Phenetics

Distance-based methods contrast with character-based methods by using an overall similarity measure between OTUs to build a tree. Character-based methods like parsimony, likelihood, and Bayesian methods fit individual character states to reconstruct or estimate a tree.

Phenetic distance methods were introduced into biosystematics in the 1960s (e.g. by Peter Sneath and Robert Sokal) for applications in what was referred to as Numerical Taxonomy. Though the term numerical taxonomy originally included cladistics methods like parsimony, it is more or less treated as a synonym of phenetics now. Historically the distance methods covered here are interrelated with classification 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. Phenetics will also appear in our discussion on classification later.

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 a large number of characters, which was thought to provide a better classification. The primary target was classification and clustering was intentionally done without recourse to evolution or phylogeny (i.e., history), ignoring these, which were viewed as unnecessary and subjective interpretations.

Many of the methods are quite fast to compute even for large numbers of OTUs and still useful for fundamentally 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 or guide trees for alignment, for example.

Phenetic methods have a number of well-known drawbacks:

1. The most obvious problem is information loss or the reduction of higher-level comparisons relative to character-based methods. Observed character states across entire OTUs are summarized as a single value. Phenetic treatments only provide information about similarity. The direct test of homology through character state congruence is not possible. Even after the tree is made ancestral similarity (symplesiomorphy) and derived similarity (synapomorphy) are not given, i.e. nodes have no attributes.
2. Underestimation of changes is acute in distance methods due to the use of pairwise distance. Typically it is less consistent than parsimony.
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. Many distance 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.

Proposed original goals or advantages:

1. Stability of classifications
2. Less subjective: many characters, equal weights, not based on "authority" alone
3. Repeatability
4. Forced people to carefully examine characters
5. More natural and informative (both proven wrong relative to cladistics methods).

Phenetic Clustering.

Metricity: 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.

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 decedents. This suggests a constant molecular clock. Most real data are not ultrametric.

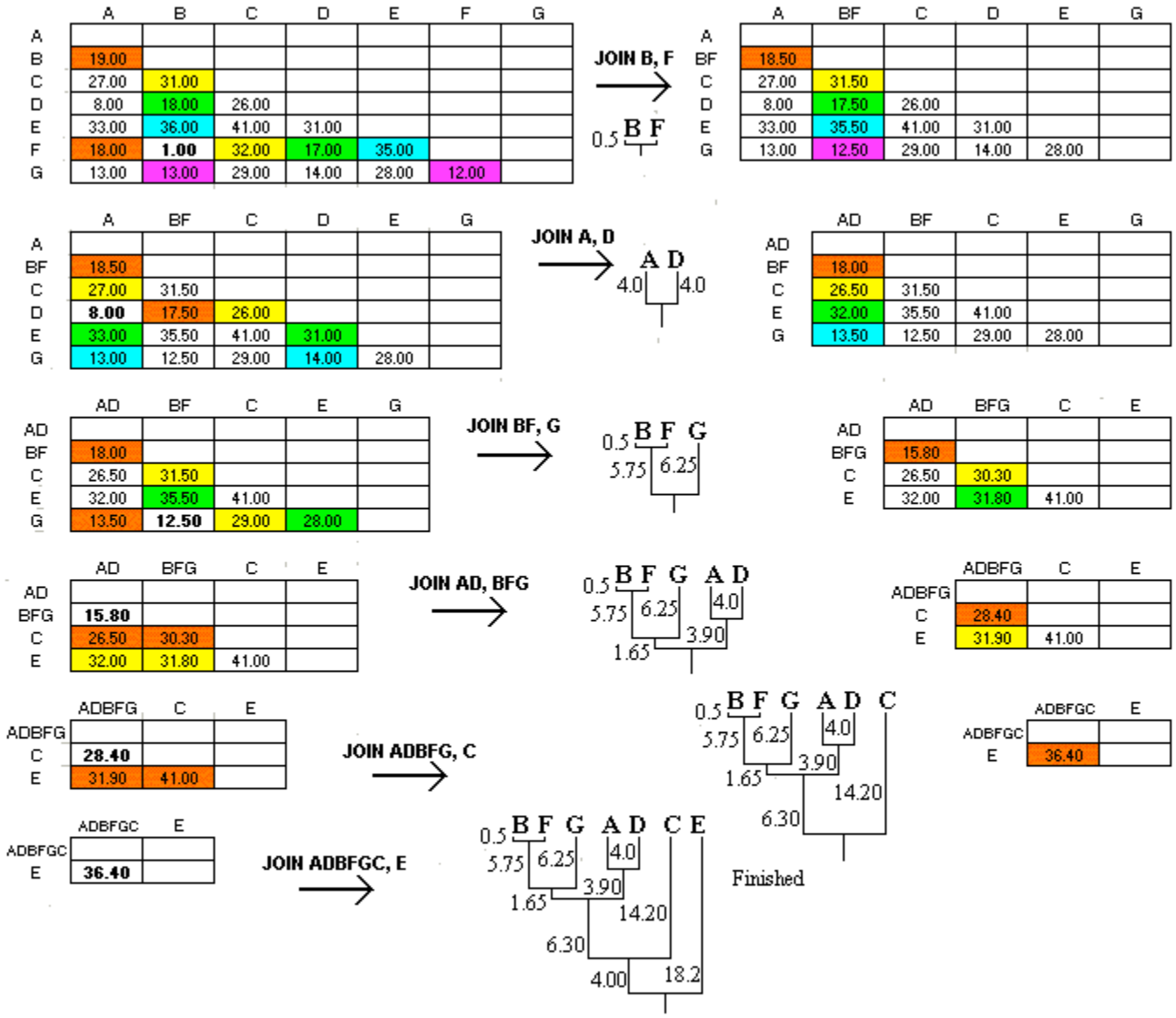
Step 1. Get a matrix of pairwise distances. Distance measures may be transformed in various ways, e.g., normalizing or scaling, redundancy removal. There are many way to calculate distance and often is it some sort go generalization of Euclidean distance, or a form of correlation coefficient.

