Feb. 10, 2020. **Phylogenetic trees II: Phenetics; distance-based algorithms**
Reading assignment: *Tree Thinking* pp 231-238

A. Introduction

Distance-based methods contrast with character-based methods by building a branching diagram (is this a phylogeny?) using an overall similarity matrix that compares OTUs pairwise. Homology-based methods like parsimony, likelihood, and Bayesian methods directly use the character matrix to reconstruct a tree instead if indirectly using a similarity matrix.

Phenetic distance methods were introduced into systematics in the 1960s (e.g. Peter Sneath and Robert Sokal) for applications in what was referred to as Numerical Taxonomy. Though the term numerical taxonomy originally included cladistic methods like parsimony, it is more or less treated as a synonym of phenetics now. Historically, the distance methods covered here are linked with classification arguments since much of the debate was between numerical-method proponents countering what they viewed as arbitrary and authoritative classifications built on a few favored character systems, which were treated with opinion-based argumentation. So phenetics will also appear in our discussion on classification later. It will also appear in our discussion about species, since one hold-out for application of distance-based methods is in so-called "phylogeographic" studies below what some consider the species level.

For proponents these were statistically and mathematically fairly well understood methods that they argued were much more objective and could be implemented by even naïve users. It was intended to involve large numbers of characters, which was thought to provide a better classification. The primary target was classification and so clustering without recourse to phylogeny, and what was viewed as unnecessary and subjective interpretations, was preferred.

Many of the methods are quite fast to compute even for large numbers of OTUs and still useful for inherently distance data like PCA data or DNA-DNA hybridization data. For reconstructing phylogenies the methods can be moderately useful as an approximation and are frequently used in combination with other methods to get starting trees to begin a large phylogenetic analysis, or guide trees to use in alignment, for example.

B. Phenetic methods have a number of well-known drawbacks for phylogenetics:

1. The most obvious is information loss by the reduction a character matrix to a distance matrix. Differences in observed character states between entire OTUs are summarized as a single value. The direct test of homology through character state congruence is not possible.

2. Underestimation of changes is acute in distance methods due to the use of pairwise distance.

3. Heterogeneous data types are problematic. In many cases we will want to combine data of different types for analyses and it isn't at all clear how the similarity of DNA sequence relates to similarity of morphology or behavioral data.

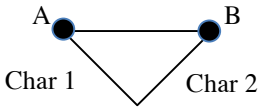
4. The distance along branches are non-independent and so problems along a given edge can be very problematic.

5. Many methods can have ties that may be arbitrarily broken such that each leads to different end results.

6. There are many methods to calculate trees and a wide variety of how distances are obtained and treated. There is often no clear biological reason to prefer one over another.

7. Phenetic methods most often use a Euclidian distance measure, which results in a path between taxa that is unrealistic in evolutionary terms. Phylogenetic methods use Manhattan distance.

What is the distance between A and B?



C. Pairwise distance matrix

There are many way to calculate distance and often is it some sort of generalization of Euclidean distance, or a form of correlation coefficient, or a genetic distance measure, or a matching coefficient as in the example below. Distance measures may be transformed in various ways, e.g. normalizing or scaling, redundancy removal.

D. Unweighted Pair Group Method with Arithmetic Averaging (UPGMA).

Fast to compute and can handle many OTUs, but it assumes ultrametric tree, which rarely holds for real data. An *agglomerative method*, it identifies the most similar cluster and joins them in decreasing order of similarity. As each OTU is joined distances are recalculated.

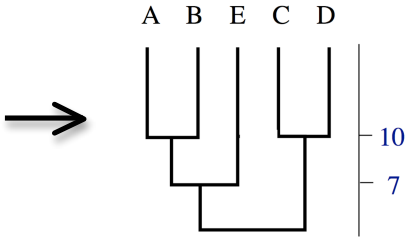
Simple example of UPGMA

Standard character matrix:

	Character											
Species	1	2	3	4	5	6	7	8	9	10	11	12
A	1	1	1	0	0	0	0	0	0	0	0	0
B	1	1	0	1	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	1	1	1	1	1	0
D	0	0	0	0	1	0	1	1	1	1	0	1
E	0	0	0	0	1	1	0	0	0	0	0	0

