Data Mining Approach for Amino Acid Sequence Classification

Dr. Sheshang Degadwala
Associate Professor, Department of Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India
sheshang13@gmail.com

Dhairya Vyas
Managing Director, Shree Drashti Infotech LLP, Vadodara, Gujarat, India
dhairyavyas@live.com

Article History

Article Submission
02 June 2021
Revised Submission
11 September 2021
Article Accepted
10 October 2021
Article Published
31 December 2021

Abstract

Computerized applications are employed all around the world, an enormous amount of data is collected. The essential information contained in large amounts of data is attracting scholars from a variety of disciplines to examine how to extract the hidden knowledge inside them. The technique of obtaining or mining usable and valuable knowledge from enormous amounts of data is known as data mining. Text mining, picture mining, sequential pattern mining, web mining, and so on are all examples of data mining fields. Sequencing mining is one of the most important technologies in this field, as it aids in the discovery of sequential connections in data. Sequence mining is used in a variety of applications, including customers' buying trends analysis, web access trends analysis, atmospheric observation, amino acid sequences, Gene sequencing, and so on. Sequence mining techniques are utilized in protein and DNA analysis for sequence alignment, pattern searching, and pattern categorization. Researchers are exhibiting an interest in the subject of amino acid sequence categorization in the field of amino acid sequence analysis. It has the ability to find recurrent patterns in homologous proteins. This study describes the numerous methods used by numerous studies to categories proteins and gives an overview of the most important sequence classification techniques.

Keywords- Data Mining, Amino Acid Sequence, Protein Family, Distance Feature

I. Introduction

Proteins are essential nutrients for human survival. They are a part of bodily tissue that may also be used as a source of energy. They work as machines that produce all living things, including viruses, bacteria, butterflies, jellyfish, plants, and humans. Around 100 trillion cells make up the human body. There are thousands of distinct proteins in each cell. Each cell performs its function as a result of these factors working together. Inside the cell, proteins are like microscopic machines.

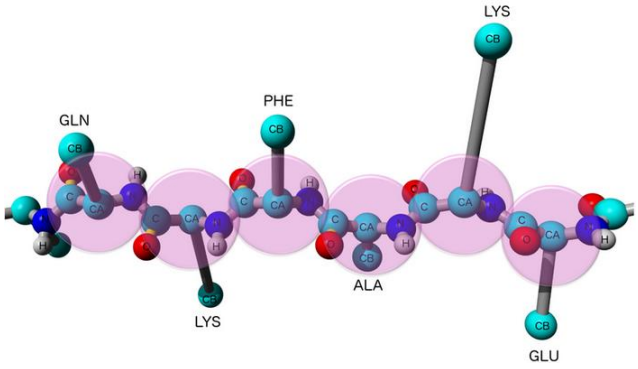

Figure. 1. Primary Structure

ISSN: 2250-0839

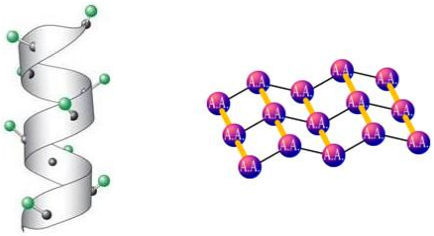

Figure. 2. Secondary Structure

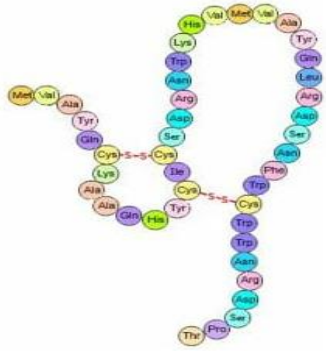

Figure. 3. Tertiary Structure

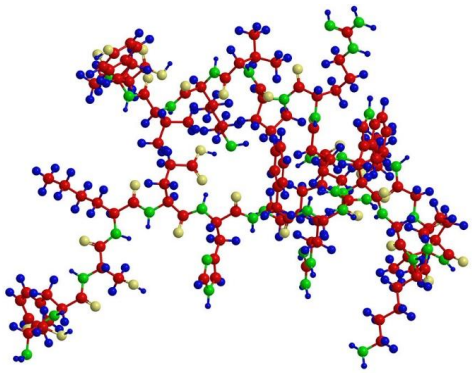

Figure. 4. Quaternary structure

Proteins are composed of tens of lots of smaller components 20 amino acids that are joined together in long chains. The 20 different types of amino acid residues that may be joined to produce a protein are A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. Proteins are also an essential element of an individual's tissues. The four components of amino structure were Primary Structural [Fig.1], Second Structure [Fig.2], Third Structure [Fig.3], and Quaternary Structure [Fig.4].

II. Related Works

Dr. Raghav Yadav and Babasaheb Satpute devised a protein arrangement that figured out the separation of each amino acid residue from the residue in [1]. Take the average of all separations of a certain amino acid residue from the initial residue at that position. Prepare a 600X 20 dataset, which means there are 20 features for each amino acid and 600 total amino acids.

Loris Nannia,n, Sheryl Brahnamc, and Alessandra Luminib suggest a novel machine learning system based on combining several amino acid descriptors extracted from various protein representations, such as position specific scoring matrix (PSSM), amino-acid sequences, and secondary structural sequences, in [4]. An ensemble of support vector machines (SVMs) is used to run the system's prediction engine, with each SVM trained on a distinct descriptor.

ISSN: 2250-0839

The authors of [5] Naveenkumar K S, Mohammed Harun Babu R, Vinayakumar R, and Soman KP primarily rely on earlier Deep learning is being used to classify proteins. To identify the amino acid structure, the properties are extracted using n-gram, Bio-vec, and Prot-v, and then categorised into distinct structures using a deep neural network to find n-dimensional vectors.

Fatima KABLI, Reda Mohamed HAMOU, and Abdelmalek AMINE provide a worldwide framework inspired by the knowledge extraction method from biological data based on association rules in [3]. This framework has three main steps: the first is to extract the descriptors, which was done using the N-Gram technique; the second is to extract the association rules between the proteins components, which was done using the Apriori algorithm; and the third is to choose the relevant rules to classify the unclassified protein. And they compared their classifier to five methods of classification on the WEKA platform, based on the validation tools, on five classes of amino acid extracted from the Uniprot data bank, and got satisfied findings that improved the performance of their protein classifier.

Wenzheng Bao, Yuehui Chen, and Dong Wang proposed a new method to predict the tertiary structure of a amino acid in [6]. The method involves extracting the protein sequence of the amino acid frequencies generalisation dipeptide information hydrophobic combination, using neural networks and flexible neural tree classifier for different the integrated structure classification model, and using neural networks and flexible neural tree classifier for different the integrated structure classification model. To verify the suggested method's efficiency, two benchmark amino acid sequence datasets (640 dataset and 1189 dataset) were chosen as the test data set. The final results suggest that our technique is effective at predicting amino acid structure.

Flexible neutral tree (FNT), a specific tree structure neutral network, has been used as the classification model in the amino acid structures' classification framework by Wenzheng Bao, Dong Wang, and Yuehui Chen in [2]. The impact factors of various feature groups, each of which plays a different role in the model, have been presented in this study. Effect Factors Scaling (IFS) algorithm has been proposed to evaluate distinct impact factors by eliminating redundant information of the selected features to some extent. The 640, 1189, and ASTRAL datasets are used as low-homology amino acid structure benchmark datasets to test the framework's performance.

In [12], Ashish Ghosh and Bijnan Parai attempted to translate the amino acid secondary structure prediction problem to a pattern classification problem and solved it using three low-cost pattern classification algorithms. For window size 11, we employed minimal distance, K-NN, and fuzzy K-NN classifiers, with minimum distance producing the best results. With window size 3, fuzzy K-NN performed better than the other two. We also changed the value of K and found that the range produced better results. Experiments were carried out with various percentages of training sets, and the results were comparable.

III. Comparative Study

TABLE I. Feature Extraction Method

Method	Advantage	Limitation
Distance [1]	-Simple Calculation -Easy to implement -batter accuracy	-Amino acid in pattern is require.
Parameters []	-Numerical output -Easy to store	-Complex to implement
Association	-Significant rules among exists ones	- Enormous number of extracted rules
Position Specific Scoring Matrix	-Easy Calculation	-Less accurate
Multi-scale local descriptor	- Describe overlapping local regions	-Large Feature Dimension

TABLE II. Classification Method

Method	Advantage	Limitation
NB [1]	Train quickly. It is simple to categories. Handles both continuous and discrete data. Streaming data is effectively handled.	Strong feature independence assumption.
NN	NN can perform tasks which linear program cannot. When element of neural network fails it continue to work.	They do not classify and cluster data, a lot of chips and a distributed run-time to train on very large datasets.
SVM	SVM is less complex. Produce very accurate classifiers. Less over fitting, robust to noise.	SVM is binary classifier, to do a multi-class classification, pair-wise classifications can be used Computationally expensive, thus runs slow
Decision Tree	+It reduces overfitting and is therefore more accurate. +Easy to Implement works with all types of data. +Multi classification Support.	-It may not work if the dependent variables considered in the model are linearly related. Therefore, one has to remove correlated variable by some other technique
KNN	-Robust to noisy training data -Effective if the training data is large	- When it comes to distance learning, it's not always apparent the sort of distance to utilize or which characteristic to employ to get the greatest results. -the cost of calculation is relatively high

