ISSN 2347 - 3983
Volume 9. No. 4, April 2021
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter18942021.pdf
https://doi.org/10.30534/ijeter/2021/18942021

Implementation of Protein Sequence Classification for Globin family using Ensemble Learning

Himanshi Agrawal[1], Neha Mehra[2], Dr.Vandan Tewari[3]
Department of Computer Science and Engineering
Shri Govindram Seksaria Institute of Technology and Science Indore (M.P.),India,
[1]himanshiagrawal08@gmail.com, [2]mehra.neha40@gmail.com
[3] vandantewari@gmail.com

ABSTRACT

Feature Extraction from protein sequence is a very important task in bioinformatics. The main focus of that work is protein sequences classification that can be used to improve drug discovery and identification of diseases for treating patients in the early stages of diagnosis. In this paper, we proposed a method which is used for feature extraction i.e. converting the protein sequence of hemoglobin in to feature vectors. The feature vectors are then given to the ensemble classifier as an input which uses various classifier to provide better result/performance as compared to any constituent learning algorithm alone.

Key words : Protein sequence, Feature extraction, classification, Ensemble learning.

1. INTRODUCTION

Bioinformatics deals with the computational management and analysis of biological information i.e. genes, genomes, proteins, cells, etc. Genome data is a complete list of nucleotides that make up of all the chromosomes of an individual or species. Genome sequence changes at every moment and every second and number of genome data are continuously increasing and the size of data varies with time. There are different types of genome data, such as protein data, DNA, SNP data. Protein data are building blocks of all organisms. Protein sequence classification is an important problem, which is helpful for determine superfamilies, characteristics, structure/function of a protein sequence i.e. unknown. The main application is to improve the drug discovery and identification of diseases for treating patients at an early stage of diagnosis. Suppose, we have obtained sequence 'S' from disease 'D' and by the method of classification and it is found that 'S' belong to superfamily fi then the combination of drugs is used for their disease D which belong to fi.

Amino acids is the structural unit of protein each amino acid is formed by the combination of one or more amino group, one or more carboxyl group and R-group linked together to the common carbon atom. A protein sequence is collection of 20 different amino acid. A protein superfamily is the group of proteins which share the common structure and functionality. Globin are superfamily which contain different families Hemoglobin (Hb), Myoglobin (Mb), Cytoglobin (Cb). First, we need to preprocess of sequences with the help of blast tool and clustal omega tool. Blast tool is a very commonly used search tool for the comparison of any DNA or protein sequence with a library or database of sequence . Blast compares a query sequence to a reference database to find the similar matching pairs [13]. After this preprocessing we apply Clustal omega tool is used for the multiple sequence alignment of large sequences protein or DNA/RNA[14]. Multiple sequence alignment approach is the order of alignment is determined by a guide-tree. It is constructed from pair-wise distances among the sequences.

1.1 Related work

Related work consist an observation as well as study of work done in the area of classification of protein sequences.

A Distance-Based Approach

Muhammad Javed Iqbal, [1] discussed the method of feature extraction for protein sequence using the distance-based. In this paper, the method of feature extraction has calculated the distance from first amino acid to each amino acid in the sequences decomposition at a different level. In this paper, the level used up to 3. So each sequence represented a feature vector of length 120. Here, the Complexity of protein sequences is increasing if it increases the level of decomposition.

B N-Gram Method

Eghbal G. Mansoori, [12] the n-gram method is used for the feature extraction from the sequence of a protein to discriminate the different superfamilies. In this paper, the feature extraction method explained a hierarchical algorithm that is built on the "topdown" approach. It constructs the feature of varying size with n-grams. It extracts the feature of size n and repeats this process until the feature extract till

(n-i)>=2. Therefore, some redundancy features are extracted then used the selection function for solving the redundancy feature. After that, the extracted features will be given to the classifier as an input.

C Long Sequence Based Approach

Yehong Chen [11] discussed the method of feature extraction for protein sequences using the Long sequence and classification of protein sequences by the deep learning neural network for the prediction of the secondary structure of the protein. This method discussed the deep learning architecture using the three layers as i. sparse autoencoder ii. Convolution feature extraction iii. Softmax classifier layer. Firstly, the BLAST tool is used in every protein sequence to extract position specific score matrix. Then, this matrix fed as an input to the auto-encoder to learn the self-taught feature. After that, these features are loaded into the convolution feature extract layer and extract a hybrid representation of protein sequences for long pssm. After that, it generated the representation for long pssm of protein sequences. After that, we generated the comprehensive representation of protein sequence and fed as an input to a Softmax classifier. And Achieve the accuracy for classification of protein is 78%.

