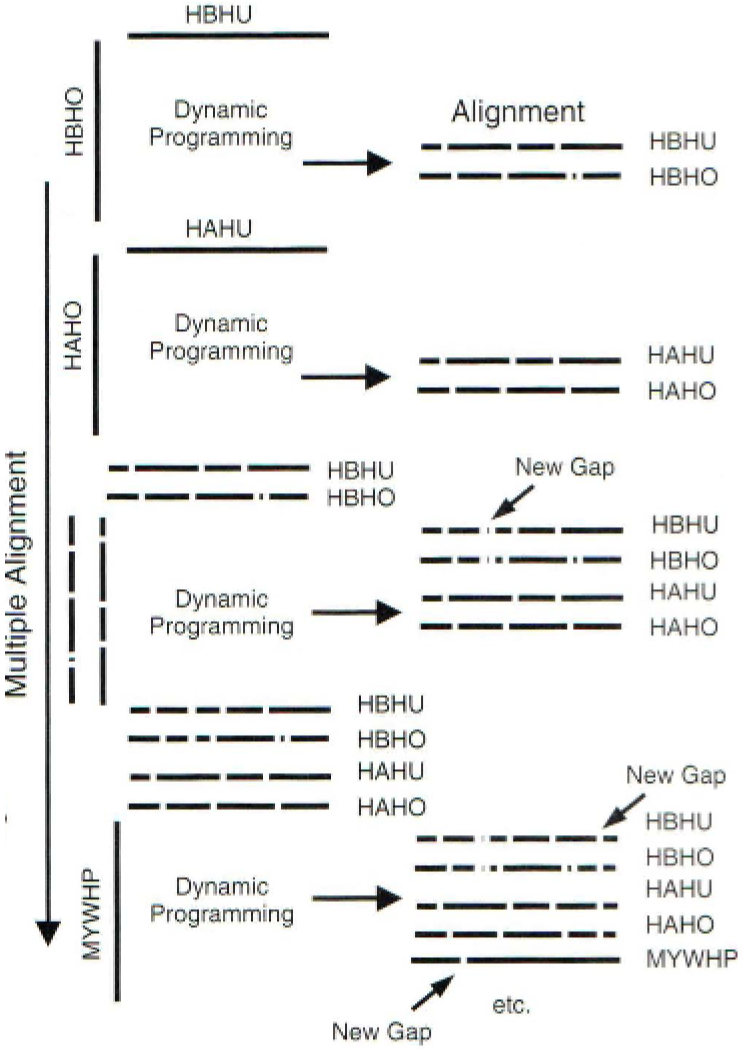
# Example Z Score Matrix



| | HAHU | HBHU | HAHO | HBHO | MYWHP | P1LHB | LGHB |
|---|---|---|---|---|---|---|---|
| HAHU | | | | | | | |
| HBHU | 21.1 | | | | | | |
| HAHO | 32.9 | 19.7 | | | | | |
| HBHO | 20.7 | 39.0 | 20.4 | | | | |
| MYWHP | 11.0 | 9.8 | 10.3 | 9.7 | | | |
| P1LHB | 9.3 | 8.6 | 9.6 | 8.4 | 7.0 | | |
| LGHB | 7.1 | 7.3 | 7.5 | 7.4 | 7.3 | 4.3 | |

**Pairwise Z-scores for comparison of each sequence pair.**
**Higher numbers mean greater similarity**

# Building The Multiple Alignment



> The first two steps are pairwise alignments.

> The third step is a comparison of profiles from the two alignments generated in steps I and 2.

>The fourth step adds a single sequence (MYWHP) to the alignment generated at step 3.

> Further sequences are added in a similar manner.

# Other MSA Algorithms

- Hierarchical not guaranteed to find optimal alignment
- TCoffee: builds a library of pairwise alignments → inputs this to a hierarchical method
- PSI-BLAST: searches DB with a single sequence, high scores are retrieved and built into a MSA
- SCANPS: Similar to PSI-BLAST uses Smith-Waterman
- STAMP: Aligns multiple protein structures vs sequences.

# Example TCoffee MSA

- Go To [http://www.ebi.ac.uk/Tools/msa/tcoffee](http://www.ebi.ac.uk/Tools/msa/tcoffee)
- Select “Use a example sequence” Then click “More options…” Then select BLOSUM
- Click Submit and then wait for the results.
- Then Select “Show Colors”. Look for good (Red) and Excellent (Blue) alignment regions. Then Select “Phylogenetic Tree”. Identify closely and distant organisms.

# MSA Lecture Exercise

- You came back from a trip to a jungle swamp after obtaining what you believe are DNA and/or protein samples of possibly known and/or unknown organisms. You want to know (1) If you have found evidence of existing or new organisms. If existing organisms, which one(s)?; (2) What part or structure of the organism's genome, if any, are we looking at?; and (3) What are related organisms ?
- The sequencing lab has provided you with a file that contains a protein sequence from the liquid sample that you gave them. The sequenced protein is contained in the file name "CS123A_Example_seq.txt" that is located in Canvas -> Files -> Module 3 Phylogenetic Trees -> Week 9 -> Slides folder.
- BLASTP the sequence to find possible best matches. In the "Organism" section type in "prokaryote" in the first window and select the (taxid:2) entry. Click on the "+ then enter and select the Rattus (taxid:10114) entry. Click "+" one last time and enter "Fish stool-associated RNA virus (taxid:2219050)". Click the BLAST button. Note the names of the top 4 "DIFFERENT" organisms. What are these organisms?
- Create and name .txt file. Get the FASTA sequence for the first 4 "DIFFERENT" matches you selected. You can get the FASTA sequence after clicking on each accession number and going to that web page. Then look for a link to the FASTA file. Click that link, then on the drop down tab in the upper left next to the word FASTA, select the "FASTA txt" option. Copy and paste the FASTA info into to the .txt file that you created and named at the start of this step.
- Copy each of the 4 FASTA sequences into your .txt file. Then do a MSA on the sequences. Use the dendrogram to determine which sequences are most closely related. Upload your answer to "which sequences are most closely related" to Canvas Lecture Exercise 2.

# Summary

- Sequence alignment is useful to identify novel and existing organisms form genomic sequences. MSA is helpful to identify homologous and conserved regions.

- BLAST & BLAST2: Performs local pairwise and multiple alignments for nucleotides, proteins, and from nucleotides to proteins and from proteins back to nucleotide. Score (S) and Expect (E) values used to help assess quality of match.

- Smith-Waterman: Uses dynamic programming to provide optimal local sequence pairwise alignment. Can be used by multiple sequence alignment (MSA) algorithms, SCANPS.

- Needleman-Wunsch: Uses dynamic programming to provide optimal global sequence pairwise alignment. Gaps can be inserted to optimal sequence scores and to make each sequence the same length. Cane used by MSA algorithms.

# Summary *(cont.)*

- Several good MSA tools: TCoffee: builds a library of pairwise alignments → inputs this to a hierarchical method.
- PSI-BLAST: searches DB with a single sequence, high scores are retrieved and built into a MSA.
- SCANPS: Similar to PSI-BLAST uses Smith-Waterman.
- STAMP: Aligns multiple protein structures vs sequences.

# Constructing Phylogenetic Trees

# Building Phylogenetic Trees From MSAs

- One Way:
  - Use the alignment score or Z score between each pair of sequences.
  - Use the similarity score between each pair of sequences.

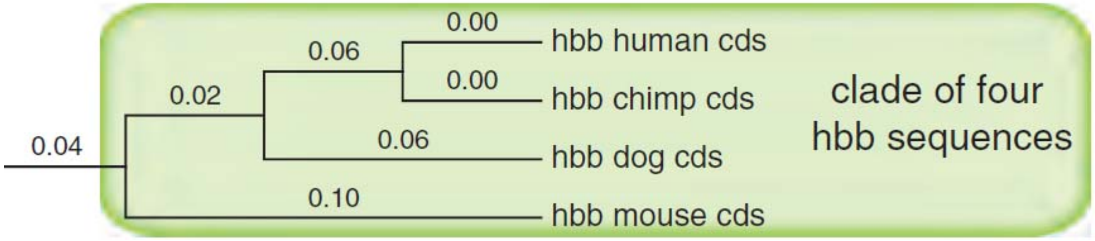
| | Seq1 | Seq2 | Seq3 | Seq4 |
|---|---|---|---|---|
| Seq1 | - | 3 | 1 | 4 |
| Seq2 | 3 | - | | |
| Seq3 | 1 | 7 | - | 3 |
| Seq4 | 4 | 0.5 | 3 | - |

  - Use hierarchical clustering techniques to build dendrogram. See previous in-lecture example.

