Predicting the Interpro accession of large protein using smaller Interpro accessions



Student: Pieter de Jong

Student number: 441583

Teacher: Martijn Herber

Course: DSLS

# Final Assignment

Project 1: InterPRO prediction

**Goal:**

The goal of this project is to predict the largest Interpro accession using the smaller interpro accessions. The largest interpro accession must cover >90% of the protein’s sequence, the other accessions are smaller interpro accessions. Protein’s without a larger accession are removed.

**Data Analysis:**

The distribution of feature length of large and small Interpro accessions are analyzed using boxplots and histograms

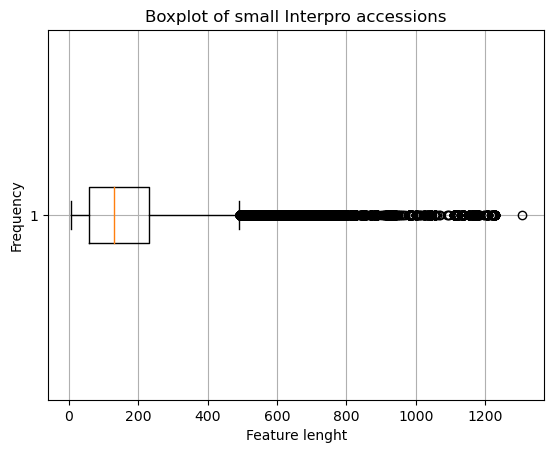


Figure 1: boxplot of small Interpro accessions

A red graph with white background

Description automatically generated

Figure 2: histogram of small Interpro accessions

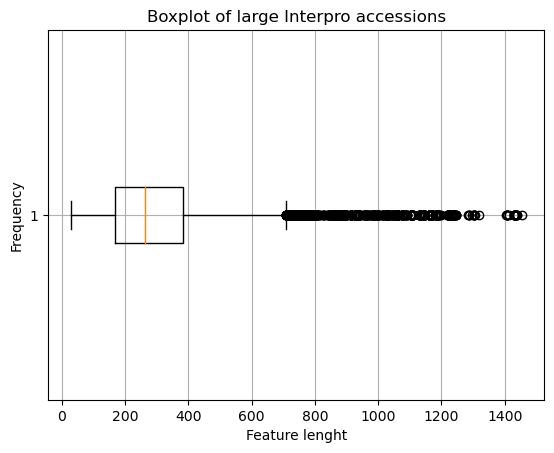


Figure 3: boxplot of large Interpro accessions

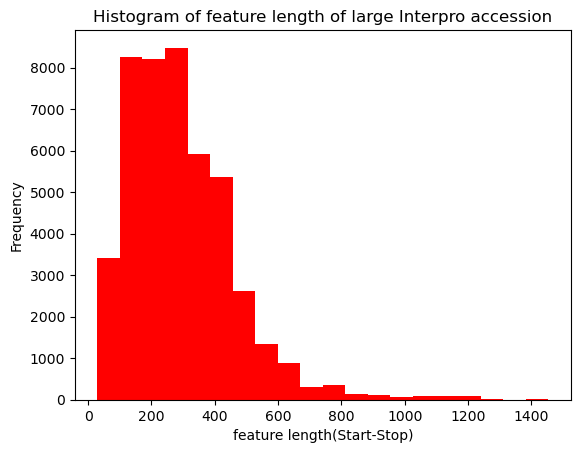


Figure 4: Histogram of large Interpro accessions

**Pre-processing:**

The data is pre-processed by taking columns 0,2,6,7,11 and renaming them "Protein\_acc", "Seq\_length", "Start", "Stop", "Interpro\_acc".

**Methods:**

To generate the boxplot and histogram, pyplot was used in ipython notebook. Data preparation and machine learning was done using DASK. Proteins were selected based on the precense a large Accession and smaller accessions. If multiple large accessions were found, the largest was taken for prediction and the others were added to the smaller accessions. For machine learning, Randomforest classifier was used. This is an ensemble method that works with large datasets. The data is devided into 2 sets, 75% for training and 25% for validation. To make the data classes suitable for sklearn, One Hot Encoding was used on the classes.

**Restults:**

The model was trained on 1 million lines of the original dataset. This resulted in an accuracy of 86%. This model is exported to a pickle file and the data used to a csv file. These files are saved as rfmodel.pkl and trainingdata.csv

A screen shot of a computer

Description automatically generated

Figure 5: Accuracy of the Random forest model using 1 million lines