Analyzing Protein Structure and Exploring the Sequence of Protein using Machine Learning Approach

Sharmin Sultana Peu
ID: 2014-2-60-113

Rajashree Roy
ID: 2014-3-60-070

A thesis submitted in partial fulfillment of requirements for the degree of Bachelor of Science and Engineering

Department of Computer Science and Engineering

East West University
Dhaka-1212, Bangladesh
April, 2018

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Rashedul Amin Tuhin, Senior Lecturer, Department of Computer Science and Engineering, East West University. We also declare that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

. . . .

(Rashedul Amin Tuhin)

Supervisor

Signature

. . . .

(Sharmin Sultana Peu)

(ID: 2014-2-60-113)

Signature

. . . .

(Rajashree Roy)

(ID: 2014-3-60-070)

Letter of Acceptance

"Analyzing Protein Structure and Exploring the Sequence of Protein using Machine Learning Approach" entitled thesis report is submitted by Sharmin Sultana Peu(ID: 2014-2-60-113), Rajashree Roy(ID: 2014-3-60-070) to the Department of Computer Science and Engineering, East West University is accepted by the department in partial fulfillment of requirements for the Award of the Degree of Bachelor of Science and Engineering on April, 2018.

Supervisor

. . . .

(Rashedul Amin Tuhin)

Senior Lecturer, Department of Computer Science and Engineering, East West University

Chairperson

. . . .

(Dr. Ahmed Wasif Reza)
Chairperson and Associate Professor
Department of Computer Science and Engineering
East West University

Abstract

Protein structure and sequence analysis is an important and essential problem. Now machine learning techniques have been widely used in bioinformatics. In this research we analyze the protein of structure and sequence and predict the class of protein sequence. Also find the accuracy of that class for different machine learning algorithm. For the data set we use the exploratory data analysis (EDA) and extracted 278866 protein features from the data set. We classify the features and measure the accuracy level of the three machine learning algorithm: Support Vector Machine (SVM), Naive Bayes and Random Forest (RF) approach for that protein sequence.

Acknowledgments

As it is true for everyone, we have also arrived at this point of achieving a goal in our life through various interactions and help from other people. However, written words are often elusive and harbor diverse interpretations even in one's mother language. Therefore, we would not like to make efforts to and best words to express our thankfulness other than simply listing those people who have contributed to this thesis in an essential way. This work was carried out in the Department of Computer Science and Engineering at East West University, Bangladesh.

First of all, we would like to express our deepest gratitude to Almighty Allah for His blessings on us. Next, our special thanks goes to our supervisor Rashedul Amin Tuhin" who gave us this opportunity and initiated us into the field of *"Analyzing* Protein Structure and Exploring the Sequence of Protein using Machine Learning Approach*"* and without whom this work would not have been possible. His encouragements, visionaries, thoughtful comments, suggestions and unforgettable support at every stage of our BSc study were simply appreciating and essential. His ability to muddle us enough to finally answer our own question correctly is something valuable what we have learned and we would try to emulate, if we get the opportunity ever. There are numerous other people too who have shown us their constant support and friendship in various ways, directly or indirectly related to our academic life. We will remember them in our hearts and hope to and a more appropriate place to acknowledge them in the future.

Sharmin Sultana Peu
April, 2018

Rajashree Roy
April 2018

Chapter 1 Introduction

Chapter 2 Related Works

Chapter 5 Results and evaluation:

Chapter 6 Future Works and Conclusion

List of Figures

Chapter 1

INTRODUCTION

1.1 Introduction

A protein is a polymeric macromolecule, and by amino acid building blocks, it is formed. It is arranged in a linear chain and by peptide bonds joined together. Protein structures play key roles to determine their functions. Many important class of biomolecule in living organisms are represent by protein. Most of the cellular processes they hold and act as structural elements, catalysis agents, signaling molecules and molecular machines of every biological system. Proteins are fundamentally involved in all cell functions. In the body every single protein has a particular function. When few proteins are involved in the bodily movement,at that time others are engaged in structural support. In functions and structures, proteins may vary. Although understanding protein structure is highly important in various areas of research (e.g., medicine and biotechnology).[1]

In cellular life, proteins are responsible for virtually every task. Cell shape and inner organization, product manufacturing, waste cleanup, and routine maintenance protein play an important role. They are the workhorse molecules of the cell and perform several sets of functionalities. Proteins act as enzymes which carry out almost all of the chemical reactions. It takes place in cells as well as by reading the genetic information form a new molecule stored in deoxyribonucleic acid (DNA). Moreover, they may bind to specific foreign particles, such as viruses and bacteria to help protect the body from antibodies. Proteins may serve as hormones, which transmit signals to coordinate biological processes between different cells, tissues, and organs, and as transcription factors that guide the differentiation of the cell and its responses to signals, and participate in the formation of tissues and muscular fiber. Determining the structure of a protein is impossible in some cases. In nature there exist protein sequences that do not implement well-defined stable three-dimensional (3D) structure under normal physiological environments. However, actively participate in molecular recognition functions. As the distinct sequence of amino acids encodes all the information to express its functions, it becomes essential, so design some effective computational methods to explore the protein sequence to its structural properties. Which is one of the major focus of the thesis. [2]

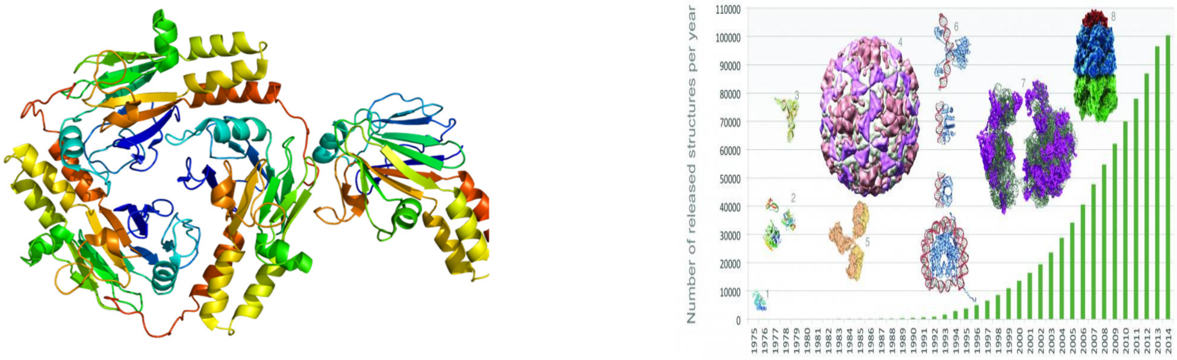


Fig1: Protein Image

Machine learning approach predicts or analyzes the protein based on known properties learned from the training data. In the area of biology, the various application uses methods which are based on machine learning algorithms. Genomics, proteomics and system biology have been utilized by these methods. In bioinformatics prediction methods supervised learning is very important. In this paper we explained how to analyze protein structure and its sequence with the help of machine learning methods.[3]

1.2 In Bioinformatics

Several methods have been developed to analyze the structure, function, and evolution of proteins. It is impossible to study all proteins experimentally. However, computational tools are used to conclude the similar proteins. By sequence, alignment proteins can be efficiently recognized in remotely related organisms. For certain properties, Genome and gene sequences can be explored by a variety of tools. There are many technical challenges such as experimental determination of the complete protein structure is an expensive, time-consuming. Additionally, in this new period, experimental methods are extremely inefficient.

