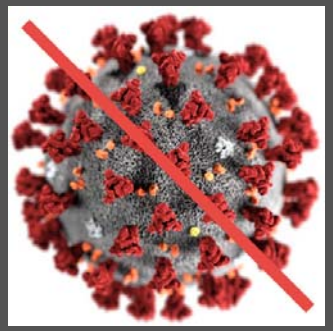
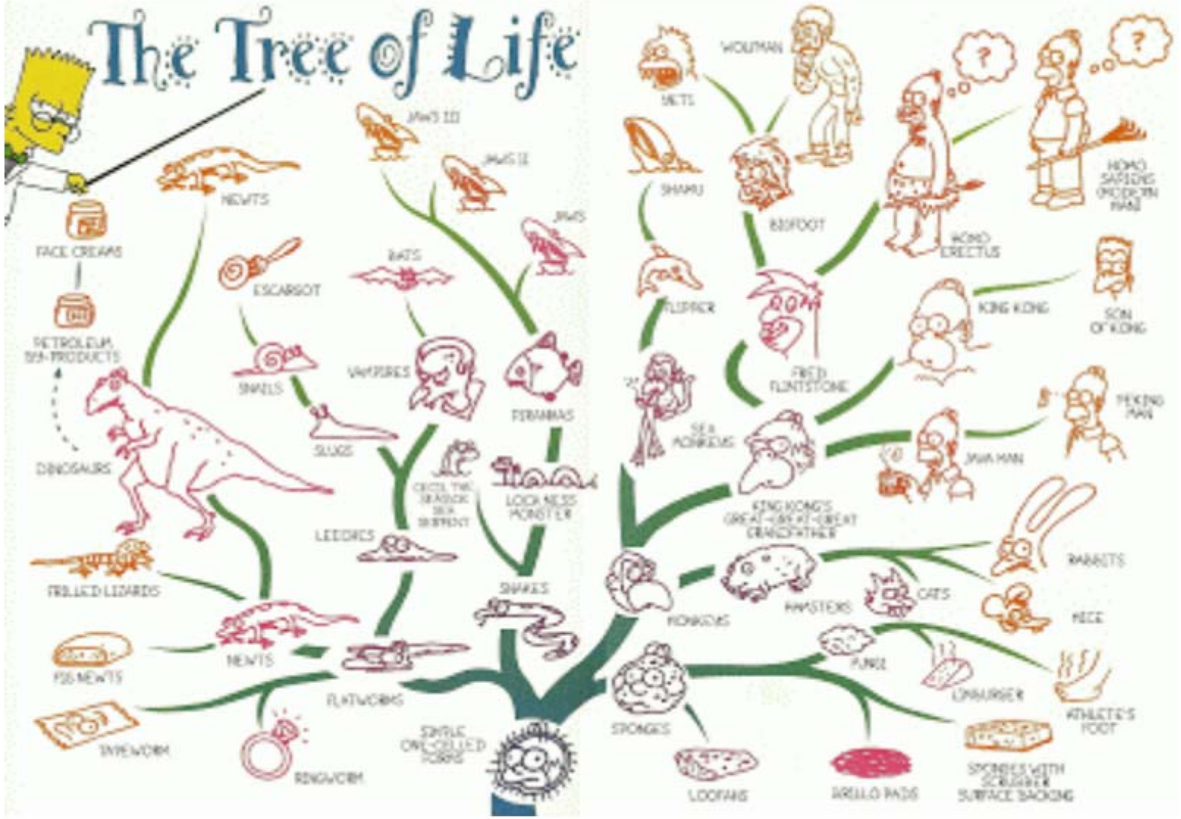


CS123A ~~Bioinformatics~~ Module 3 – Week 10 – Presentation 1

Leonard Wesley

Computer Science Dept

San Jose State Univ



Bart Simpson's Tree of Life

© Matt Groening

Agenda

- Quiz 2 Study Guide
- Weekly Feedback RE Replacing Final Exam With Project
- Phylogenetic Trees
  - Continue Hierarchical Clustering & UPGMA Tree Building Example
  - Distance measures, Pros & Cons
  - Molecular Clocks

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.

