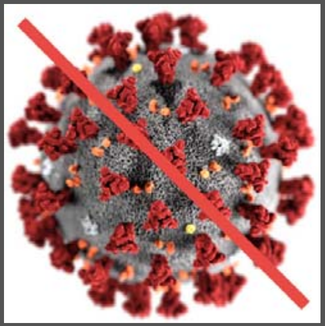


CS123A Bioinformatics Module 3 – Week 9 – Presentation 2

Leonard Wesley

Computer Science Dept

San Jose State Univ

"Just work till midnight, you need to relax too"

Agenda

- Phylogenetic Trees
  - Section **“Molecular Phylogeny: Properties of Trees”** starting on page 259 in textbook
  - Hierarchical Clustering & UPGMA Tree Building Example
  - Distance measures, Pros & Cons
  - Molecular Clocks

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.

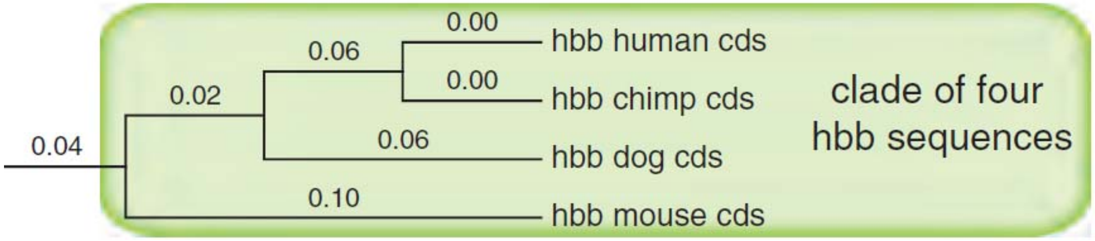
	Seq1	Seq2	Seq3	Seq4
Seq1	-	3	1	4
Seq2	3	-		
Seq3	1	7	-	3
Seq4	4	0.5	3	-

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

Some Definitions First

Two Types Of Information In Phylogenetic Trees

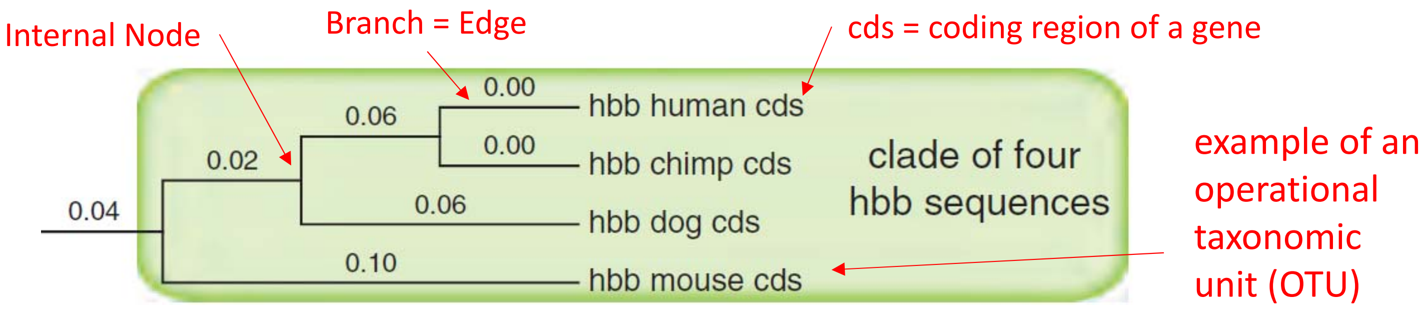
- Topology: *Defines the relationships of the proteins (or other objects) that are represented in the tree. For example, the topology in the tree shows the common ancestor of two homologous protein sequences.*



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

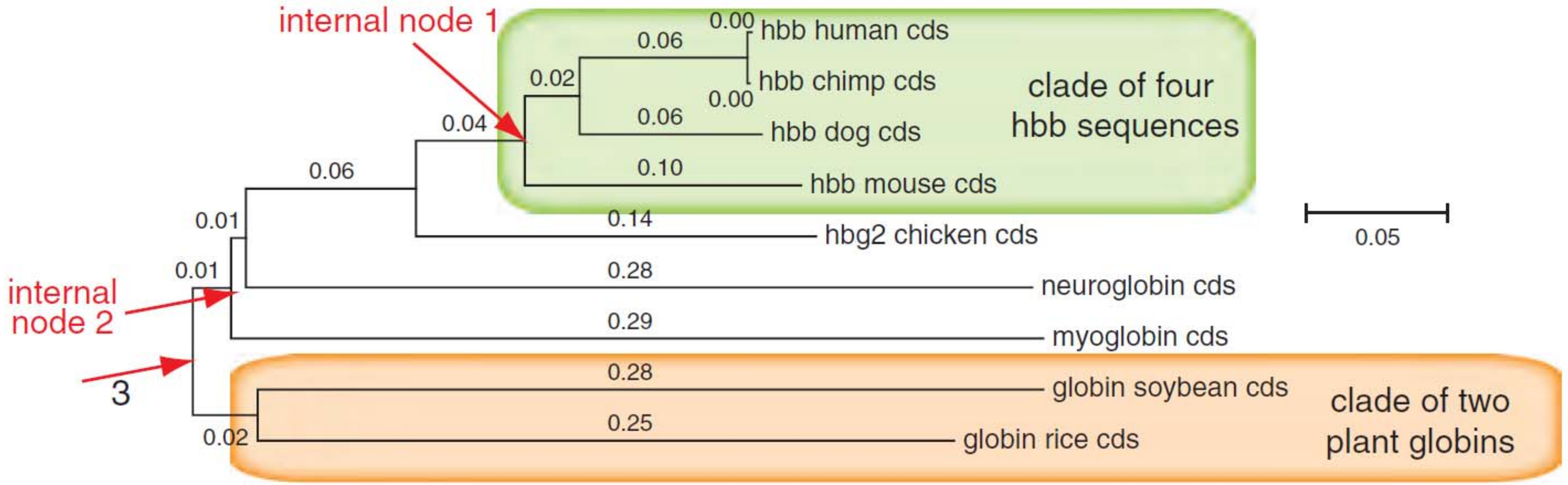
Parts Of A Phylogenetic Tree: Branches & Nodes

- 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)



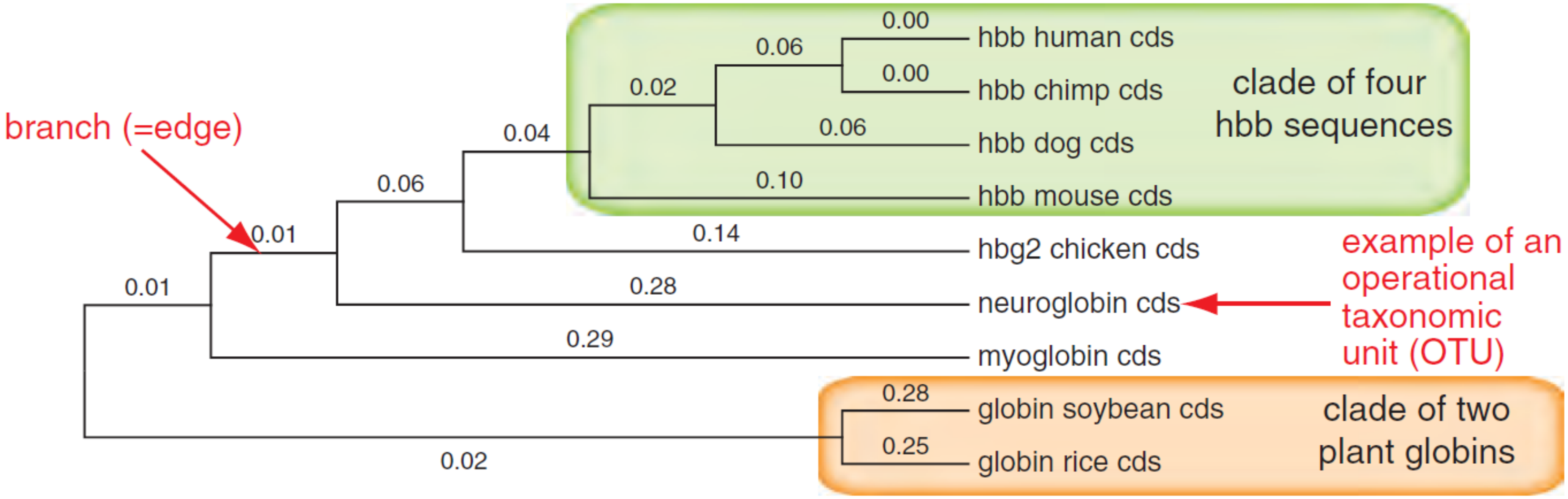
Several Ways To Build A Phylogenetic Tree

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



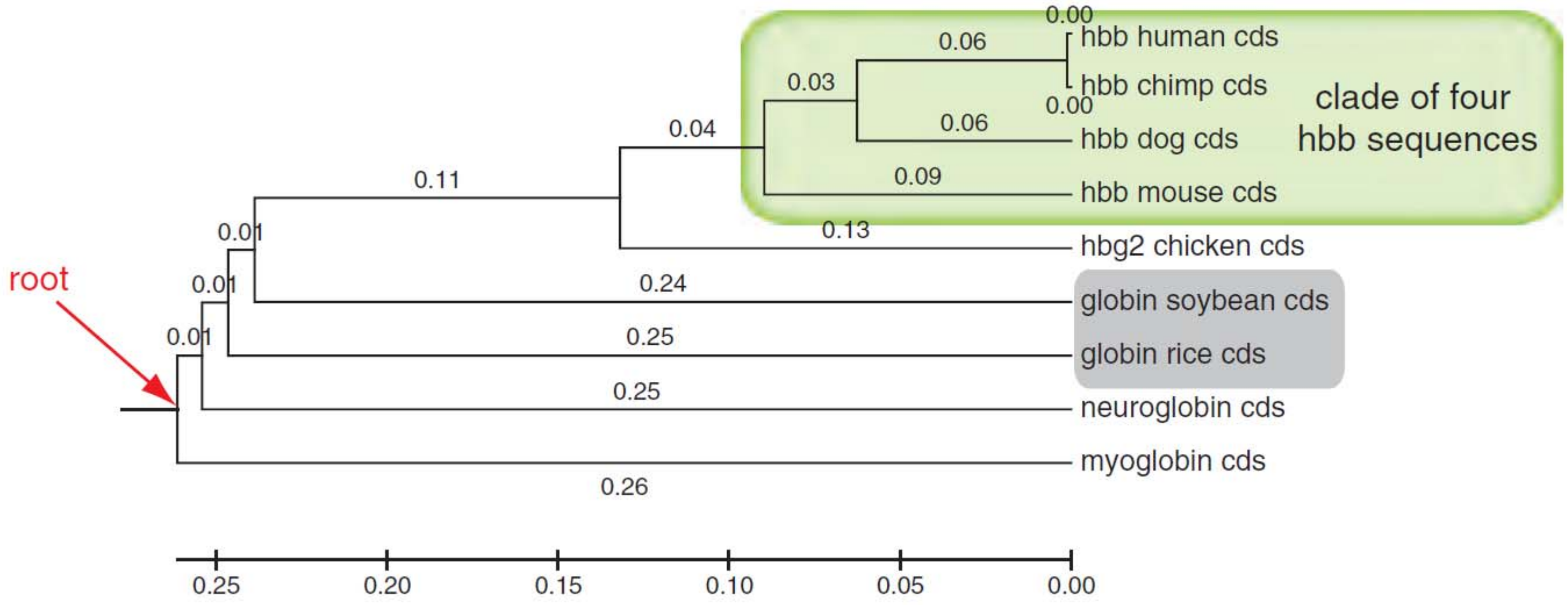
Several Ways To Build A Phylogenetic Tree *(cont.)*

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



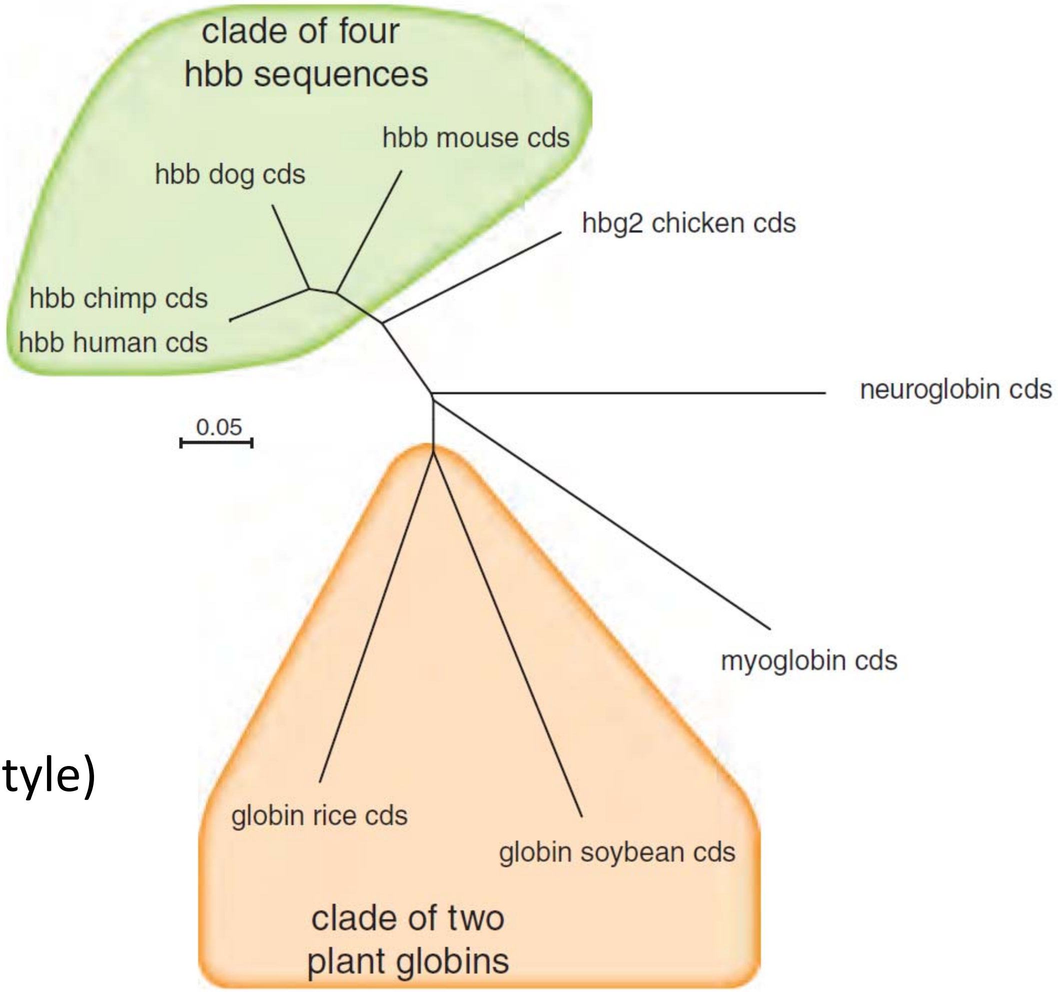
Several Ways To Build A Phylogenetic Tree *(cont.)*

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



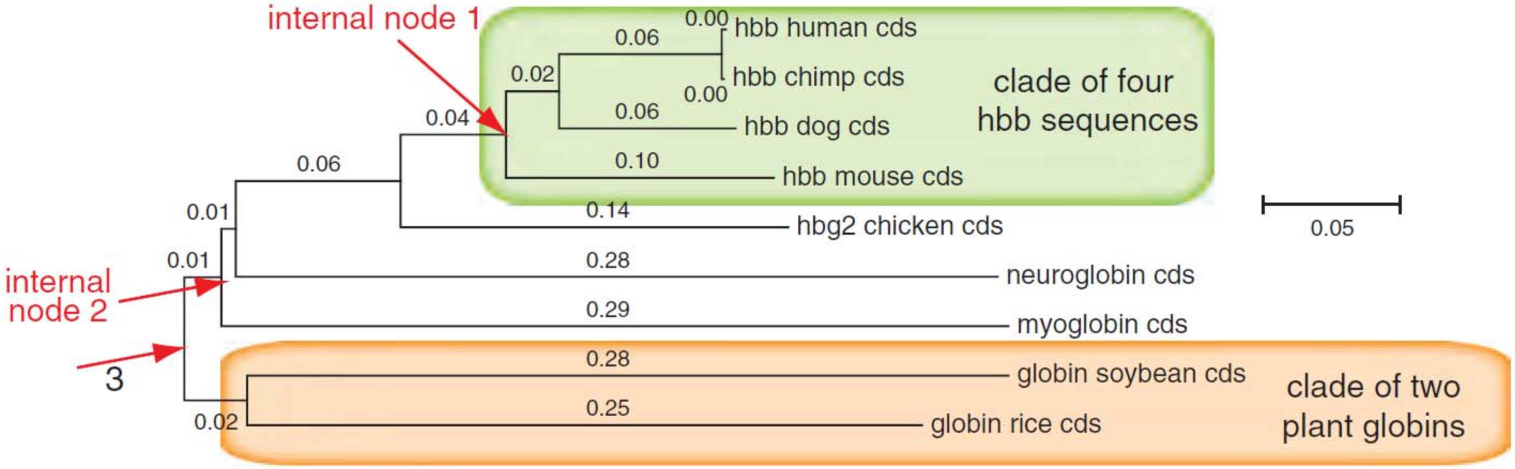
Several Ways To Build A Phylogenetic Tree *(cont.)*

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



Branch Lengths Should Be Defined For Every Tree.

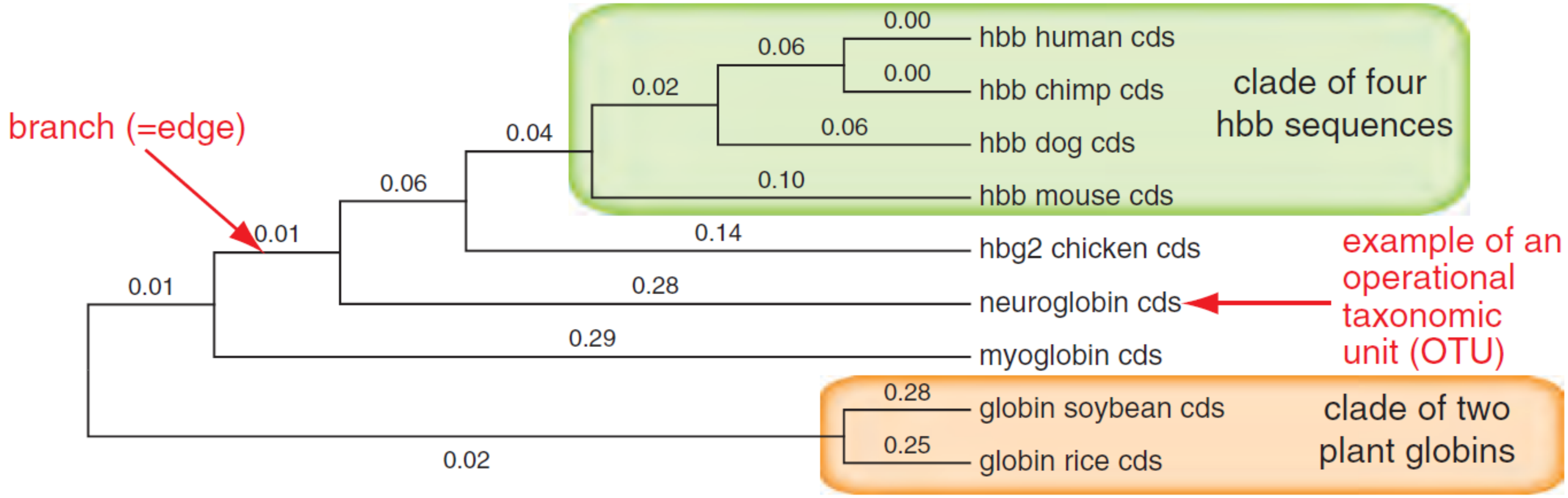
- In some trees, the branch length represents the number of nucleotide or amino acid changes that have occurred in that branch. For example, in the figure below, scale bars are given, and the branch lengths are in units of base differences per site.



- 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

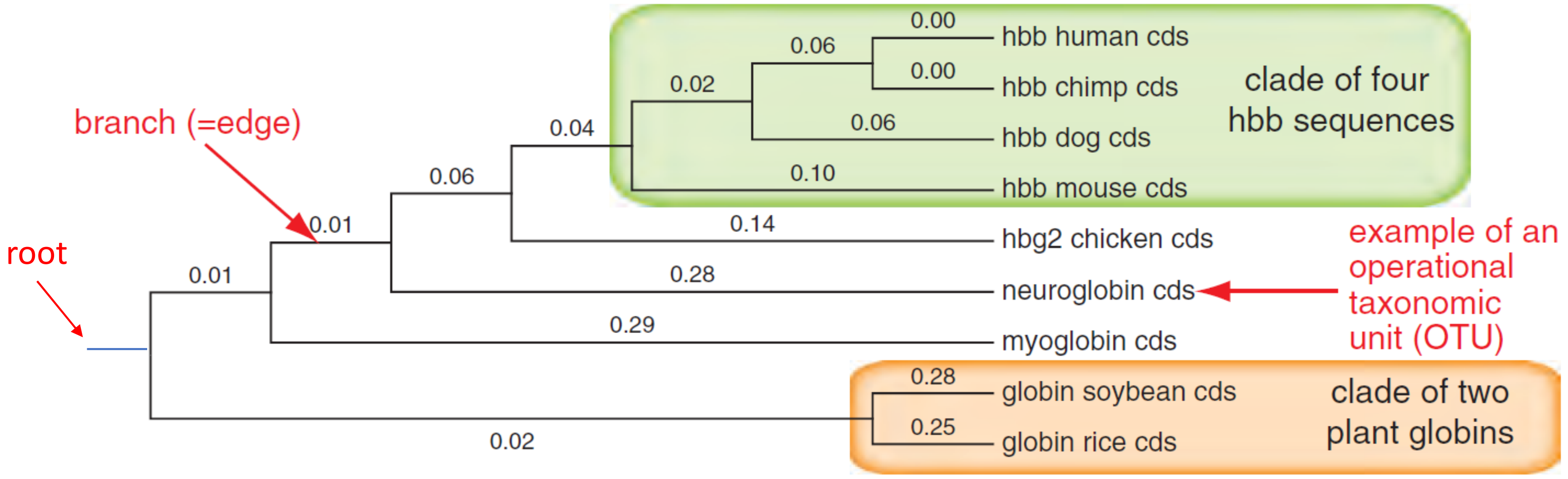
- In the figure below, the branches are unscaled. This implies that they are not proportional to the number of changes.
- This form of presenting a tree (called a cladogram) has the advantage of aligning the OTUs neatly in a vertical column. This may be especially useful if the tree has many dozens of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

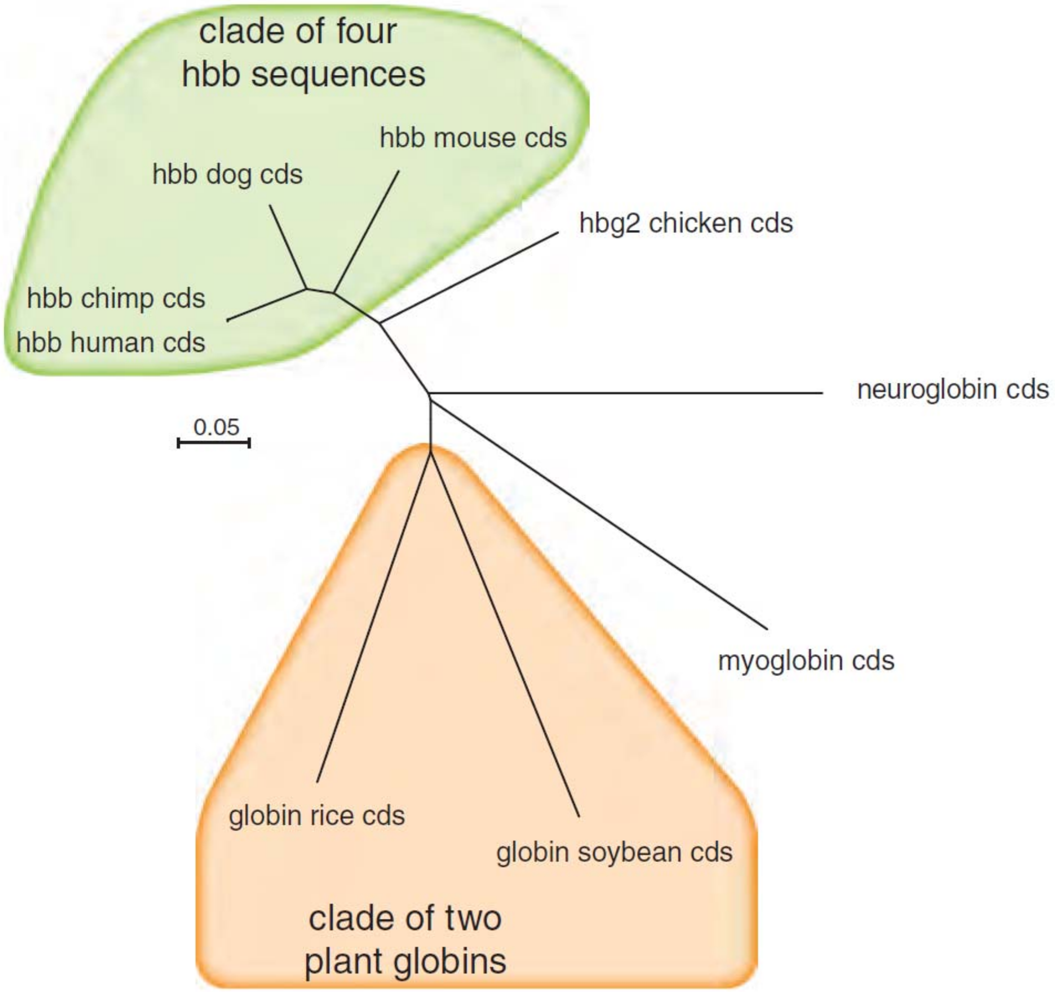


Unrooted Trees

- 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.

