ISA Project:
Parameterless clustering by dynamic tree-cutting

Simon Lehmann Knudsen
simkn15@student.sdu.dk
ECTS: 10
04/09-2017 - 31/01-2018
5th semester in Computer Science

The suggested ISA shall provide the candidate Simon Lehmann Knudsen insights into clustering and investigate parameterless clustering by dynamic tree-cutting.

Background

Clustering, or cluster analysis, is a way of grouping a set of objects such that a cluster of objects is more similar to each other than those of another cluster. Clustering can help with the description of patterns of similarities and differences in a data set. There are many different tools to do cluster analysis, which all intend to find an optimal clustering depending on a set of criteria. Given a set of criteria and parameters, we can analyze the data and discover the clusters. The resulting clusters can all be either feasible or infeasible.

Method

We want to investigate if it is possible to analyze clusters by dynamic tree-cutting. Cluster validity indices, also called clustering quality measures, are an objective criteria for judging the quality of a clustering and/or compare a clustering against a gold standard. They also help to decide whether a clustering is feasible for a given dataset and they assist to detect an optimal parameter set resulting in the best possible clustering for a given tool. Dynamic tree-cutting allows each branch to be cut when the quality of the cluster is feasible, resulting in an optimal cluster analysis. The final tree may have branches cut on different levels.

We intend to use TransClust as the basis by extending it to a hierarchical clustering approach in R. For each split in the dendrogram we create a randomly sampled dataset following the similarity distribution as the original clustering

and use the total edit costs of TransClust to assess the cost difference between the random dataset vs. the actual dataset. The maximal difference will indicate the optimal split through the clustering. We want to apply this to sequenced based protein homology dataset for which it is already shown that basically each protein family would require its own parameter set which is inherently difficult to achieve with standard clustering methods.