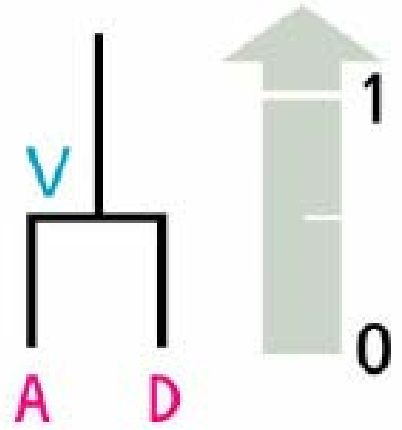
- The distance is linear with evolutionary time.

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)

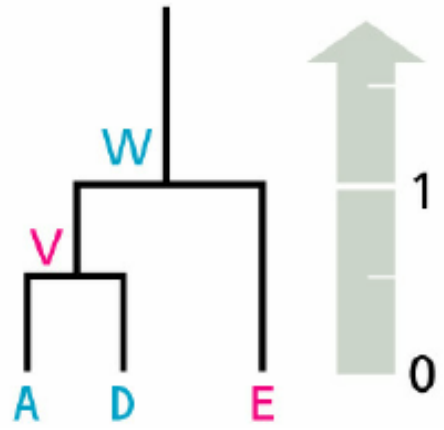
d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6



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.

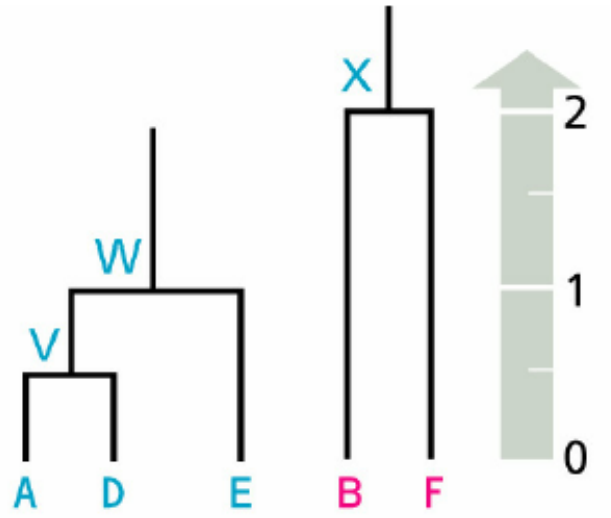
d_{ij}	B	C	E	F	V
B	–	8	6	4	6
C		–	8	8	8
E			–	6	2
F				–	6



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.

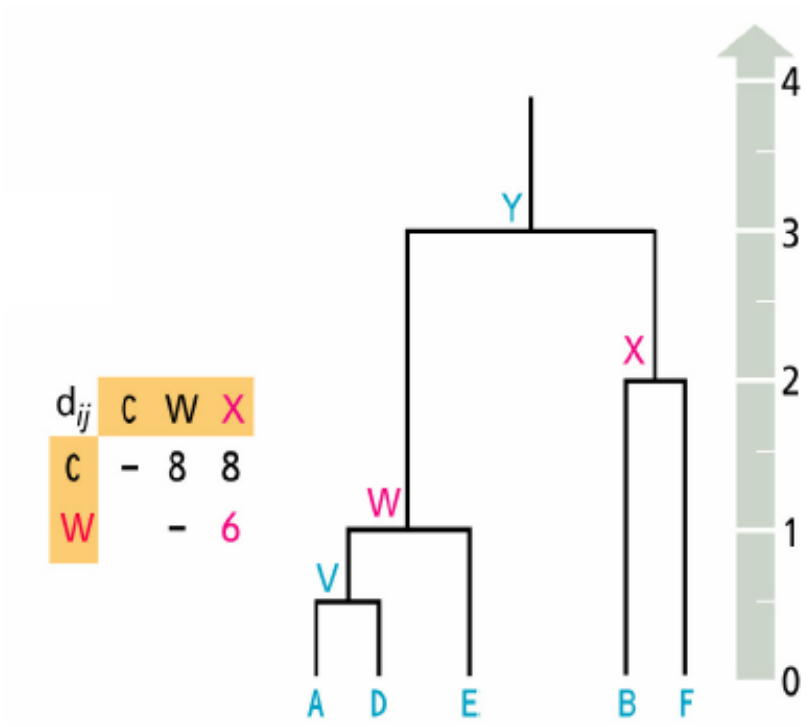
d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6