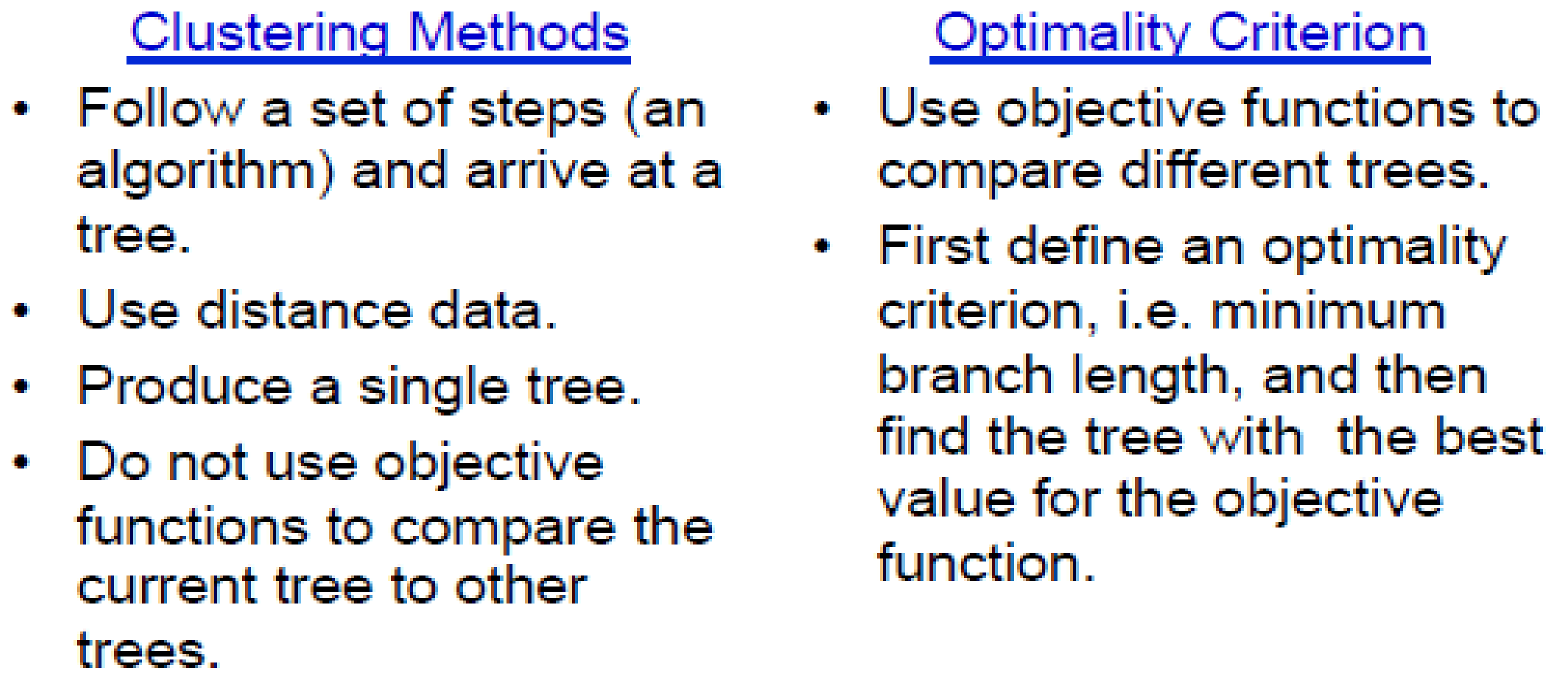
- NOTE: Mutational Saturation – after one sequence site mutates, subsequent mutations cannot render it any “more” mutated/divergent.

- 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 a distance tree. Inserts branches between pairs of closest neighbors and terminals in the 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

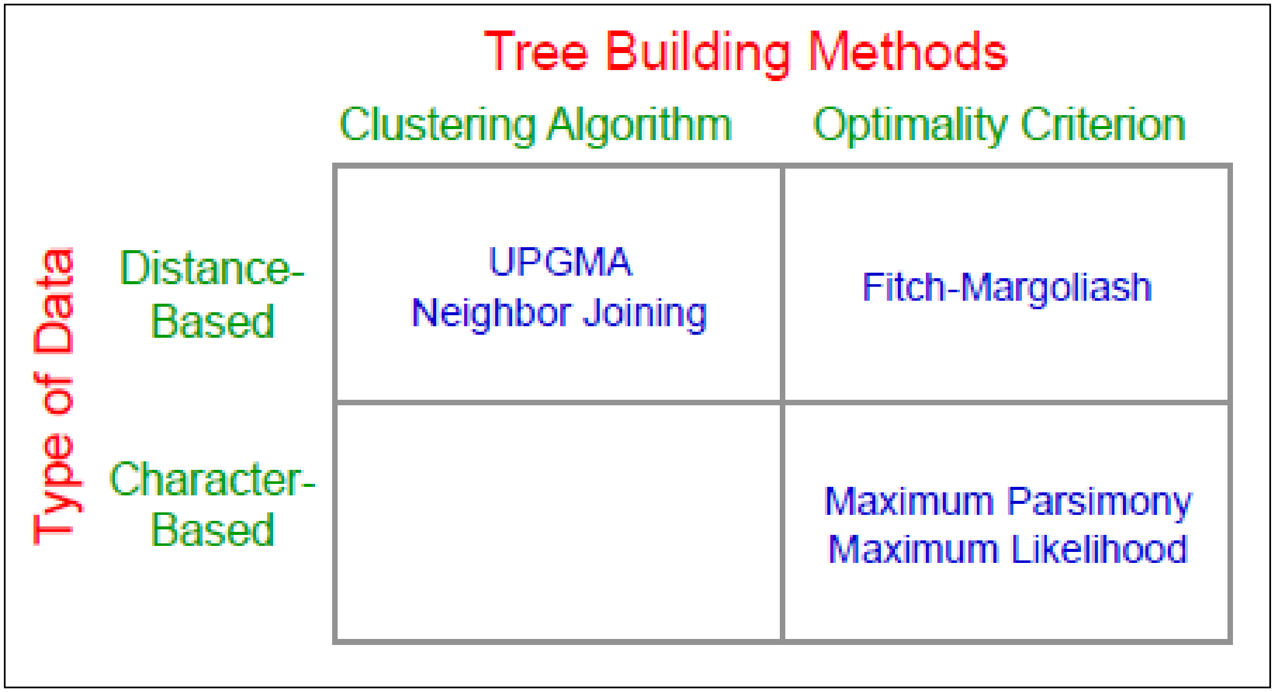
- Speed
- Robustness, with parameterization, can be made less or more sensitive to variations in sequences.
- Ability to reconstruct trees for very large numbers (thousands) of sequences.
- Most clustering methods reconstruct phylogenetic trees for a set of sequences on the basis of their pairwuse evolutionary 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

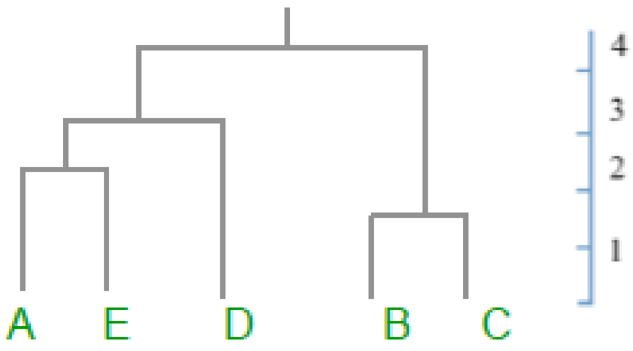


		Tree Building Methods	
		Clustering Algorithm	Optimality Criterion
Type of Data	Distance-Based	UPGMA Neighbor Joining	Fitch-Margoliash
	Character-Based		Maximum Parsimony Maximum Likelihood

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.

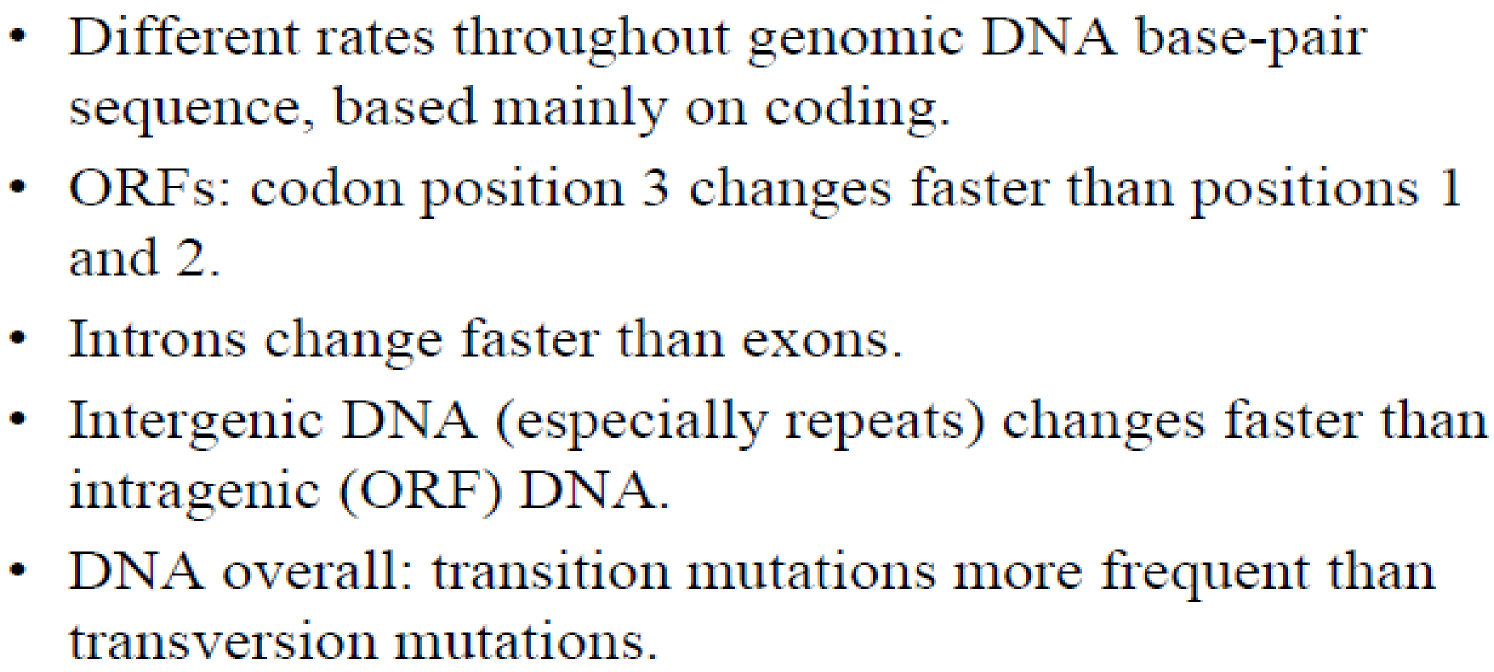
- The edge lengths are determined by the difference in the heights of the nodes at the top and bottom of an edge.

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.

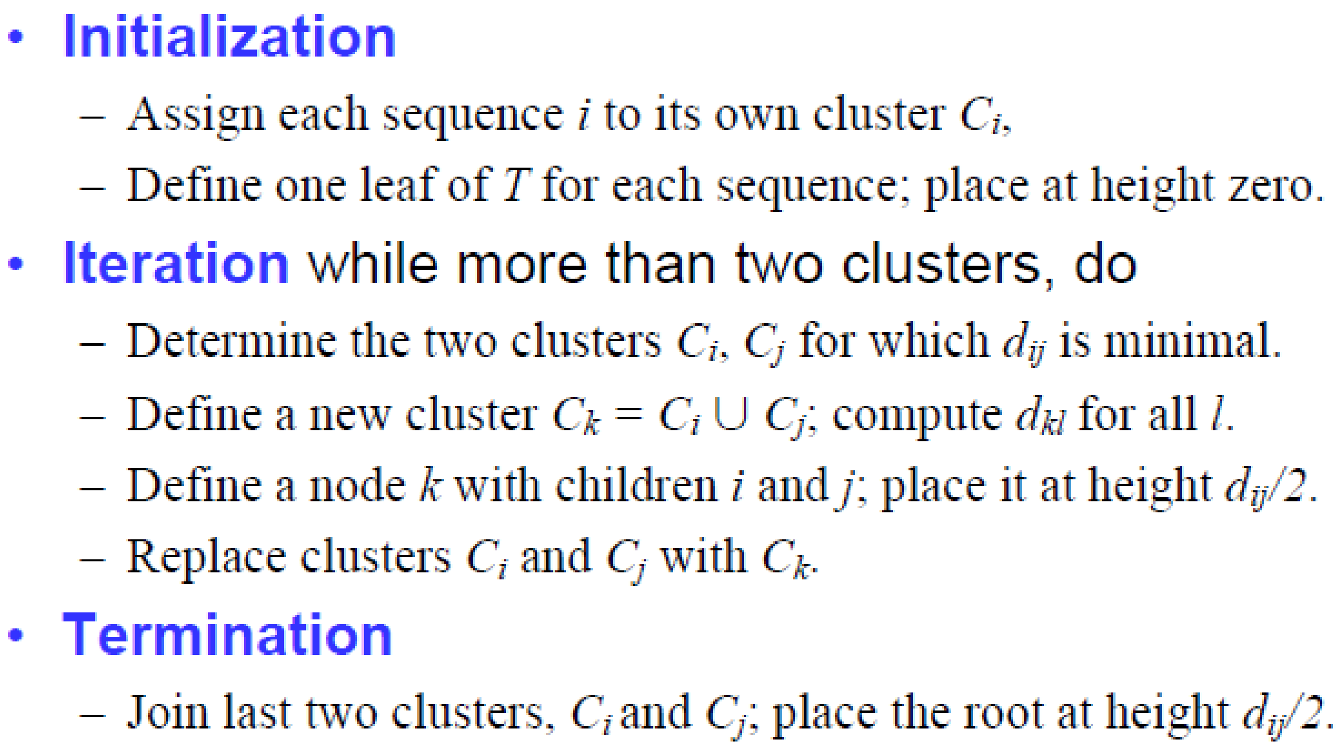
- The distance is linear with evolutionary time.

Rates Of Evolutionary Change



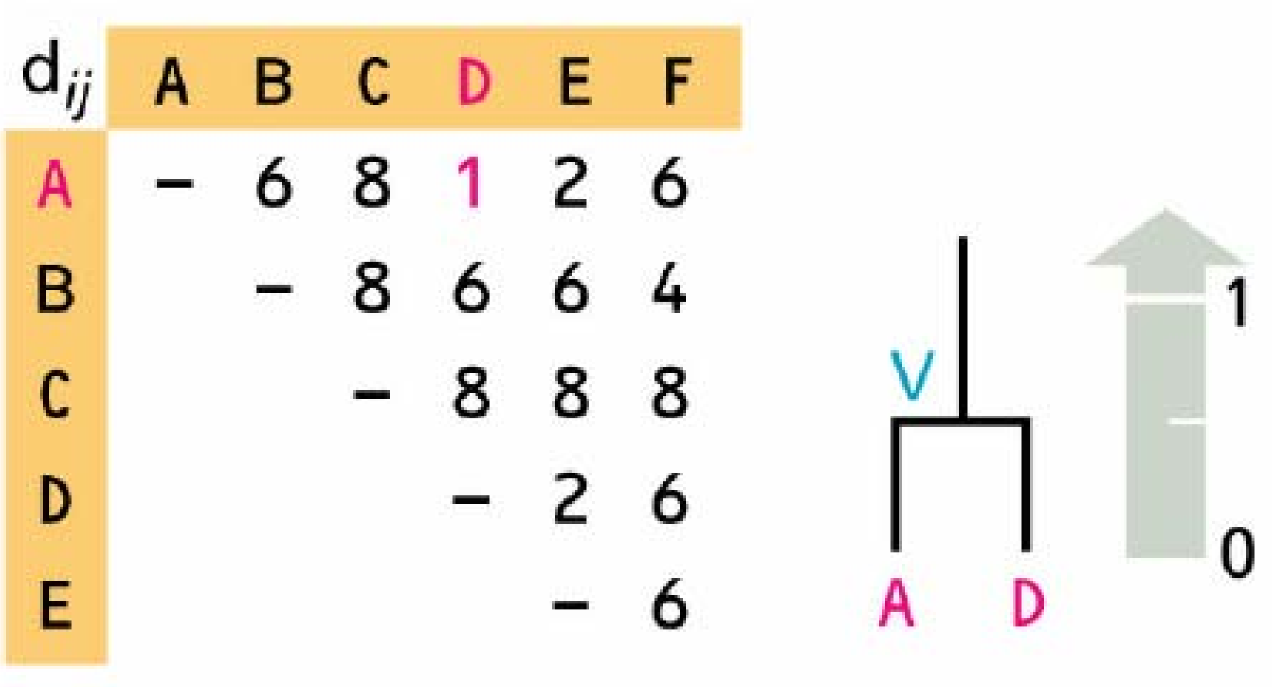
- Different rates throughout genomic DNA base-pair sequence, based mainly on coding.
- ORFs: codon position 3 changes faster than positions 1 and 2.
- Introns change faster than exons.
- Intergenic DNA (especially repeats) changes faster than intragenic (ORF) DNA.
- DNA overall: transition mutations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

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)

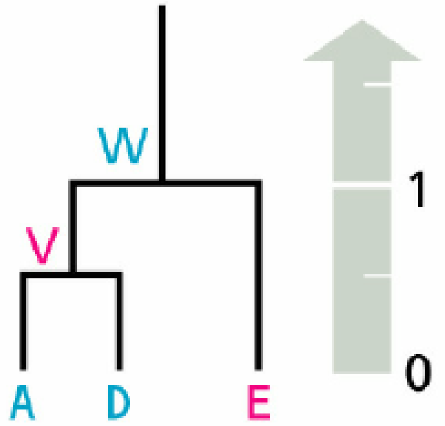


d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

UPGMA: Example (2nd Iteration)

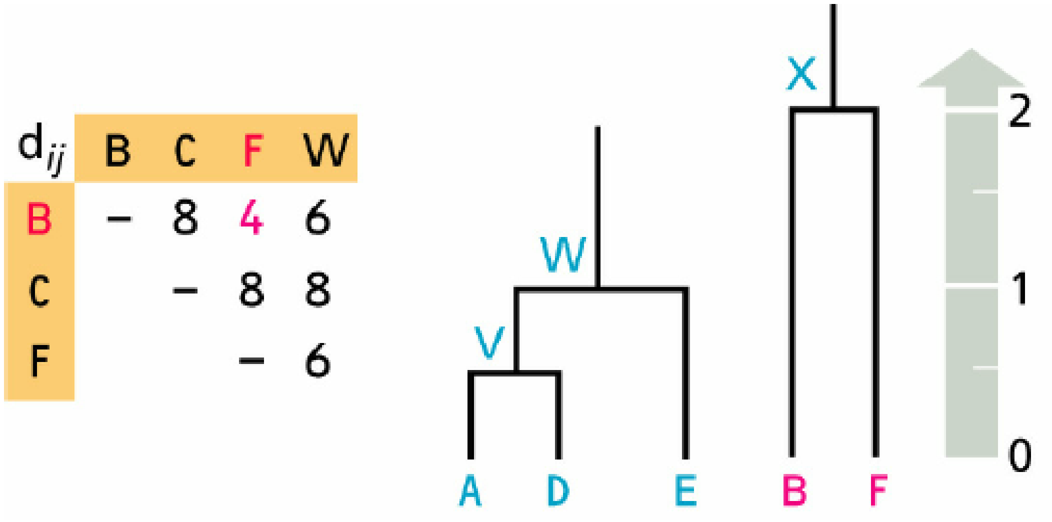
The table of distances is updated to reflect the average distances from V to the other sequences.
V and E are the closest and are combined to create a new cluster W of height 1 in T.

