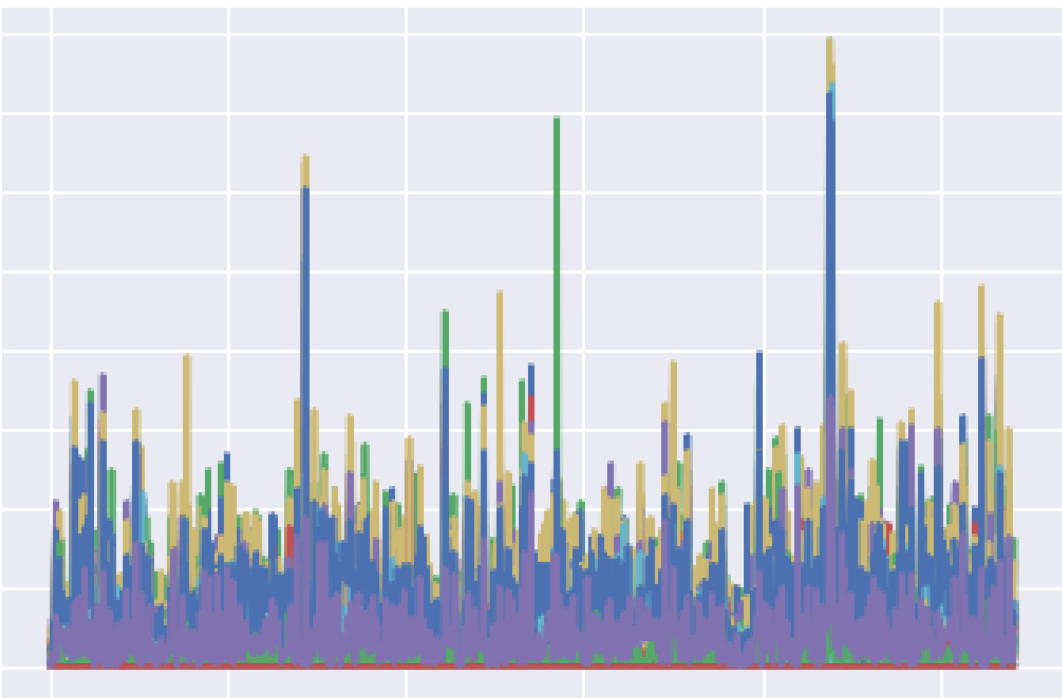
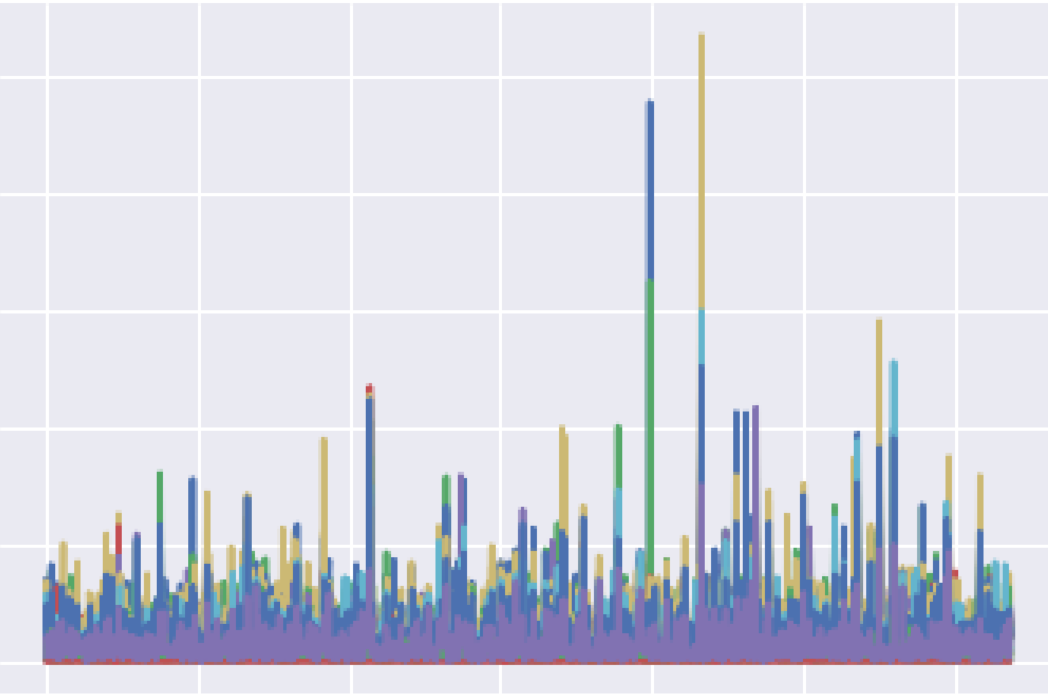
10/12/2017 – first run, below results are invalid.

A quick plot of non-TM negative samples:



Non-TM Positive samples:



Above graphs are distribution of data features, each bar represents a count of a particular protein/proteins combination. The difference in their distribution indicate a significant difference between Negative and Positive data samples.

I quickly realised the distribution include the commenting lines in the fasta files. Which provide hints to whether a sample contains signal peptide or not.

Further accuracy check reveals unrealistic 100% prediction results, which lead to further investigation.

