# Protein Sequence Classification

Jakub Duda, Soulaimane Salehddine, Mariana Lourenço

May 28, 2024

## Introduction

This report summarizes our exploratory project which aims to classify protein sequences from two families—globin and zinc finger—based on their dinucleotide composition using machine learning techniques. The objective is to transform the sequence data into a tabular format and predict the respective protein family using various machine learning classifiers.

## Methods

We analyzed two protein sequence families provided in FASTA files: globin and zinc finger. Globins are heme-containing proteins involved in oxygen binding/transport, while zinc-finger proteins interact with DNA, RNA, and other molecules. The sequences were processed to extract dinucleotide composition for classification.

### Data Collection:

Two protein sequence families, globin and zinc finger, were provided in FASTA files. Globins are heme-containing proteins involved in oxygen transport, while zinc-finger proteins interact with DNA, RNA, and other molecules.

## Classification

In the classification pipeline, we used Scikit-learn to test three machine learning algorithms: Support Vector Machine (SVM), Random Forest, and Naive Bayes. Stratified k-fold cross-validation with 10 folds was employed to evaluate the performance of these models, ensuring balanced class representation across the folds. The models were evaluated using metrics such as accuracy, precision, recall, and F1-score, with the mean and standard deviation of these metrics calculated across all folds.

## Tools and Packages

- **Biopython:** For reading and parsing FASTA files.
- **Pandas:** For data manipulation and tabular representation.
- **Scikit-learn:** For building and evaluating machine learning models.

## Results

The results showed that the Random Forest classifier outperformed the others, achieving the highest average F1-score, precision, recall, and accuracy. This suggests that Random Forest is well-suited for capturing the patterns in the k-mer compositions of protein sequences. SVM also performed well, while Naive Bayes showed relatively lower performance, indicating it may be less effective for this feature representation.

| | | accuracy | precision | recall | f1 |
|---|---|---|---|---|---|
| SVM | mean | 0.993769 | 0.993933 | 0.993769 | 0.993754 |
| | std | 0.004845 | 0.004659 | 0.004845 | 0.004848 |
| Random Forest | mean | 0.995148 | 0.995312 | 0.995148 | 0.995132 |
| | std | 0.006236 | 0.006029 | 0.006236 | 0.006265 |
| Naive Bayes | mean | 0.961231 | 0.967829 | 0.961231 | 0.962566 |
| | std | 0.013602 | 0.010131 | 0.013602 | 0.012873 |

Figure 1: Protein Classification

## Difficulties

Several difficulties were encountered during the project, including ensuring correct parsing of the FASTA files, efficiently generating and counting k-mers for large sequences, and selecting appropriate algorithms and tuning their hyperparameters for optimal performance.

## Conclusion

In conclusion, this study demonstrated that protein sequences can be effectively classified into their respective families based on k-mer composition using machine learning. The Random Forest classifier showed the best performance, highlighting its capability to handle the complexity of protein sequence data. Future work could explore other feature representations and advanced machine learning techniques to further improve classification accuracy.