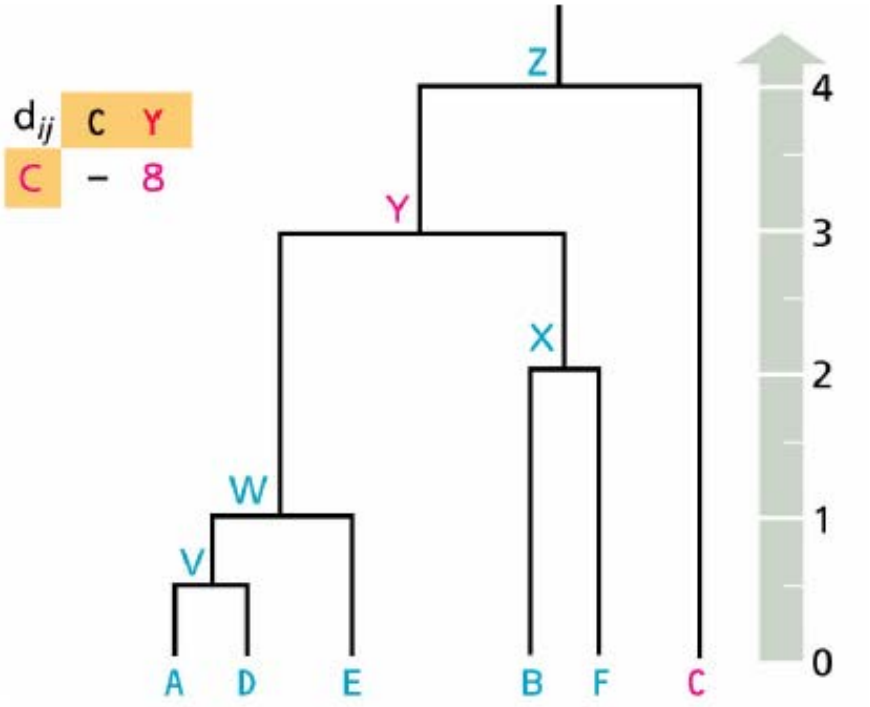
UPGMA: Example (4th Iteration)

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)



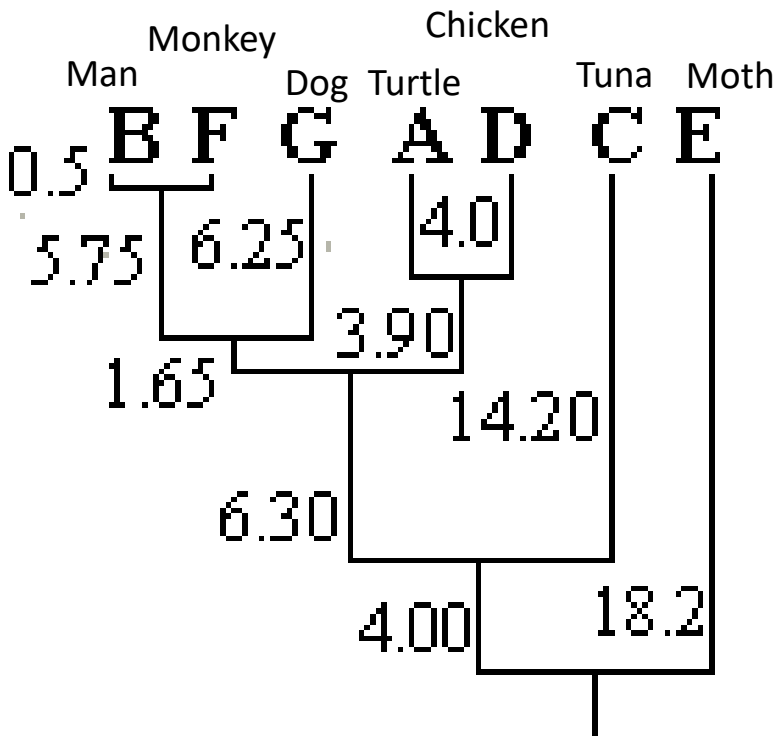
The remaining 2 sequences, C and Y of distance 8 are combined to create a new cluster Z of height 4 in T.

