Advanced Bioinformatics Final Report: Hierarchical Clustering of Protein sequences utilizing minimum energy cover pruning.

Brendan Benshoof

April 30, 2014

1 Overview

When searching for a topic upon which to base our semester project, we found that while multiple clustering approaches had been applied to subsets of the available protein sequences. Upon discussion of methods to prune the large all-to-all graph that would be needed to cluster larger sets of proteins then previously clustered. The intent to only remove edges that did not provide useful information to later clustering caused me to connect the problem with an existing problem in sensor networks, which was to minimize the number (and range of) connections between battery powered wireless sensors. We decided to attempt to apply the technique to a graph of protein sequences.

We broke the project into four parts: Fetcher, Graphifier, Clusterizer and Visualization. The names were chosen to give each group (and code repository) a useful name. As we approached the end of the semester, Olga deemed it responsible to add an additional section to act as validation for our model.

1.1 Fetcher

The Fetcher was planned to be a collection of scripts for the retrieval of protein sequences and meta-information for use in the graph generation and visualization processes. In addition, as we did not expect to be able to run on the entire PDB they were tasked with generating an initial small subset of proteins for which there existed a ground truth of clustering for validation of our process. In general, the only one of these goals achieved was to write a script to assign a cath-id to a given protein. I implemented a script to translate PDBNR to XML format for use by the clustering group. Once the PDBNR was converted to XML, we saw no other contributions for this group.

1.2 Graphifier

Given the sequences of every protein in the PDBNR, we needed to generate a graph of inter-sequence similarity. This required an alignment and similarity scoring of every sequence against every other sequence.

1.3 Clusterizer

1.4 Visualization

1.5 Validation

2 Graph Pruning and Hierarchical Clustering

The aim of your individual contribution /What you planned to do

3 Accomplishments

4 Results