D New Online Hierarchical Based Approach

Abdollah Dehzangi, [3] discussed the method for extracting the feature of the structure and evolution for protein fold recognition using the local feature segmentation. In this paper, protein sequences are divided into small-small segments and then extract the distribution and autocovariance of the features from each segment.. The global feature means a sequence-based occurrence method to extract structure and evolutionary features for protein fold recognition. This method extracts the semi-occurrence feature gather straightforwardly from PSSM for the summation of the substitution likelihood of amino acid and SPINE-M for normalized the likelihood of the auxiliary structure element. Complexity is increases with increasing the number of folds and therefore more discriminatory information is required.

E Dual Similarity Based Approach

Neha Bharilla, [6] discussed the method for feature extraction of the local and global features of protein sequences. In this paper, extract features with the help of six related group for each protein sequence by measuring both the local and global similarities. The global similarity measure is determined with the feasibility of particular amino acid corresponding to each sequences with a different position in the superfamily. The local similarity measures by evaluating the weight according to the specific position of each amino acid and then assign the weight of amino acid to the respective with six replacement group. These six replacement groups corresponding to the classes of 20 amino acids such as Aliphatic, Cyclic, Aromatic, Basic, Acidic and their amides, Hydroxyl or Sulfur selenium-containing.

F Protein Interaction Based Approach

Zhehuan Zaho [4] discussed the feature extraction method using the protein-protein interaction with the deep learning network. In this method train the unlabeled data based on the auto-encoders algorithm using greedy layer-wise technique and initialize the deep multilayer neural network with parameters and train this network using the gradient descent algorithm.

Many researchers are solving the method for feature extraction of the protein sequence but they have some issues of the existing method. This paper proposed a method for the feature extraction of the protein sequence.

In this paper, the feature extraction strategy is clarified in section2 which is utilized for the extraction of features of the globin family. Once the features are extracted and after that features are given to the ensemble classifier as an input.

The rest of the paper is organized as follows, in section 3 briefly described the concept of feature extraction of the globin family and section 4 described the method of classification and experimental results are discussed in section 5 and conclusion in section 6.

2 FUNDAMENTAL OF FEATURE EXTRACTION

Feature extraction means that convert protein sequence in to numerical value i.e. property of an sequence. The main task of protein sequence is to encode it into a feature vector and then feature vector given to any machine learning algorithm. It extracts only six permissible features concurring to each protein grouping.

2.1 Calculate the occurrence of particular amino acid in each sequences and total length of each sequence. And then calculate the recurrence of specific amino acid in each arrangement of protein.

$$fij = \frac{(occu)ij}{N}$$

Where, N is the overall number of amino acid in the individual sequence of super family, (fij) is the frequency of particular acid in the ith sequence and jth amino acid and (occurij) is the sum of particular amino acid in the each sequence.

2.2 Encoding the protein sequence

The amino acids with in the sequence share the auxiliary relationship with each other. The encoding of the amino acids is completed concurring to the belongingness of each amino acid to the particular gather and allocates the expansion of frequencies of amino acids to the individual group as described in preliminaries e1 = {G,A,V,L,I} e2 =

{S,C,T,M} e3 = {P} e4= {F,Y,W} e5 = {H,R,K} e6 = {D,E,N,Q}.

3 FEATURE EXTRACTION

Feature extraction strategies are utilized to transform the sequential order into a feature vector which represented the properties of the sequence. The protein sequence can be any length. To begin with, the protein sequences are pre-processed with the assistance of Blast tool and Clustal omega. Blast tool is used to measure the sequence similarity. Blast-p is used for the preprocessing of protein sequence in the alignment. The alignments for minimizing the evolutionary distance or maximizing the similarity between the two sequences are compared with query database. After that, the clustal omega tool is used for the multiple sequence alignment. Then, the sequences are given into feature extraction method. The feature extraction methods extract only six admissible features according to each protein sequence. There are representing the protein sequence is given below in Table1.

Table 1: Protein Sequence

S.No.	Sequence				
1	G	H	K	N	C
2	T	M	Q	P	S
3	V	L	I	Y	W
4	D	E	A	I	P

Protein sequence is represented in Table1. For example G occur 1 time in sequence 1 and total length of sequence is 5 then frequency of G is 1/5. Thus in the same way, the frequency is calculate for other amino acid. Now, the feature vector is calculated according to the six group and put the value into individual group by the addition of frequency of amino acid to the individual group.

