Bioinformatics II:

Distance based method to build a phylogenetic tree

Dr Manaf A Guma

University Of Anbar- college of applied sciences-Heet.

Department of chemistry

1

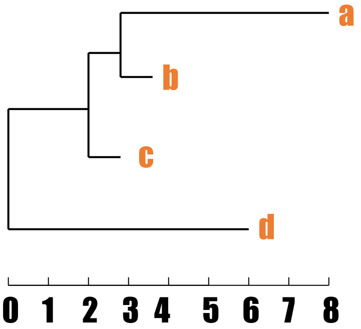
What are the common Methods to build a phylogenetic tree?

1. Distance-based measures: (UPGMA and NJ methods)
2. Character Based Methods: (Parsimony: straightforward method and Maximum likelihood). They are statistical measures.
3. Additional Method (Quartets Based and Disc Covering) (we are NOT going to study).

2

Distance (ultrametric) Matrices

$$\begin{array}{c|cccc} a & 0 & & & \\ b & 6 & 0 & & \\ c & 7 & 3 & 0 & \\ d & 14 & 10 & 9 & 0 \\ \hline & a & b & c & d \end{array}$$



3

Distance-based measures

- How do Distance-based measures work?
- Only pairwise distance between the sequences are considered.
- If the sequences are long, we care only about the tree structure but not the ancestor sequences.
- The sequences can be computed by methods based on sequences alignment such as UPGMA (do you remember !)
- The (**unweighted pair group method with arithmetic mean**) is a simple methods based on collecting a hierarchical clustering of data.

4

What does unweighted term indicate?

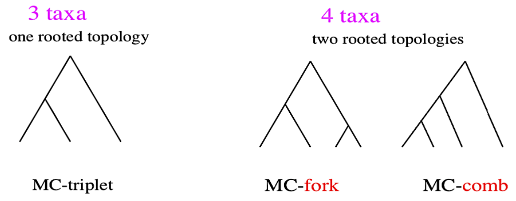
- It indicates that that all distances contribute equally to each average that is computed and does not refer to the math by which it is achieved.
- What does it reflect?
- The UPGMA algorithm constructs a rooted tree dendrogram which is a tree diagram, especially one showing classification relationships that reflects the structure present in a pairwise similarity or dis—similarity matrix.

5

What UPGMA based on?

- It is based on a Molecular clock.
- What is a Molecular clock MC?
- It is a technique assumes that the genome is mutated regularly with time.
- It undergoes two factors:

1. Rooted trees
2. Equal distance from root to all leaves.



- So, UPGMA estimates the time when two species are diverged from their ancestor.
- At each step, the nearest two clusters are combined into a higher-level cluster.

6

What does the UPGMA algorithm produce?

- The UPGMA algorithm produces rooted dendrograms and requires a constant-rate assumption - that is, it assumes an ultrametric tree in which the distances from the root to every branch tip are equal.
- When the tips are molecular data (*i.e.*, DNA, RNA and protein) sampled at the same time.
- It show the differences between the estimated distances and the real distances in an evolutionary tree.

7

Solve out the following estimated (experimental) matrix using UPGMA

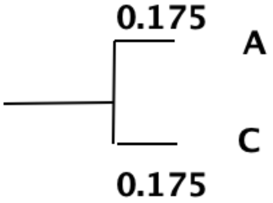
- We will follow specific steps to calculate the real matrix.
- These steps represents the UPGMA.
- The goal is to find the difference between the estimated (experimental) matrix and the calculated matrix.
- Note: this is can be applied for too many species but we will need a to use computerized software to solve it out.

8

Step 1

- To find the difference between the estimated (experimental) tree and the real tree we will work out on the following estimated matrix M:
- Choose the species that have the smallest distances.
- Here: A and C (0.35)
- Draw a node in the middle of the distances AC/2= 0.35/2=0.175.

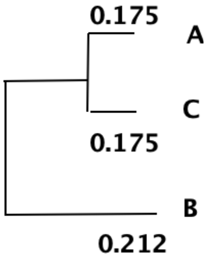
	A	B	C
B	0.4		
C	0.35	0.45	
D	0.6	0.7	0.55



Step 2

- Now, we are going to use AC as a 1 matrix .
- Then we will have a distance between A-C and B:
- (AB+BC)/2= 0.4+0.55/2= 0.425
- (AD+CD)/2= 0.6+0.55/2= 0.575
- Thus, from the new M, the smallest distance between the node A-C and B =0.5425 which allow a new node between A-C-B
- The length of the branch between A-C and B – mid of the distance
- AC-B/2= 0.425/2= 0.212 therefore the tree will be

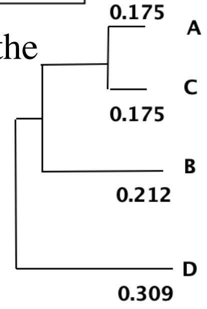
	A-C	B
B	0.425	
D	0.575	0.7



Step 3

- Also, we will have a new M, consist of A-C-B a one group with D
- The distance between A-C-B and D is calculated as:
- (BD+AD+CD)/3= 0.7+0.6+0.55/3 = 0.617
- D will be the last branch that added to the tree. To calculate the distance ACB-D/2= 0.617/2= 0.309

	B-A-C
D	0.617



Step 4

- So, the calculated matrix (which represents the distance of the evolutionary tree)by UPGMA is not equal to the estimated Matrix which as follow:
- so, the UPGMA fails to achieve the (experimental) results.

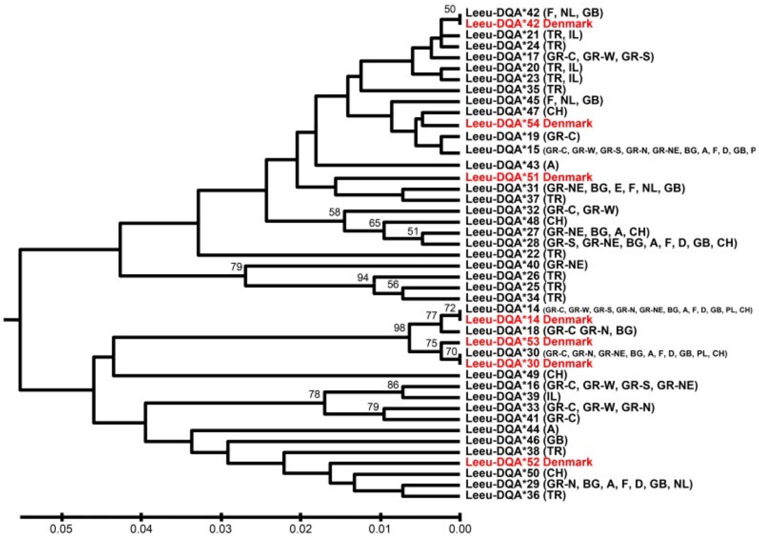
Calculated M

	A	B	C
B	0.425		
C	0.175	0.425	
D	0.617	0.617	0.617

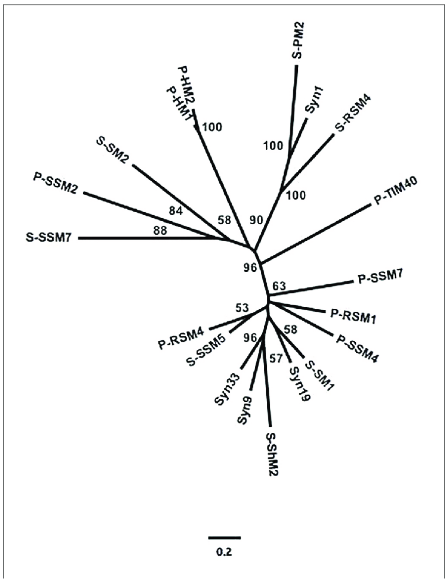
Estimated (experimental) **M**

	A	B	C
B	0.4		
C	0.35	0.45	
D	0.6	0.7	0.55

An example of a tree using UPGMA matrix



13



Neighbor joining NJ ?

- It is another method to draw a phylogenetic tree.
- It considers the distances matrices to calculate the divergence the instead of Molecular clock.
- It is presented as as star *. (see the figure).
- Each pairwise is calculated and evaluated if it can be joined or not.
- The length of each branch is calculated.
- The smallest pairwise value is the nearest neighbor.

14