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	

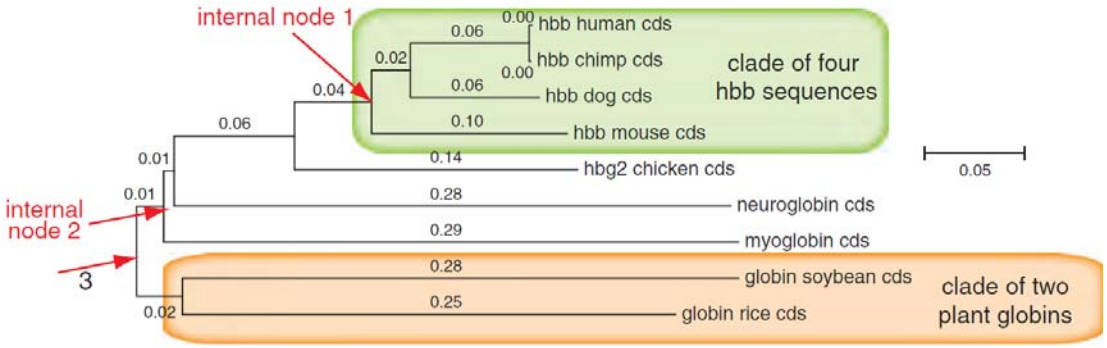
Answer To In-Class Exercise

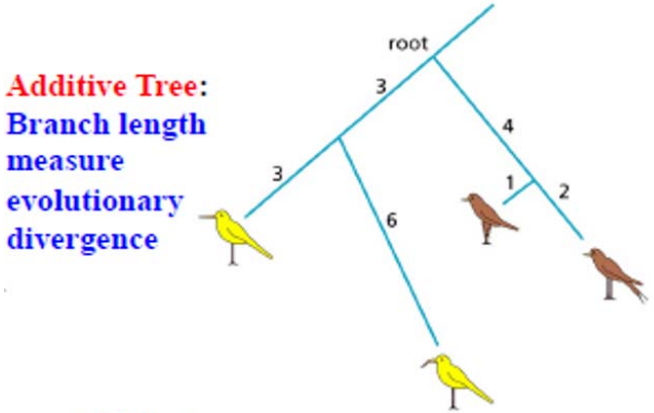


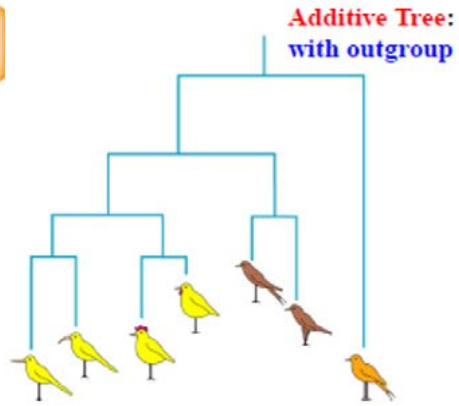
Phylogenetic Trees
Called By Several Other Name(s)

Rectangle/ Additive Tree Styles

(Type of Neighbor Joining Trees)

