IB 200A Principals of Phylogenetic Systematics *Spring 2012*

Phenetics

Distance vs character states:

1. Distance measures include the number of nucleotide or amino-acid substitutions between molecular sequences.

2. Character state measures specify the nucleotide or amino acid at a particular site, or the presence or absence of a deletion or an insertion.

3. What we are trying to assess is “special similarity.”

The fundamental positions of numerical taxonomy:

1. The greater the content of information in the taxa of a classification and the more characters on which it is based, the better a given classification will be.

2. *A priori*, every character is of equal weight in creating natural taxa.

3. Overall similarity between any two entities is a function of their individual similarities in each of the many characters in which they are being compared.

4. Distinct taxa can be recognized because correlations of characters differ in the groups.

5. Phylogenetic inferences can be made from the taxonomic structures of a group and from character correlations, given certain assumptions about evolutionary pathways and mechanisms.

6. Taxonomy is viewed and practiced as an empirical science.

7. Classifications are based on phenetic similarity.

Phenetic methodology

1. Phenetic classification starts with the collection of raw measurement data on the chosen set of morphs, called *Operational Taxonomic Units*, or *OTU*'s.

2. The data may then be transformed in some manner to remove redundancy, and to normalize the values to a common scale. Commonly used transformations include simple normalization, principle component analysis, and factor analysis.

3. A measure of similarity or dissimilarity (or both) is computed for each pair of OTU's. This is usually a generalized distance measure, based on a generalization of Euclidean distance, or some form of correlation coefficient.

4. A clustering method is used to group OTU's that are most similar. The agglomerative clustering algorithms are popular for this purpose. These method involve joining the two most similar OTU's into a new, compound, OTU, and recomputing the similarity measure for the resulting cluster. This is repeated until all OTU's have been merged into one. The different agglomerative clustering methods differ mainly in the way that the new similarity measures are calculated.

SINGLE Single-link method – brings together the closest elements.

UPGMA Unweighted pair-group method, arithmetic average - the average distance between elements of each cluster (weighted by number of elements).

(AB) and C+(DE) = (55+2x90)/3 = 78.33

WPGMA Weighted pair-group method, arithmetic average (not weighted by number of elements). (AB) and C+(DE) = (55+90)/2 = 72.5

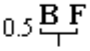
WPGMC Weighted pair-group method, centroid average (assumes dissimilarity).

WPGMS Weighted pair-group method, Spearman's average (assumes correlation).

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

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

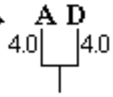
JOIN B, F



	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	8.00	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	**8.00**	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

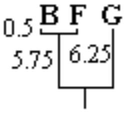
JOIN A, D



	AD	BF	C	E	G
AD					
BF	18.00				
C	26.50	31.50			
E	32.00	35.50	41.00		
G	13.50	12.50	29.00	28.00	

	AD	BF	C	E	G
AD					
BF	18.00				
C	26.50	31.50			
E	32.00	35.50	41.00		
G	13.50	**12.50**	29.00	28.00	

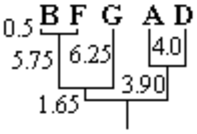
JOIN BF, G



	AD	BFG	C	E
AD				
BFG	15.80			
C	26.50	30.30		
E	32.00	31.80	41.00	

	AD	BFG	C	E
AD				
BFG	**15.80**			
C	26.50	30.30		
E	32.00	31.80	41.00	

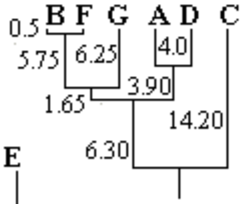
JOIN AD, BFG



	ADBFG	C	E
ADBFG			
C	28.40		
E	31.90	41.00	

	ADBFG	C	E
ADBFG			
C	**28.40**		
E	31.90	41.00	

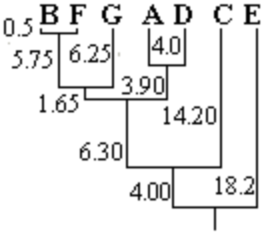
JOIN ADBFG, C



	ADBFGC	E
ADBFGC		
E	36.40	

	ADBFGC	E
ADBFGC		
E	**36.40**	

JOIN ADBFGC, E



Finished

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

Neighbor-joining - The neighbor-joining (NJ) method is used to estimate phylogenetic trees. While the method is based on the idea of parsimony, the NJ method does not attempt to obtain the shortest possible tree for a set of data. Rather, it attempts to find a tree that is usually close to the true phylogenetic tree. Analyses of simulated datasets found the NJ method to yield more accurate trees than either UPGMA or parsimony in most of the models.