Sequence profiling tools can find restriction enzyme sites, open reading frames in nucleotide sequences, and predicted secondary structures. Bioinformatics is now essential for the analysis of genes and proteins. From artificial intelligence and robotics to genome analysis bioinformatics is now used to mean rather different things. The term was applied to the computational manipulation and analysis of biological *sequence* data (DNA and protein) but now tends also to be used to embrace the manipulation and analysis of 3D structural data. The development of rapid DNA sequencing techniques and the corresponding progress in computer-based technologies molecular biology has witnessed an information revolution, which is allowing us to cope with this information overflow in increasingly efficient ways. The term that was created to contain computer applications in biological sciences is *bioinformatics*. [4]

1.3 Exploratory data in protein analyzing

Exploratory data analysis (EDA) summarize the main characteristics and also analyze the data. Different Bioinformatics problems can be solved by using EDA. In 2D and 3D modeling of biomolecules, EDA is used, and it analyzes their structures and sequences. The structures sequence of intra-molecular substances, carry all different genetically information. EDA also finds patterns of data. [5]

1.4 Machine Learning Method for protein structure analysis

Machine learning is a subfield of computer science, and it learn from data, rather than the programmed instructions. Some techniques used for machine learning that are Naïve Bayes, Support Vector Machine, Random Forest, Artificial Neural Networks, K Nearest Neighbor and Deep learning. To solve different type of problems in bioinformatics Machine learning techniques are used. One of the important goals followed by bioinformatics and theoretical chemistry is protein structure prediction. It is highly important in biotechnology and medicine. Different machine learning algorithms and methods have been utilized in various domains like genomics, proteomics and systems biology. [3][6].

1.5 Our Contribution

Here we are trying to analyze the structure of protein and sequence generation by using different machine learning approach. Based on a different algorithm we can find the better accuracy of the different protein sequence.

The Key contribution of our work:
We use EDA for the data sets and clean the data sets to classify the features. We also classify the class of the sequences using machine learning approach.

1.6 Organization of this book

We organized this book as In Chapter I presents the Introduction part,32 Chapter II presents some related works, Chapter III describes The exploratory data analysis, presents the evaluation The experimental results and discussions are presented in Chapter IV, The results and evaluation is discussions are presented in Chapter V, Chapter VI is future work, and the last Chapter VII is the conclusion.

1.7 Conclusion

In this present study, machine learning can be applied to protein structure and explore the sequence of the protein. Using machine learning algorithm, we can classify the protein. Several methods are there to do that.

Chapter 2

RELATED WORK

2.1 Introduction

Recently some method and system are proposed to analysis the protein and sequence of protein using machine learning method. In this section, we have talked about those existing methods. protein is very vital element for our body structure. So we have discussed the related works to find the limitation of those technologies and introduced some advanced features which will help us a lot.

2.2 Related Works

2.2.1 Sugar-binding residue prediction system from protein sequences using support vector machine

In biological processes, the interactions between proteins and sugar chains play essential roles such as intercellular communication, immunity, and cellular recognition. Using machine learning algorithms, numerous methods have been proposed for protein-sugar binding site prediction. In that paper, they were classifying sugars into acidic and nonacidic sugars and showed that their binding sites have different amino acid occurrence frequencies. From this result, they developed sugar-binding residue predictors committed to the two classes of sugars: an acid sugar binding predictor and a nonacidic sugar binding predictor. For predicting of Amino acid sequences they used only Support vector machine as a machine learning algorithm and evaluated the performance of the predictors using five-fold cross-validation. Though, they are not effective to learn various properties of binding site residues caused by various interactions between proteins and sugars. For acidic and nonacidic sugar-binding, showed the best performance in the prediction of sugar-binding residues and nonacidic sugar-binding residues. This approach is particularly effective when the difference in the sequence features between the groups is large. To get the better performance, they can use other different machine learning algorithms. In few years, some of the researchers use methods utilizing glycan arrays have been developed as high throughput solutions, enabling researchers to obtain data on in vitro interactions between multiple sugar chains and proteins. Docking simulation is also a prediction method for sugar-binding residues based on their tertiary structures. To implement this method, many protein-ligand docking programs, and molecular simulations are often employed. [9]

2.2.2 Protein Structure and Sequence Similarity on the Accuracy of Machine-Learning Scoring Functions for Binding Affinity Prediction

This study is focused on structure-based protein-ligand binding affinity prediction. Scoring Functions (SFs) are used to predict the strength of the non-covalent interaction as the binding affinity between two molecules (protein-protein or protein-DNA) after they have been docked.

In this study, they categorized the SFs into two broad classes: classical SFs and machine-learning SFs. This study shows that RF-Score is substantially more predictive than traditional X-Score when using training sets generated with sequence similarity instead of structural similarity. That means that RF-Score is substantially more predictive than X-Score. By increasing the training set size, the number of highly similar Proteins will also increase. So RF will produce a better result. Moreover, RF-Score-v3 outperformed X-Score even with an averagely sized dataset.[11][12]

2.2.3 Protein Residue Contacts and Prediction Methods

Many machine learning algorithms have been applied to predict protein residue contacts. Initial approaches to ab initio contact prediction used artificial neural networks genetic algorithm, random forest, hidden Markov model, and support vector machines. Now by using of deep learning architectures with and without including correlated mutation information. Deep learning methods give the best results. These machine learning-based methods use a wide range of features as input including features related to the local window of the residues, information about the residue type, and the protein itself. This includes features like secondary structure, sequence profiles, solvent accessibility, mutual information of sequence profiles, residue type information, sequence separation length between the residues under consideration, and pairwise information between all the residues involved.[18]

2.2.4 Prediction and characterization of human ageing-related proteins by using machine learning:

There is a huge impact on human health for ageing. More than three hundred genes have been related to human ageing. Using state-of-the-art machine learning methods, they classified human proteins as ageing-related or non-ageing-related.

They applied supervised machine learning for analysis the ageing-related and non-ageing-related proteins.

Support-vector machine (SVM), k-nearest neighbor (KNN), and decision tree classifiers were used for predicting ageing-related genes of the nematode on the GenAge database. To classify human DNA repair genes as ageing-related or non-ageing-related, they used naïve Bayes classifier and J48 decision tree. Also applied three state-of-the-art machine learning tools, XGBoost, logistic regression and support-vector machine, to classify human proteins as ageing-related or non-ageing-related

The final prediction was to measure the importance of a given protein in of the human ageing process and to identify new ageing-related protein candidates.[19]

2.2.5 Deep Learning Architectures for Protein Structure Prediction

In bioinformatics to predict the Protein structure prediction, machine learning techniques have been widely used. In their paper they used deep learning methods to develop a new area of research in machine learning, and show great success in diverse areas of signal and information processing studies. In this article, they provide a brief review of recent development and application of deep learning methods for protein structure prediction. They determine the 3D structure of a protein from its amino acid sequence. They face many problems to predict protein 3D structure directly from its amino acid sequence because there was no known template structure. They predict protein properties such as secondary structure, residue-residue contacts, and disorder regions that can be used to facilitate 3D structure prediction.

2.2.6 Impact of genetic variation on three-dimensional structure and function of proteins

The PDB (Protein Data Bank) contains atomic level three-dimensional (3D) structures of biological macromolecules (proteins, DNA, RNA). Knowledge of the 3D structure of the gene product can help to understand its function and role in disease. Here they observe the structural and functional changes caused by single amino acid differences, including changes in enzyme activity, aggregation propensity, structural stability, binding, and dissociation, the context of large assemblies.

Knowledge of the 3D structure of a gene product is helpful in predicting and understanding

both function and role in disease. The goal of this study is to improve the understanding of the relationship between point mutations and experimentally observed consequences in 3D.Single Nucleotide Variations (SNVs) represent the most common genetic variations observed in humans. In this study, they analysis small dataset to understand the consequences of a particular genetic variation at the level of the fixed protein.[15]

2.3 Conclusion

In this book we analysis protein structure and sequence exploration of protein by some machine learning algorithm with the protein data set.

Chapter 3

EXPLORATORY DATA ANALYSIS

3.1 Important for Bioinformatics

Various Bioinformatics problems can be solved with the help of EDA. It generates the new strategy and new hypotheses from different biomedical data. In 2D and 3D modeling of biomolecules and analyze their structures and sequences it can also be used. The structures of biomolecules, 3D models, the sequence of intra-molecular substances, angles between successive bonds, angles with side chains, they all carry different genetical information. EDA is used to find patterns in such data like disease control, analysis the junks in RNA and DNA in early embryo developments.

3.2 In Case of protein

Protein is consist of amino acid residues. Proteins are large biomolecules. The sequence of amino acids and their 3D structures carry information about metabolic reactions, DNA replication, etc. EDA is a Protein-Coding gene. EDA methods are applied to provide insights into the function of a protein. By analyzing the sequence of proteins, DNA, RNA, we may assume genetic behaviors, patient's genetic suitability for certain drugs can be determined by examining the genetic behaviors.

Proteins are characterized by different typologies of structures (structural, geometrical, energy). Most of these features are similar to a protein family. Using EDA features can identify proteins that belong to a family, as well as define the boundaries among families. Some features are redundant. Sometimes they could generate noise in identifying which variables are essential. If they are related or not to a function of a protein family, then we defined an original approach to analyzing protein features for defining their relationships and peculiarities within protein families. In R environment a multistep approach has been mainly performed for getting-cleaning data, exploratory data analysis and predictive modeling for classification.[22]

Chapter 4

PROPOSED MODEL

4.1 Introduction

Exploratory data analysis (EDA) analyze the datasets and summarize their main characteristics. In statistics and data science areas it is used to analyze the dataset and also search for models which describe the data well. This method deals with different visual models for analysis and predicting the dataset. From initial data analysis (IDA)_EDA is different that focuses on the quality of data, Quality of measurements, handling missing values and making appropriate transformations of variables and so on. EDA can be measured by many different Tools. This is very useful for calculating EDA.

4.2 Exploration of protein sequence and prediction analysis

To experiment the topic, we use EDA to analyze the data set. We use python language for EDA. Then we analyze the data and verify the protein sequence classification we use machine learning approach. We apply the Naïve Bayes approach to explore the sequence of the protein from the protein classification. In our data set, there are some missing values, so we deduct some rows from our data set. Here we also count the count distribution and also selects some features from the data sets. However, for prediction of protein structure and sequence analysis, we apply Naive Bayes approach.

4.3 Algorithm we used

4.3.1 Naive Bayes

Based on Bayes Theorem it is a classification technique with an assumption of independence among predictors. A Naive Bayes classifier assumes the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if the color is red, the shape is round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as Naive.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is highly sophisticated classification methods.

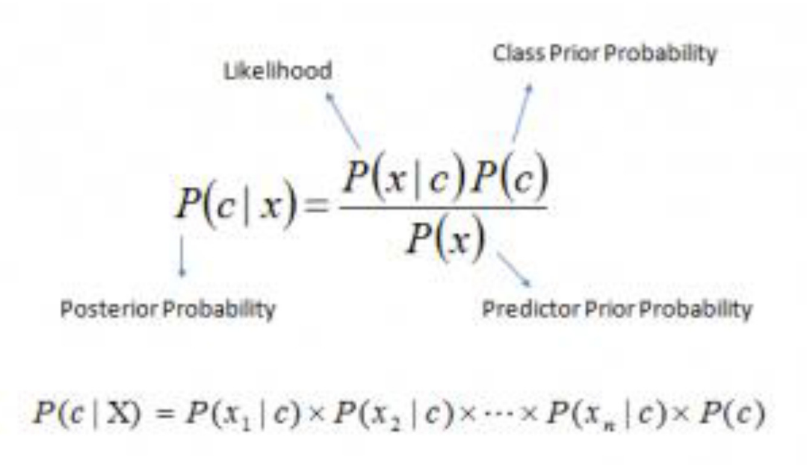
Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c), the equation is :

□ *P*(*c*|*x*) is the after probability of *class* (c, *target*) given *predictor* (x, *attributes*).

□ *P*(*c*) is the previous probability of *class*.

□ *P*(*x*|*c*) is the likelihood which is the probability of *predictor class that are given*.

□ *P*(*x*) is the predict the previous probability .



$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig 2: Naïve Bayes equation

In our study, we analyze our data set by using Naïve Bayes classifier. Moreover, experiments with the data available for protein, its show that our prediction method gives a relatively high level of accuracy. Some challenges and possibilities for future developments are also discussed. We apply here a Naive Bayes model to provide probabilistic predictions.

4.3.2 Support vector machine

A support vector machine (SVM) builds a set of hyperplanes in an infinite-dimensional space. By SVM many analysis can be done like Classification, regression, outliers detection, etc. SVM find the "maximum-margin hyperplane" that divides the group of points .

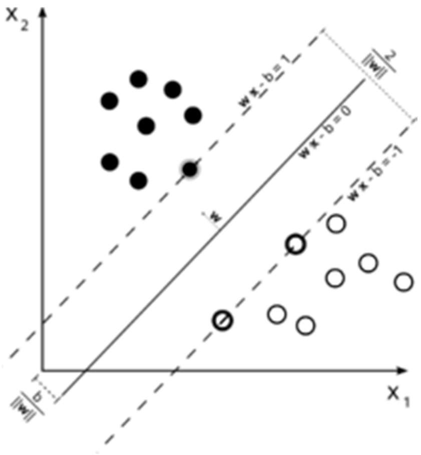


Fig 3: Support Vector Machine

Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors. Any hyperplane can be written as the set of points .

4.3.3 Random Forest (RF)

Random forest algorithm can use both for classification and the regression problems. In machine learning algorithm Random Forest is a flexible and easy to use. Even without hyper-parameter tuning, it gives the a great result most of the time. It is a supervised classification algorithm. This algorithm creates the forest with some trees. in many cases it gives the high accuracy results. In the forest, higher the number of trees gives the high accuracy results. It can handle the missing values. It also works for the categorical values.

Chapter 5

RESULTS AND EVALUATION

5.1 Exploration of protein sequence

After exploration of data, we found this classification, and we also find some null values after filtering.

Here Number of records after processing the data.

278866 is the number of records in the final filtered dataset

5.2 Count Distribution

HYDROLASE	46336
TRANSFERASE	36424
OXIDOREDUCTASE	34321
IMMUNE SYSTEM	15615
LYASE	11682
HYDROLASE / HYDROLASE INHIBITOR	11218
TRANSCRIPTION	8919
VIRAL PROTEIN	8495
TRANSPORT PROTEIN	8371
VIRUS	6972
SIGNALING PROTEIN	6469
ISOMERASE	6356
LIGASE	4964
MEMBRANE PROTEIN	4891
PROTEIN BINDING	4884
STRUCTURAL PROTEIN	4335
CHAPERONE	4156
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	3548
SUGAR BINDING PROTEIN	3474
DNA BINDING PROTEIN	3199
PHOTOSYNTHESIS	3139
ELECTRON TRANSPORT	3047
TRANSFERASE / TRANSFERASE INHIBITOR	3032
METAL BINDING PROTEIN	3023
CELL ADHESION	2999

UNKNOWN FUNCTION	2842
PROTEIN TRANSPORT	2674
TOXIN	2626
CELL CYCLE	2496
RNA BINDING PROTEIN	1969
MEMBRANE PROTEIN, IMMUNE SYSTEM	1
FATTY ACID-BINDING	1
TELOKIN-LIKE PROTEIN	1
Antibiotic, plant protein	1
VIRAL PROTEIN, SUGAR BINDING PROTEIN	1
GLYCOGEN METABOLISM	1
PLP-BINDING PROTEIN	1
CURAREMIMETIC PROTEIN	1
Hormone activator	1
TRANSPORT BINDING	1
RNASE-2	1
Viral protein, Transcription	1
HYDROLASE(ACTING ON LINEAR AMIDES)	1
FERREDOXIN	1
apoptosis regulator	1
Structural genomics, structural protein	1
DNA BINDING PROTEIN, SUGAR BINDING PROTEIN	1
SUGAR BINDING PROTEIN	1
MEMBRANE PROTEIN, DE NOVO PROTEIN	1
CALMODULIN-BINDING	1
TRANSFERASE(GLUCOSYL TRANSFERASE)	1
COMPLEX (HYDROLASE /PRODUCT)	1
Biosynthetic Protein, Structural Protein	1
C-TYPE LECTIN	1
LATE PROTEIN	1
Iron Binding Protein	1
HYPOTENSIVE HORMONE	1
CONDENSING ENZYMES	1
SRC HOMOLOGY 2 DOMAIN	1
CARBOXYL METHYLESTERASE	1
Name: classification, Length:4468, dtype: int64	

Fig 4: Count Distribution

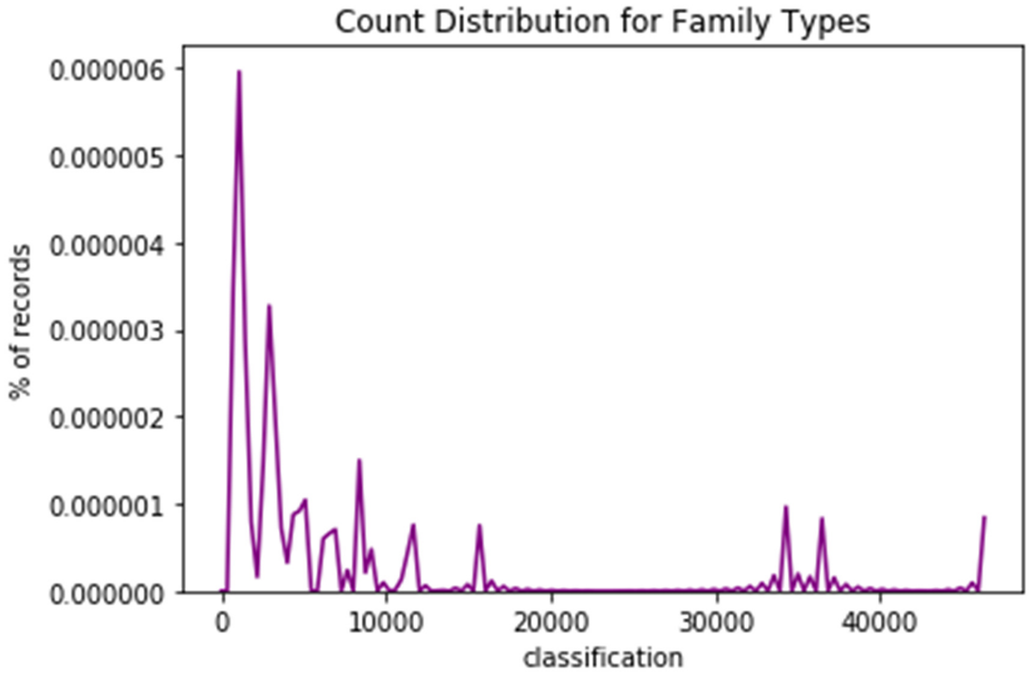
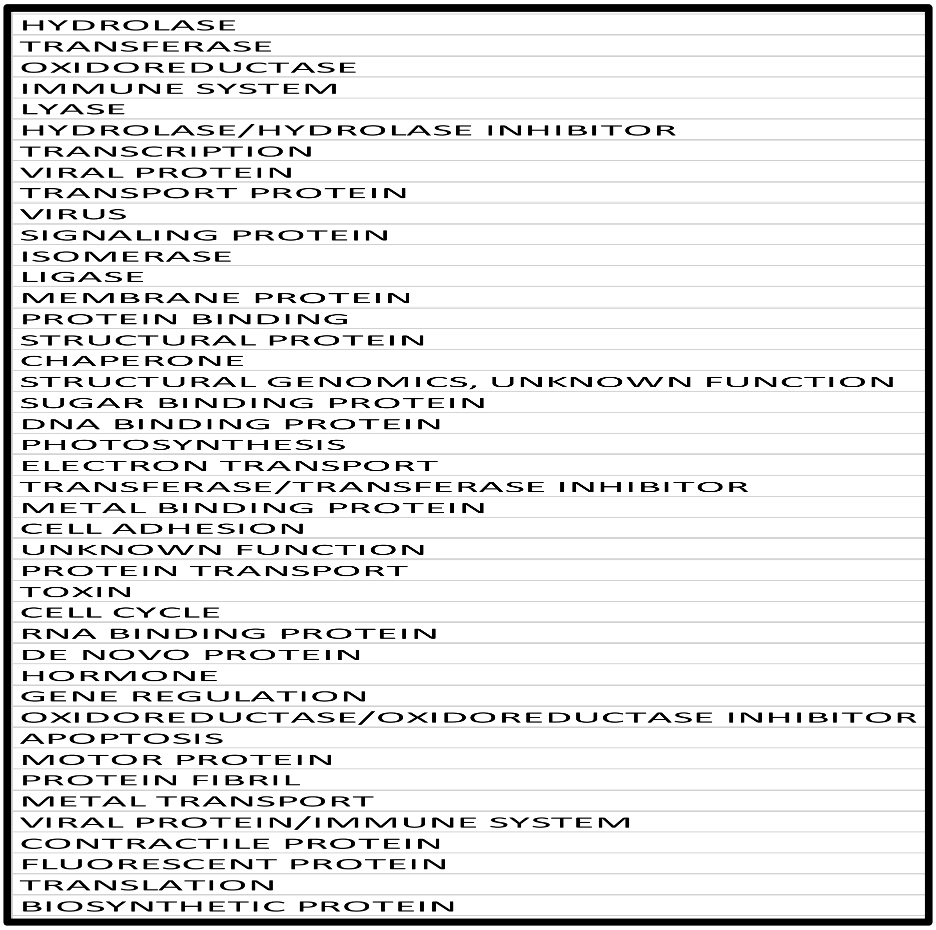


Fig 5: Count Distribution Family Type

5.3 Classification type after filtering



HYDROLASE
TRANSFERASE
OXIDOREDUCTASE
IMMUNE SYSTEM
LYASE
HYDROLASE/HYDROLASE INHIBITOR
TRANSCRIPTION
VIRAL PROTEIN
TRANSPORT PROTEIN
VIRUS
SIGNALING PROTEIN
ISOMERASE
LIGASE
MEMBRANE PROTEIN
PROTEIN BINDING
STRUCTURAL PROTEIN
CHAPERONE
STRUCTURAL GENOMICS, UNKNOWN FUNCTION
SUGAR BINDING PROTEIN
DNA BINDING PROTEIN
PHOTOSYNTHESIS
ELECTRON TRANSPORT
TRANSFERASE/TRANSFERASE INHIBITOR
METAL BINDING PROTEIN
CELL ADHESION
UNKNOWN FUNCTION
PROTEIN TRANSPORT
TOXIN
CELL CYCLE
RNA BINDING PROTEIN
DE NOVO PROTEIN
HORMONE
GENE REGULATION
OXIDOREDUCTASE/OXIDOREDUCTASE INHIBITOR
APOPTOSIS
MOTOR PROTEIN
PROTEIN FIBRIL
METAL TRANSPORT
VIRAL PROTEIN/IMMUNE SYSTEM
CONTRACTILE PROTEIN
FLUORESCENT PROTEIN
TRANSLATION
BIOSYNTHETIC PROTEIN

Fig 6: Classification Type

5.4 Prediction analysis

Precision = (TP/(TP+FP))

Recall= (TP/(TP+FN))

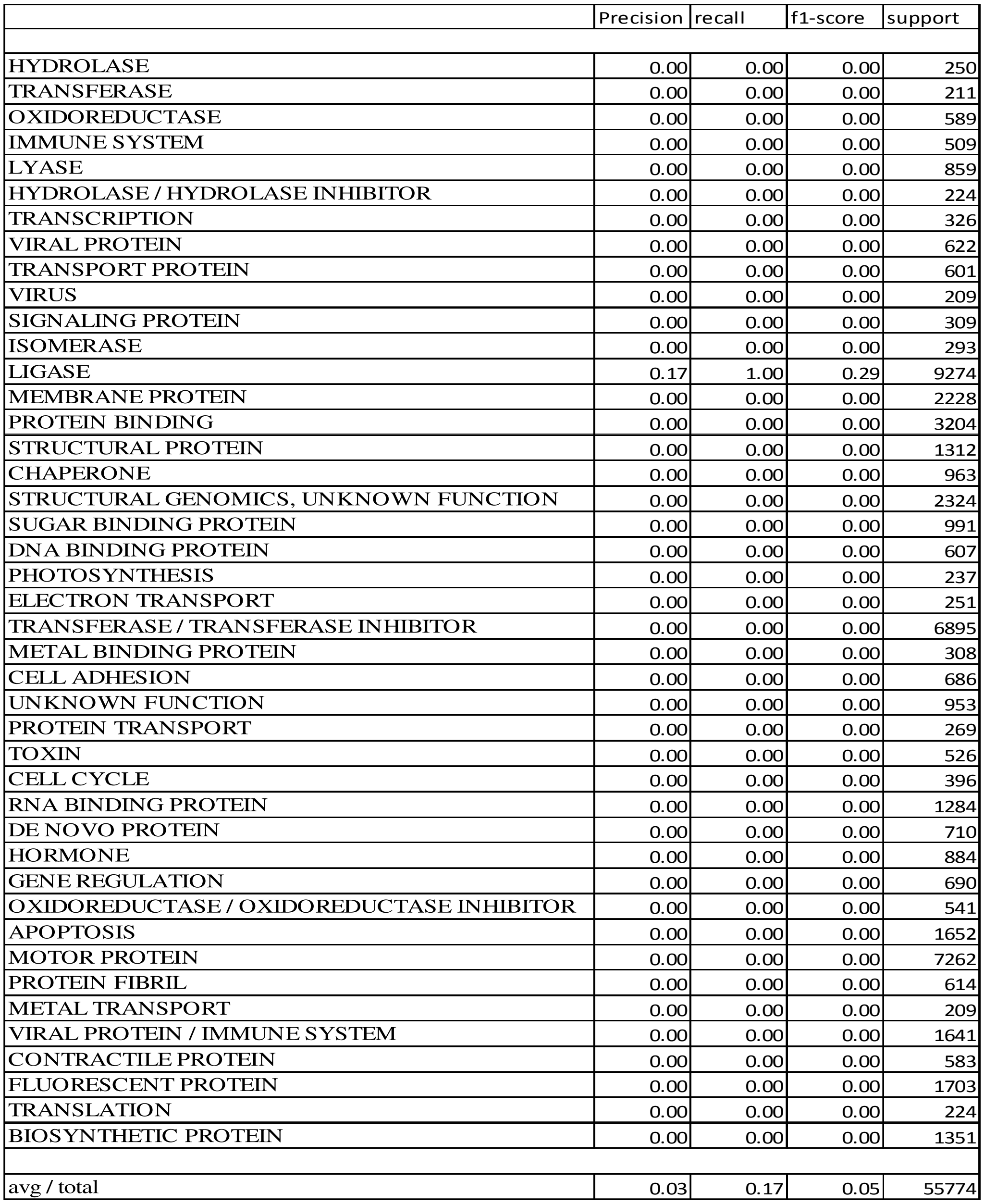
$F1Score = 2 * (Precision * Recall)/(Precision + Recall)$

Where TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative, The confusion matrix refers the risks and gains as well as cost and benefit in a classification model.

5.5 Classification Report

5.5.1 For SVM (Support Vector Machine)

To experiment this, we use python language.And we get 16.62% accuracy for SVM.

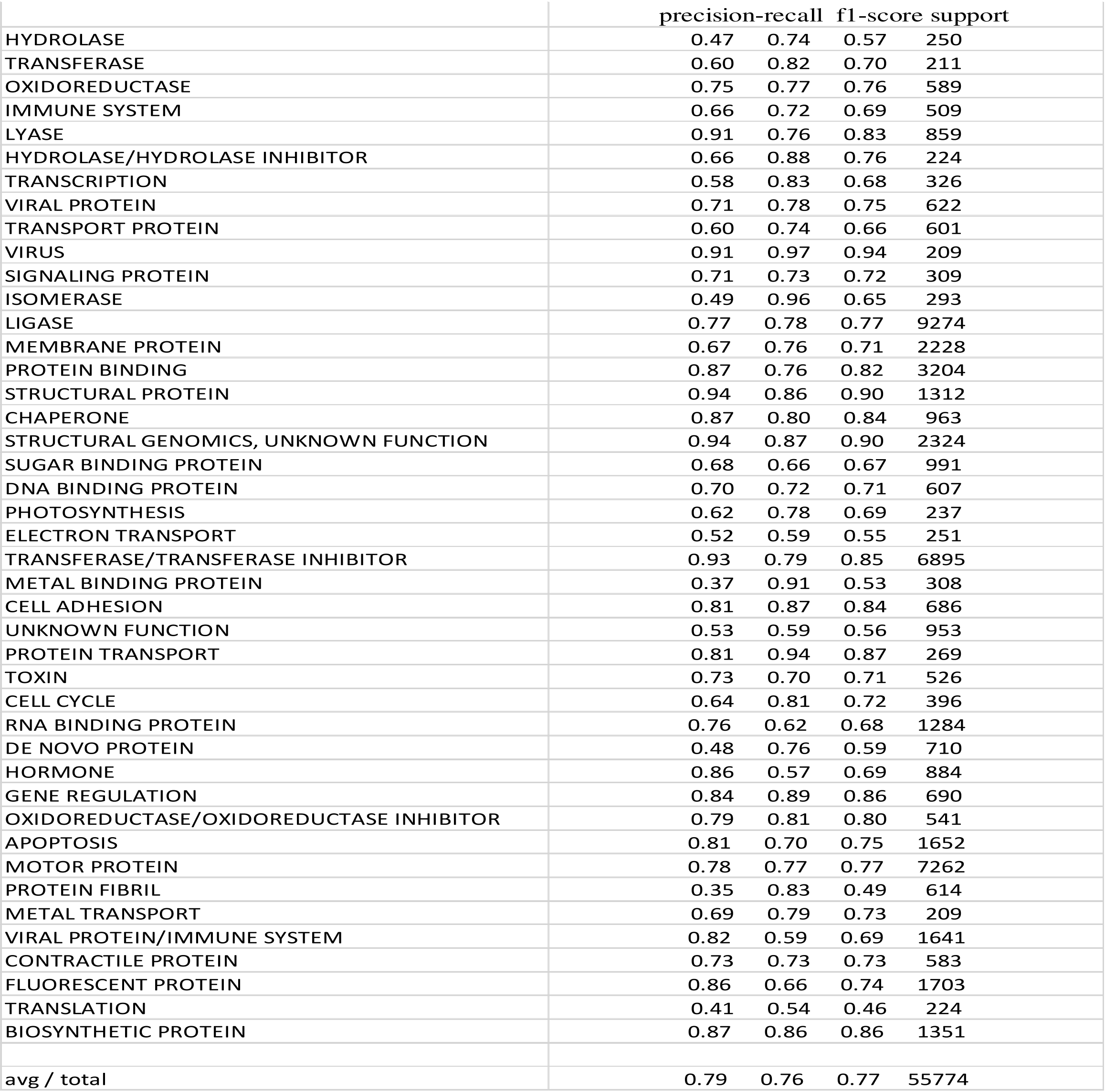


	Precision	recall	f1-score	support
HYDROLASE	0.00	0.00	0.00	250
TRANSFERASE	0.00	0.00	0.00	211
OXIDOREDUCTASE	0.00	0.00	0.00	589
IMMUNE SYSTEM	0.00	0.00	0.00	509
LYASE	0.00	0.00	0.00	859
HYDROLASE / HYDROLASE INHIBITOR	0.00	0.00	0.00	224
TRANSCRIPTION	0.00	0.00	0.00	326
VIRAL PROTEIN	0.00	0.00	0.00	622
TRANSPORT PROTEIN	0.00	0.00	0.00	601
VIRUS	0.00	0.00	0.00	209
SIGNALING PROTEIN	0.00	0.00	0.00	309
ISOMERASE	0.00	0.00	0.00	293
LIGASE	0.17	1.00	0.29	9274
MEMBRANE PROTEIN	0.00	0.00	0.00	2228
PROTEIN BINDING	0.00	0.00	0.00	3204
STRUCTURAL PROTEIN	0.00	0.00	0.00	1312
CHAPERONE	0.00	0.00	0.00	963
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	0.00	0.00	0.00	2324
SUGAR BINDING PROTEIN	0.00	0.00	0.00	991
DNA BINDING PROTEIN	0.00	0.00	0.00	607
PHOTOSYNTHESIS	0.00	0.00	0.00	237
ELECTRON TRANSPORT	0.00	0.00	0.00	251
TRANSFERASE / TRANSFERASE INHIBITOR	0.00	0.00	0.00	6895
METAL BINDING PROTEIN	0.00	0.00	0.00	308
CELL ADHESION	0.00	0.00	0.00	686
UNKNOWN FUNCTION	0.00	0.00	0.00	953
PROTEIN TRANSPORT	0.00	0.00	0.00	269
TOXIN	0.00	0.00	0.00	526
CELL CYCLE	0.00	0.00	0.00	396
RNA BINDING PROTEIN	0.00	0.00	0.00	1284
DE NOVO PROTEIN	0.00	0.00	0.00	710
HORMONE	0.00	0.00	0.00	884
GENE REGULATION	0.00	0.00	0.00	690
OXIDOREDUCTASE / OXIDOREDUCTASE INHIBITOR	0.00	0.00	0.00	541
APOPTOSIS	0.00	0.00	0.00	1652
MOTOR PROTEIN	0.00	0.00	0.00	7262
PROTEIN FIBRIL	0.00	0.00	0.00	614
METAL TRANSPORT	0.00	0.00	0.00	209
VIRAL PROTEIN / IMMUNE SYSTEM	0.00	0.00	0.00	1641
CONTRACTILE PROTEIN	0.00	0.00	0.00	583
FLUORESCENT PROTEIN	0.00	0.00	0.00	1703
TRANSLATION	0.00	0.00	0.00	224
BIOSYNTHETIC PROTEIN	0.00	0.00	0.00	1351
avg / total	0.03	0.17	0.05	55774

Fig 7: Confusion Matrix For SVM

5.5.2 For Naive Bayes

Here we use Naïve Bayes approach, and we get 76.18% accuracy.

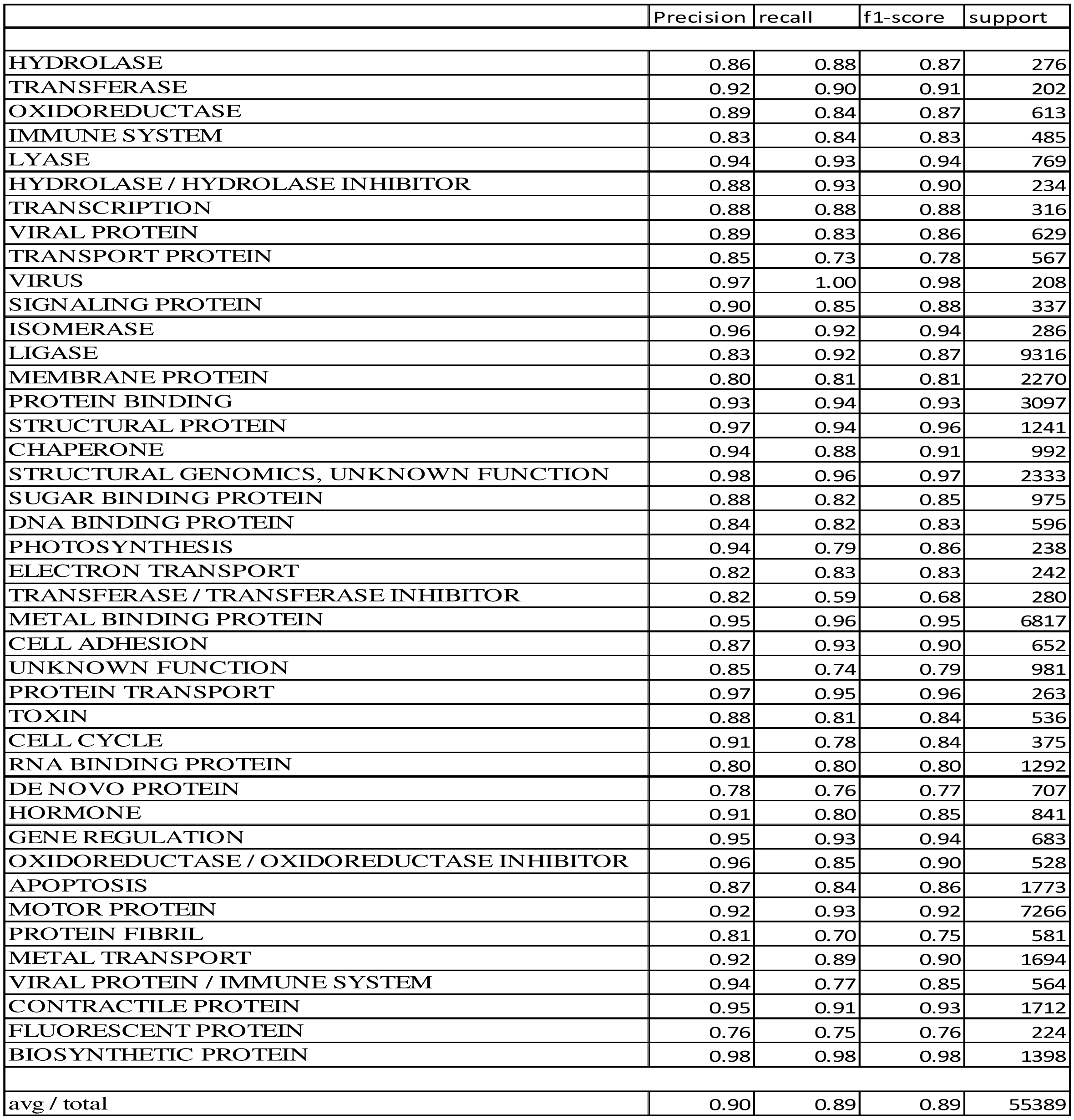


	precision-	recall	f1-score	support
HYDROLASE	0.47	0.74	0.57	250
TRANSFERASE	0.60	0.82	0.70	211
OXIDOREDUCTASE	0.75	0.77	0.76	589
IMMUNE SYSTEM	0.66	0.72	0.69	509
LYASE	0.91	0.76	0.83	859
HYDROLASE/HYDROLASE INHIBITOR	0.66	0.88	0.76	224
TRANSCRIPTION	0.58	0.83	0.68	326
VIRAL PROTEIN	0.71	0.78	0.75	622
TRANSPORT PROTEIN	0.60	0.74	0.66	601
VIRUS	0.91	0.97	0.94	209
SIGNALING PROTEIN	0.71	0.73	0.72	309
ISOMERASE	0.49	0.96	0.65	293
LIGASE	0.77	0.78	0.77	9274
MEMBRANE PROTEIN	0.67	0.76	0.71	2228
PROTEIN BINDING	0.87	0.76	0.82	3204
STRUCTURAL PROTEIN	0.94	0.86	0.90	1312
CHAPERONE	0.87	0.80	0.84	963
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	0.94	0.87	0.90	2324
SUGAR BINDING PROTEIN	0.68	0.66	0.67	991
DNA BINDING PROTEIN	0.70	0.72	0.71	607
PHOTOSYNTHESIS	0.62	0.78	0.69	237
ELECTRON TRANSPORT	0.52	0.59	0.55	251
TRANSFERASE/TRANSFERASE INHIBITOR	0.93	0.79	0.85	6895
METAL BINDING PROTEIN	0.37	0.91	0.53	308
CELL ADHESION	0.81	0.87	0.84	686
UNKNOWN FUNCTION	0.53	0.59	0.56	953
PROTEIN TRANSPORT	0.81	0.94	0.87	269
TOXIN	0.73	0.70	0.71	526
CELL CYCLE	0.64	0.81	0.72	396
RNA BINDING PROTEIN	0.76	0.62	0.68	1284
DE NOVO PROTEIN	0.48	0.76	0.59	710
HORMONE	0.86	0.57	0.69	884
GENE REGULATION	0.84	0.89	0.86	690
OXIDOREDUCTASE/OXIDOREDUCTASE INHIBITOR	0.79	0.81	0.80	541
APOPTOSIS	0.81	0.70	0.75	1652
MOTOR PROTEIN	0.78	0.77	0.77	7262
PROTEIN FIBRIL	0.35	0.83	0.49	614
METAL TRANSPORT	0.69	0.79	0.73	209
VIRAL PROTEIN/IMMUNE SYSTEM	0.82	0.59	0.69	1641
CONTRACTILE PROTEIN	0.73	0.73	0.73	583
FLUORESCENT PROTEIN	0.86	0.66	0.74	1703
TRANSLATION	0.41	0.54	0.46	224
BIOSYNTHETIC PROTEIN	0.87	0.86	0.86	1351
avg / total	0.79	0.76	0.77	55774

Fig 8: Confusion Matrix For Naive Bayes

5.5.3 For RF (Random Forest)

Here we get the 89% accuracy, which is better than the Naïve Bayes approach.



	Precision	recall	f1-score	support
HYDROLASE	0.86	0.88	0.87	276
TRANSFERASE	0.92	0.90	0.91	202
OXIDOREDUCTASE	0.89	0.84	0.87	613
IMMUNE SYSTEM	0.83	0.84	0.83	485
LYASE	0.94	0.93	0.94	769
HYDROLASE / HYDROLASE INHIBITOR	0.88	0.93	0.90	234
TRANSCRIPTION	0.88	0.88	0.88	316
VIRAL PROTEIN	0.89	0.83	0.86	629
TRANSPORT PROTEIN	0.85	0.73	0.78	567
VIRUS	0.97	1.00	0.98	208
SIGNALING PROTEIN	0.90	0.85	0.88	337
ISOMERASE	0.96	0.92	0.94	286
LIGASE	0.83	0.92	0.87	9316
MEMBRANE PROTEIN	0.80	0.81	0.81	2270
PROTEIN BINDING	0.93	0.94	0.93	3097
STRUCTURAL PROTEIN	0.97	0.94	0.96	1241
CHAPERONE	0.94	0.88	0.91	992
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	0.98	0.96	0.97	2333
SUGAR BINDING PROTEIN	0.88	0.82	0.85	975
DNA BINDING PROTEIN	0.84	0.82	0.83	596
PHOTOSYNTHESIS	0.94	0.79	0.86	238
ELECTRON TRANSPORT	0.82	0.83	0.83	242
TRANSFERASE / TRANSFERASE INHIBITOR	0.82	0.59	0.68	280
METAL BINDING PROTEIN	0.95	0.96	0.95	6817
CELL ADHESION	0.87	0.93	0.90	652
UNKNOWN FUNCTION	0.85	0.74	0.79	981
PROTEIN TRANSPORT	0.97	0.95	0.96	263
TOXIN	0.88	0.81	0.84	536
CELL CYCLE	0.91	0.78	0.84	375
RNA BINDING PROTEIN	0.80	0.80	0.80	1292
DE NOVO PROTEIN	0.78	0.76	0.77	707
HORMONE	0.91	0.80	0.85	841
GENE REGULATION	0.95	0.93	0.94	683
OXIDOREDUCTASE / OXIDOREDUCTASE INHIBITOR	0.96	0.85	0.90	528
APOPTOSIS	0.87	0.84	0.86	1773
MOTOR PROTEIN	0.92	0.93	0.92	7266
PROTEIN FIBRIL	0.81	0.70	0.75	581
METAL TRANSPORT	0.92	0.89	0.90	1694
VIRAL PROTEIN / IMMUNE SYSTEM	0.94	0.77	0.85	564
CONTRACTILE PROTEIN	0.95	0.91	0.93	1712
FLUORESCENT PROTEIN	0.76	0.75	0.76	224
BIOSYNTHETIC PROTEIN	0.98	0.98	0.98	1398
avg / total	0.90	0.89	0.89	55389