**See PDF version of handout for worked out examples*

Unweighted Pair Group Method with Arithmetic mean (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.

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. 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.

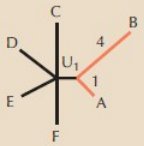
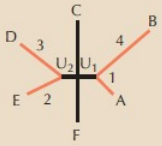
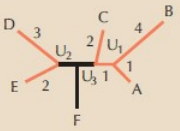
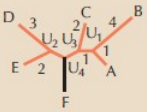
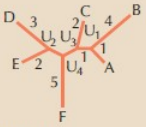
APPLICATION OF UPGMA CLUSTERING METHOD ON SELECTED CYTOCHROME C DATA TO CALCULATE PHYLOGENETIC RELATIONS



Key: the boldface number on the left side indicates the smallest entry (closest match), and directs which entries are to be joined. The height of the new branch is 1/2 times this smallest value. The matrix is reduced as the entries are joined; cells of one color on the left are combined (averaged) to form the new entries (same color) on the right.

From: http://www.nmsr.org/upgma.htm

TABLE 27.11. Neighbor-joining example

	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5
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	U_1 C D E C 3 D 6 7 E 5 6 5 F 7 8 9 8	U_1 C U_2 C 3 U_2 3 4 F 7 8 6	U_2 U_3 U_3 2 F 6 6	U_4 F 5
Step 1					
S calculations S_x = (sum all D_x)/(N – 2), where N is the # of OTUs in the set.	S_A = (5+4+7+6+8)/4 = 7.5 S_B = (5+7+10+9+11)/4 = 10.5 S_C = (4+7+7+6+8)/4 = 8 S_D = (7+10+7+5+9)/4 = 9.5 S_E = (6+9+6+5+8)/4 = 8.5 S_F = (8+11+8+9+8)/4 = 11	S_{U_1} = (3+6+5+7)/3 = 7 S_C = (3+7+6=8)/3 = 8 S_D = (6+7+5+9)/3 = 9 S_E = (5+6+5+8)/3 = 8 S_F = (7+8+9+8)/3 = 10.6	S_{U_1} = (3+3+7)/2 = 6.5 S_C = (3+4+8)/2 = 7.5 S_{U_2} = (3+4+6)/2 = 6.5 S_F = (7+8+6)/2 = 10.5	S_{U_2} = (2+6)/1 = 8 S_{U_3} = (2+6)/1 = 8 S_F = (6+6)/1 = 12	Because N – 2 = 0, we cannot do this calculation.
Step 2					
Calculate pair with smallest (M), where $M_{ij} = D_{ij} - S_i - S_j$.	Smallest are M_{AB} = 5 – 7.5 – 10.5 = –13 M_{DE} = 5 – 9.5 – 8.5 = –13 Choose one of these (AB here).	Smallest is M_{CU_1} = 3 – 7 – 8 = –12 M_{DE} = 5 – 9 – 8 = –12 Choose one of these (DE here).	Smallest is M_{CU_1} = 3 – 6.5 – 7.5 = –11	Smallest is M_{U_2F} = 6 – 8 – 12 = –14 M_{U_3F} = 6 – 8 – 12 = –14 $M_{U_2U_3}$ = 2 – 8 – 8 = –14 Choose one of these ($M_{U_2U_3}$ here).	
Step 3					
Create a node (U) that joins pair with lowest M_{ij} such that $S_{IU} = D_{ij}/2 + (S_i - S_j)/2$.	U_1 joins A and B: $S_{AU_1} = D_{AB}/2 + (S_A - S_B)/2 = 1$ $S_{BU_1} = D_{AB}/2 + (S_B - S_A)/2 = 4$	U_2 joins D and E: $S_{DU_2} = D_{DE}/2 + (S_D - S_E)/2 = 3$ $S_{EU_2} = D_{DE}/2 + (S_E - S_D)/2 = 2$	U_3 joins C and U_1: $S_{CU_3} = D_{CU_1}/2 + (S_C - S_{U_1})/2 = 2$ $S_{U_1U_3} = D_{CU_1}/2 + (S_{U_1} - S_C)/2 = 1$	U_4 joins U_2 and U_3: $S_{U_2U_4} = D_{U_2U_3}/2 + (S_{U_2} - S_{U_3})/2 = 1$ $S_{U_3U_4} = D_{U_2U_3}/2 + (S_{U_3} - S_{U_2})/2 = 1$.	For last pair, connect U_4 and F with branch length = 5.
Step 4					
Join i and j according to S above and make all other taxa in form of a star. Branches in black are of unknown length. Branches in red are of known length.					
Step 5					**Comments**
Calculate new distance matrix of all other taxa to U with $D_{xU} = D_{ix} + D_{jx} - D_{ij}$, where i and j are those selected from above.					Note this is the same tree we started with (drawn in unrooted form here).

From http://www.icp.ucl.ac.be/~opperd/private/upgma.html.