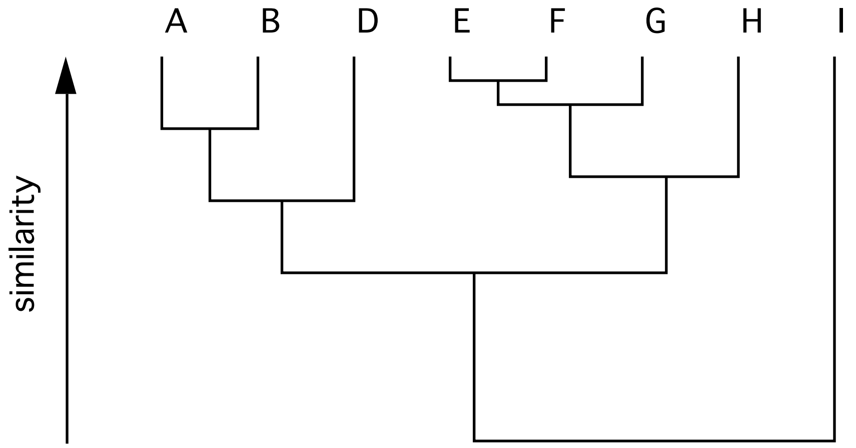
CLUSTER ANALYSIS

Used to extract of patterns of hierarchical structure from distance or similarity matrix.

Results expressed as dendrogram linking objects (some times called OTUs- operational taxonomic units) or more rarely variables.

eg.

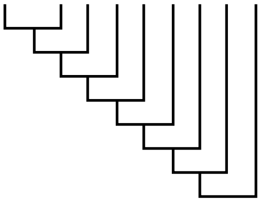


Algorithms are either agglomerative (making big clusters from little ones) or divisive (splitting big clusters into little ones). Methods that follow are all agglomerative.

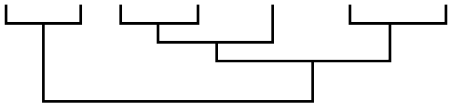
A. Single Linkage. OTU joins cluster based on its highest similarity with any one member of the cluster. Sometimes called "Nearest Neighbor" — widely used in evolutionary studies.

Space distorting- space in vicinity of clusters seems to contract as more OTUs join cluster. Leads to chaining

e.g.



B. Complete Linkage. OTU joins cluster based on its lowest similarity with any one member of the cluster.



Space distorting. Space in cluster appears to expand due to the way things join. Not used much.

Rest of agglomerative methods are all based on multiple linkage joins. Are all space conserving.

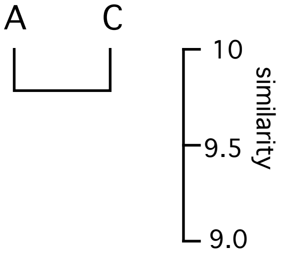
C. WPGM. Weighted pair group method. Group joins cluster based on its average similarity with previous members of cluster. Averages are weighted.

Similarity with last member to join cluster is weighted equally with that of all previous members of cluster.

e.g. Similarity matrix:

	A	C	E	F
A	-	9.78	9.34	8.55
C	9.78	-	9.5	8.71
E	9.34	9.5	-	9.02
F	8.55	8.71	9.02	-

Step 1: join A to C as they have the greatest similarity



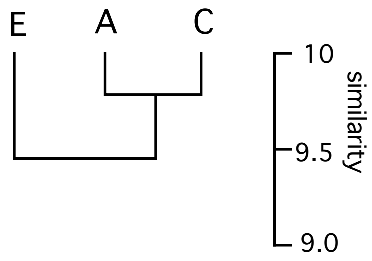
Step 2: treat AC as a single unit and recalculate similarities

E vs AC = (9.34 + 9.5)/2 = 9.42
F vs AC = (8.55 + 8.71)/2 = 8.63

	AC	E	F
AC	-	9.42	8.63
E	9.42	-	9.02
F	8.63	9.02	-

note: similarity of F & E are not effected by previous cluster.

Step 3: join E with AC as it has the greatest similarity

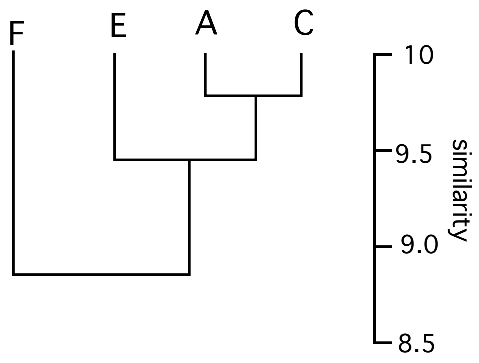


Step 4: treat ACE as a single unit and recalculate

F vs ACE = (9.02 + 8.63)/2 = 8.83

	ACE	F
ACE	-	8.83
F	8.83	-

Step 5: join F with ACE



Stop when all OTUs are clustered

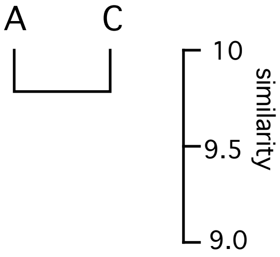
D. UPGM. Unweighted pair group method. OTU joins clusters based on average similarity with previous cluster. Averages are unweighted.

UPGM is the most common clustering method because it minimizes the amount of distortion between dendrogram and similarity matrix.

e.g. Similarity matrix:

	A	C	E	F
A	-	9.78	9.34	8.55
C	9.78	-	9.5	8.71
E	9.34	9.5	-	9.02
F	8.55	8.71	9.02	-

Step 1: join A to C as they have the greatest similarity



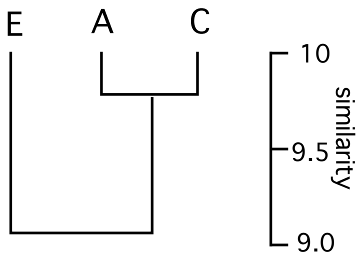
Step 2: treat AC as a single unit and recalculate similarities

E vs AC = (9.34 + 9.5)/2 = 9.42
F vs AC = (8.55 + 8.71)/2 = 8.63

	AC	E	F
AC	-	9.42	8.63
E	9.42	-	9.02
F	8.63	9.02	-

note: similarity of F & E are not effected by previous cluster.

Step 3: join E with AC as it has the greatest similarity

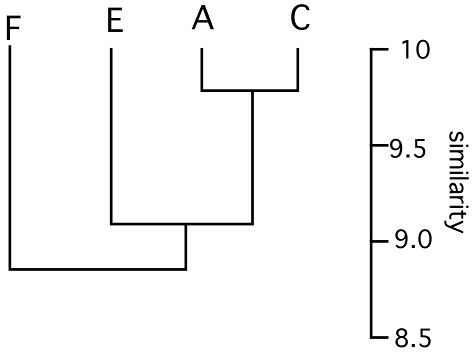


Step 4: treat ACE as a single unit and recalculate

F vs ACE = (8.55 + 8.71+ 9.02)/3 = 8.76

	ACE	F
ACE	-	8.76
F	8.76	-

Step 5: join F with ACE



Stop when all OTUs are clustered

- Ward's Method (sometimes called minimum variance)

Based on the least amount of increase in the sum of squared deviations from cluster means. Must cluster using squared Euclidean distances.

Basic precept: calculate error sum of squares (ESS) between each pair of OTUs- then join together the two that gives the minimum ESS. An ESS for any given cluster is calculated as:

$$ESS = \sum_{i=1}^{m}\left[\sum_{k=1}^{p} x_{ik}^2 - \frac{1}{p}\left(\sum x_{ik}\right)^2\right] \text{ where ...}$$

i = character
k = OTU
m = number of characters
p = number of OTUs in cluster

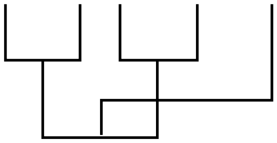
Difficult to compute without computers.

Beware- some data may produce "stray branches" that do not connect to other clusters. If this happens, try an other technique e.g. UPGM

- Centroid Method

Uses unweighted (or weighted) averages (or medians) typically on distance measures (squared Euclidean distances) between variables.

note the centroid may move with the addition of new clusters- may lead to reversals



Divisive methods

Not used much. Attempt to minimize the variance within groups and maximize the variance between groups.

Phenon Line

Can have from 1 to n different clusters. What's the cut off (phenon line)? Can there be any significant number of clusters for a given data set? No. Must make a subjective decision for the cutoff line.

Clustering is NOT a statistical procedure- there is no way of telling if data are more or less clustered than would be expected from a random population.