
Assignment 8 Cluster analysis of sequence data

1. Re-create the state sequence object `biofam.seq` and the matrix `dOM` of pairwise OM dissimilarities based on the properties matrix considered in the previous assignment.

2. Create the hierarchical cluster tree object once with the Ward method and once with WPGMA (McQuitty) method, and using the sequence weights in each case (Tip: retrieve the weights with `attr(biofam.seq, "weight")`.)

3. Display both hierarchical trees side by side.

4. Select the three-cluster solution from the Ward analysis, and label the clusters by looking at the I-plots by cluster.

5. Make and comment the silhouette plot of the retained solution. (Tip: Use the `silhouette` function of the `cluster` package.)

6. Look at the partition quality measures returned for the Ward-three-cluster solution by the `wcClusterQuality` function (library `WeightedCluster`) when specifying the sequence weights. Compare the average silhouette with the value obtained with `summary(silhouette(...))` which does not account for weights.

7. Study with logistic regressions how the cluster membership is related to the sex, birthyear, and the language of the questionnaire.

8. In order to make a PAM clustering, first examine the evolution (values and plot) of the quality indicators for the solutions for $k = 2, \ldots, 20$. What value k do you retain?

9. Label the clusters of the PAM solution for the chosen k and plot the mean time spent in the states by cluster.