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



UPGMA: Example (3rd 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.

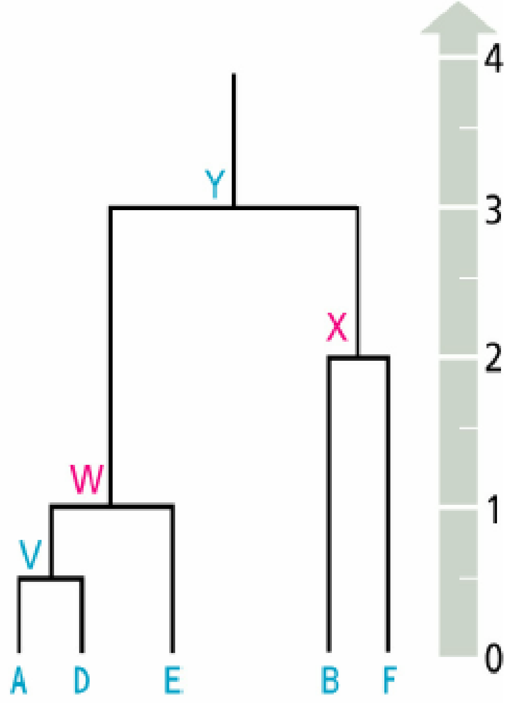


d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6

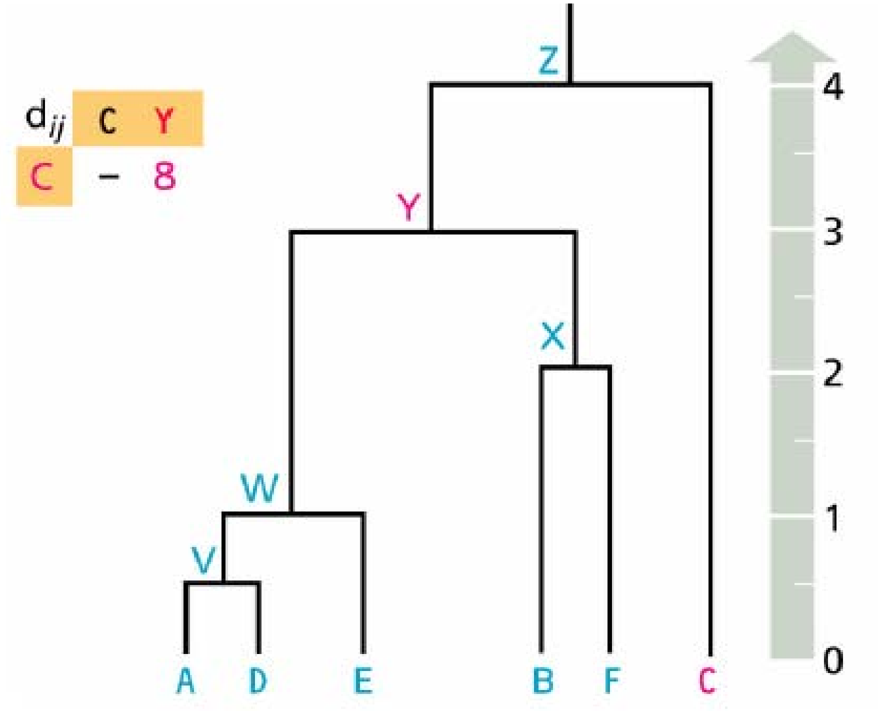
UPGMA: Example (4th Iteration)

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)

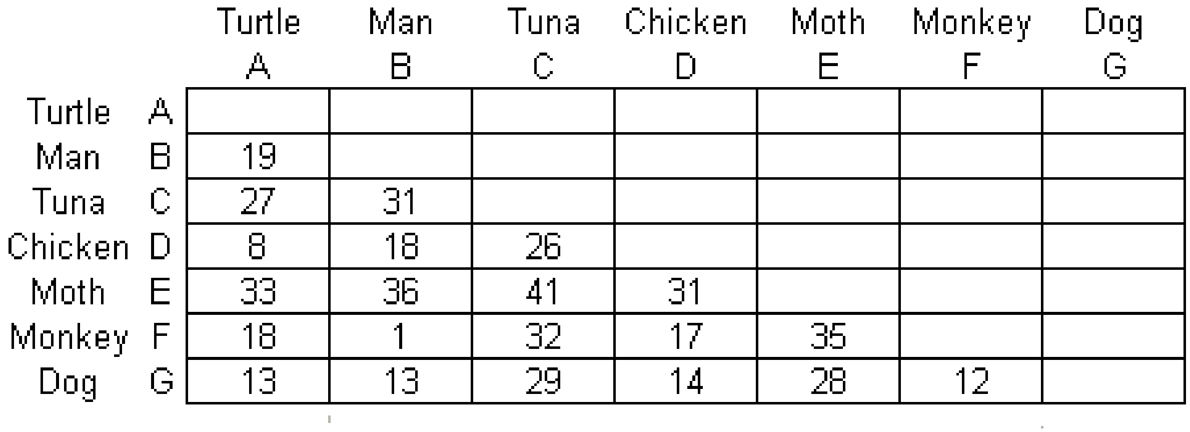


d_{ij}	C	Y
C	–	8

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

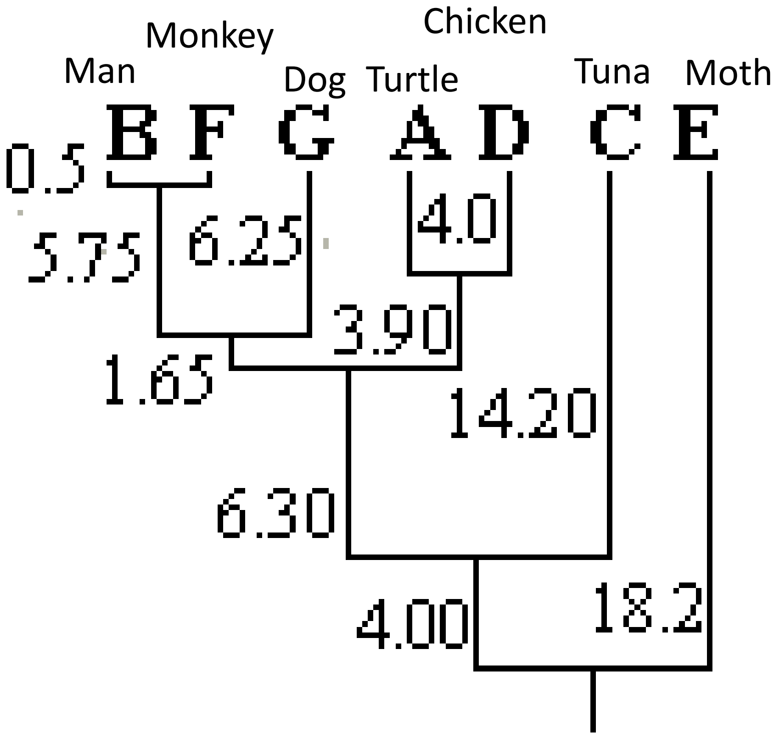
In-Class Lecture Exercise

- 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 A	Man B	Tuna C	Chicken D	Moth E	Monkey F	Dog 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	

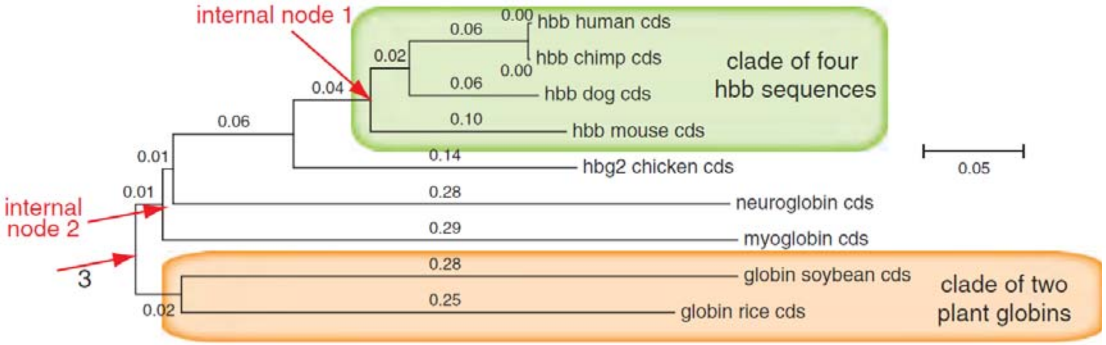
Answer To In-Class Exercise

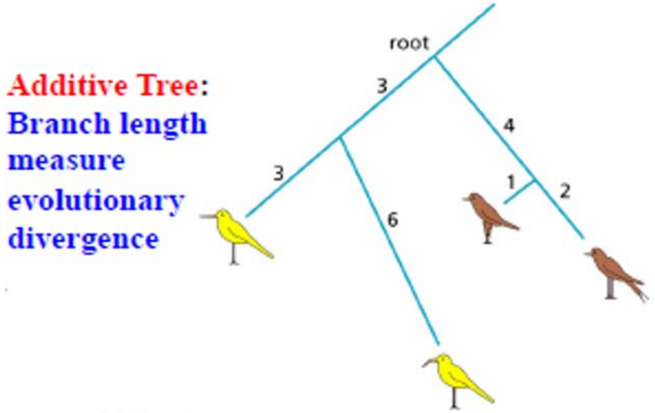


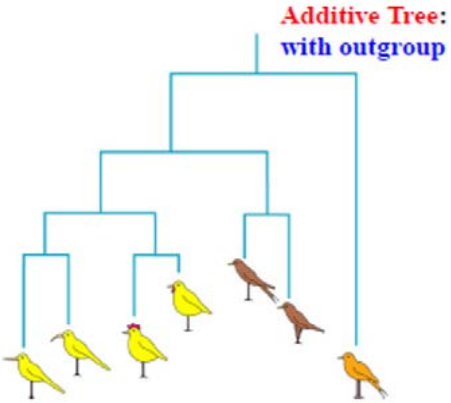
Phylogenetic Trees
Called By Several Other Name(s)

Rectangle/ Additive Tree Styles

(Type of Neighbor Joining Trees)

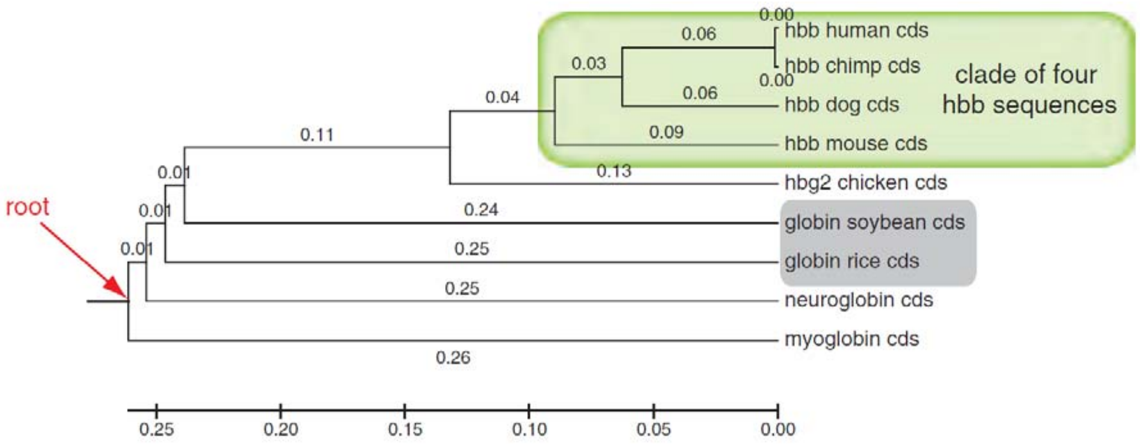


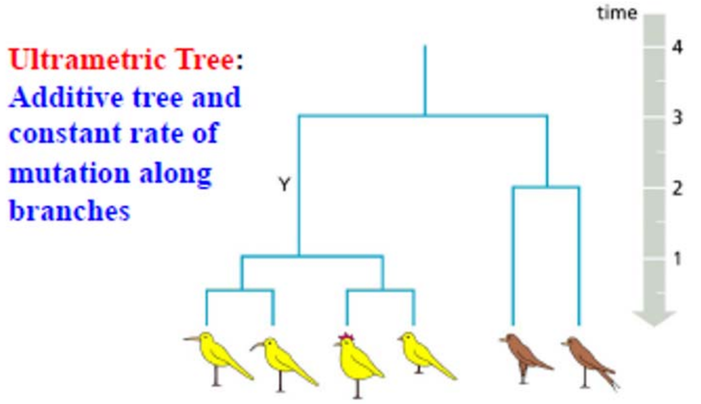




Topology/Ultrametric Trees

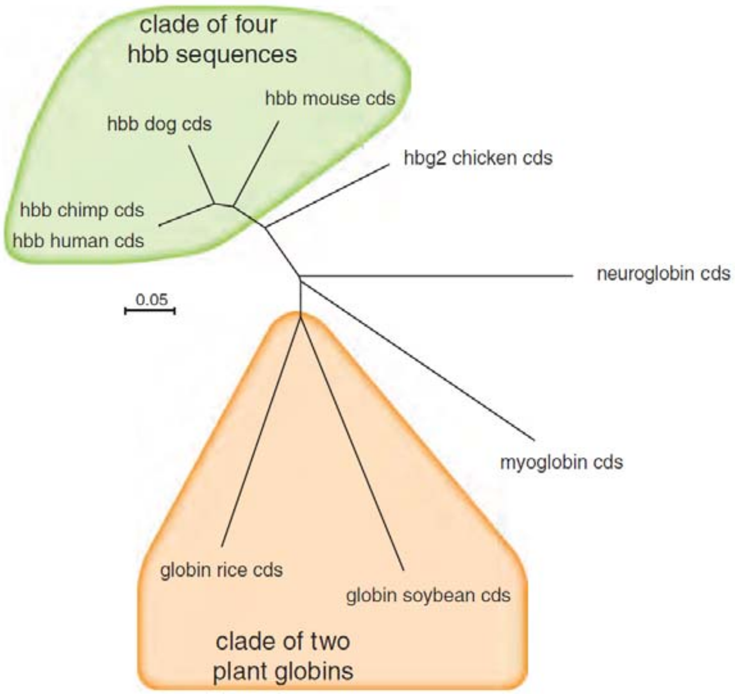
(Type of Neighbor Joining Trees)

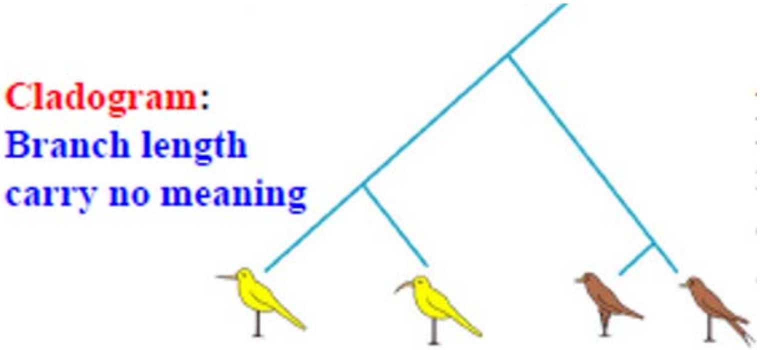



Ultrametric Tree: Additive tree and constant rate of mutation along branches

Produced via UPGMA algorithm

Cladogram/Unrooted Tree




Cladogram:
Branch length
carry no meaning

Limitations Of Distance-Based Methods

- A distance-based phylogenetic tree is derived from the pairwise distance of aligned sequences and not from the original sequence data.

- The distance information may not contain all the sequence information.

Character-Based Approaches

- Sometimes we do not have a distance metric between the species we are interested in.

- What we might have instead, are observable features.

- We use the observable features to build the tree. These trees are called Character-Based trees

Properties Of Character-Based Trees

- The building of the tree is based on morphological features and not on distances.

- Examples of morphological features:
  - – has feathers
  - – has a backbone
  - – has a certain amino acid at a certain position in the sequence
  - – whether or not a certain protein regulates another protein.

Some Information About Pairwise Distances

- Many tree building methods are based on the notion of distances between pairs of sequences.

- What is the nature (i.e., type) of distance measures used to build trees, and how are they calculated to arrive at a numeric (i.e., quantitative) value in the distance tables we have used so far?

Fraction Of Pair Difference: One Type Of Distance

- Let $d_{i,j}$, the distance between two sequences i and j, be the fraction of sites in a sequence that are different *(presupposing an alignment of two sequences)*.

