Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-020-00528-3

ORIGINAL RESEARCH

Protein class prediction based on Count Vectorizer and long short term memory

S. R. Mani Sekhar[1] · G. M. Siddesh[1] · Mithun Raj[1] · Sunilkumar S. Manvi[2]

Received: 24 December 2019 / Accepted: 28 September 2020

Abstract Proteins class and function prediction is one of the most significant task in computational bioinformatics. The information about the protein functions and class plays a vital role in understanding biological cells and has a great impact on human life in factors such as personalized medicine. The technical advancement in the areas of biological aspects and understanding of biological processes results in features and characteristics of important Proteins. Prediction of amino acid sequence involves prediction of amino sequence folding and its structures from the primary sequence obtained. In this work, Machine learning prediction algorithms have applied for protein class prediction. This method takes consideration of macromolecules of biological significances. Later the solution focuses on the understanding of different protein family, subsequently classify the protein family type sequence. This is achieved through machine learning algorithms Naive Bayes (NB) and Random forest (RF) algorithms with count vectorized feature and LSTM. These algorithms are used to classify the protein family on its protein sequence. Finally, result shows that LSTM predicts the protein class more accurately than the RF, and NB algorithm. LSTM achieves an accuracy of 96% whereas RF & NB with an accuracy of 91% and 86%.

Keywords Protein · Protein–protein interactions · Naïve bayes · Features · Random forest · Machine learning · LSTM

1 Introduction

All living organisms are composed of cells, behind the functioning of the cells Proteins play a major role due to their important aspects in biological activity and also it is very important to understand their protein functionality. The importance of proteins and its functions in understanding how biological activities can be activated at the molecular level. This kind of understanding helps in development of personalized medicine, betterment of crops and therapeutic interventions and also supports in understanding the technical aspects of biological entities and computer systems. With also overwhelming growth of proteins with unidentified functions. Due to this circumstances it is very difficult to manually identify and predict the functionality and group them to protein family. Many methods have been proposed in order to characterize the protein functionality and essential protein prediction. These kind of techniques are based on fundamental information about proteins that might be depending on their amino acid sequence and also using tools such as Basic Local Alignment Search Tool (BLAST). The sequence of a protein structure [1] determines the characteristics that includes sub-cellular localization and structural information with its functionality. Features based on their data type i.e., contextual data and Protein–Protein Interactions (PPI) data. In the advanced technologies for prediction activities use these features based contextual data and PPI data. In the recent times we have access to interesting information

✉ S. R. Mani Sekhar
manisekharsr@gmail.com

[1] Department of Information Science and Engineering, Ramaiah Institute of Technology, Bengalore, India

[2] School of Computing and Information Technology, REVA University, Bengalore, Karnataka, India

Published online: 11 October 2020

about the proteins and have very effective models and techniques for precise protein function prediction.

With the help of technical advancements helps us to understand the biological process is improving day by day and new features explaining are emerging regularly. Important issue here is to focus how we can build effective models by understanding the biological information and its characteristics and features. The observations may not always help us in classification of proteins whose functionalities may not be identified due to lack of biological information.

The paper is organized as follows. Section II brief about the related work in protein class prediction. Next Section III illustrates the proposed machine learning model. Subsequently, Section IV focuses on the Results and discussion part. Finally, Section V provides a brief discuss on Conclusion and future work.

2 Related work

In this Section, information about the literature survey on the protein sequence prediction, protein class, and protein–protein interaction prediction is discussed.

Extracting co-evolutionary features [2] from protein sequences for predicting protein–protein interaction, deals with the study of proteins having their functionality by interacting with other proteins. Also, provide a brief about the information on molecular mechanisms of biological process that includes DNA regulations, complexity of the protein and cell signaling. This explains about the constraints involved with extracting features of the co-evolutionary features obtained from the information of sequences.

[3] Uses classification system for protein sequences. This system has three main steps: the pre-processing phase consists of extracting the descriptors, using the N-gram technique, extracting the rules of association between the protein components, applying the apriori algorithm; selecting the relevant rules for classifying the unclassified protein as a third step. The approach solves the problem of representation of protein sequences based on the n-gram technique and successful association rules and sorting approaches. And also deals to support develop an unknown protein classification system.

Bankapur et al. [4] developed a computational model for feature extraction using SkipXGram and character embedded features techniques for 25PDB and FC699 dataset. Later the combined features are modeled using different machine learning algorithms, subsequently the system is evaluated on 2-layers protein sequence. The accuracy can be further increase by incorporating Count Vectorizer.

[5] focus on the assessing part of protein structure prediction named as "estimating the accuracy of structure prediction models (EMA)". They deal with the training of Convolutional Neural Networks (CNN) for identification of IDDT and GDT TS kind, subsequently helps in exploring relative solvent using different dataset.

In this [6] work author has proposed a novel deep neural network for protein class prediction. The system considers three different input types profile: features with HHNlits and PSI-BLAST and sequence features for class prediction. Subsequently deep network is integrated with residual unit for effective protein secondary structure prediction.

When dealing with a distance based encoding of feature extraction with respect to bioinformatics, approximately around 20,000–25,000 gens were identified in Deoxyribonucleic acid (DNA) of human and these information is stored in the database for future study. These sequencing data related to DNA, RNA and proteins are increasing every day. The building block of the protein includes 20 essential amino acids. The distanced-based feature-encoding technique [7] provides the distance between each amino acid from its first sequence of amino acid and these features are then used for proposing method called sequence encoding of decomposition and this can be represented up to length of 120 feature dimensions ranges to $n \times 120$, where n is denoted by number of maximum sequences taken and that can be composed to level 4 and the total number of features are around 160.

A PPI in Bioinformatics covers important aspects for intercellular signaling ways, complex structures [8] and also various biochemical processes PPIs. This kind of experimentation is quite expensive and can be seen many drawbacks. Proteins are made up of biomolecules as well as macromolecules contain long chain of amino acids [9, 10].

3 Proposed model

In this work prediction of protein class is depend upon the different machine learning algorithms like NB [11] and RF [12]. Figure 1, illustrates the architecture diagram of proposed solution of protein class prediction. The libraries and

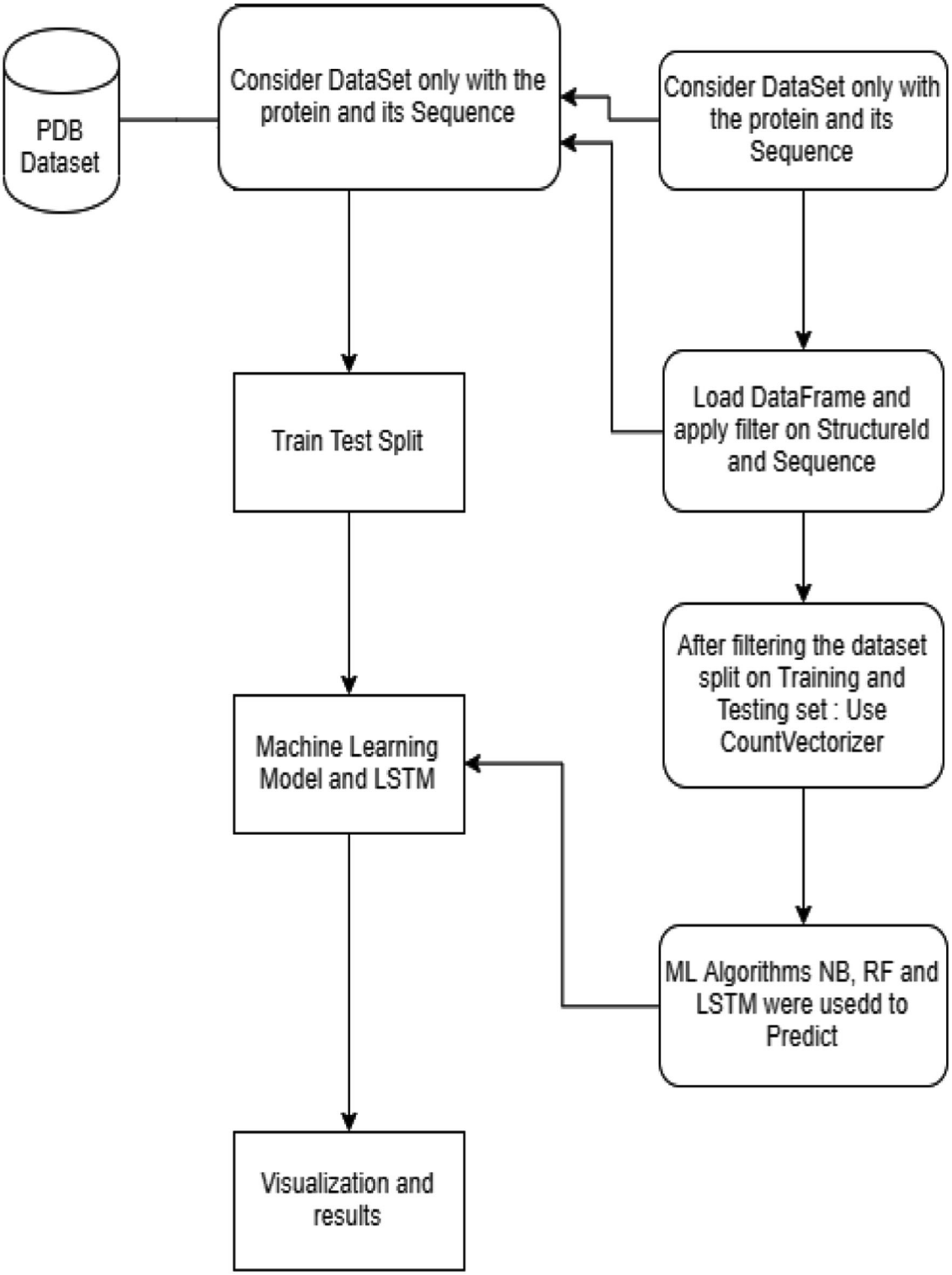

Fig. 1 Proposed Machine learning model for class distribution

protein data bank (PDB) datasets [13] are imported. As the protein sequence type is of a string, the model incorporate a library called Count Vectorizer. It is a feature extractor that is used in Natural Language Processing machine learning models. Thereafter, once the data loaded in the data frames it is processed and only the required field is taken in to consideration. Due to the presence of class type in the dataset, it helps the model to learn a patter for a specific class. Subsequently allows the dataset to be filter. Here 80% of data is used in a training set whereas 20% of data is used in testing set.

Once splitting of data is done Count Vectorizer is utilized to create a dictionary composed from the training dataset. This type extracts individual characters to gain features. In a protein, it"s not the individual amino acid that gives the identification of its chain. There are secondary and tertiary structures that are formed via the bonds of amino acids in the present sequence. As a result, using ngram_range of (4, 4) seems to be perfect choice for feature extraction. This procedure will extract the different subsets of length 4 allowing amino acids [9, 10] that interact with each other.

Subsequently, the extracted features use a Machine learning models, resulting in the protein class prediction the algorithm used here are NB [10] & RF [11]. Algorithm 1 illustrate the computation process for protein structure prediction by integrating count vectorized features and Naïve Bayes methodologies [14], whereas algorithm 2 uses a Random Forest Technique with count vectorized features for protein structure prediction.

Algorithm 1: Protein Class prediction by integrating count vectorized feature and Naïve Bayes technique.

Step 1: A = {$attr_1$,$attr_2$,$attr_3$,..,$attr_n$ }, where $attr_i$ are the attributes of table
$attr_n$ -> class attribute
$clss_1$,$clss_2$,…,$clss_m$ € $attr_n$ where $clss_i$ are distinct class values
Step 2: ∀ $attr_i$ $1 \leq I < n$ find Prob ($attr_i$|$clss_j$) $1 \leq j \leq m$
Step 3: Let X = {x_{n1},x_{n2},..,x_{ni}} be the labelled Sample
Step 4: Prob (y|X) = Prob (X|y) Prob (y) / Prob (X)
Step 5: Find Prob (X|$clss_i$) , where $clss_i$ € $attr_n$ and $1 \leq I \leq m$
Let Q be the set of values found in this step
Step 6: Prob (X|$clss_i$) = max {Q}
Step 7: if any other labelled samples left go to step 3
Step 8: Stop

Algorithm 2: Protein Class prediction by integrating count vectorized feature and Random forest technique.

