Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7723 - 7731
Received 25 April 2021; Accepted 08 May 2021.

Protein Classification Model Using Supervised Machine Learning Algorithms

B. Pravallika[1], G.Shivani[2], MD.Mutharif Ali[3], P.Navya[4]

[1] Assistant Professor

[1,2,3,4] Department of Information Technology,

[1,2,3,4] Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India.

Abstract-

The protein database helped the life science community to study different diseases and come with new drugs and solutions that help human survival. The constantly-growing PDB is a reflection of the research that is happening in laboratories across the world. This can make it both exciting and challenging to use the database in research and education. Structures are available for many of the proteins and nucleic acids involved in the central processes of life, so you can go to the PDB archive to find structures for ribosomes, oncogenes, drug targets, and even whole viruses. However, it can be a challenge to find the information that you need, since the PDB archives so many different structures. You will often find multiple structures for a given molecule, or partial structures, or structures that have been modified or inactivated from their native form.

Keywords: Machine Learning, Protein, Classifier, Supervised Algorithms, Multinomial NB.

I. INTRODUCTION

The study of proteins is a longstanding problem in the research community. The life-making proteins are amino acids connected and fold into three-dimensional structures that define their functions. For the formulation of new medicines, diseases study, and protein engineering fields, the knowledge of these structures is critical. The three-dimensional structures of the proteins in experimentation are very expensive and time-consuming.

Alternatively, a computer can be used for producing hundreds of miles of potential 3D structures for each protein and determining the most appropriate predictions. For protein classification, different types of approaches were used. Initial experiments employed methods of clustering. Not all of these methods are consistently effective. More recently, the usage of ML algorithms like the Random Forest Classifier, supporting vector machines (SVM), and other characteristics used to estimate the quality of the protein model is also a new approach. One of the most important methods of predicting the best models of a protein class was machine learning (ML). Vectors capturing every typical structure along a wide array of protein acid chains have been designed with enormous effort. New algorithms and new hardware are available, which significantly improve the provision of protein models, but still, improve prediction accuracy.

The problem definition is to Analyze the Protein Data Bank(PDB) molecules structures and the Sequences of Proteins by performing Machine Learning Algorithms to classify based on data molecules expert system that is used to predict and identify protein labels. The functionality of various sources of data is used to classify based on various formation scenarios to predict its best data for drugs. The functionality of various sources of data is used to classify based on various formation scenarios to predict its best data for drugs. Change is the only constant and inevitable for projects that pave way for social reform. This project is as expandable as its area of application. Protein is one such field where constant improvement is the need of the hour.

II. EXPERT SYSTEM

CorrectPrediction of protein class is of greatimportance as ithelps thelife science communityto study about different diseases and come with new drugs and solution that help the humansurvival. Many methods were used for the same purpose. The first method to be used for proteinprediction is the Artificial Neural Network method. Machine learning was used for protein classprediction. Another machine learning model that is used to predict the protein class is RandomForest.

The Limitations of the system are thenumberofclasslabelsareingreaternumberthenaccuracywillbedecrease and if we makeuse thedifferentfeaturesthentheaccuracywillbeimproved. By keeping these form of strategies it will get functional.

III.RELATED WORK

Jianlin Cheng and his Team introduced with current biology machine learning methods which is an bioinformatics, computational, and systems where it can be widely used. Here, they viewed the development in protein structure prediction machine learning methods, Structural biology, and bioinformatics, one of the most fundamental challenges. The prediction of the protein structure is a complex, often degraded, four-level problem.1-D prediction of primary amino acid sequence structural characteristics;2-D forecasting of the relationship between amino acids, 3-D projection of the structures of the tertiary protein, and 4-D projection of the complex multi-protein quaternary structure. Arrange of controlled and uncontrolled learning processes have been put in place over the years to solve such problems and have contributed significantly to the development of advanced protein structure forecasts. Theydiscusses the development and implementation in the prediction of 1-D, 2-D, 3-D, and 4-DProteinStructures, of hideous Markov Models, neural networks, and support vector systems.

ShokoufehMirzaei and his team worked on functionality of a protein which determined by its structure, and effective methods are required to determine protein structures to promote scientific and medical science. Because existing methods of test structure determination have a high price tag, calculation prediction is highly desirable. The computer techniques produce several 3D structures known as decoys because of a protein sequence. However, the best decoys can be chosen because only a handful of end-user can handle them. Thusly, score capacities are fundamental to the imitations determination. They join quantifiable characteristics into a solitary pointer of bait quality. The current markup highlights don't reliably choose the best baits. Unfortunately, AI strategies give extraordinary conceivable outcomes to improve the fake rating. There paper contains two essential score capacities for AI, for example, the comparability between the figure structure and the trial structure without knowing the last mentioned. They utilize different estimations to contrast these scoring capacities with three bleeding edge scores. That was a first attempt to use the same non-redundant dataset to compare various score functions and the same features. The results showed that the addition of information can be significantly higher than the method used.

The precise score function is an important element for predicting the effective protein structure. Where Hongyi Zhou and Jeffrey Skolnick they both created a summed-up, distance-subordinate all-nuclear factual potential to address this significant issue that has not been settled. The new GOAP(A Generalized Orientation-Dependent , All-Atom Statistical Potential for Protein Structure Prediction) is subject to the general direction of the planes in the matching of communications that are associated with every weighty particle. GOAP represents a generalization of past orientation-dependent potential which takes only representative atoms or lateral and polar atomic blocks into consideration. GOAP is split into contributions from distances and angles. For the distance-dependent component of GOAP, the ideal DFIRE gas reference status is used. Out of the 11 frequently used decoys, 226 of the native structures were tested and decoyed as the best against the 127 by DFIRE. Decoy sets with the same structure as native ones (still under ~4.0A~THE) and with ROSETTA as initio decoy sets are a major improvement. FIRE or only RWplus' lateral guide was poor in these two sorts of decoys. Although OPUS-atomic PSP's contact potential depends on a block (that recognizes 196targets), it remains 15 percent worse than GOAP (DFIRE, RWplUS, and dDFIRE). GOAP, therefore, promises progress concerning statistical potential base to do knowledge and all atoms structure.

GianlucaPollastri and his Team had states that secondary structure predictions for protein structure and function for several methods are increasingly being used. This version of the program consists of SSPro's secondary classification in three categories and (b) SSPO8's first version in 8 categories which was produced by b. The first is a two-way recurring architectural envelope and a major irredundant training system. These profiles are derived PSIBLAST. The results are described in three test sets, in which approximately 78 percent of the SSpro achieved a sustained performance. The have reported disarray designs, contrast PSI-BLAST, and BLAST profiles, and assess suitable upgrades in execution of a protein structures.

IV. PROPOSED METHODOLOGY

This project proposes the use of the multinomial NB algorithm and the decision-tree algorithm in the protein classification model using supervised machine learning algorithms. The algorite size and number of technical indicators used are compared by parameter. Precision has been calculated for each algorithm. There are mainly four phases to the proposed architecture of the work carried out: function extraction from the data set, supervised classification of the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7723 - 7731
Received 25 April 2021; Accepted 08 May 2021.

training data set, supervision of classification of test dataset, and evaluation of results. The Advantages are More class labels are used in the proposed system than the existing system, Only relevant features are extracted to get better accuracy, The accuracy for three algorithms is computed and visualized for the same.

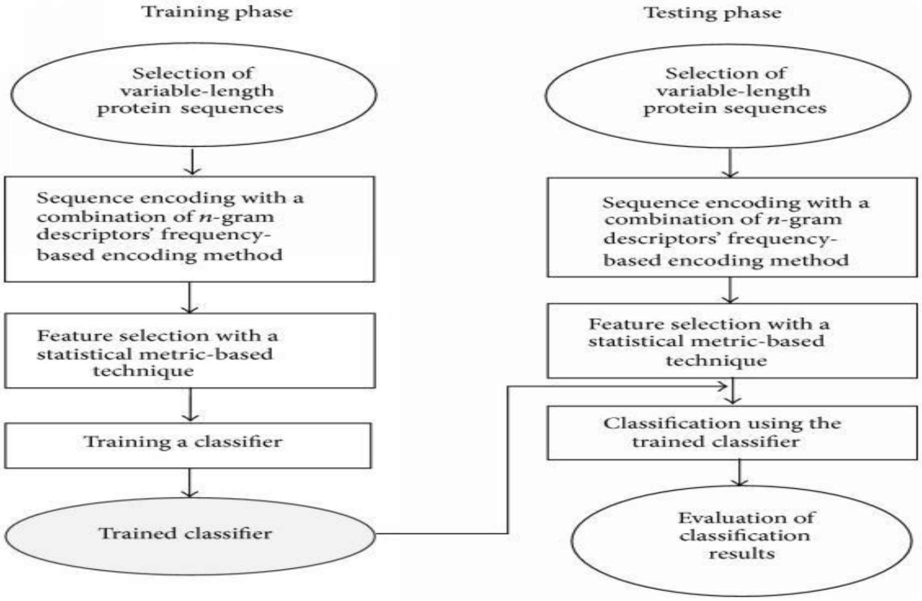

Training phase
Selection of variable-length protein sequences
Sequence encoding with a combination of n-gram descriptors' frequency-based encoding method
Feature selection with a statistical metric-based technique
Training a classifier
Trained classifier
Testing phase
Selection of variable-length protein sequences
Sequence encoding with a combination of n-gram descriptors' frequency-based encoding method
Feature selection with a statistical metric-based technique
Classification using the trained classifier
Evaluation of classification results

Fig 1: Training and Testing Classification of the Data

V. IMPLEMENTATION AND METHODOLOGY

Supervised Classification (Training & Testing Dataset):

The data has been divided into two parts i.e. training and testing data in the 70:30 ratios. Learning algorithms have been applied to the training data and based on the learning; predictions are made on the test data set.

By applying Multinomial NB, Random Forest, and Decision Tree Algorithm to use best accuracy level to predict the drugs**.**

Data Functionality:

The data will varies based on various protein molecular structures which will classify on various forming sequences with Structure ID as mentioned in-sample data set.

The Prediction on Train and Test data will be targeted on classification attribute which counts on accurate molecular methodology

Source of Data

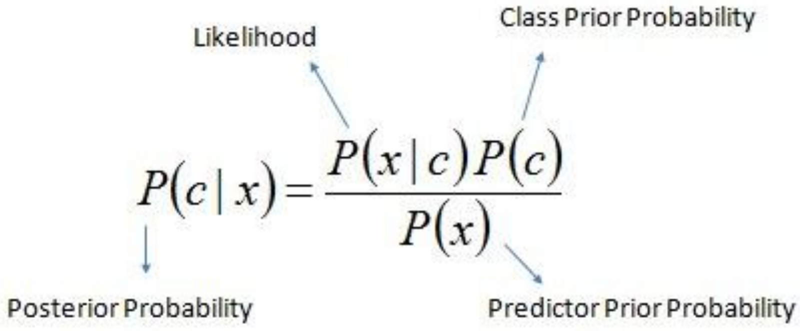
The Source of the Data is Protein Data Bank (www.wwpdb.org) and the Kaggle website.

Algorithm 1: Multinomial Naive Bayes Algorithm

The Multinomial Algorithm show will classify the Naïve Bayes. This will be considered as on Bayes Hypothesis, where the Naïve says that highlight within the data set is commonly free. The event of one highlight does not influence the

Received 25 April 2021; Accepted 08 May 2021.

likelihood of the event of the other highlight.



$$P(c\,|\,\mathrm{X}) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

Fig. 2: Multinomial Naïve Bayes Algorithm Functionality.

