Leonardo Hügens (201705764) | Hynek Šabacký (202311458) | Lohras Karimi (202311453)

Group Assignment 3 Report

Approach

The approach we used was to read the amino-acid sequences in the zincfinger.fasta and globin.fasta files into dictionaries of frequencies of pairs of amino acids, along with the identifiers of the sequences, and also creating a “class” label that indicates to with type of sequences they correspond to: 0 for zincfinger and 1 for globins. Then, we used three machine learning models: scalar vector machine (svm), Random Forests and NaiveBayes. We used a 10-fold stratified cross-validation procedure, and evaluated those models according to three metrics: recall, precision and f1-score. For each metric, we recorded it’s average and standard deviation.

Difficulties

A difficulty we had was to deal with the presence of 3 letters that are not amino-acids: “X”, “Z” and “B”. These are place-holders to indicate that in those places there is ambiguity of amino acids.

Tools and Packages

The tools and packages we used were: argparse, numpy, pandas and sklearn.

Results

	recall	precision	f1-score
SVM	**0.9861**	**0.9861**	**0.9861**
Random Forests	0.9583	0.9583	0.9583
NaiveBayes	0.9514	0.9514	0.9514

Conclusion

We find that scalar vector machines work best as a model to discriminate amino acid sequences between zincfinger or globin sequences, with a recall value of 0.9861.