Distance matrix (simple matching metric -- there are many other metrics)

	A	B	C	D	E
A	--				
B	10	--			
C	3	3	--		
D	3	3	10	--	
E	7	7	6	6	--

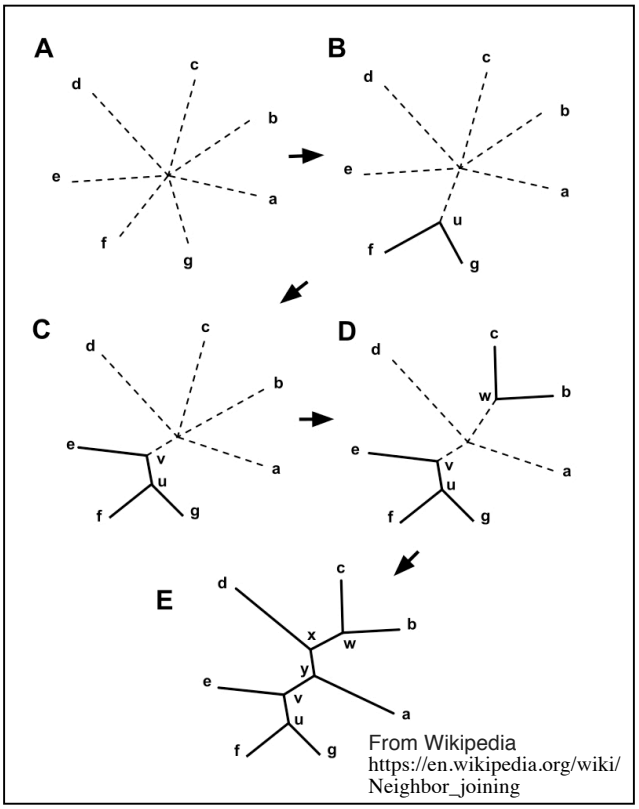


☞ As an aside, what would be the parsimony tree for this matrix? Where would the characters map? There is no homoplasy so you should be able to see this by applying tree thinking. Good practice!

Note that in addition to average linkage as used in UPGMA there other kinds of linkage, notably single linkage (similarity between an unplaced element and existing clusters is whichever cluster contains the most similar element) and complete linkage (similarity between an unplaced element and existing clusters is whichever cluster contains the most different element).

E. Neighbor Joining (NJ).

Also fast to compute and able to handle many OTUs. Doesn't need an ultrametric tree, but does assume an additive tree. A *divisive method*, the algorithm starts with a matrix of distances among the OTUs and a completely unresolved tree – star phylogeny or bush. The pair of OTUs that will most greatly reduce the overall distance is found, merged and the matrix is reduced. This continues until all OTUs are joined.



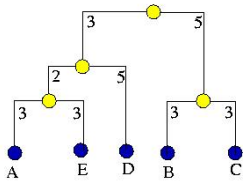
F. Some relevant terms:

Metricity: refers to a property of a distance measure: 1. An element's distance to itself is zero. 2. Distance between elements is greater than zero. 3. Distances between any two elements are symmetrical. 4. Satisfies the triangle inequality. Many measures are not metric.

Ultrametric tree: A special case of an additive tree (where the distance between OTUs or nodes is the sum of the branches) where the distance from any node to the tip is the same in all descendants. This suggests a constant molecular clock. Most real data are not ultrametric, but sometimes we are tempted to force them to be, in order to get time estimates from phylogenies. More on this later in the class.

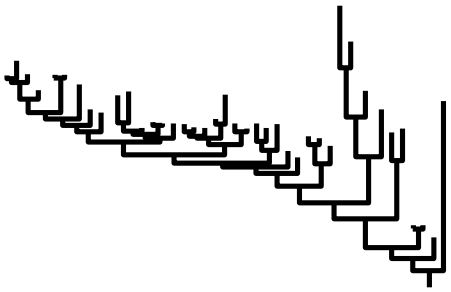
Ultrametric matrix and its tree:

	A	B	C	D	E
A		16	16	10	6
B			6	16	16
C				16	16
D					10
E					



from: http://www.diku.dk/~pawel/comp-bio/ev_trees/intro/intro/ultrametric.html

A non-ultrametric tree:



Answer from page 2: