Harvard-MIT Division of Health Sciences and Technology
HST.508: Genomics and Computational Biology

Hierarchical Methods

Hierarchical clustering is an iterative procedure in which n data points are partitioned into groups which may vary from a single cluster containing all n points, to n clusters each containing a single point. Hierarchical clustering techniques can be divided into "agglomerative" and "divisive" methods. In the former, clusters initially containing one element each are successively fused to generate larger clusters. At each step, the clusters to be fused are those that are, according to some predefined metric, most similar ("closest") to each other. In the latter, a large cluster is divided into successively smaller clusters. Both fusions and divisions are irreversible.

Complete Linkage Clustering

Defining Feature: Distance between groups is defined as that of the furthest pair of individuals, where a pair consists of one member from each group.

Method: Complete linkage clustering is an agglomerative method. Therefore, the two clusters with the lowest average distance are joined together to form the new cluster.

Example:
Cluster elements 1 through 5 whose inter-element distances are represented in the following *distance matrix*.

	1	***2***	***3***	***4***	***5***
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Solution:

Repeatedly execute the following procedure:
i) identify the "closest" elements in the distance matrix
ii) fuse them into a cluster
iii) compute the new distance matrix

i) In the distance matrix above, individuals 1 and 2 are closest, so they are joined to form a two-member cluster.

ii) *The distances between this cluster and the remaining elements in the distance matrix are computed as follows:*

$d(12)3 = max[d13,d23] = d13 = 6$

$d(12)4 = max[d14,d24] = d14 = 10$

$d(12)5 = max[d15,d25] = d15 = 9$

iii) Record these distances in the new distance matrix:

	(12)	3	4	5
(12)	0			
3	6	0		
4	10	4	0	
5	9	5	3	0

Repeat:

i) the smallest entry in the new distance matrix is that for elements 4 and 5, so these form a new cluster

ii) compute the new distances:
d(12)3 = 6
d(12)(45) = max[d(12)4, d(12)5] = 10
d(45)3 = max[d34,d35] = 5

iii) record these in the new distance matrix:

iv)

	(12)	3	(45)
(12)	0		
3	6	0	
(45)	10	5	0

Repeat:

i) the smallest distance is now d(45)3 = 5, so element three is added to the cluster containing elements 4 and 5.

ii) etc.

Producing a dendrogram of these results:

The partitions produced at each stage are as follows:

Stage	Groups
P5	[1], [2], [3], [4], [5]
P4	[1 2], [3], [4], [5]
P3	[1 2], [3], [4 5]
P2	[1 2], [3 4 5]
P1	[1 2 3 4 5]

Corresponding dendogram: