A picture containing wheel

Description automatically generated

Haris Ramzan | Bioinformatics | June 5, 2020

Project Report

protein family classification through machine learning

# Problem statement

Although we have many computational resources in terms of memory, but our main nuisance is time taken by the traditional database system to process large amount of biological data. When we consider the traditional approaches, different databases are grouping the protein families, but they are very time consuming and does not deal according to the user choice. To deal with described problem machine learning technique is proposed which will better classify the protein families [2].

# Introduction

In recent years, the field of bioinformatics is growing rapidly as biologists are generating more and more data and new techniques are also developed to save the time and efforts as well. This is the new field of research that requires variety of skills to handle the storage, analysis and handling biological information by considering the data protection rules and regulations. Different algorithms and techniques need to be developed to properly gain the insights and provide explanation of sequence-based data. The term sequence in protein describes the arrangement of amino acids in the sequence that construct proteins. These bunch of sequences are stored on multiple databases and every database store classification of protein sequence. Multiple tools and methods have been developed by scientists to classify the protein sequence [1]. We need better machine learning technique to enhance and fast processing of sequences and perform prediction based on sequences. There are several existing traditional data mining approaches that are used to classify the protein sequence but there is a need to develop various classification models which classifies the protein sequence according to the user’s requirements. As we know there is a need to identify the proteins having same functions therefore new classification models for predicting protein families will benefit and speed up the process of making the groups of protein families having similar functions. We will be performing feature extraction based on sequence information and implementing machine learning model to better predict the family of protein as compared to traditional approaches. Our work will be focused on the three main components named as preprocessing of the dataset, feature extraction from the sequences and implementation of machine learning model for prediction of protein family to which sequence belongs. Features will be extracted from sequences by using different libraries which will be help in predicting protein family by using machine learning model. By using the machine learning technique there will be a significant decrease in time taken to classify the true protein family [3].

# Literature Review

Various strategies have been produced for general sign characterizations dependent on the measurable hypothesis in the previous decades, for example, choice trees, factual procedures, support vector machine (SVM), and neural systems (NN) utilized the word division strategy for include extraction on the protein succession and afterward used the SVM for classification. Another useful approach followed by [4] in feature extraction through sequence is n-gram technique where each set of n-gram is a feature, commonly 3 n-gram is used due to time and saving memory. In [4] protein arrangement datasets with ten-super families (classes) were gotten from the PIR databases and signified as PIR1 and PIR2, individually. There are 949 protein successions tests in PIR1 and 534 protein arrangements tests in PIR2.

High level feature extraction is done by considering the local and global similarity of sequences. To fulfill the purpose of local and global similarity 2-gram encoding and 6-letter exchange group methods are used. The principle of minimum description is also applied for calculating motifs significance [1].

In [5] three phase model was proposed to classify unknown protein families to known families. In the first process noisy sequences data is removed which will result in better accuracy and less consumption of time. On the second phase necessary features are extracted like molecular weights and residues count and ranking is applied on the features. Finally, neighborhood analysis is performed to classify the family of protein.

# Proposed solution

We are supposed to facilitate the process of grouping proteins with similar functions by providing the suitable model for the classification of protein family prediction through machine learning. Hopefully, our work will speed up the process of classifying protein families as compared to traditional database approaches [2]. For this purpose, we aim to design such a machine learning model which predicts the family class either hydrolyze, transfers or sequence belongs to another class will be treated as others. Our approach will narrow down the cost of resources as well as time taken by analyzing sequences and other structural information.

# Feature selection

As our dataset is based on protein sequences, we had to extract the feature which are reliable for our machine learning model. For this purpose, I have used ProtParam library to extract different structural and sequence properties from sequences. There are several features are extracted some of them are listed below to understand which kind of features are selected:

* instability index
* Molecular weight
* Aromaticity
* isoelectric point
* Helix
* Turn
* Sheet
* Amino Acids percentage

Some features are selected from main dataset which was providing structural information of sequence against structure id which are density percentage, density Matthews and resolution.

# Model building

As we know machine learning models are based on the input data divided as train and test set. Train data is used to train the machine learning model and validation of results is done by test data. After preprocessing and selection of features train-test divide is done for model building and validation of the model and before feeding the data to train the model numerical values are scaled for better generalization. XGBOOST models are implemented using sklearn which is well known library for Machine learning.

# Evaluation metric

To checkout the results of machine learning model different evaluation matrices are provided in sklearn matrix to verify the model generalization. I have used accuracy and F1 score as evaluation matric to get a idea about how model is performing. By considering the results of evaluation matrices it is concluded that XGBOOST performed better than SVM.

# References

|  |  |
| --- | --- |
| [1] | Dr.S.Vijayarani and M. S. , "PROTEIN SEQUENCE CLASSIFICATION IN DATA MINING," *International Journal of Information Technology, Modeling and Computing,* vol. 2, p. 8, 2014. |
| [2] | R. vijay, "A case study on Pfam dataset to classify protein families.," 2019. |
| [3] | Rafsanjani, Muhammad; ahmed, sajid; Fareed, Dewan MD; shatabda, Swakkhar; sharma, Alok; Dehzangi, Abdollah;, "PyFeat: A Python-based Effective Feature Generation Tool for DNA, RNA, and Protein Sequences," 2019. |
| [4] | T. Huang, C. Jiuwen and L. Xiong, "Protein Sequence Classification with Improved Extreme Learning Machine Algorithms," *JOUR,* 2014. |
| [5] | S. Suprativ and C. Rituparna, "A Brief Review of Data Mining Application Involving Protein Sequence Classification". |