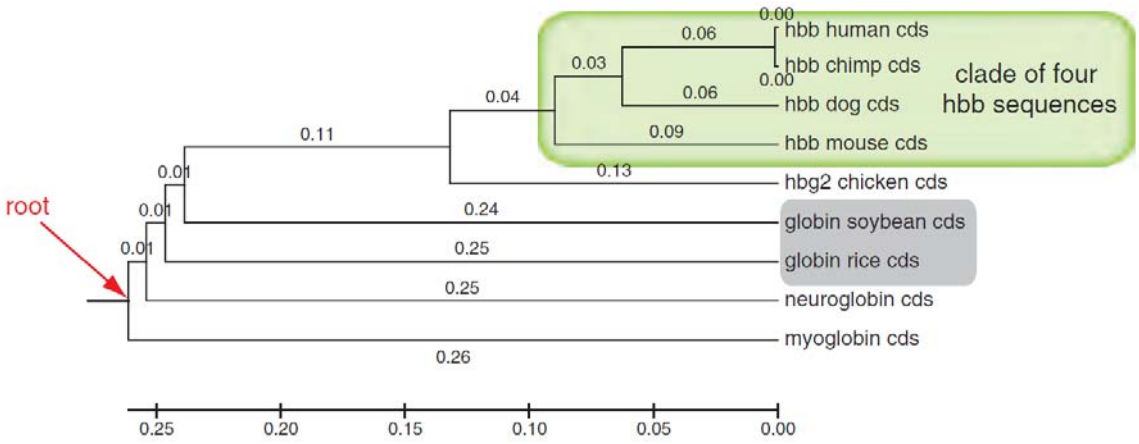


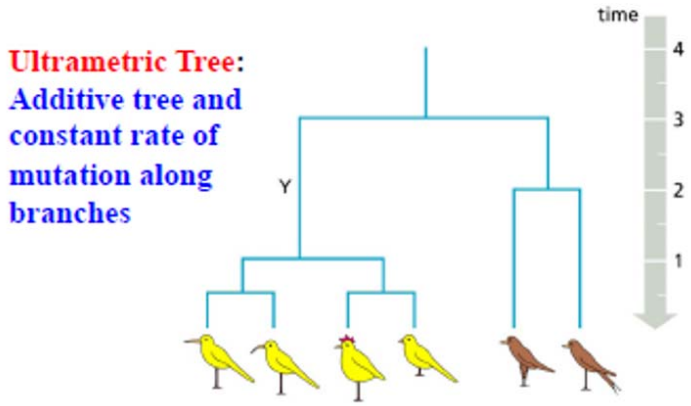




Topology/Ultrametric Trees

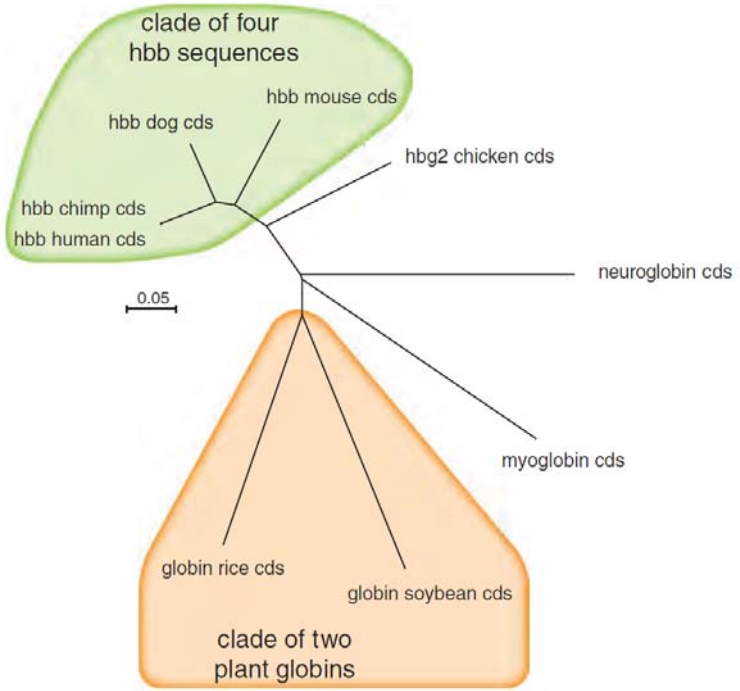
(Type of Neighbor Joining Trees)

