Homework 4 Report

Tom Bedor

My program is a reasonably close implementation of GLIMMER as described in the original paper. The main difference is that it does not take frame position (ie position 1,2, or 3 in the codon) into account for scoring the kmers.

In the ORF generation step, the sequence is parsed through in each of the three reading frames. An ORF is defined as the sequence between stop codons. In the pruning step, each ORF that is completely contained in some other ORF is removed. This is an n\*n step and is time consuming.

Next, the training set is created by counting all ORF’s with a length of longer than 1,000 as genes. If no ORF’s of this length are found, then the strand is assumed to be the reverse compliment of the coding strand, and no genes are predicted. A background set is composed of all ORF’s with a length of less than 150.

K-mers are then counted, for k=1 to k=9, for both training and background set. Gene scores are calculated with the formulas given in the 1998 GLIMMER paper.

For validation I applied the algorithm to the test set. 97% of the long ORF’s were counted as genes, compared with about 27% of the rest. I couldn’t make much sense of the provided predictions.200 file, as the ORF’s corresponding to the indices didn’t seem to correspond to either stop or start codons.