Fig 9: Confusion Matrix For RF

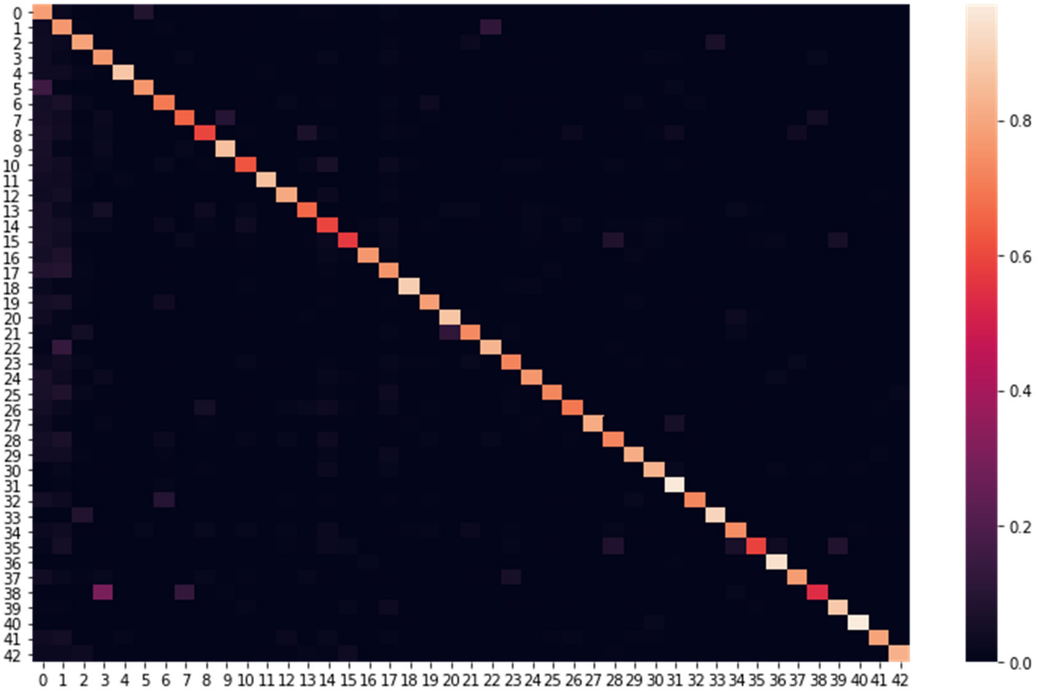


Fig 10: Graphical representation of confusion matrix

So finally we can see that Random forest gives us most balanced result.

Algorithm	Accuracy (%)
Support Vector Machine	16.62
Naïve Bayes	76.18
Random Forest	89

Chapter 6

FUTURE WORKS AND CONCLUSION

6.1 Future work

Here we only use the Support Vector Machine, Naïve Bayes approach and RF, though RF gives us a better accuracy than the Naïve Bayes approach and Support Vector Machine. However, in future to analyse the structure of the protein and explore the protein sequence many different machine learning approach can be applied. Knowledge discovery is our future work.

6.2 Conclusion

In the previous sections, machine learning methods have been used comprehensively in the field of protein structure and sequence analysis. Protein. Protein structures play key roles to determine their functions and sequences. Proteins represent the most important class of biomolecules. In this brief review. In our work we are trying to predict class of the sequence of protein using machine learning algorithm and get a better accuracy according to three of machine learning approach and found better accuracy in RF.

BIBLIOGRAPHY

[1].Protein. (2017). [online] Available at: https://en.wikipedia.org/wiki/Protein [Accessed 21 Apr. 2018].

[2] : Protein Structure and Function Prediction Using Machine Learning Methods. (2014). [ebook] Hemalatha N. Available at: https://pdfs.semanticscholar.org/c18f/52e409fee72ee25263458eea8b85d4045820.pdf [Accessed 5 Oct. 2014].

[3] Machine Learning Approaches for the Prediction of Protein Sequences. 2018.

[4][online] Available at: https://en.wikipedia.org/wiki/Protein.

[5] [5]] 2018. [Online]. Available: https://en.wikipedia.org/wiki/Exploratory_data_analysis. [Accessed: 11- Apr- 2018].

[6] Feature selection on a dataset of protein families: from exploratory data analysis to statistical variable importance", 2018

[7] https://onlinelibrary.wiley.com/doi/full/10.1002/prot.25499/" Structural effects of point mutations in proteins,10 April,2018

[8][online] Available at:https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25505

[9] Development of a sugar-binding residue prediction system from protein sequences using support vector machine." 23 October 2016. (http://creativecommons.org/licenses/by-nc-nd/4.0/).

[10] Improving intermolecular contact prediction through protein-protein interaction prediction using evolutionary analysis with expectation-maximization. 2018.

[11] "The Impact of Protein Structure and Sequence Similarity on the Accuracy of Machine-Learning Scoring Functions for Binding Affinity Prediction", 2018.

[12] y. Li, "Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein–Ligand Interactions", 2018.

[13] Predicting functionally important residues from sequence conservation. John A. Capra and Mona Singh, 2018.

[14] A statistical model for improved membrane protein expression using sequence-derived features. Shyam M. Saladi, Nauman Javed, Axel Müller, and William M. Clemons, Jr., 2018.

[15] Impact of genetic variation on three dimensional structure and function of proteins", 2017.

[16] Systematic Identification of Machine-Learning Models Aimed to Classify Critical Residues for Protein Function from Protein Structure", 2017.

[17] Structural Class Classification of 3D Protein Structure Based on Multi-view 2D Images", 2018.

[18] Protein Residue Contacts and Prediction Methods", 2016.

[19] "Prediction and characterization of human ageing-related proteins by using machine learning", Csaba Kerepes, 2018.

[20] https://en.wikipedia.org/wiki/Exploratory_data_analysis

[21] Improving intermolecular contact prediction through protein-protein interaction prediction using coevolutionary analysis with expectation-maximization. (2018). [ebook] Miguel Correa Marrero1, Richard G.H Immink2,3, Dick de Ridder1, and Aalt D.J van Dijk1,3,4. Available at: http://dx.doi.org/10.1101/254789. [Accessed 8 Jan. 2018]

[22] ***Ritchie DW (February 2008). "Recent progress and future directions in protein-protein docking". Current Protein & Peptide Science. 9 (1): 1–15. doi:10.2174/138920308783565741. PMID 18336319.*** **[23] https://proteinstructures.com/, date 24.10.2017**