The algorithm used to find NJ trees is similar to that of the distance Wagner procedure. The algorithm starts with a matrix of distances among the OTUs and a completely unresolved phylogenetic tree – star phylogeny. The closest pair of OTUs is found, merged into a new HTU and the original pair of OTUs is deleted until the matrix is reduced to a single HTU and the tree is fully resolved. The "closest pair" of OTUs is the pair of OTUs (or HTUs) where their merging produces the maximum reduction in the unresolved distance matrix. In NJ trees are constructed by linking together the two OTUs or HTUs that are the closest mutual "neighbors". The distance Wagner algorithm differs in the way "closest" is defined and in the way the distance between a new node and the existing nodes is computed.

The clustering diagram is then converted into a classification by selecting a cut level for each taxonomic rank, and identifying the clusters that are distinct at the cut level as the taxa of that rank.

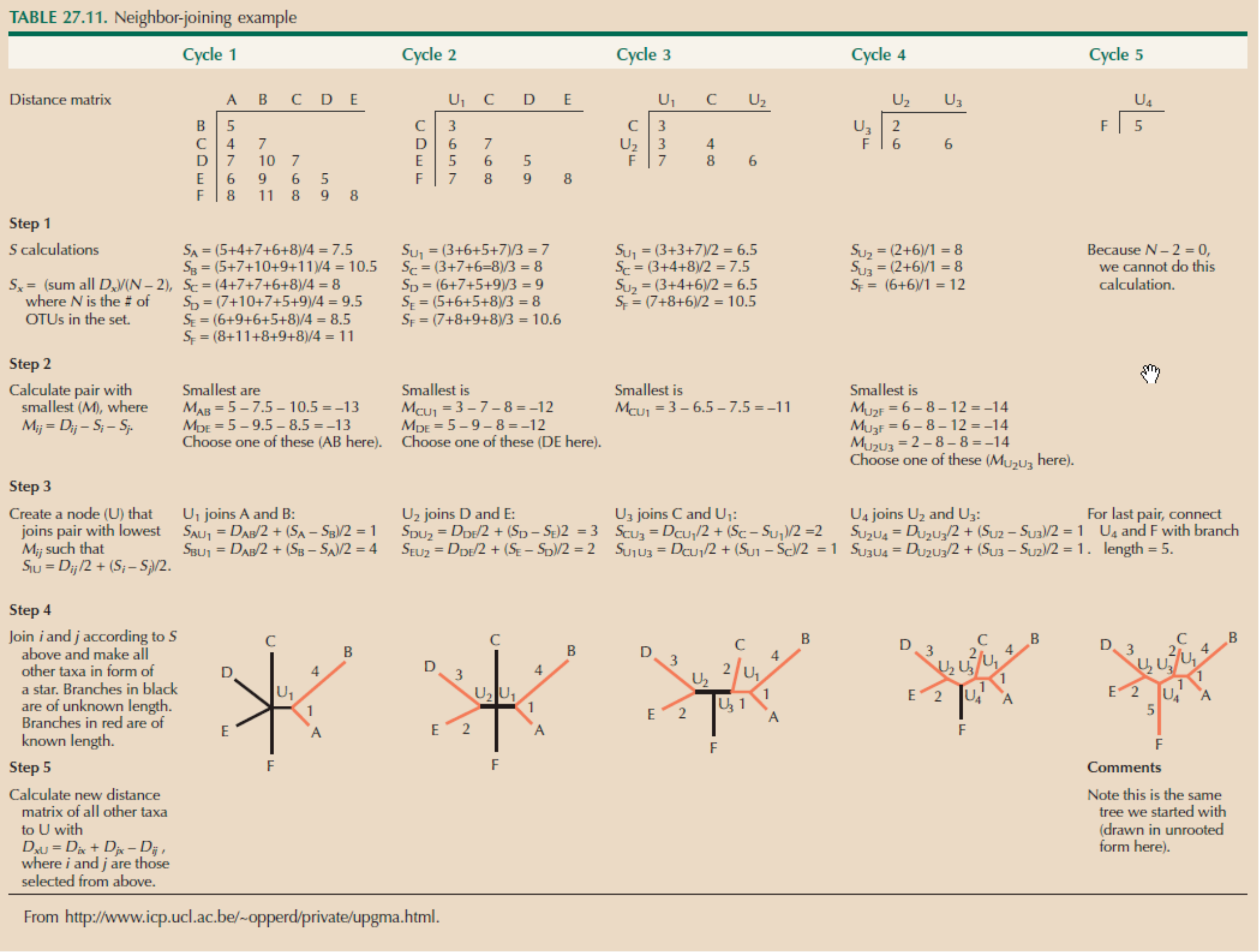


TABLE 27.11. Neighbor-joining example

	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: $S_{CU_3} = D_{CU_1}/2 + (S_C – S_{U_1})/2 =2$ $S_{U_1U_3} = D_{CU_1}/2 + (S_{U1} – S_C)/2 = 1$	U_4 joins U_2 and U_3: $S_{U_2U_4} = D_{U_2U_3}/2 + (S_{U2} – S_{U3})/2 = 1$ $S_{U_3U_4} = D_{U_2U_3}/2 + (S_{U3} – S_{U2})/2 = 1$.	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.

The separation of phenetic from phylogenetic considerations in taxonomic procedure

1. The available fossil record is so fragmentary that the phylogeny of the vast majority of taxa is unknown. Evolutionary branching sequences must be inferred largely from phenetic relationships among existing organisms.

2. Phenetic classification is possible for all groups. By contrast, cladistic classification, based on branching sequences, requires historical inferences about the direction of evolution in a group of organisms.

3. Even when fossil evidence is available, this evidence itself must first be interpreted in a strictly phenetic manner with the exception that a time scale is given in addition, which may restrict certain interpretations of the phylogeny-since the criteria for choosing the ancestral forms in a phylogeny are phenetic and are based on the phenetic relationship between putative ancestor and descendant.

