Ryerson University

Department of Electrical & Computer Engineering
Faculty of Engineering & Architectural Science

Course Title:	Computations in Genetics
Course Number:	BME 808
Semester:	Winter 2021

Instructor:	Dr. Prathap Siddavaatam
TA:	Amirreza Rezvantalab

Assignment/Lab Title:	Final Project - Protein Fold Recognition Using Support Vector Machines

Due Date:	Monday, April 5th, 2021
Submission Date:	Monday, April 5th, 2021

Student LAST NAME	**Student FIRST NAME**	**Student Number**	**Section**	**Signature**
Nowicka	Otylia	500750409	03	O.N
Mullen	Andrew	500787631	02	A.M

*By signing above you attest that you have contributed to this written lab report and confirm that all work you have contributed to this lab report is your own work. Any suspicion of copying or plagiarism in this work will result in an investigation of Academic Misconduct and may result in a "0" on the work, an "F" in the course, or possibly more severe penalties, as well as a Disciplinary Notice on your academic record under the Student Code of Academic Conduct, which can be found online at: http://www.ryerson.ca/senate/current/pol60.pdf

Table of Contents

1. Introduction

1.1 Background

Protein folds is the process by which the shape of a protein chain is translated into its three-dimensional structure [2]. The overall shape of a protein depends on the level of the protein structure. The stages of protein folding are the primary structure, secondary structure, tertiary structure and quaternary structure. Each level of structure increases the complexity of the protein structure [2]. Primary structure is the simplest level of protein structure and involves the sequence of amino acids within a polypeptide chain. Each chain is assembled in a specific order, different from each other. Any change within the DNA of the gene that encodes the protein, can affect the function and structure of the protein [2]. The secondary structure refers to the local folded structures that interact with the R groups. Alpha-helix and beta-pleated sheets are the most common types of secondary structures that are held together by hydrogen bonds [2]. The tertiary structure is a three-dimensional structure of the polypeptide, that is caused by the interactions between the amino acids of the proteins and it's R groups. The quaternary structure consists of the first three structures and is composed of a single polypeptide [2]. In some cases the proteins are made up of more than one polypeptide chain. When these multiple polypeptide chains come together, they form what is called a subunit and this is what gives the protein it's quaternary structure [2].

1.2 Motivation

Using support vector machines allows the entire amino acid sequence to be used to be able to predict the three-dimensional structure of a protein [1]. Extracting different features from the protein sequences is one of the approaches that will be used to be able to predict protein folding. This is an important attribute to have as it will be able to predict the functionality and behaviour of an unknown sequence, based on the result of the prediction [2].

2. Related Works

2.1 Components

The "ProFold: Protein Fold Classification with Additional Structural Features and a Novel Ensemble Classifier" article was used as a guideline to implement the classification of the protein folds. Feature extraction was done on the SCOP database and amino acid frequencies as well as physicochemical properties were extracted [1].

3. Methodology

3.1 Libraries

3.1.1 Python 3.5

Python 3.5 is a general coding language that will be used, as it is used in many types of programming and software developments [4]. Unlike other programming languages, Python 3.5 emphasizes code readability, that allows the use of English words instead of punctuation, which makes it one of the simplest programming languages.

3.1.2 Sklearn

Sklearn will be used, as it is one of the most useful libraries for machine learning in Python. Sklearn contains many different tools and different types of classification, regression and clustering statistical models for machine learning [3].

3.1.3 NumPy

NumPy is a Python library that will be used to provide a powerful data structure. NumPy is very useful for being able to work with arrays, as well as mathematical functions and logical operations [4].

3.2 Software Platforms

3.2.1 Jupyter notebook

Jupyter Notebook will be the software platform used. Jupyter Notebook is an open source web application that enables users to create and share documents that contain live codes, equations and texts [5]. Jupyter Notebook supports multiple different programming languages such as Python 3.5.

3.3 Database

3.3.1 SCOP 2: Structural Classification of Proteins

The database called Structural Classification of Proteins will be used. This database contains detailed and comprehensive descriptions of the structural information from all proteins that are known [6]. The database allows for browse by structural class of alpha, beta and small protein structural classification, as well as by the protein type.

4. Results

Features selected for this project were individual amino acid frequencies, as well as the frequency of the physicochemical attributes which are hydrophobicity, van der waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility. These physicochemical properties all contribute to the overall shape of the proteins and will help predict the secondary structures of the protein. In these results 0.0, 1.0, 2.0, 3.0, 4.0 correspond to alpha, beta, alpha/beta, alpha + beta, and other small proteins.

Predicted	0.0	1.0	2.0	3.0	4.0	All
Actual						
0.0	908	33	205	216	20	1382
1.0	58	855	223	280	11	1427
2.0	80	111	1088	263	0	1542
3.0	235	368	630	758	48	2039
4.0	42	21	3	31	218	315
All	1323	1388	2149	1548	297	6705

Figure 1: Confusion matrix numerical values

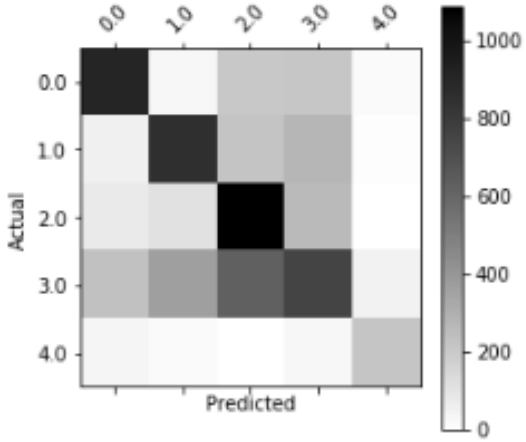

Figure 2: Confusion matrix visual representation

	precision	recall	f1-score	support
0.0	0.686	0.657	0.671	1382
1.0	0.616	0.599	0.607	1427
2.0	0.506	0.706	0.590	1542
3.0	0.490	0.372	0.423	2039
4.0	0.734	0.692	0.712	315
accuracy			0.571	6705
macro avg	0.606	0.605	0.601	6705
weighted avg	0.572	0.571	0.565	6705

Figure 3: Performance metrics for the model

5. Discussion and Conclusion

For this project we have proposed a protein fold classification system which uses a support vector machine classifier to predict five different classes of protein fold. These classes correspond to alpha, beta, alpha/beta, alpha+beta and other small proteins. The support vector machine trained had an accuracy of 57.1% and was trained with a polynomial kernel of degree 5. The first group of features chosen were amino acid frequency. As the types of amino acids indicate the structure of proteins, knowing which ones occur at which frequencies is a key component. Furthering this idea, is that the physicochemical properties of amino acids are what determines the primary structure. A study done by Chen et al. demonstrated the efficacability of grouping amino acids based on their physicochemical properties for use in an ensemble model using support vector machines. This feature extraction method classified amino acids into three groups for eight different physicochemical properties. By looking at the confusion matrix and the performance metrics for this model you can see that the model performed the worst on the alpha+beta structures followed. Since alpha+beta proteins are a combination of two of the other classes, which makes sense as the different patterns are repeated in a random order. Alpha/beta proteins have two distinct portions of alpha and beta so they are separable in some manner. In future work, extracting the features using a sliding window method could be done to see how the features change with different segments over the length of the amino acid sequence. This will help give specific positional information and potentially better help identify protein structure.

6. References

[1] https://www.hindawi.com/journals/bmri/2016/6802832/
[2]Schaeffer, R Dustin, and Valerie Daggett. “Protein folds and protein folding.” *Protein engineering, design & selection : PEDS*vol. 24,1-2 (2011): 11-9. doi:10.1093/protein/gzq096
[3] https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
[4]Harris, Charles R et al. “Array programming with NumPy.” *Nature*vol. 585,7825 (2020): 357-362. doi:10.1038/s41586-020-2649-2
[5] Bernstein, Matthew N., et al. “Jupyter Notebook-Based Tools for Building Structured Datasets from the Sequence Read Archive.” *F1000Research*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/pmc/articles/PMC7445559/.
[6] https://scop.mrc-lmb.cam.ac.uk