Table 2: Encoding Of Each Protein Sequence

S.No.	E1	E2	E3	E4	E5	E6
1	0.2	0.2	0	0	0.4	0.2
2	0	0.6	0.2	0	0	0.2
3	0.6	0	0	0.4	0	0
4	0.4	0	0.2	0	0	0.4

These feature vectors are given to the classifiers as an input to determine the performance of algorithm. The classifiers are used for determine the performance and comparing the results with SVM, Naïve bayes, decision tree, Adaboost with naïve bayes, Adaboost with decision tree and Ensemble classifiers.

4 CLASSIFICATION

Classification is the systematic arrangement in groups or categories according to established criteria or classes such as binary and multiclass classification. In multiclass classification classify the thing into two or more classes and binary classification classifies the thing into two classes. There are some classification techniques used for protein sequence to extract some features from the sequences and these features are depending on the functional and structural properties and also classification techniques used for classifying the superfamilies and subfamily. Therefore, given an obscure protein, the task is to classify the protein into known superfamilies. This will help to predict the function and structure of the obscure protein. There are various method for protein sequence classification is described below.

4.1 Method of sequence classification

1. Feature-based classification: This method is used for classify the sequence but before classified the sequence we need convert the sequences into a feature vector and after that used classification methods.
2. Distance-based classification: This method is used for determine the quality of the classification by using the measures similarity between sequences.
3. Model-based classification: This method is used for classify the sequence by using the Hidden markov model(HMM) and other statistical model.

There are various classification technique are available for classify the protein sequence but we have used the ensemble learning. Ensemble learning method to check accuracy of all the classifiers on the given input and then output of a classifier is considered, which have maximum accuracy with the help of voting classifier.

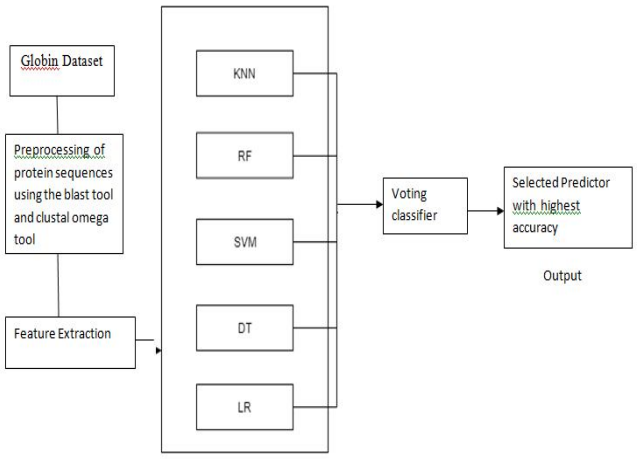

Figure 1. Architecture of protein sequence classification

5 EXPERIMENTAL RESULTS

In this section, tests are performed on the globin dataset which is taken from the uniprot database.

The dataset is divided into two parts, that is training and testing whereas in the 60-40 and 70-30 ratio. The accuracy is measured of each run and then average calculate for all the run. The results are reported accordingly.

5.1 Performance Parameters

We performs the execution of algorithm with these two parameters, that is Accuracy and Computational Time as described below:

A Accuracy:

It is proportional of correct classification from total no. of sequences.

$$\text{Accuracy} = \frac{\text{correctly classified}}{\text{total no. of sequences}} \text{ X100}$$

B Total Computational Time (TCT) :

It defined as the sum of the time required for feature extraction (FET) and classification time (CT).

$$TCT = FET+CT$$

Table- 3: Performance evaluation with different classifier

S.No.	Classifier name	Accuracy		TCT (mins)	
		70-30	60-40	70-30	60-40
1	Naïve Bayes	56.48	54.94	1.8	1.6
2	SVM	61.3	60	26	24
3	Decision tree	55.75	54.94	1.7	1.6
4	Adaboost with naïve bayes	56.25	54.94	6.2	5.583
5	Adaboost with decision	55.3	54.94	1.7	1.63
6	Ensemble	76.25	75.17	40	33

In table3, the result shows of the proposed approach for the globin dataset, we determine the performance of the proposed approach in terms of Accuracy and total computation time. By using the ensemble classifier maximum accuracy is recorded for 60-40 partition is 75.17% and 70-30 partition is 76.25% for the globin family and also maximum total computation time by the ensemble classifier.