A: A T G G C T A A G T T
B: A T G G C T A A G T T

(# diff. sites = 0 / length of seq. =11) x 100%
= 0% $\therefore$ $d_{i,j}$ = 0

A: A T G G G T A - G T T
B: A T C G C T A A G T T

(# diff. sites = 3 / length of seq. =11) x 100%
= 27.3% $\therefore$ $d_{i,j}$ = 27.3

Fraction Of Pair Difference Distance Table

	A	B	C	D	E
A	--				
B	27.3	--			
C	5	34	--		
D	33.7	3.9	18	--	
E	12	44	11	55	--

Fraction Of Pair Difference (f): Good For Small Fractions

- For two unrelated sequences, random substitutions will cause f to approach the fraction of difference expected by chance. However, we want distances to become larger as f tends to this value.

- RECALL: Random sequences of A, C, T, and G = 25% similarity or 75% difference by chance. That is, we want f to rapidly ↑ as the % difference approaches what one would expect comparing two unrelated sequences having random substitutions.

Some Disadvantages Of Pair-Wise Distances

- Pairwise distance data tend to underestimate the path-distance between taxa on a phylogram.
- Pairwise distances effectively "cut corners" in a manner analogous to geographic distance: the distance between two cities may be 100 miles "as the crow flies," but a traveler may actually be obligated to travel 120 miles because of the layout of roads, the terrain, stops along the way, etc.
- Between pairs of taxa, some character changes that took place in ancestral lineages will be undetectable, because later changes have erased the evidence (often called multiple hits and back mutations in sequence data).
- This problem is common to all phylogenetic estimation, but it is particularly acute for distance methods

Jukes-Cantor Model For DNA Behaves In This Manner

$$d_{i,j} = -\frac{3}{4}\log(1-(4f\ /3))$$

$d_{i,j}$ will tend to ∞ as $\boldsymbol{f}$ tends toward 75%.

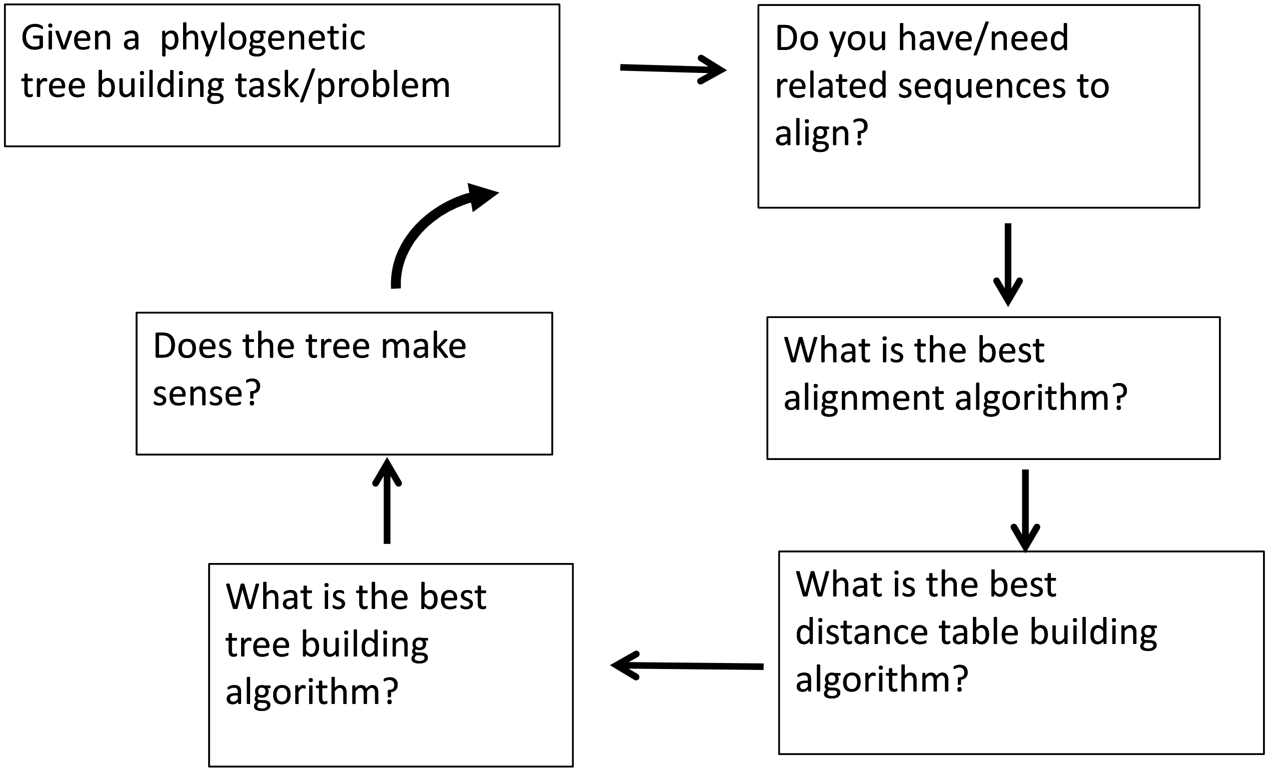
Slightly different model for amino acids, codons, … etc.

Alternative Distance Measures

- Degree of homology
- Combination of Z-scores and fraction of difference
- UniFrac* method tuned for microbes.

* Catherine Lozupone1 and Rob Knight2, UniFrac: a New Phylogenetic Method for Comparing Microbial Communities, *Appl. Environ. Microbiol. December 2005 vol. 71 no. 12 8228-8235*

Be Mindful …



Given a phylogenetic tree building task/problem

Do you have/need related sequences to align?

What is the best alignment algorithm?

What is the best distance table building algorithm?

What is the best tree building algorithm?

Does the tree make sense?

Molecular Clocks

- UPGMA produces a special kind of rooted tree. Edge lengths viewed as time that is measured by a *molecular clock* that ticks at a constant rate.

- UPGMA ASSUMPTION: All genetic changes (mutations/substitutions/… etc.) happen at the same rate (i.e., x mutations or substitutions …etc. / tick of the clock).

- That is, the sum of the times down any path of a UPGMA generated tree to leaf nodes is the same, whatever the choice of path.

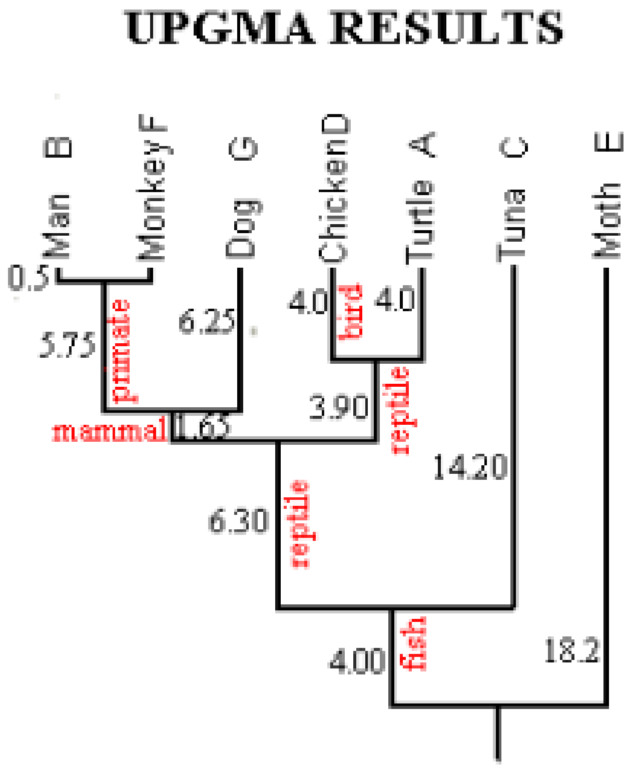
UPGMA Caution

- If distance data is not reflective of a constant evolutionary clock, then a UPGMA tree will not be correct.

- A test for correctness is to determine if we have *ultrametric* conditions. A distance $d_{i,j}$ is ultrametric if for any triplet of sequences x^i, x^j, x^k, the distances $d_{i,j}$, $d_{j,k}$, $d_{i,k}$ are either all equal, or two are equal and the remaining one is smaller.

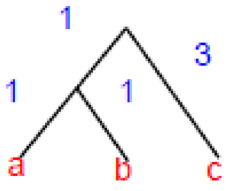
- This condition holds for distances derived from a tree with a molecular clock.

Real UPGMA vs Real Ultrametric Tree



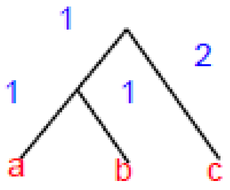
Ultrametric vs Molecular Clock

Ultrametric doesn't imply molecular clock



	a	b	c
a	0	2	5
b		0	5
c			0

Molecular clock implies ultrametric



	a	b	c
a	0	2	4
b		0	4
c			0

Additivity & Neighbor-Joining (NJ)

- With UPGMA we also assumed another property called additivity: Given a tree, its edge lengths are additive if the distance between any pair of leaves is the sum of the lengths of the edges on the path connecting them. This is automatically built in with UPGMA.

- However, the molecular clock property can fail, i.e., the evolutionary clock (e.g. mutations/substitutions) do not happen at a constant rate. NJ to the rescue.

Neighbor-Joining (NJ) Next Lecture