Approach

The goal of this project is to classify protein sequences into two categories: Globin and Zinc Finger proteins. The steps done to achieve this are the following:

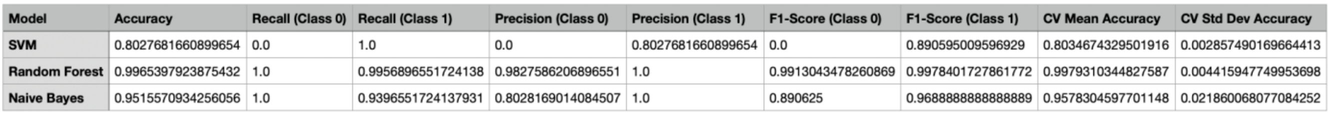
1. **Reading Protein Sequences:** We read protein sequences from FASTA files and store them in a dictionary.
2. **Feature Extraction:** The sequences are processed to generate all possible 2-mers of amino acids, after that we calculate the FFPs for each sequence.
3. **Data Preparation:** The FFPs are compiled into a DataFrame, and a class label (Globin=1, ZincFinger=0) is added to each sequence.
4. **Model Training and Evaluation:** We do a train test split on the dataset and then use three machine learning models: SVMs, Random Forests and Naive Bayes.

Tools and Packages Used

- **Libraries:**
  - **Pandas:** For data manipulation and creation of DataFrames.
  - **Scikit-learn:** For machine learning models, cross-validation, and evaluation metrics.
  - **System Module:** For command-line argument parsing.

Main Results

The performance of the classifiers was evaluated using the following metrics: accuracy, recall, precision, and F1-score, as well as cross-validation mean accuracy and standard deviation. The results are the following:



Model	Accuracy	Recall (Class 0)	Recall (Class 1)	Precision (Class 0)	Precision (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)	CV Mean Accuracy	CV Std Dev Accuracy
SVM	0.8027681660899654	0.0	1.0	0.0	0.8027681660899654	0.0	0.890595009596929	0.8034674329501916	0.002857490169664413
Random Forest	0.9965397923875432	1.0	0.9956896551724138	0.9827586206896551	1.0	0.9913043478260869	0.9978401727861772	0.9979310344827587	0.004415947749953698
Naive Bayes	0.9515570934256056	1.0	0.9396551724137931	0.8028169014084507	1.0	0.890625	0.9688888888888889	0.9578304597701148	0.021860068077084252

Conclusion

We successfully used machine learning models to classify protein sequences and concluded that Random Forests were the best model out of all beucase it performed the best on all metrics except CV mean standard deviation.