Input: Attribute & Labelled sample
Output: Class to which labelled sample belongs to
Step 1: A = {$atrr_1$,$atrr_2$,$atrr_3$,..,$atrr_n$ }, where $atrr_i$ are the attributes of table
$atrr_n$ -> Class attribute
$clss_1$,$clss_2$,…,$clss_m$ € $attr_n$ where ci are distinct class values
Step 2: Build bootstrap table
Step 3: Select "k" random features, where k<<n
Step 4: find the best split attribute "b" where b € k
Step 5: Build decision tree by making "b" as root
Step 6: Repeat from step 3 for creating further trees. Collection of all these trees is random forest
Step 7: for the given sample X = {X_{ni}} ,and $1 \leq i \leq n$
Find prediction from each tree in the random forest
Step 8: X € $clss_i$ depending on which $clss_i$ has got highest votes in step 7
Step 9: Stop

Algorithm 3, discuss the LSTM architecture, here hidden layers give output as weight parameters such as previous hidden state and cell state ct. Input (it), Output (ot) and forget (ft) gates are responsible for changes in the cell state and hidden state (ht). Equations 1 to 6 discuss about the LSTM process.

$$ft = (wf.\ xt + uf.ht - 1) \quad (1)$$

$$it = (wi.\ xt + uf.ht - 1) \quad (2)$$

$$c \sim t = \tan h\ (wc.xt + uc.ht - 1) \quad (3)$$

$$ot = (wo.xt + uo.ht - 1) \quad (4)$$

$$ct = ft\ x\ ct - 1 + it\ x\ c \sim t \quad (5)$$

$$ht = ot\ x\ \tanh(ct) \quad (6)$$

Here "x" is element wise product and " = " is the sigmoid function. Thereafter "u" and "w" are different weight parameters for each hidden layer in the LSTM. There is no weight parameter consideration after the hidden layers.

Algorithm 3: Protein Class prediction by LSTM

Input: **X** (current input), **H** (previous hidden state) and **C** (previous memory state)
Output: **H** (current hidden state) and **C** (current memory state)
Step 1: First, the previous hidden state and the current input get concatenated. We"ll call it combine.

$$Ct = Ct\text{-}1 * ft$$

Ct = Current memory state at time step 1
Step 2: Combine get inserted into the layer of forgetting. This layer removes data that is not relevant.
Step 3: Using combine to create a candidate sheet. The candidate retains values that can be added to the cell state.
Step 4: Combine gets fed into the output layer as well. This layer decides that candidate information should be transferred to the new cell state.
Step 5: Using those vectors and the previous cell state, the cell state is determined after computing the forgotten layer, candidate layer, and input layer.
Step 6: Then the output is measured.
Step 7: Multiplying the output point wise and the new cell condition gives us the new hidden state.

$$Ct = Ct + (It * C`t)$$

But it"s going to be a filtered version. So we apply Tanh to Ct and then we do element wise multiplication with the O output window, that"s our current hidden Ht state.

$$Ht = Tanh(Ct)$$

Step 8: We proceed to the next step these two Ct and Ht and repeat the same process.
Step 9: Stop

A popular variant of RNN is LSTM [13–17]. It consists of memory to store the previously processed data along with repeated sequences to process the long sequences.

4 Data set

This study uses Kaggle domain for structural protein sequences dataset. The dataset is downloaded from: https://www.kaggle.com/shahir/protein-data-set#pdb_data_seq.csv. This protein dataset is derived from the PDB Protein Research Collaborator for Structural Bioinformatics (RCSB) [13]. The dataset is divided into two sub-sets-the first part contains data on protein meta which includes details on protein identification, methods of extraction, etc. and the second part contains sequences of protein structure. The two datasets were organized using the protein"s "structureID" attribute. The first dataset is made up of 1,41,000 rows and 14 columns while the second dataset is made up of 4,67,000 rows and 5 columns. Later, this raw dataset is filtered by removing empty or unnecessary fields. Finally, we have considered only first 10,000 rows for more effective computation.

