What is Hierarchical Clustering?

Introduction

Hierarchical clustering is an agglomerative (top down) clustering method. As its name suggests, the idea of this method is to build a hierarchy of clusters, showing relations between the individual members and merging clusters of data based on similarity.

In the first step of clustering, the algorithm will look for the two most similar data points and merge them to create a new "pseudo-datapoint", which represents the average of the two merged datapoints. Each iterative step takes the next two closest datapoints (or pseduo-datapoints) and merges them. This process is generally continued until there is one large cluster containing all the original datapoints. Hierarchical clustering results in a "tree", showing the relationship of all of the original points.

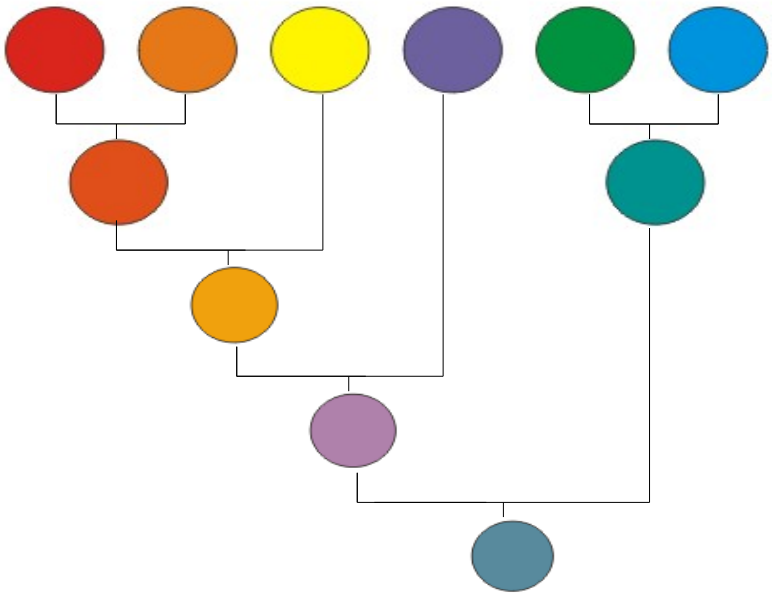
An example of agglomerative clustering using colours

The example below demonstrates the agglomerative clustering using colours.

In the first stages, the red and orange are combined to make reddish-orange. Green and blue are similar to each other and are merged to form aqua.

Next, the reddish-orange and yellow are the closest colours, and create a light orange. Now we have three remaining colours: light orange, the original purple and aqua. The hierarchical clustering algorithm is only interested in finding the closest relationship at each stage regardless of the degree of similarity. So in this case, the light orange and purple are more similar than the any other combination.

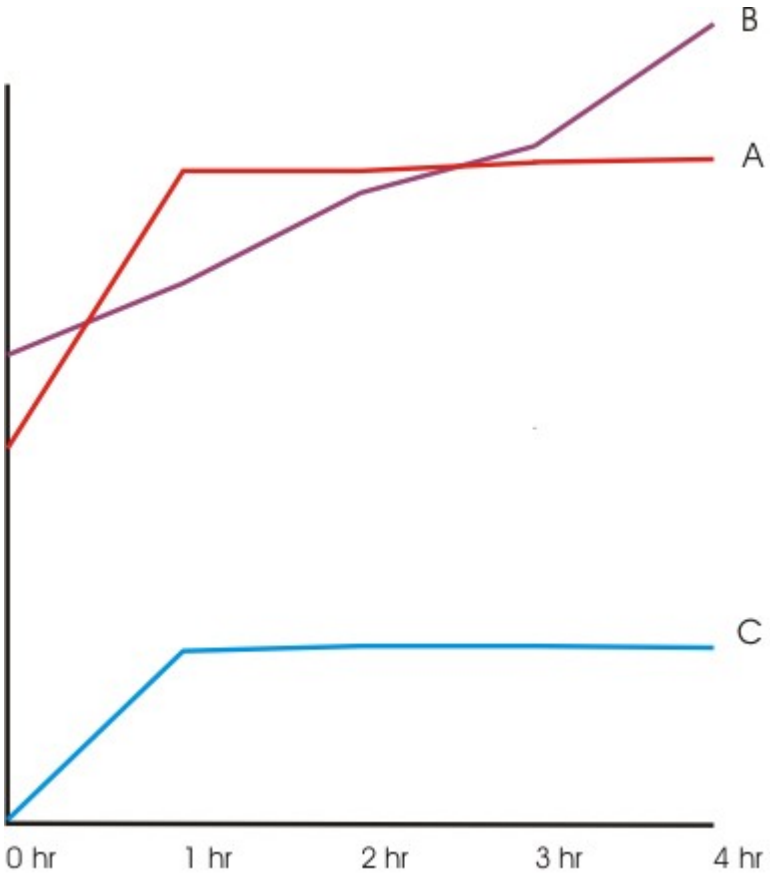
We are now left with the orange-purple mixture and aqua. They are now the most similar samples left and so form a final cluster. The resulting tree shows the overall similarity of the original samples (this is the hierarchy).



Distance Metrics

In order for the clustering algorithm to work, we need to have some means by which similarity is judged. This is generally called a distance measurement. Two commonly used metrics for measuring correlation (similarity/distance) are the Euclidean and the Pearson correlations. The type of correlation metric used depends largely on what it is that you are trying to measure.

Take a look at the following example. In the case of gene-expression data (microarrays), which pair of genes from the graph below would be considered the most similar in terms of expression?



Most likely, you felt genes A and C were behaving most similarly. Despite an overall difference in the overall amount of expression, the trends between the two genes is nearly identical.

The Pearson correlation metric would agree with you in this case as it looks more at trends than overall distances. Were you to use a Euclidean distance measurement, genes A and B would most likely have been found to be more similar because at each time point the two genes were much closer together than they were to C.

You must decide based on what you are measuring which metric is most suitable to your situation. For most gene expression experiments you will likely find Pearson correlations to be more appropriate.