# Some Definitions First

# Two Types Of Information In Phylogenetic Trees

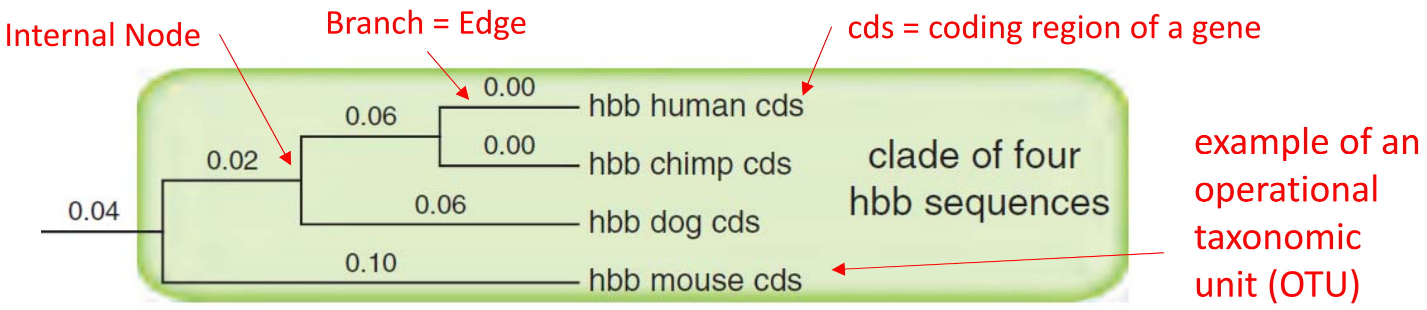
- Topology: *Defines the relationships of the proteins (or other objects) that are represented in the tree. For example, the topology in the tree shows the common ancestor of two homologous protein sequences.*



- BRANCH LENGTH: *The branch lengths sometimes (but not always) reflect the degree of relatedness of the objects in the tree.*

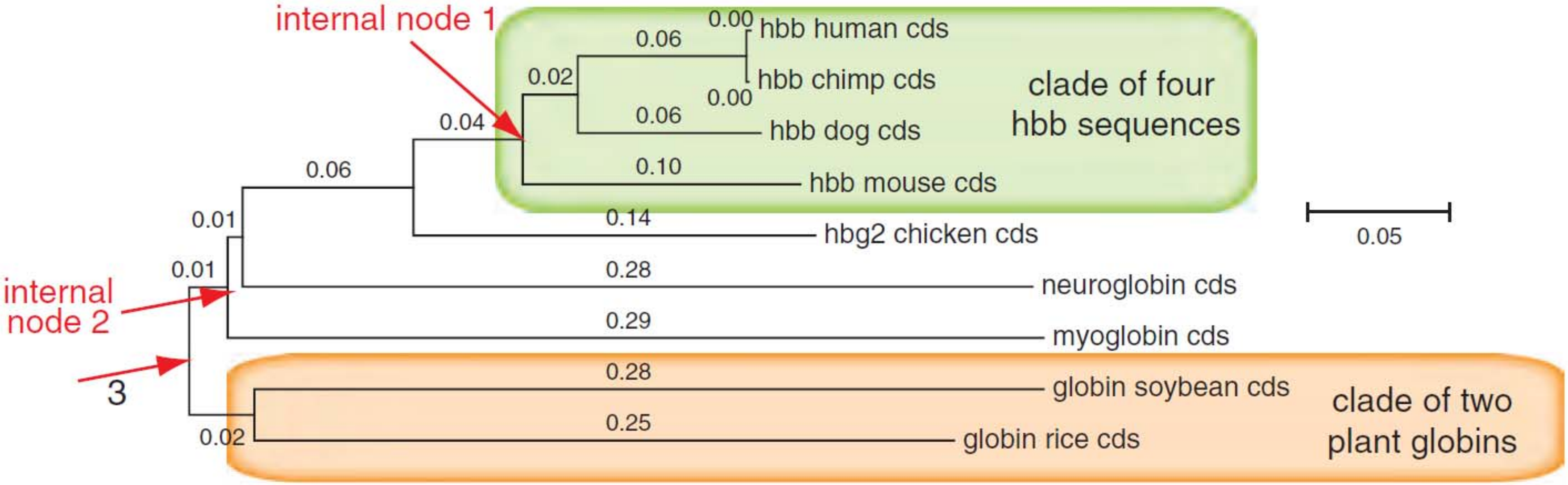
# Parts Of A Phylogenetic Tree: Branches & Nodes

- Only one branch (also called an edge) connects any two nodes. Nodes represent the taxonomic units (taxa or taxons), and is the intersection or terminating point of two or more branches. Taxa will typically be DNA or protein sequences.
- An operational taxonomic unit (OTU) is an extant taxon present at an external node or leaf.  OTUs are the available nucleic acid or protein sequences that we are analyzing in a tree.
- The internal nodes represent ancestral sequences that we can infer but can only very rarely observe (as in the case of sequencing DNA from extinct organisms)



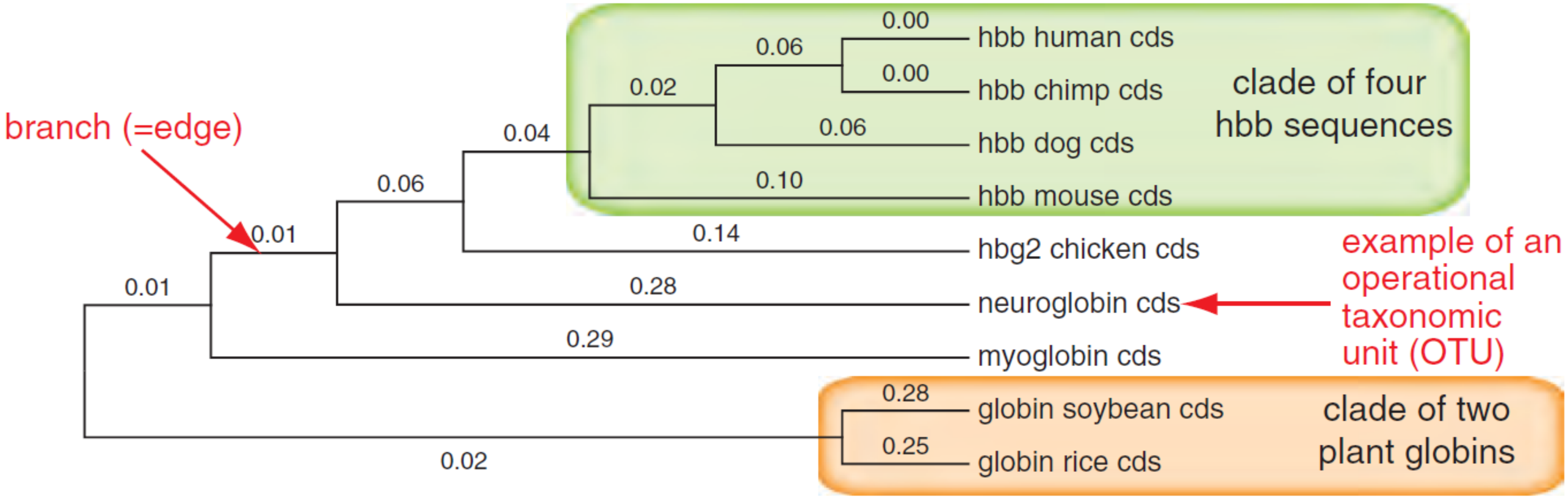
# Several Ways To Build A Phylogenetic Tree

- Nine globin coding sequences: neighbor-joining tree (rectangular tree style)