Initially, combining the two datasets into a single dataset by merging them into the attribute "structureID." The lines without labels and sequences will be dropped or deleted after merging. First, since the dataset includes various macromolecules, only protein is selected based on the "macromoleculeTypex" attribute for further processing. The database is made up of many different types of biologically important macromolecules. Many information files are protein records. For DNA being the precursor to RNA, which is the biomolecules that interact directly in biological pathways and cycles when translated. The Fig. 2 shows the sample dataset. It contains the following feature like "structureid", "chainid", "sequence", "residueCount", and the "macromoleculeType". Here "structureid" and "chained" used as in identifier for a given structure and chain. Thereafter, "sequence" shows the amino acids pattern for the specific ID. Whereas, "residualCount" shows the number of residual for the given ID. Finally, "macromoleculeType" display the type of ID such as DNA, RNA, Protein.

Figure 3, focus on the post-process dataset. Once the data loaded into two separate dataframes, a filter and a join must be performed to get the data together. Begin with filtering the datasets where the classification [15] is equal to "Protein" followed by removing all other variables other than "structureId" and "sequence" for the loaded data.

Proteins are typically based around one or several functions that are determined by their type of family. For example, we can have a Hydrolase group protein that focuses on catalyzing hydrolysis (breaking bonds by adding water) to help promote the breakdown of protein chains or other molecules. Additionally, only 10 common classes of proteins are used, based on the row count. Figure 4, shows the abundance of each of the 10 protein groups picked the classes of protein family contains "hydrolase" protein contains frequency or the dataset belogs to 9500 proteins of hydrolase, whereas "transferace" have the frequency of 7500 protein dataset, similarly "oxidoreductase" have the protein frequency in the dataset around 4500, next set of protein family "immune system" has frequency of 4000 protein dataset, "transcription" has the dataset around 2500 protein family, "signalling" protein has around 2000 data set that belongs to this category, and protein classes such has "lyase", "transport" protein, "protein binding", and "structural genomics", has around 1800–2000 family of proteins that belongs to the top ten classes selected.

As it is important to translate these categorical values into binary form. For this, LabelBinarizer converts the string labels into one warm representation. The values are assigned 1 in one warm representation if that value is

Fig. 2 Sample protein dataset [13]

	structureId	chainId	sequence	residueCount	macromoleculeType
2	100D	A	CCGGCGC(	20	DNA/RNA Hybrid
3	100D	B	CCGGCGC(	20	DNA/RNA Hybrid
4	101D	A	CGCGAATT	24	DNA
5	101D	B	CGCGAATT	24	DNA
6	101M	A	MVLSEGEV	154	Protein
7	102D	A	CGCAAATT	24	DNA
8	102D	B	CGCAAATT	24	DNA
9	102L	A	MNIFEMLF	165	Protein
10	102M	A	MVLSEGEV	154	Protein
11	103D	A	GTGGAAT(	24	DNA
12	103D	B	GTGGAAT(	24	DNA
13	103L	A	MNIFEMLF	167	Protein
14	103M	A	MVLSEGEV	154	Protein
15	104D	A	CGCGTATA	24	DNA/RNA Hybrid

	structureId	sequence
4	101M	MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR...
7	102L	MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAAKSE...
8	102M	MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR...
11	103L	MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNSLDAAK...
12	103M	MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR...

Fig. 3 Post-process dataset with Structure ID and Sequence

otherwise allocated to the value 0. Later, tokenizer method from the Keras library is used for more pre-processing of sequences, which converts each character in the sequence into a number. Each sequence's length is also made standardized for accurate processing. A maximum of 256 characters are used here.

5 Results and discussion

The proposed model uses a protein dataset which contain the 8000 sequences with different protein ID"s. The experiments are performing on the different rows count like 2000 rows, 4000 rows, 6000 rows and 8000 rows.