4. From the point ofview of biology in general, it is probably of more interest to describe the overall similarity of organisms than their branching sequences. If the classifications are to have predictive value, it is evident that those based on overall similarity will be most predictive.

The advantages of numerical taxonomy

1. Numerical taxonomy has the power to integrate data from a variety of sources, such as morphology, physiology, chemistry, affinities between DNA strands, amino acid sequences of proteins, and more. This is very difficult to do by conventional taxonomy.

2. Through the automation of large portions of the taxonomic process, greater efficiency is promoted. Thus, less highly skilled workers or automata can do much taxonomic work.

3. The data coded in numerical form can be integrated with existing electronic data processing systems in taxonomic institutions and used for the creation of descriptions, keys, catalogs, maps, and other documents.

4. Being quantitative, the methods provide greater discrimination along the spectrum of taxonomic differences and are more sensitive in delimiting taxa.

5. The creation of explicit data tables for numerical taxonomy has already forced workers in this field to use more and better-described characters.

6. A fundamental advantage of numerical taxonomy has been the reexamination of the principles of taxonomy and of the purposes of classification. This has benefited taxonomy in general, and has led to the posing of some fundamental questions.

7. Numerical taxonomy has led to the reinterpretation of a number of biological concepts and to the posing of new biological and evolutionary questions.

Problems of estimating phenetic relationships

There are four major problems of phenetic classifications.

1. Incongruence between classifications based on different parts of the body or different life history stages.

2. Differences in estimates of relationships produced by different similarity coefficients.

3. Differences in interpreting relationships produced by different clustering methods.

4. Possible effects of parallelism and convergence on taxonomic judgments based on estimates of phenetic relationships.