Report of final assignment (Programming 3)

Author: Azadeh Pirzadeh

Topic: make a model that predicts the largest feature of protein, based on small features of that protein.

In this python code I used Dask to be able to work on big data so, I split the data into 20 partitions.

In the first step, I removed the noise and calculated the length of each feature of the proteins, If the length was more than 90% of the protein length I categorized it as a long feature and used the column "class_p" to identify the long features and small features.

In the next step, I removed proteins that have no small and no longer features which mean just keeping the proteins that have at least one long and one small feature.

Then I divided the Dask data frame into two separate Dask data frames: one with small features of each protein and another with large features of each protein. Because some proteins have more than one large feature so I kept just the largest of the long features.

After that, I made a table that contains protein_accession as each row and interpro_accession as each column and the value of each cell included the counts of that feature in that special protein, then merged the table and the data frame of the largest feature base on protein_accession.

Finally, I divided the merged table into a train and test set and used it in the random forest model.

As a result, I got 33% accuracy on 100.000 lines of “all_bacilli.tsv” and 44.3% on 1.000.000 lines of that. Those accuracies were calculated in the average result of 200 trees. You can see the second result in the result.txt. I could not have enough time to run it on the whole file or use the structure of the server and clients to improve the speed of the run.

I believe that it is not a very good result but it can be a base for working more on this model. In the next step by using some Ensembling methods and changing the parameters, we can improve the accuracy.