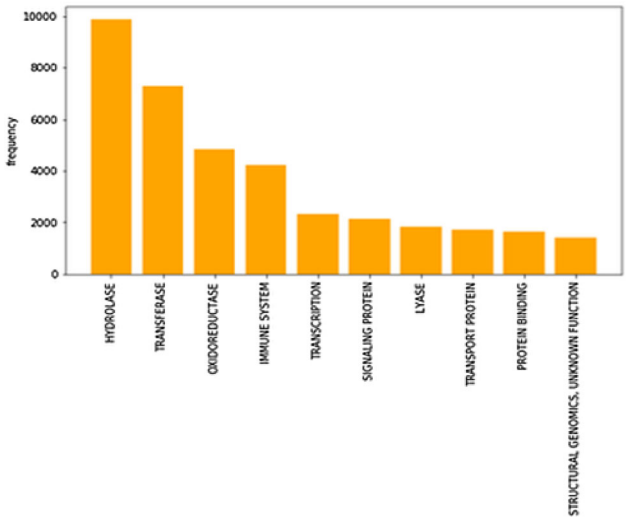

Fig. 4 Top ten protein classes

Figure 5, illustrates dataset consists of 2000 rows and count class classification is considered above 100 and Hydrolase, oxidoreductase types are present and accuracy is achieved 87% with NB Algorithm and 89% with RF Algorithm.

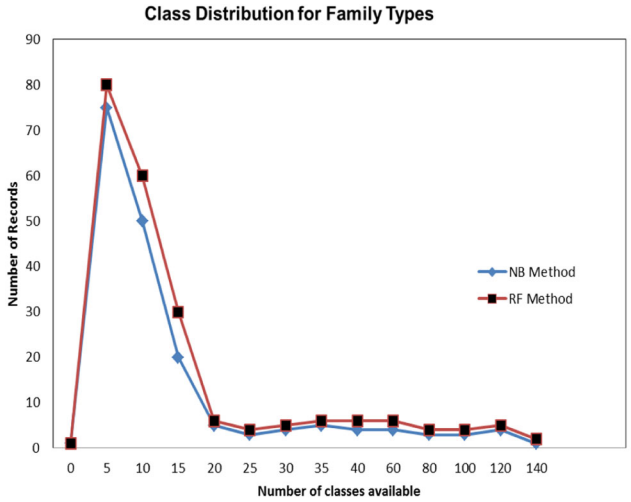

Fig. 5 Protein class prediction from 2000 rows

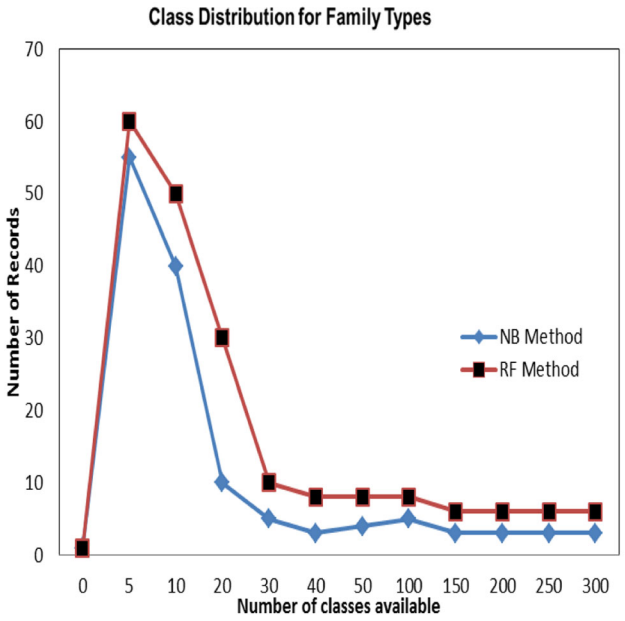

Fig. 6 Protein class prediction from 4000 rows

Figure 6, dataset consists of 4000 rows and count class classification is considered above 100 and Hydrolase, oxidoreductase, hydrolase/hydrolase inhibitor, transferase, electron transport types present, after filtering 963 rows are present and accuracy is achieved around 86% with NB and 88% with RF.

Figure 7, contains 6000 rows and count class classification is considered above 100 and Hydrolase, oxidoreductase, hydrolase/hydrolase inhibitor, transferase, electron transport lyase types present, after filtering 1691 rows are present and accuracy is achieved around 86% and with the count classification within 400, 91% accuracy achieved with NB and 92% with RF.

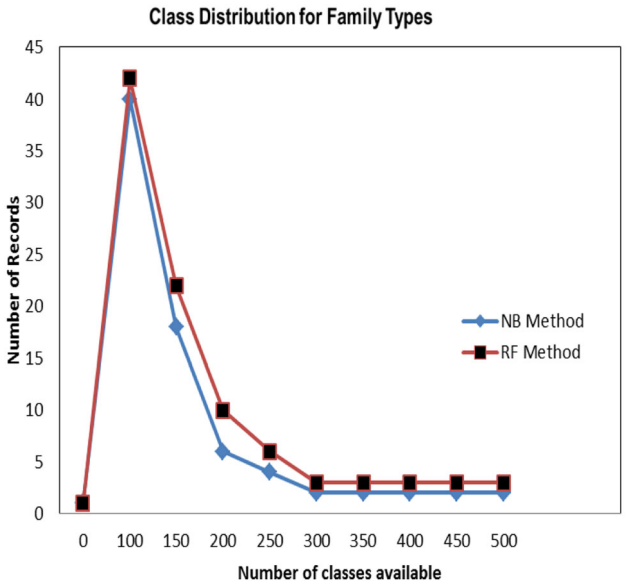

Fig. 7 Protein class prediction from 6000 rows

Finally, Fig. 8 dataset consists of 8000 rows and count classification is considered within 500 and Hydrolase, oxidoreductase, hydrolase/hydrolase inhibitor, Transferase, electron transport types present, after filtering 2095 rows are present and accuracy is achieved around 86% with NB and 91% with RF.

Figure 9, represents the overall comparison for NB, RF and LSTM and from the observations for 2000 rows of data NB predicts around 83%, whereas RF predicts 90% and LSTM shows the performance of 92%. For 4000 rows NB predicts 86%, RF accurately 91% and LSTM predicts classes about 91%. Dataset of 6000 rows NB, RF and LSTM predicts the class 91,90,94% respectively.8000 rows

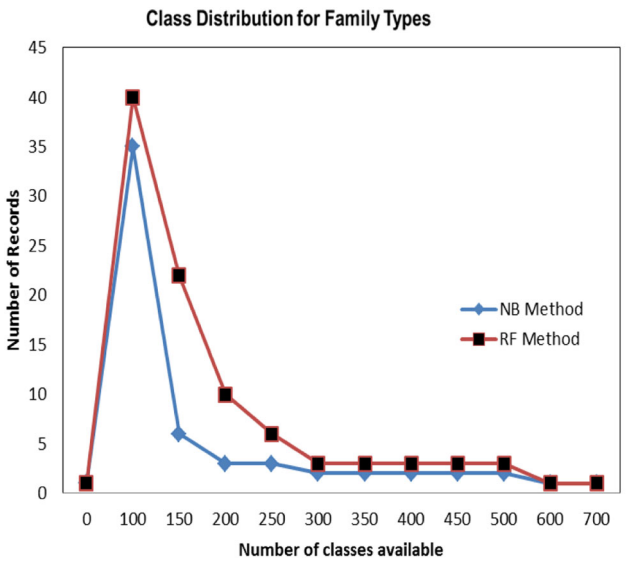

Fig. 8 Protein class prediction from 8000 rows

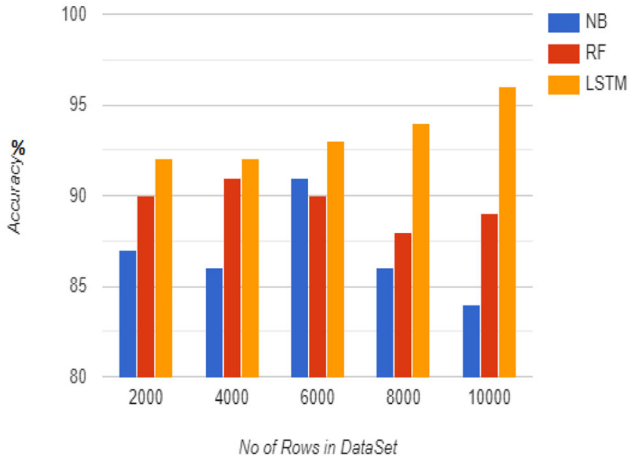

Fig. 9 Protein class prediction using NB, RF and LSTM

of protein sequence predictions for NB, RF and LSTM shows 86%, 88% and 94% respectively. Finally, on computation on 10,000 rows LSTM predicts the protein class of 96% whereas NB predicts 84% and RF predictions shows 88%.

6 Conclusion and future work

The information about the protein functions plays a vital role in understanding biological cells and has a great impact on human life in factors such as personalized medicine, betterment of crops and therapeutic interventions. The proposed work uses Machine Learning techniques NB & RF for protein class predicts from their amino acid sequences with count vectorized feature. The model performance analysis with Machine Learning techniques such as Naïve Bayes and random forest for prediction of protein class from their amino acid sequences. With the assistance of Naïve Bayes and Random Forest Classifier Protein Structure, with a dataset of 8000 rows, achieved an accuracy of 86% and 91% respectively, with the help of 4 vectors. The result shows that Random Forest performs better than Naïve Bayes. Further in LSTM observations, the model is trained on 25,350 samples and validated on 4474 samples and the accuracy is found to be 96%. In future accuracy can be achieved more with consideration of more than 4 vectors with the help of other factors such as pH, Molecular weight and other components may yield more information on family group and also to include more characters to allow for higher interaction between the amino acids.

References

1. Pauling L, Corey RB, Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 37:205
2. Rehman HU, Azam N, Yao J, Benso A (2017) A three-way approach for protein function classification. PLoS ONE 12(2):0171702
3. Kabli F, Hamou RM, Amine A (2017) New classification system for protein sequences. In 2017 First International Conference on Embedded and Distributed Systems (EDiS), IEEE. Oran, Algeria, pp. 1–6
4. Bankapur, Sanjay, and Nagamma Patil (2018) Protein Secondary Structural Class Prediction Using Effective Feature Modeling and Machine Learning Techniques. In 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE pp.18–21
5. Lima, Emerson Correia, Fábio Lima Custódio, Gregório Kappaun Rocha, and Laurent E. Dardenne (2018) Estimating Protein Structure Prediction Models Quality Using Convolutional Neural Networks. In 2018 International Joint Conference on Neural Networks (IJCNN), IEEE pp. 1–6
6. Fang, Chao, Yi Shang, and Dong Xu. (2017) A New Deep Neighbor Residual Network for Protein Secondary Structure Prediction. In 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE pp. 66–71
7. Iqbal MJ, Faye I, Said AM, Samir BB (2014) Data mining of protein sequences with amino acid position-based feature encoding technique. In: Herawan T, Deris MM, Abawajy J (eds) Proceedings of the First International Conference on Advanced Data and Information Engineering. Springer, Singapore
8. Anfinsen C (1972) The formation and stabilization of protein structure. Biochem J 128:737
9. Dictionary (2019) Amino. https://www.dictionary.com/. Accessed 25 March 2019
10. Amino acid, [Online]. Available: https://en.wikipedia.org/. Accessed 22 May 2015
11. Robles V, Larrañaga P, Peña JM, Menasalvas E, Pérez MS, Herves V, Wasilewska A (2004) Bayesian network multi-classifiers for protein secondary structure prediction. Artif Intell Med 31:117
12. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
13. Protein data bank. Availabe https://www.kaggle.com/shahir/protein-data-set#pdb_data_seq.csv
14. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
15. Hawkins J, Boden M (2005) The Applicability of recurrent neural networks for biological sequence analysis. IEEE/ACM Trans Comput Biol Bioinform 2(3):243–253
16. Jain G, Sharma M, Agarwal B (2019) Optimizing semantic LSTM for spam detection. Int J Inf Technol 11:239–250
17. Chhachhiya D, Sharma A, Gupta M (2019) Designing optimal architecture of recurrent neural network (LSTM) with particle swarm optimization technique specifically for educational dataset. Int J Inf Technol 11(1):159–163