IV. Proposed Methodology

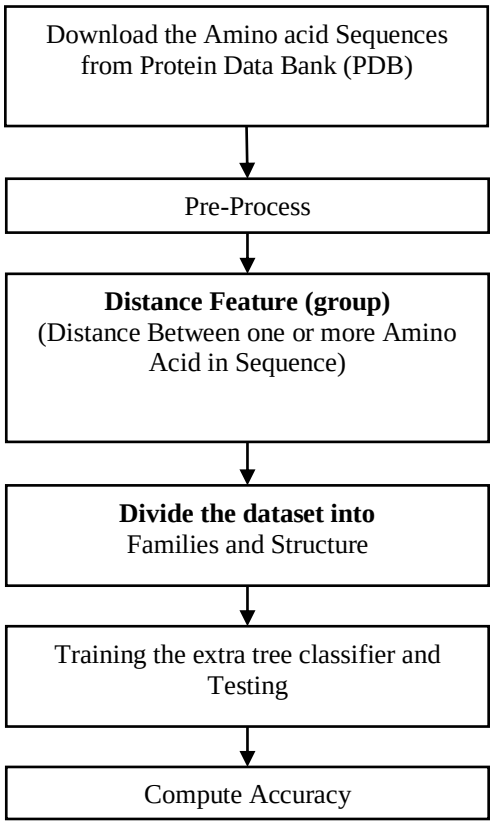

Figure. 5. Proposed Flow

1st Step: data on amino acid sequences

Download the whole FUPF class dataset.

2nd Step: Prepare the data

We'll choose sequences from the table and sample them evenly.

3rd Step: Obtain N-gram

Patterns of 1-1, 2-2, and 3-3 pair amino acid repeat count.

4th Step: Labelling

All FUPF classes are labelled using four classes.

5th Step: Train/Test Data The SVM, KNN ,NB and extra tree classifier are used to train and test data.

6th Step Result

A. Dataset

The sequencing of numerous amino acids, as well as their characteristics, are maintained in databases, according to researchers. Protein Data Bank (PDB), UniProt, Swiss-Prot, SCOP, and other well-known databases are examples. These databases can be used to extract amino acid sequences.

The following is an example of a amino acid sequence for the beta chain of haemoglobin:

"PEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL DNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALA"

B. Modified Distance features Extraction

For each amino acid sequence in each family, characteristics were extracted. Calculate the distance among each amino acid residue and the first residue in each amino acids sequence. Formerly income the middling of all detachments from the first residue for each amino acid residue. For example, if a residue 'K' appears 21 times in a sequence, we will compute 21 distances for 'K', i.e. for each occurrence of 'K', its distance from the first residue will be computed. The feature value for 'K' will then be the mean of those 21 distances. As there are 20 distinct amino acid residues that occur many times in a sequence, we will acquire twenty feature values for each amino acid sequence. Prepare a dataset with a size of 600 X 20 (i.e. 20 characteristics per amino acid and 600 total proteins).

C. Extra Tree Classification

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique that generates a classifier performance by combining the results of numerous contra decision forests collected in a "forest." Apart from how well the decision tree algorithm in the forest were created, it is theoretically similar to a Classification Algorithm.

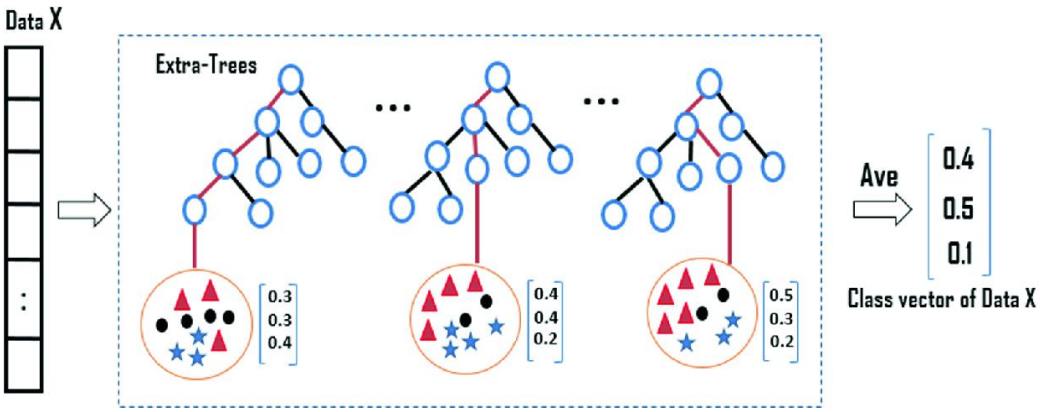

Figure. 6. Extra Tree Classification Step

The Forest of Extra Trees' Choice The original training sample is used to generate all of the trees. The tree would then be provided a representative selection of k characteristics from the features and functionality at every test node, in which it must pick the best feature to divide the data as per a statistical criterion (typically the Gini Index). This random selection of traits is used to generate several power tree structure.