Algorithm2 : Random Forest Algorithm

The Random Forest Algorithm is collected of distinctive choice trees, each with the same hubs, but utilizing distinctive information that leads to distinctive takes-off. It consolidates the choices of different decision trees in arrange to discover a reply, which speaks to the normal of all these decision trees.

$$Gini=1-\sum_{i=1}^{C}(p_i)^2$$

$$Entropy=\sum_{i=1}^{C}-p_i*\log_2(p_i)$$

Fig. 3: Random Forest Algorithm Classification Statistics.

Algorithm3 :Decision Tree Algorithm

A Decision Tree is a Supervised learning method that can be utilized for both classification and Relapse issues, but for the most part, it is preferred for tracking Classification issues. It could be a Tree-Structured classifier, where inside hubs speak to the highlights of a dataset, branches speak to the choice rules and each leaf node speaks to the result.

Entropy(s)= -P(yes) $\log_2$ P(yes) – P(no) $\log_2$ P (no)
Information Gain= Entropy (S) - [(Weighted Average) * Entropy (each feature)]

Received 25 April 2021; Accepted 08 May 2021.

VI. RESULT AND DISCUSSIONS

1. Graph for Class Label Distribution

The datasets will vary based on statistical analysis in each phase of Protein Classes of Data because the data used to be predicted on various classifications of protein sequences. The technical analysis of the graph will define the protein enzymes and their counts.

Protein Classes	Datasets Counts
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

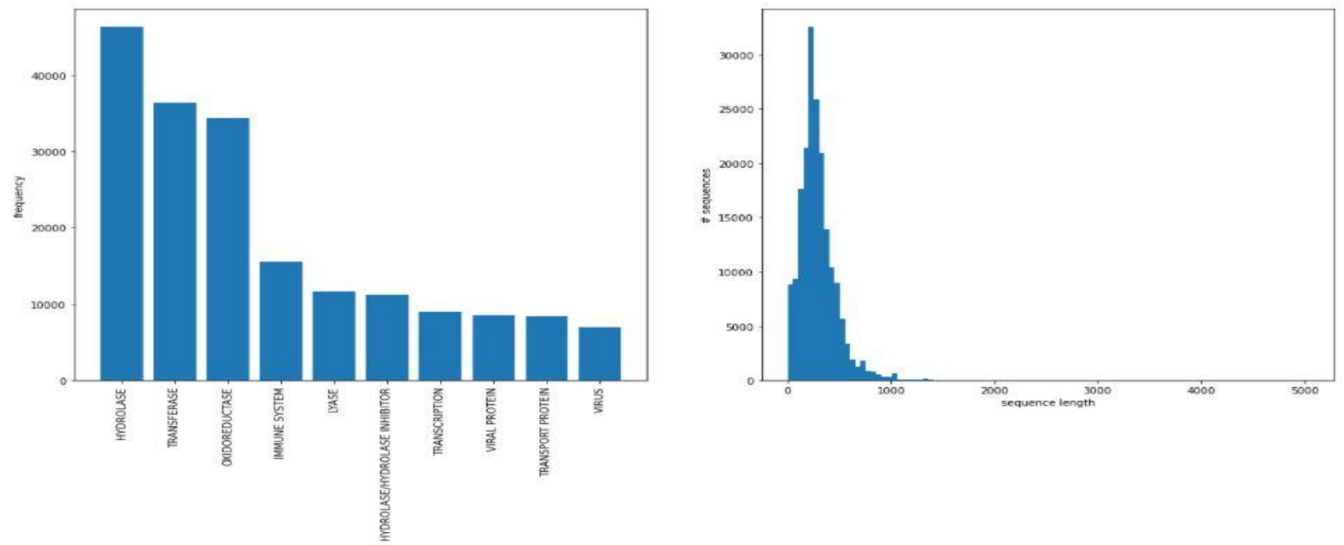


Fig. 4.Protein Class Enzymes Graphical Analysis.

2. Performance of Different Algorithms

The Supervised Algorithm predictions will vary based on the formulae and statistics as respective systematic parameters of the data which used taken in Training and Testing the datasets. The below graph of the status of each algorithm that classifies the prediction accuracy will change based on the required calculation of data. As Multinomial, Random Forest and Decision Tree will consolidate by its functional parameters.