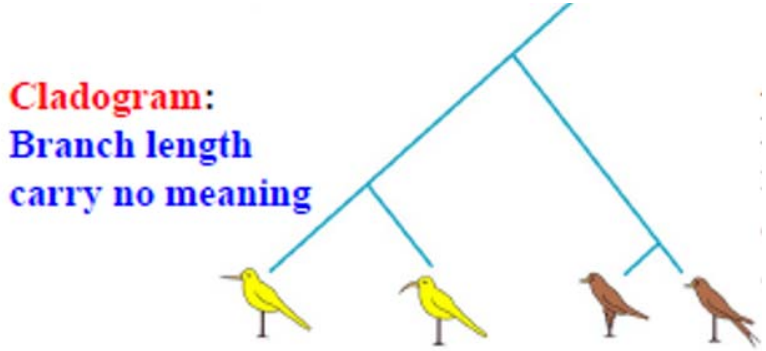




Produced via
UPGMA algorithm

Cladogram/Unrooted Tree





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 $\uparrow$ 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).

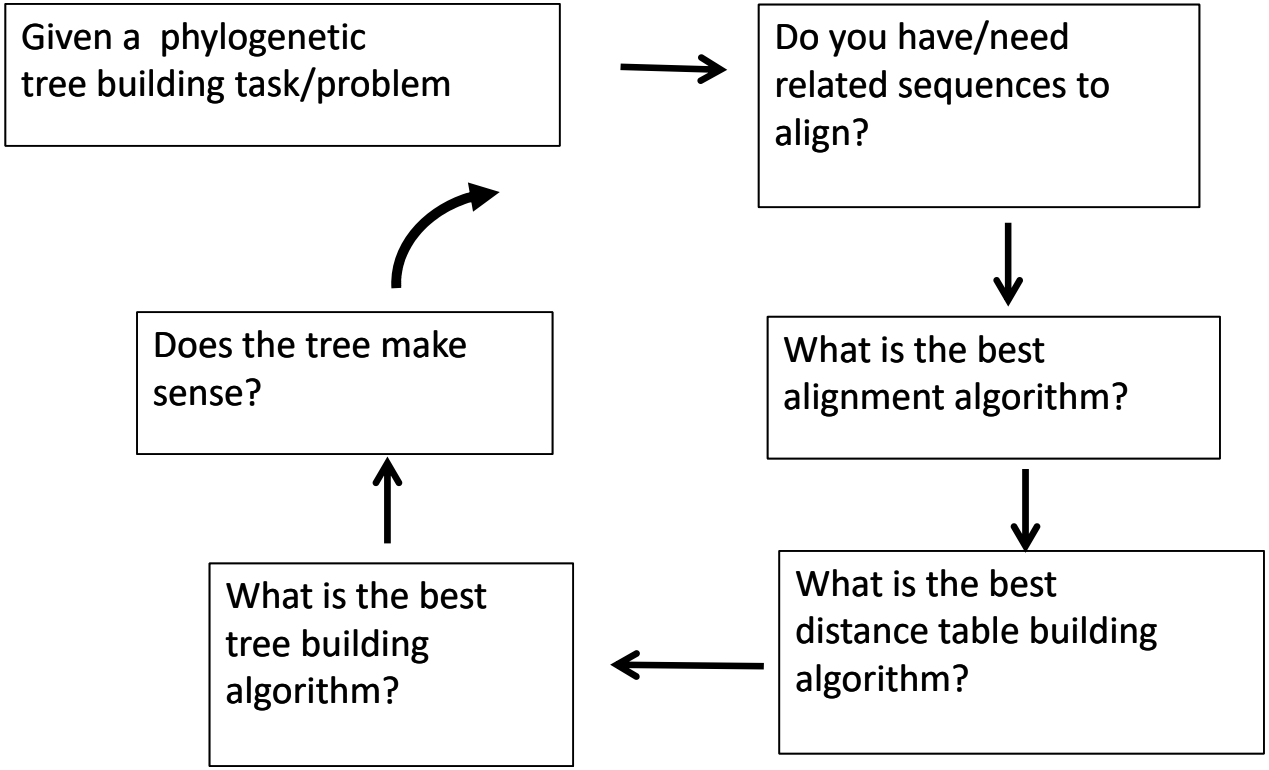
- This problem is common to all phylogenetic estimation, but it is particularly acute for distance methods

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 …



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.

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