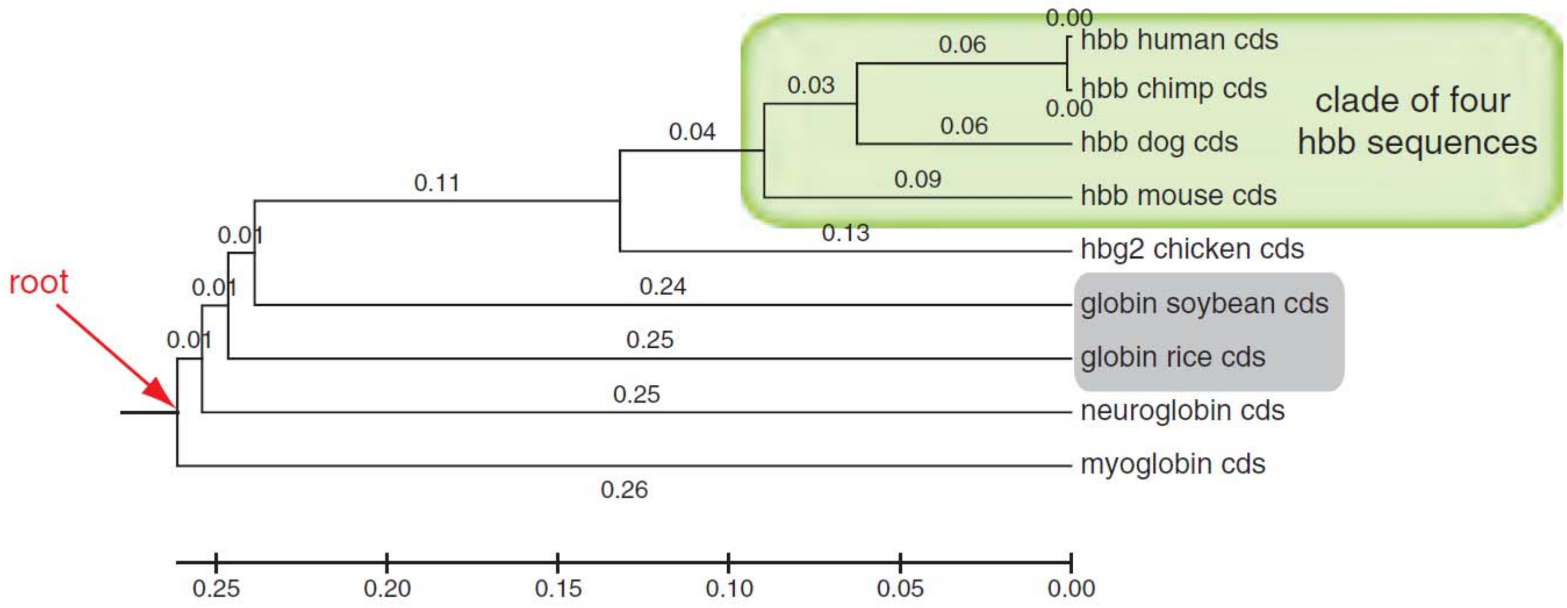
# Several Ways To Build A Phylogenetic Tree *(cont.)*

- Nine globin coding sequences: neighbor-joining tree ("topology only" tree style)



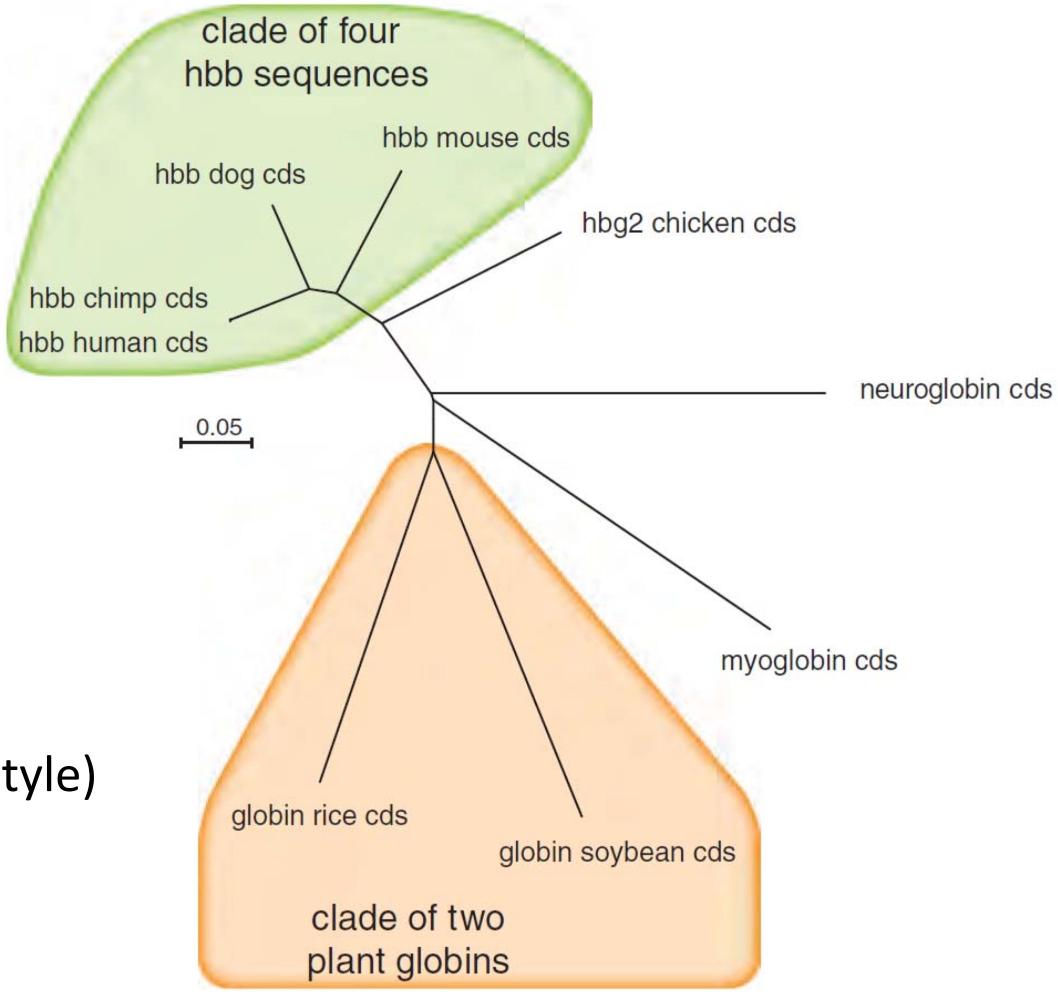
# Several Ways To Build A Phylogenetic Tree *(cont.)*

- Nine globin coding sequences: UPGMA (unweighted pair group method of arithmetic averages) tree.



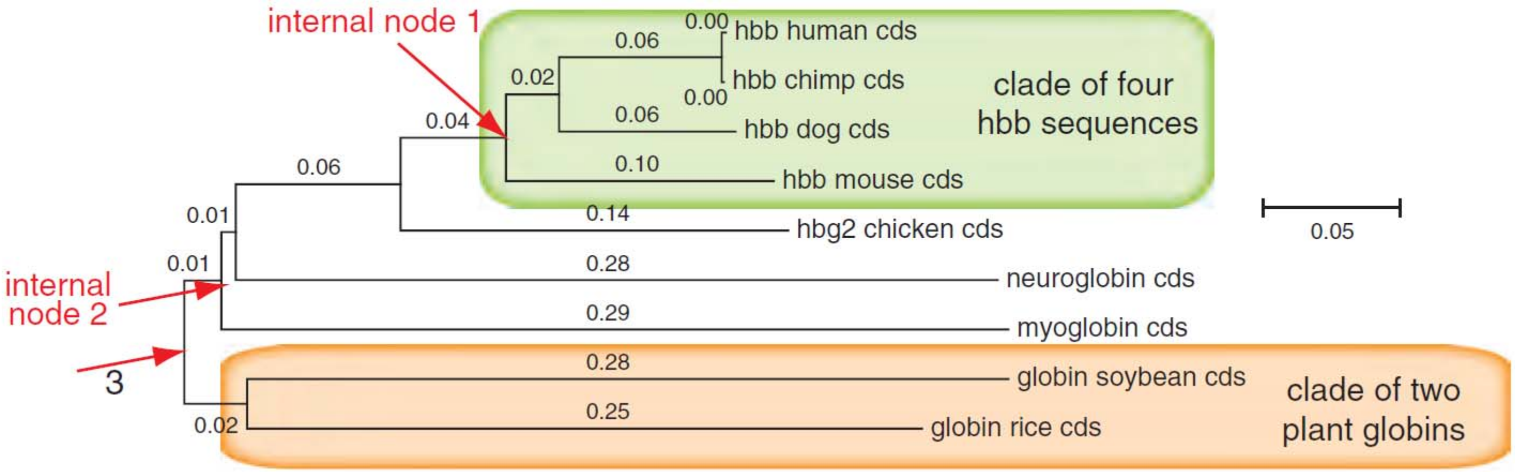
# Several Ways To Build A Phylogenetic Tree *(cont.)*

- Nine globin coding sequences: neighbor-joining tree (radial tree style)



# Branch Lengths Should Be Defined For Every Tree.

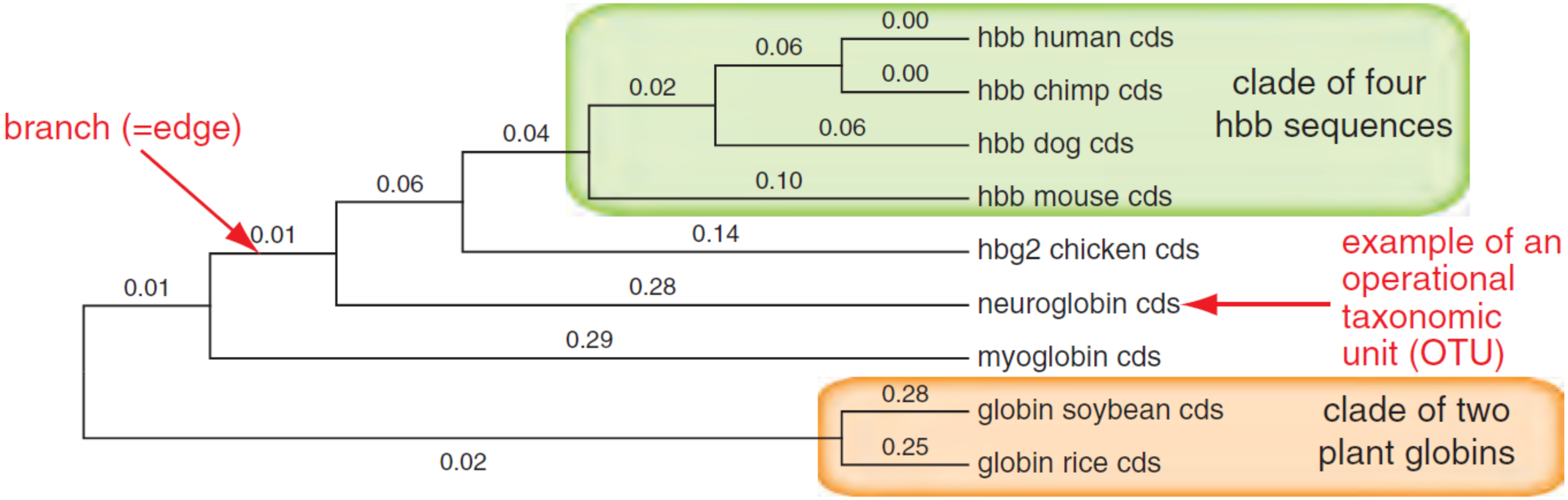
- In some trees, the branch length represents the number of nucleotide or amino acid changes that have occurred in that branch. For example, in the figure below, scale bars are given, and the branch lengths are in units of base differences per site.



- This format (called a phylogram) has the helpful feature of conveying a clear visual idea of the relatedness of different proteins within the tree.

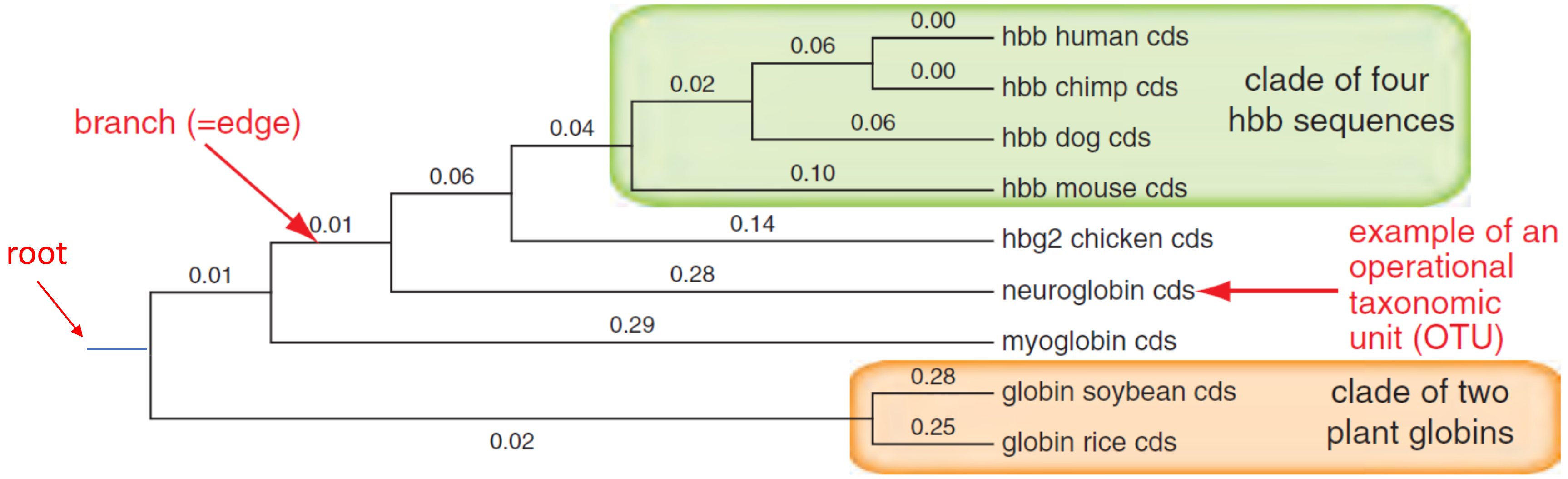
# Unscaled Branch Lengths

- In the figure below, the branches are unscaled. This implies that they are not proportional to the number of changes.
- This form of presenting a tree (called a cladogram) has the advantage of aligning the OTUs neatly in a vertical column. This may be especially useful if the tree has many dozens of OTUs.



