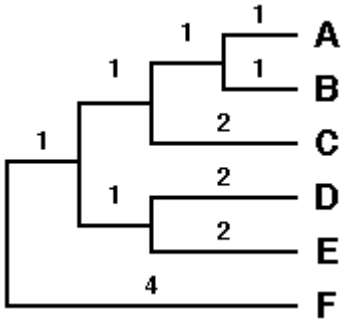
Construction of a distance tree using clustering with the Unweighted Pair Group Method with Arithmatic Mean (UPGMA).

The UPGMA is the simplest method of tree construction. It was originally developed for constructing taxonomic phenograms, i.e. trees that reflect the phenotypic similarities between OTUs, but it can also be used to construct phylogenetic trees if the rates of evolution are approximately constant among the different lineages. For this purpose the number of observed nucleotide or amino-acid substitutions can be used. UPGMA employs a sequential clustering algorithm, in which local topological relationships are identifeid in order of similarity, and the phylogenetic tree is build in a stepwise manner. We first identify from among all the OTUs the two OTUs that are most similar to each other and then treat these as a new single OTU. Such a OTU is referred to as a composite OTU. Subsequently from among the new group of OTUs we identify the pair with the highest similarity, and so on, until we are left with only two UTUs.

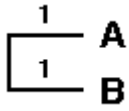
Suppose we have the following tree consisting of 6 OTUs:



The pairwise evolutionary distances are given by the following distance matrix:

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

We now cluster the pair of OTUs with the smallest distance, being A and B, that are separated a distance of 2. The branching point is positioned at a distance of 2 / 2 = 1 substitution. We thus constuct a subtree as follows:



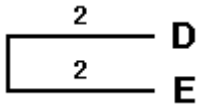
Following the first clustering A and B are considered as a single composite OTU(A,B) and we now calculate the new distance matrix as follows:

```
dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8
```

In other words the distance between a simple OTU and a composite OTU is the average of the distances between the simple OTU and the constituent simple OTUs of the composite OTU. Then a new distance matrix is recalculated using the newly calculated distances and the whole cycle is being repeated:

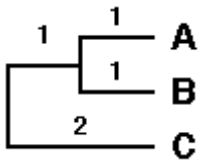
Second cycle

	A,B	C	D	E
C	4			
D	6	6		
E	6	6	**4**	
F	8	8	8	8



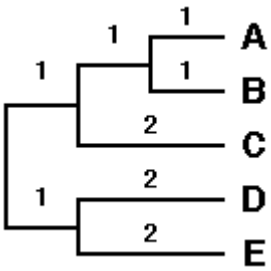
Third cycle

	A,B	C	D,E
C	**4**		
D,E	6	6	
F	8	8	8



Fourth cycle

	AB,C	D,E
D,E	**6**	
F	8	8



Fifth cycle

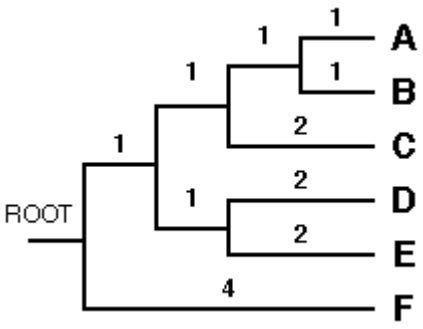
The final step consists of clustering the last OTU, F, with the composite OTU.

	ABC,DE

F	8

Although this method leads essentially to an unrooted tree, UPGMA assumes equal rates of mutation along all the branches, as the model of evolution used. The theoretical root, therefore, must be equidistant from all OTUs. We can here thus apply the method of mid-point rooting. The root of the entire tree is then positioned at dist (ABCDE),F / 2 = 4.

The final tree as inferred by using the UPGMA method is shown below.



So now we have reconstructed the phylogenetic tree using the UPGMA method. As you can see we have obtained the original phylogenetic tree we started with.

However, there are some pitfalls:

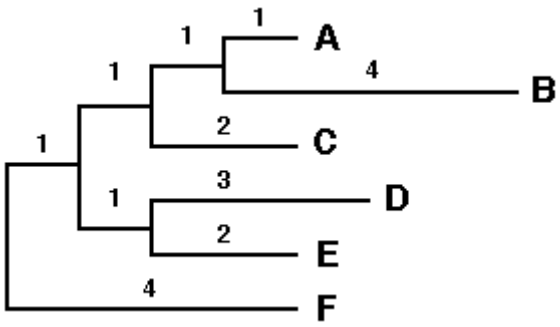
- the UPGMA clustering method is *very sensitive* to unequal evolutionar rates. This means that when one of the OTUs has incorporated more mutations over time, than the other OTU, one may end up with a tree that has the wrong topology.
- Clustering works only if the data are *ultrametric*
- Ultrametric distances are defined by the satisfaction of the '*three-point condition*'.

What is the three-point condition?

For any three taxa: dist AC <= max (distAB, distBC) or in words: the two greatest distances are equal, or UPGMA assumes that the evolutionary rate is the same for all branches

If the assumption of rate constancy among lineages does not hold UPGMA may give an erroneous topology. This is illustarted in te following example:

Suppose the you have the following tree:



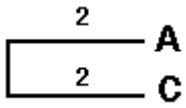
Since the divergence of A and B, B has accumulated mutations at a much higher rate than A. The Three-point criterion is violated ! e.g. distBD <= max (distBA,distAD) or,
10 <= max (5,7) = False

The reconstruction of the evolutionary history uses the following distance matrix:

	A	B	C	D	E

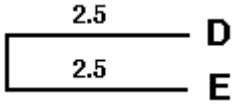
B	5				
C	**4**	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

We now cluster the pair of OTUs with the smallest distance, being A and C, that are separated a distance of 4. The branching point is positioned at a distance of 4 / 2 = 2 substitutions. We thus constuct a subtree as follows:



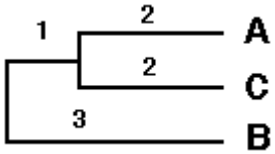
Second cycle

	A,C	B	D	E
B	4			
D	7	10		
E	6	9	**5**	
F	8	11	8	9



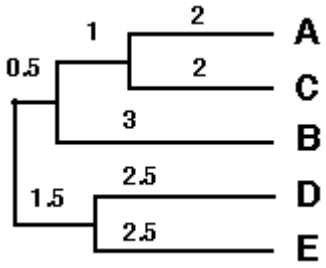
Third cycle

	A,C	B	D,E
B	6		
D,E	6.5	9.5	
F	8	11	8.5



Fourth cycle

	AC,B	D,E
D,E	**8**	
F	9.5	9.5

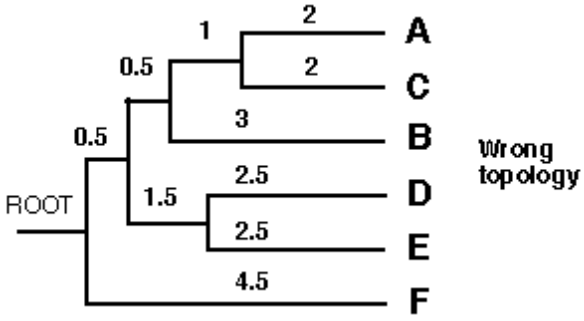


Fifth cycle

The final step consists of clustering the last OTU, F, with the composite OTU, ABCDE.

	ABC,DE
F	**9**

When the original, correct, tree and the final tree are compared it is obvious that we end up with a tree that has the wrong topology.



***Conclusion*: The unequal rates of mutation has led to a completely**

different tree topology.

Last updated: 9 September 1997.

---

created by :Fred Opperdoes