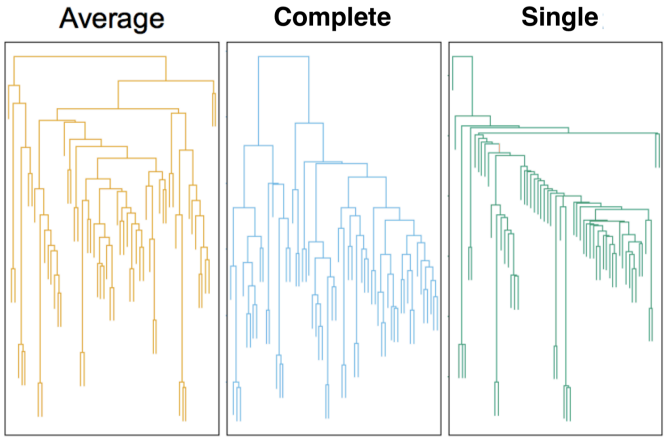
CS466 Project Information (4 credits)

Team member: Yuwei Chen (yuwei3)

Project Name: Comparison of hierarchical clustering algorithms (single linkage, complete linkage and centroid) and k-means.

Goal:

Inspired by course lecture about "Clustering III", I decide to work on algorithm comparison of hierarchical clustering and partitional clustering.



Mouse tumor data from [Hastie *et al.*]

In the sea of DNA and protein sequences, a majority of these sequence remain unknown to people. To analyze tremendous sequences efficiently, it is helpful to group these biological sequences which are somewhat related.

Different clustering algorithms will produce different results on the same data. Even with the same data, partitional algorithm such as k-means may produce different cluster every time it runs.

Besides, when using k-means for clustering, people need to specify the number of clusters based on heuristic beforehand. This can be a hard decision when few background knowledge is available on dataset. This problem can be eased when using agglomerative clustering, because the generated tree may correspond to a meaningful taxonomic structure of data.

I would like to investigate how far apart two clustering results are of the same

data using different clustering algorithm, and compare the results of hierarchical clustering using different proximity methods, i.e. single linkage, complete linkage and centroid linkage.

Data: QIIME datafiles 13_8
(http://qiime.org/home_static/dataFiles.html)