Table 4: Performance evaluation with different classifier

S.No.	Classifier name	Accuracy		TCT (mins)	
		70-30	60-40	70-30	60-40
1	Naïve Bayes	56.63	56.35	2	1.55
2	SVM	33.2	32.96	53.5	50
3	Decision tree	58.3	57.51	2.3	2
4	Adaboost with naïve bayes	58.87	58.27	27	26
5	Adaboost with decision	56.25	55.75	2.4	2
6	Ensemble	76.68	76.37	77.6	75

using bootstrapping method

In table4, the result shows of the proposed approach for the globin dataset, we determine the performance of the proposed approach in terms of Accuracy and total computation time. By using the ensemble classifier with bootstrapping sampling method maximum accuracy is recorded for 60-40 partition is 76.37% and 70-30 partition is 76.68% for the globin family and also maximum total computation time by the ensemble classifier.

6. CONCLUSION

In this paper, we presented ensemble learning method to classify the protein sequence. By implementing the method of feature extraction, It extracts the six admissible features for particular protein sequences. These features are extracted then it is given to the ensemble classifier as an input in which SVM, KNN, decision tree, Naïve bayes classifier, random forest after that apply voting classifier are used and 76% accuracy has been achieved. It can be used to improve the drug discovery and identification of diseases for treating patients at an early stage of diagnosis. The proposed approach can be further enhanced using the transfer learning because in this a model that has been trained to perform a specific task is being reused as a starting for another similar task. Transfer

learning is faster due to training time is reduced. So it will take less time for classify the new protein sequence.

REFRENCES

[1] M.J.Iqbal, I.Faye, B.Samir, "**A Distance-Based Feature Encoding Technique for Protein Sequence Classification in Bioinformatics**," in IEEE International Conference on Computational Intelligence and Cybernetics, Yogyakarta, Indonesia, 2013, pp. 1-5.

[2] A.Dehzangi, K. Paliwal, J. Lyons, and A. Sattar, **"A Segmentation-Based Method to Extract Structural and Evolutionary Features for Protein Fold Recognition**," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 11, no. 3, 2014, pp. 510-519.

[3] M.K.F. Mhamdi, "**New online hierarchical feature extraction algorithm for classification of protein**," in 25th International Workshop on Database and Expert Systems Applications, Munich, Germany, 2014, pp. 10-14.

[4] Z.Zhao,Z.Yang,H.Lin,J.Wang,and S.Gao, "**A protein protein interaction extraction approach based on deep neural network**," in International Journal of Data Mining and Bioinformatics, Vol. 15, No. 2, 2016, pp. 145-64.

[5] K. Yan, Y. Xu, X. Fang, C. Zheng, and B. Liu, "**Protein fold recognition based on sparse representation based classification**," in Artificial intelligence in medicine, Vol.79, 2017 pp. 1-8.

[6] N. Bharill and A. Tiwari, "**A Novel Technique of Feature Extraction Based on Local and Global Similarity Measure for Protein Classification**," in International on Bioinformatics Models, Methods and Algorithms, vol. 3, 2015, pp. 219-224.

[7] G. Mansoori, J. Zolghadri, and D. Katebi, "**Protein Superfamily Classification Using Fuzzy Rule-Based Classifier**," in IEEE Transaction on nanobioscience, VOL. 8, NO. 1, 2009, pp. 92-99.

[8] S. Abdulkadhar, G. Murugesan, J. Natarajan*,"**Recurrent Convolution Neural Networks for classification of protein-protein interaction articles from biomedical literature**," in 2017 Third IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2017, pp. 192-197.

[9] A. Bihari, S. Tripathi, "**Protein Classification using Machine Learning and Statistical Techniques: A Comparative Analysis**", January 2019, pp. 1-19.

[10] N. Mehra, A. Tiwari," **A Computational Analysis of Protein Sequences for Cyclophilin Superfamily using Feature Extraction**," in 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 1953-1957 .

[11] Y. Chen, "**Long Sequence feature extraction based on deep learning neural network for protein secondary structure prediction**," in IEEE3rd Information Technology and Mechatronics Engineering Conference, Chongqing, China, 2010, pp. 843847.

[12] G. Mansoori∗, J. Zolghadri, and D. Katebi, " **Protein Superfamily Classification Using Fuzzy Rule-Based Classifier**," in IEEE Transaction on nano-bioscience, VOL. 8, NO. 1, 2009, pp. 92-99.

[13] S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "**Basic local alignment search tool**". J. Mol. Biol. 2015, pp. 403-410.

[14] Sievers, Fabian, et al. "**Fast, scalable generation of high quality protein multiple sequence alignments using Clustal Omega**," Molecular systems biology 7.1 vol.539, 2011, pp.1-6.