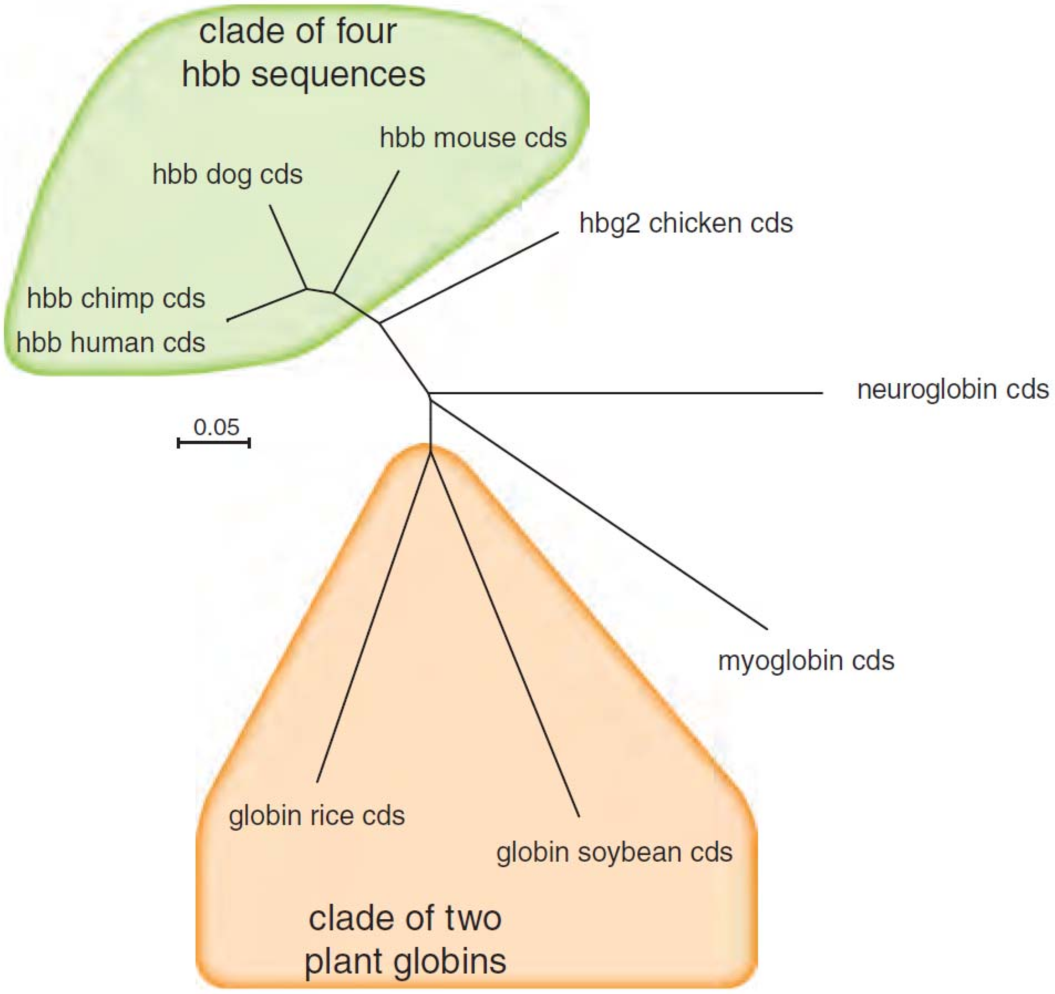
# Tree Roots

- If one assumes a constant molecular clock, then time and distance are proportional.
- The direction of time moves from oldest (at the root) to newest (at the OTUs). Often the root is not known today, and some tree-making algorithms do not provide conjectures about placement of a root.



# Unrooted Trees

- The alternative to a rooted tree is an unrooted tree, shown below. An unrooted tree specifies the relationships among the OTUs.
- However, it does not define the evolutionary path completely or make assumptions about common ancestors.
- If a tree is unrooted you may choose to add a root. The two main ways to do this are by specifying an outgroup and by midpoint rooting.



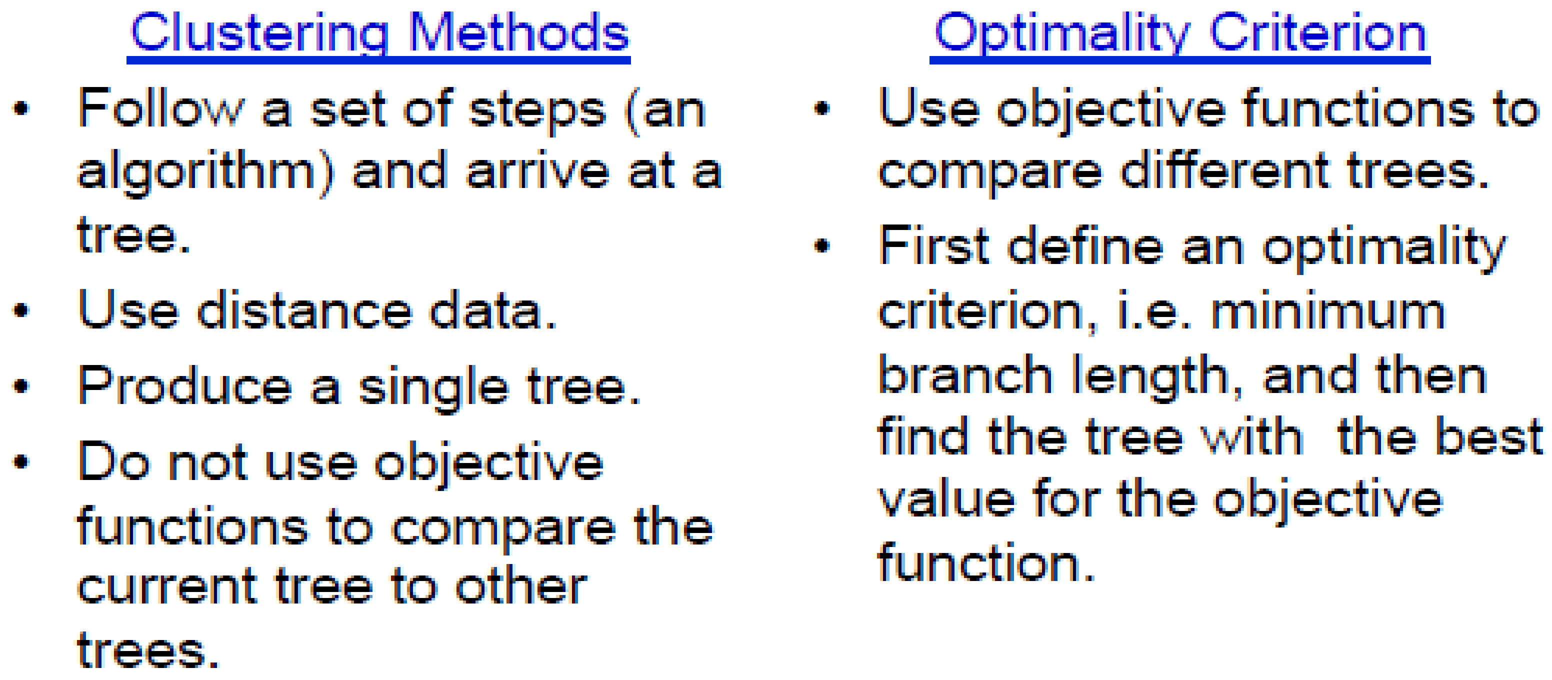
# Distance-Based Phylogenetic Tree Building

- Distance-Based Method: Use the distance between aligned sequences to derive trees.
- NOTE: Mutational Saturation – after one sequence site mutates, subsequent mutations cannot render it any “more” mutated/divergent.
- Subsequent mutations can make sequences equal again. (e.g., valine → isoleucine → valine)

# Phylogenetic Distance Algorithms

- UPGMA: Un-weighted pair group method with arithmetic mean. A clustering method, joins branches based on distance between pairs and average of joined pairs.
- NJ: Neighbor Joining Uses applied with a distance tree. Inserts branches between pairs of closest neighbors and terminals in tree.
- FM: Fitch Margoliash Maximizes fit of observed pair wise distances to a tree by minimizing the squared deviation of all possible observed distances.
- ME: Minimum Evolution Tries to find shortest tree that is consistent with path lengths measured in a manner similar to FM.

# Two Main Types Of Tree Building Methods



## Clustering Methods

- Follow a set of steps (an algorithm) and arrive at a tree.
- Use distance data.
- Produce a single tree.
- Do not use objective functions to compare the current tree to other trees.

## Optimality Criterion

- Use objective functions to compare different trees.
- First define an optimality criterion, i.e. minimum branch length, and then find the tree with the best value for the objective function.

# Strength Of Clustering

- Speed

- Robustness, with parameterization, can be made less or more sensitive to variations in sequences.

- Ability to reconstruct trees for very large numbers (thousands) of sequences.

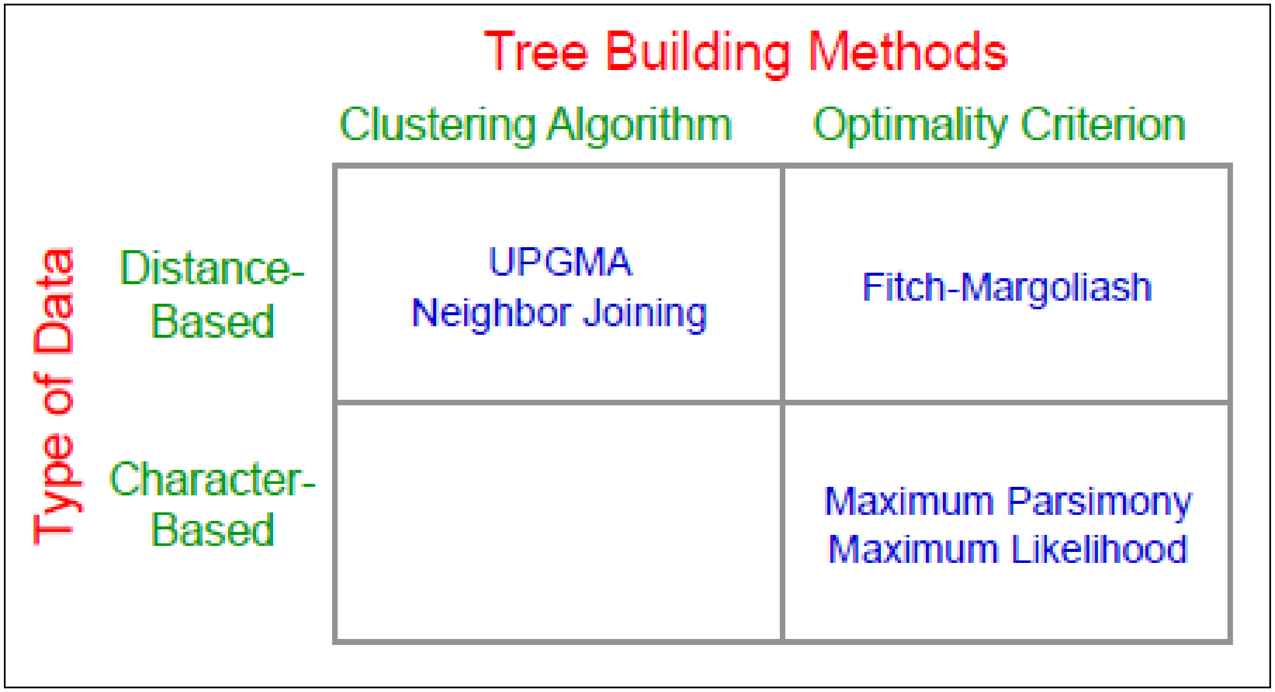
- Most clustering methods reconstruct phylogenetic trees for a set of sequences on the basis of their pairwuse evolutionary distances.

# Strength Of Optimality-Based Methods

- Can be more accurate if you have a good objective function and substitution data.

- Can be used to compare trees.

# Classification Of Tree Building Methods

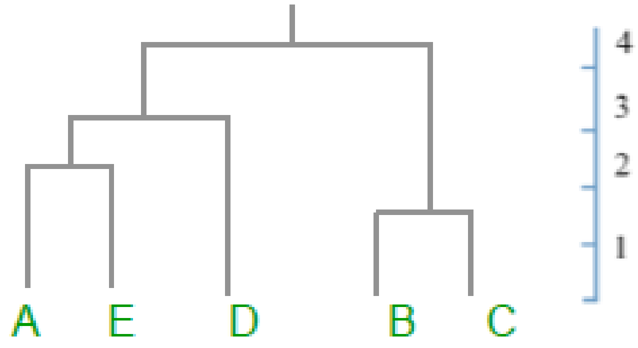


| | | Tree Building Methods | |
|---|---|---|---|
| | | Clustering Algorithm | Optimality Criterion |
| Type of Data | Distance-Based | UPGMA<br>Neighbor Joining | Fitch-Margoliash |
| | Character-Based | | Maximum Parsimony<br>Maximum Likelihood |

# Distance-Based Method

- Given: an *n*× *n* matrix *M*, where *M(i,j)* is the distance between objects *i* and *j*
- Build an edge-weighted tree such that the distances between leaves *i* and *j* correspond to *M(i,j)*

|   | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | 8 | 8 | 5 | 3 |
| B |  | 0 | 3 | 8 | 8 |
| C |  |  | 0 | 8 | 8 |
| D |  |  |  | 0 | 5 |
| E |  |  |  |  | 0 |



# UPGMA

## Un-weighted pair group method with arithmetic mean

# UPGMA: Un-Weighted pair group method with arithmetic mean

- Clusters sequences at each stage of amalgamating two operational taxonomic units (OTUs) and at the same time creating a new node in the tree.

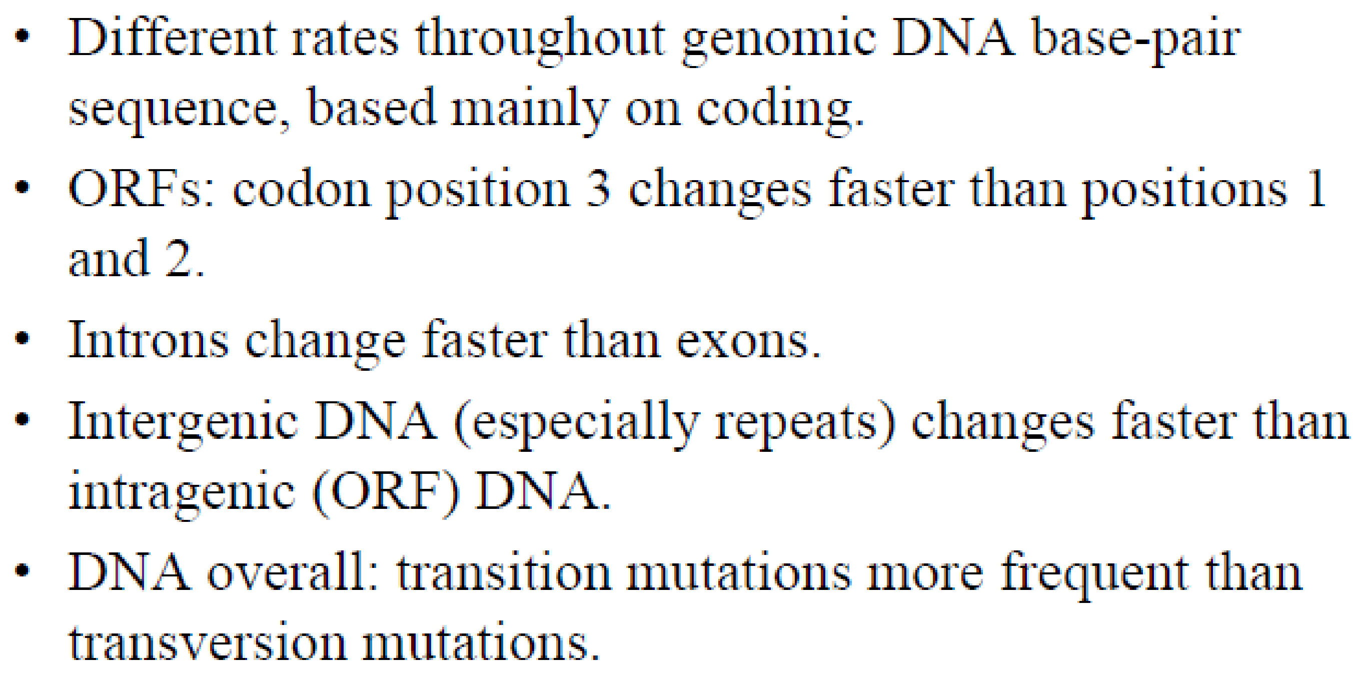
- The edge lengths are determined by the difference in the heights of the nodes at the top and bottom of an edge.

# The Molecular Clock

- UPGMA assumes that:
    - – the gene/amino acid substitution rate is constant, in other words: divergence of sequences is assumed to occur at the same rate at all points in the tree.
    - Known as the Molecular Clock.

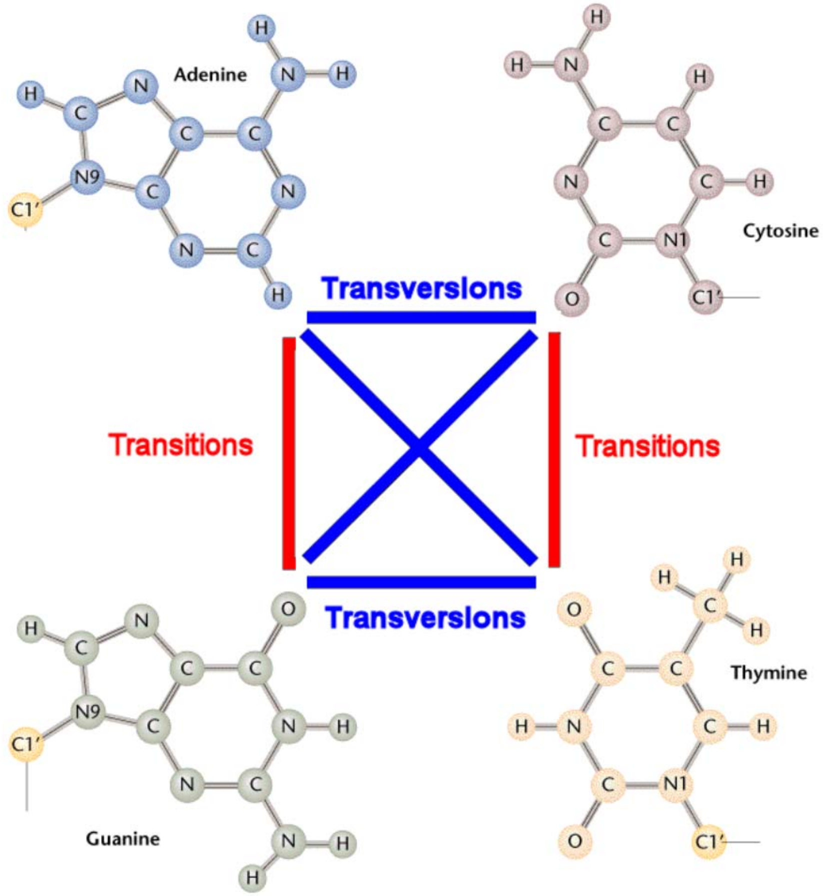
- The distance is linear with evolutionary time.

# Rates Of Evolutionary Change



- Different rates throughout genomic DNA base-pair sequence, based mainly on coding.
- ORFs: codon position 3 changes faster than positions 1 and 2.
- Introns change faster than exons.
- Intergenic DNA (especially repeats) changes faster than intragenic (ORF) DNA.
- DNA overall: transition mutations more frequent than transversion mutations.

# Transition vs Transversion Mutations



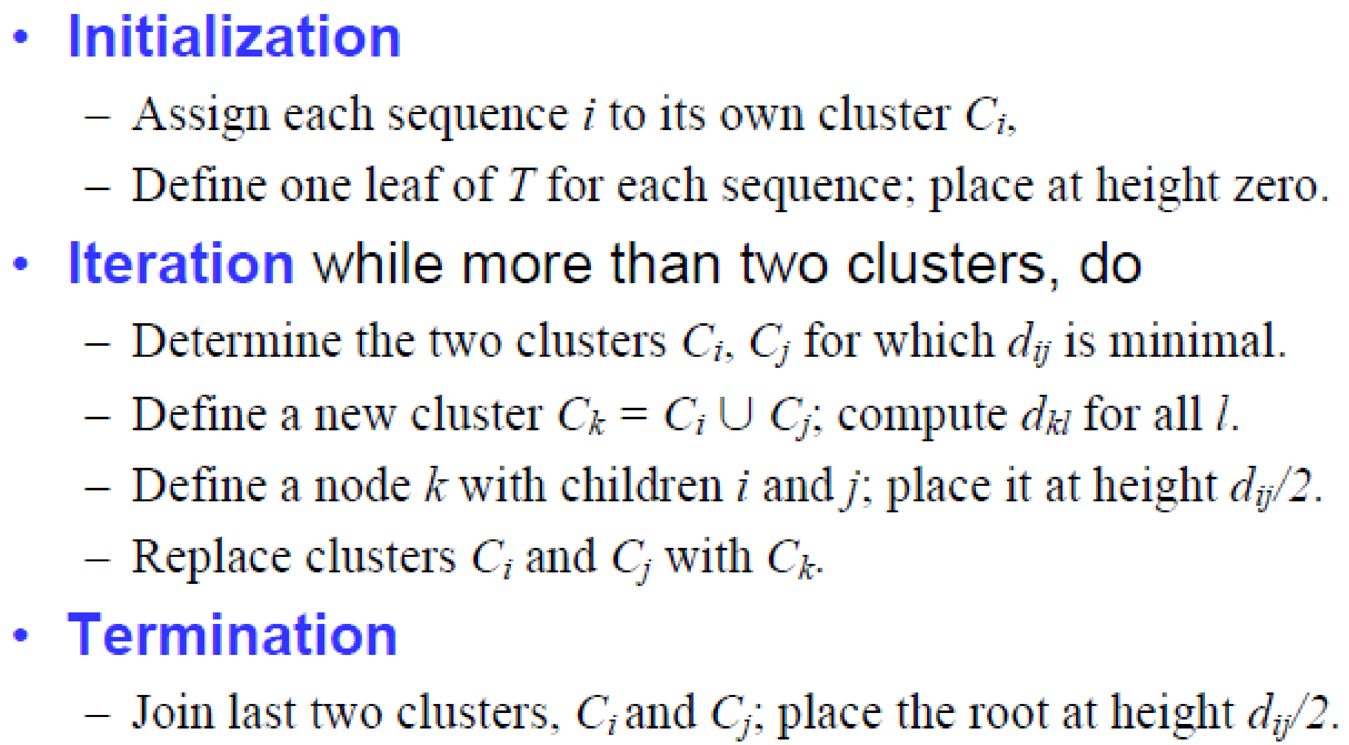
Transition *versus* Transversion mutations

# UPGMA Algorithm

- The algorithm iteratively picks two clusters and merges them, thus creating a new node in the tree.
- The average distance between two clusters is determined by:

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i,\, q \in C_j} d_{pq}, \text{ where } C_i \text{ and } C_j \text{ are clusters.}$$

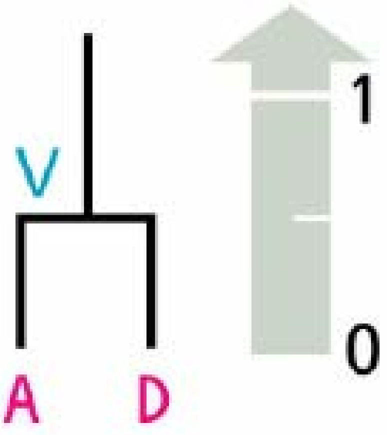
# UPGMA Algorithm



- **Initialization**
  - Assign each sequence $i$ to its own cluster $C_i$,
  - Define one leaf of $T$ for each sequence; place at height zero.
- **Iteration** while more than two clusters, do
  - Determine the two clusters $C_i$, $C_j$ for which $d_{ij}$ is minimal.
  - Define a new cluster $C_k = C_i \cup C_j$; compute $d_{kl}$ for all $l$.
  - Define a node $k$ with children $i$ and $j$; place it at height $d_{ij}/2$.
  - Replace clusters $C_i$ and $C_j$ with $C_k$.
- **Termination**
  - Join last two clusters, $C_i$ and $C_j$; place the root at height $d_{ij}/2$.

# UPGMA: Example (1st Iteration)

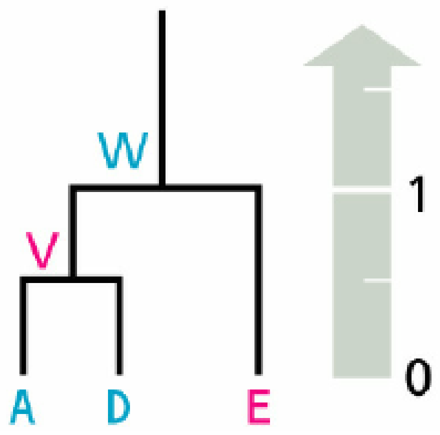
| $d_{ij}$ | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | – | 6 | 8 | 1 | 2 | 6 |
| B |  | – | 8 | 6 | 6 | 4 |
| C |  |  | – | 8 | 8 | 8 |
| D |  |  |  | – | 2 | 6 |
| E |  |  |  |  | – | 6 |



# UPGMA: Example (2nd Iteration)

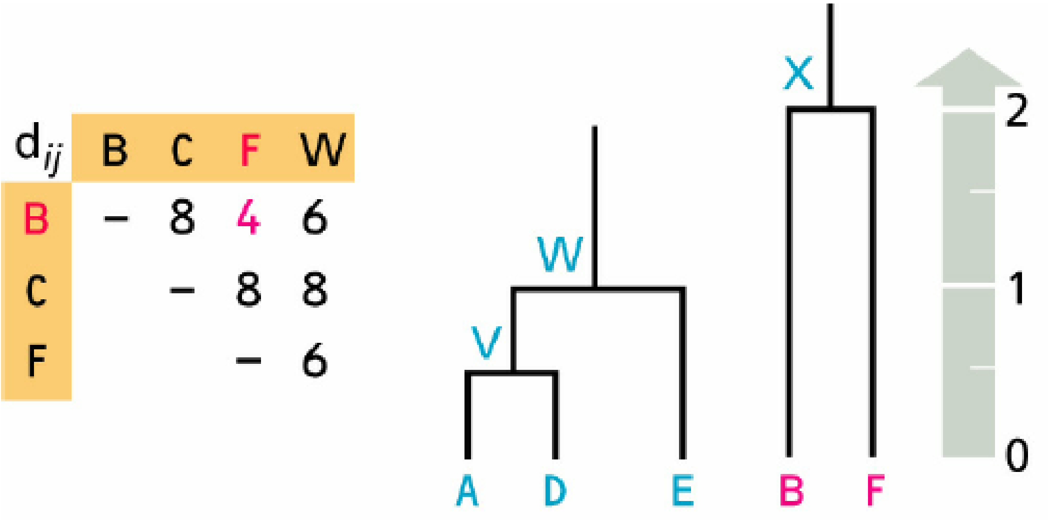
The table of distances is updated to reflect the average distances from V to the other sequences.
V and E are the closest and are combined to create a new cluster W of height 1 in T.

| $d_{ij}$ | B | C | E | F | V |
|---|---|---|---|---|---|
| B | – | 8 | 6 | 4 | 6 |
| C | | – | 8 | 8 | 8 |
| E | | | – | 6 | 2 |
| F | | | | – | 6 |



# UPGMA: Example ($3^{rd}$ Iteration)

After updating the table of distances, B and F are the closest sequences and are combined to create a new cluster X of height 2 in T.

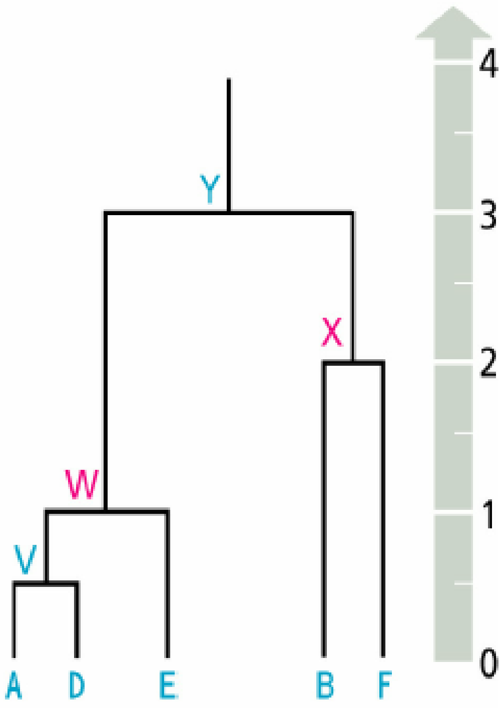


| $d_{ij}$ | B | C | F | W |
|---|---|---|---|---|
| B | – | 8 | 4 | 6 |
| C | | – | 8 | 8 |
| F | | | – | 6 |

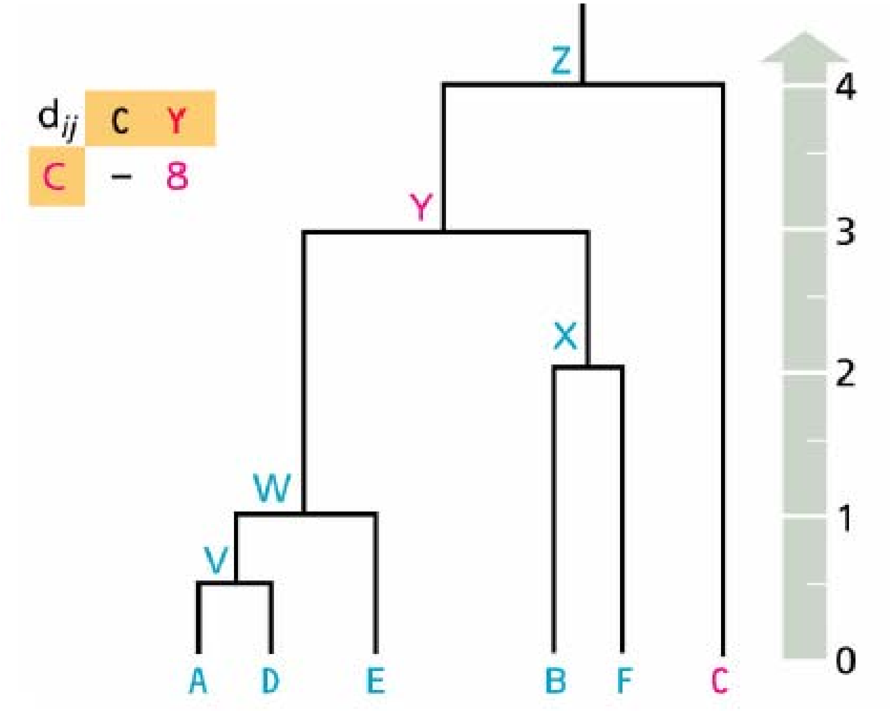
# UPGMA: Example (4th Iteration)

Once more the table is updated. W and X are the closest sequences and are combined to create a new cluster Y of height 3 in T.

| $d_{ij}$ | C | W | X |
|---|---|---|---|
| C | - | 8 | 8 |
| W | | - | 6 |



# UPGMA: Example (Completion)

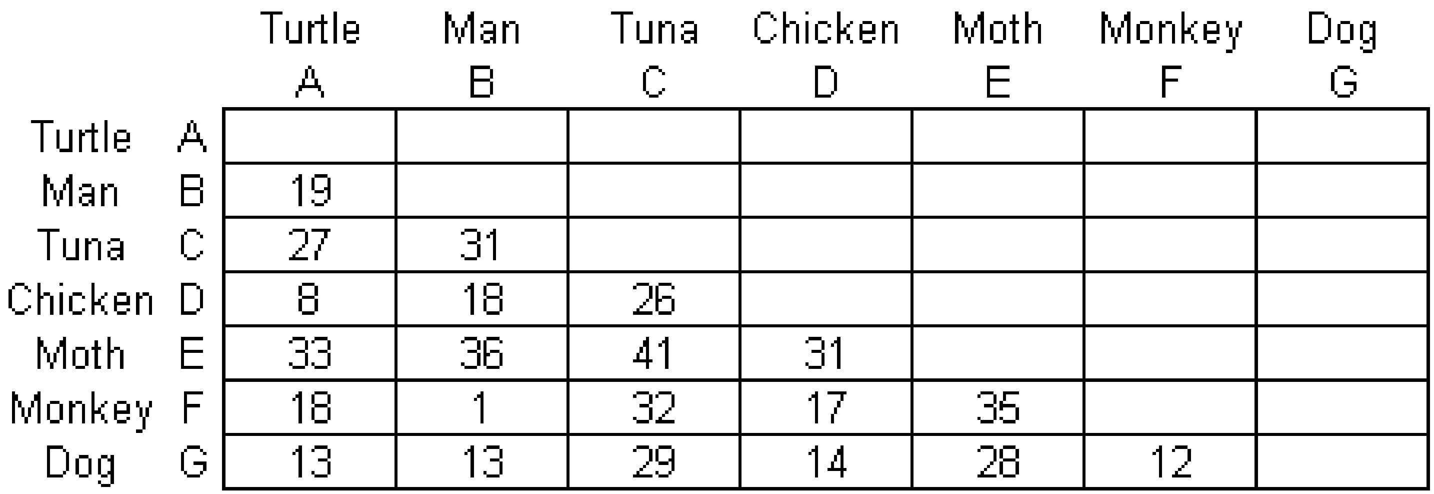


| $d_{ij}$ | C | Y |
|---|---|---|
| C | – | 8 |

The remaining 2 sequences, C and Y of distance 8 are combined to create a new cluster Z of height 4 in T.

# In-Class Lecture Exercise

- Build the Phylogenetic tree using the UPGMA method and the following distance table. Must show work in submission. Does the tree you build make sense? Explain. Answer on next slide.



|  |  | Turtle | Man | Tuna | Chicken | Moth | Monkey | Dog |
|---|---|---|---|---|---|---|---|---|
|  |  | A | B | C | D | E | F | G |
| Turtle | A |  |  |  |  |  |  |  |
| Man | B | 19 |  |  |  |  |  |  |
| Tuna | C | 27 | 31 |  |  |  |  |  |
| Chicken | D | 8 | 18 | 26 |  |  |  |  |
| Moth | E | 33 | 36 | 41 | 31 |  |  |  |
| Monkey | F | 18 | 1 | 32 | 17 | 35 |  |  |
| Dog | G | 13 | 13 | 29 | 14 | 28 | 12 |  |

# Answer To In-Class Exercise

# UPGMA Can Be Used To Root an Unrooted Tree Next Class