Multinomial Naïve Bayes Algorithm Accuracy :89.63%

Random Forest Algorithm Accuracy : 33.31%

Decision Tree Algorithm Accuracy : 92.58%

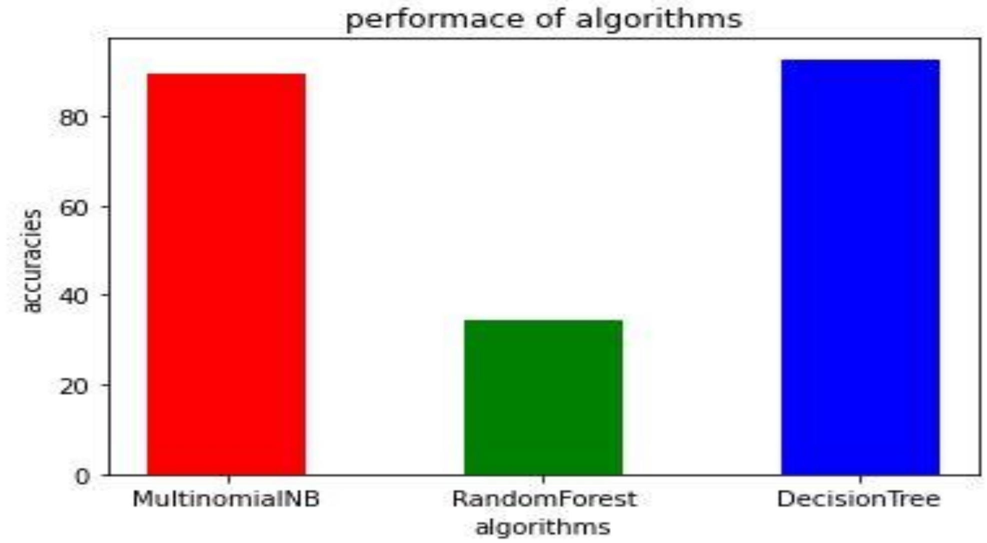
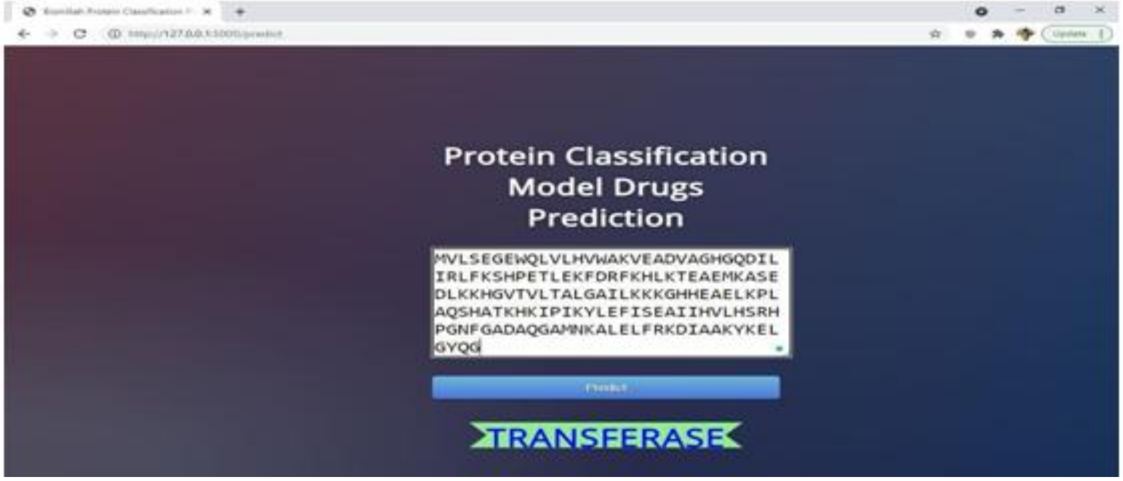


Fig. 5. Classification Accuracy Levels of Algorithms.

3. Protein Classification Sample Drug Prediction

The Protein Sequences are used to predict based on Count Vectorizer to give a count ID for the formation in Machine Learning Model Building Analyzer. The datasets will predict the protein classes Enzymes assign to their respective functioning. The below classification is build based on Multinomial Naïve Bayes Algorithm to predict efficiently and accurately with protein sequences results. The User Interface will function on its respective data.



Protein Classification
Model Drugs
Prediction

MVLSEGEWQLVLHVWAKVEADVAGHGQDIL
IRLFKSHPETLEKFDRFKHLKTEAEMKASE
DLKKHGVTVLTALGAILKKKGHHEAELKPL
AQSHATKHKIPIKYLEFISEAIIHVLHSRH
PGNFGADAQGAMNKALELFRKDIAAKYKEL
GYQG

Fig.User Interface of Protein Sequence Classifications.

VII. CONCLUSION AND FUTURE ENHENCEMENT

In this paper, Random Forest, Tree Classifier Decision and Multinomial Naive Bayes algorithms have been used to forecast protein labels. We have taken into account the heterogeneity of data sets and new solutions for machine learning when building a research framework for protein prediction. The results show that the Random Forest Algorithm exceeds all the other algorithms for large datasets in terms of accuracy, and when the data set is smaller to almost half the original, the Naïve Bayes Algorithm shows the best accuracy results. The decrease in the number of technical indicators will also reduce the precision of the protein class predictions of each algorithm.

This analysis will give good results in protein class prediction by using this system we can able to predict any type of protein class with high accuracy, then if anyone wants to check a particular class of protein they can follow this analysis and can get good results in his research.

VIII. REFERENCES

[1] AlbertsB.etal.(2002), The Shape and Structure of Proteins, Molecular Biology of the Cell; Fourth Edition. New York and London: GarlandScience.ISBN0-8153-3218-1.

[2] Mirzaei S, T.Sidi, C.Keasar, and S.Crivelli (2016),Purely Structural Protein Scoring Functions Using Support Vector Machine andEnsemble Learning, IEEE/ACM Transactions.

[3] ZhouH.,andSkolnickJ.(2011),GOAP:aGeneralizedOrientation Dependent,All-Atom Statistical Potential for Protein Structure Prediction, Biophysical Journal, vol. 101, no. 8,pp.2043-2052.

[4] KhouryG.etal.,(2014),WeFold:ACoo petitionforProteinStructurePrediction.Proteins: Structure, Function, and Bioinformatics;82(9):1850-1868,doi:10.1002/prot.24538.

[5] Faraggi E. and Kloczkowski A. (2014), A global machine learning based scoring functionfor protein structure

prediction, Proteins: Structure, Function, and Bioinformatics, vol. 82, no. 5,pp.752-759.

[6] Computational Biology and Bioinformatics, Volume: PP Issue : 99. DOI:10.1109/ TCBB.2016.2 602269.

[7] Pollastri, Gianluca, et al., (2002), Improving the prediction of protein secondary structurein three and eight classes using recurrent neural networks and profiles. "Proteins: Structure, Function, and Bioinformatics47.2:228-235.

[8] ZarembaW.etal.(2014),Recurrent Neural Network Regularization, a Xivpreprintar Xiv: 1409.23 29.

[9] KeasarCetal.,(2017).“An Assessment of WeFold : A Framework of International Collaborative Pipelines for ProteinStructurePrediction.”Submitted to Scientific Reports.

[10] A.C.M.May.Towardmoremeaningfulhierarchicalclassificationofaminoacidsscoring functions. Protein Engineering,12:707-712,1999.

[11] Marco TulioRibeiro Sameer Singh Carlos Guestrin 2016, Introduction to LocalInterpretable Model-AgnosticExplanations(LIME)A techniquetoexplainthepredictionsofanymachine learning classifier. O’Reily Learning Platform,https://www.oreilly.com/learning/introduction-to-local-interpretablemodel-agnostic-explanations-lime,Lastaccessed12/24/2018