BIOINFORMATICS - REPORT GROUP-ASSIGNMENT 3

Group members:

- Diogo Ferreira - 201805258 - MCC
- Sara Rescalli - 202210943 - Mobilidade
- Telmo Ribeiro - 201805124 – MCC

The goal of this assignment is to create a predictive model for binary classification of protein sequences from the globin and zinc finger families by employing the Feature Frequency Profiles (FFP) approach.

DATA PRE-PROCESSING

To convert the protein sequences into a suitable tabular representation, we have developed a function called "*generate_FFP_dataframe(fileA, fileB, k)*", that takes as input the name of two FASTA files and creates a pandas dataframe with the FFP values for all the sequences present in the input files. The function utilizes the *"get_all_kmers(k)"* function to generate all possible k-mers of the protein alphabet, thanks to the *product* function from the *itertools* library. The resulting k-mers are used as column names in the dataframe. Furthermore, the function iteratively calls the "*get_kmers_frequency(sequence, k)*" function, which takes a sequence as input and returns a dictionary containing the frequency of each k-mer in the sequence normalized by the total number of 2-mers in the sequence. Each resulting dictionary is added as a row to the dataframe. Moreover, an additional column labeled "Class" was added to the dataframe ("zinc-finger" = 1, "globin"= 0 – we choose as a positive label the less numerous class).

CLASSIFICATION MODELS

After preprocessing the data, a predictive model for the binary classification problem was developed. The objective is to predict the type of protein based on the FFP data. In order to accomplish this, we have applied three popular machine learning algorithms: Random Forests, Support Vector Machines (SVM), and Naive Bayes. The implementation of these algorithms is done using the *sklearn* library. The performance of these models is tested with Stratified 10-fold cross-validation using four metrics: accuracy, recall, precision, F1-score. The code provides a dataframe that presents the average and standard deviation across the 10 folds for all the metrics and all the applied machine learning algorithms.

	mean_acc	mean_rec	mean_prec	mean_f1	std_acc	std_rec	std_prec	std_f1
RF	**0.99792**	**1.00000**	**0.98977**	**0.99479**	**0.00317**	**0.00000**	**0.01562**	**0.00795**
SVM	0.99584	0.98940	0.98952	0.98940	0.00555	0.01618	0.01601	0.01422
NB	0.95568	1.00000	0.82235	0.90078	0.02204	0.00000	0.07145	0.04423

Tab. 1- Performance of the 3 machine learning algorithms used for classification ("zinc-finger" = 1, "globin"= 0).

CONCLUSION

The results reveal that RF outperformed SVM and NB in terms of mean accuracy, precision, recall, and F1 score. In particular, RF achieved the highest mean accuracy of 0.99792, indicating that has the highest proportion of correct predictions on average. Moreover, RF and NB achieved perfect recall, indicating that they correctly identified all instances of the zinc-finger family. RF has also the highest mean precision, indicating a low proportion of false positives (few misclassified globin). These results are a little unexpected because it means that for RF and NB it's more difficult to correctly classify the most numerous class (the globin family) than the less numerous one (the zinc-finger family). Nevertheless, the F1 scores suggest that RF has the best overall balance between precision and recall, and NB the worst. Looking at the standard deviations, it is possible to see that the RF is also the algorithm with the lower variability and so the highest consistency across the different runs of the cross validation.
It's possible to conclude that the RF succeeds very well in classifying the family of the proposed sequences.