To step up and play feature extraction using this natural vegetation, the standardised total slight drop in the arithmetical criteria used for the judgement of characteristic of divide (Hdi Indicator if the Gini Coefficient is used in the judgement of characteristic of split) is calculated for each characteristic during the building projects of the forest. This number is known as the Gini Significance of the feature. To carry out feature extraction, each feature is rated by Gini Significance in decreasing order, and the client selects the top k variety of options available on their choices.

V. Results and Analysis

	A	B	C	D
1	Entry	Entry nam	Length	Sequence Protein families
2	O95696	BRD1_HUI	1058	MRRKGRCHRGSAARHPSSPCSVKHSPTRETLTYAQAQRMVEIEIEGRLHRISIFDPLEIILEDDLTAQEMSECNSNKENSERPPVCLRTKRHKNNRVKKKNEALPSAHGTPASASALPEPKVRIVEYSPPSAPRRPPVYYKFIE
3	Q9NPI1	BRD7_HUI	651	MGKKHKKHKHSDKHLYEEYVEKPLKLVLKVGGNEVTELSTGSSGHDSSLFEDKNDHDKHKDRKRKKRKKGEKQIPGEEKGRKRRRVKEDKKKRDRDRVENEAEKDLQCHAPVRLDLPPEKPLTSSLAKQEEVEQTPLQEALN
4	O60885	BRD4_HUI	1362	MSAESGPGTRLRNLPVMGDGLETSQMSTTQAQAQPQPANAASTNPPPPETSNPNKPKRQTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTN
5	Q58F21	BRDT_HUI	947	MSLPSRQTAIIVNPPPPEYINTKKNGRLTNQLQYLQKVVLKDLWKHSFSWPFQRPVDAVKLQLPDYYTIIKNPMDLNTIKKRLENKYYAKASECIEDFNTMFSNCYLYNKPGDDIVLMAQALEKLFMQKLSQMPQEEQVV
6	Q15059	BRD3_HUI	726	MSTATTVAPAGIPATPGPVNPPPPEVSNPSKPGRKTNQLQYMQNVVVKTLWKHQFAWPFYQPVDAIKLNLPDYHKIIKNPMDMGTIKKRLENNYYWSASECMQDFNTMFTNCYIYNKPTDDIVLMAQALEKIFLQKV
7	Q9UIF8	BAZ2B_HU	2168	MESGERLF WAL family
8	P25440	BRD2_HUI	801	MLQNVTPHNKLPGEGNAGLLGLGPEAAAPGKRIRKPSLLYEGFESPTMASVPALQLTPANPPPPEVSNPKKPGRVTNQLQYLHKVVMKALWKHQFAWPFRQPVDAVKLGLPDYHKIIKQPMDMGTIKRRLENNYYWA
9	Q9NRL2	BAZ1A_HU	1556	MPLLHRKF WAL family
10	Q9UIF9	BAZ2A_HU	1905	MEMEANE WAL family
11	Q12830	BPTF_HUM	3046	MRGRRGR PBTF family
12	Q9NSI6	BRWD1_H	2320	MAEPSSARRPVPLIESELYFLIARYLSAGPCRRAAQVLVQELEQYQLLPKRLDWEGNEHNRSYEELVLSNKHVAPDHLLQICQRIGPMLDKEIPPSISRVTSLLGAGRQSLLRTAKDCRHTVWKGSAFAALHRGRPPEMPVN
13	P55201	BRPF1_HU	1214	MGVDFDVKTFCHNLRATKPPYECPVETCRKVYKSYSGIEYHLYHYDHDNPPPPQQTPLRKHKKKGRQSRPANKQSPSPSEVSQSPGREVMSYAQAQRMVEVDLHGRVHRISIFDNLDVVSEDEEAPEEAPENGSNKEN
14	Q9UIG0	BAZ1B_HU	1483	MAPLLGRI WAL family, BAZ1B subfamily
15	Q9H0E9	BRD8_HUI	1235	MATGTGKHKLLSTGPTEPWSIREKLCLASSVMRSGDQNWVSVSRAIKPFAEPGRPPDWFSQKHCASQYSELLETTETPKRKRGEKGEVVETVEDVIVRKLTAERVEELKKVIKETQERYRRLKRDAELIQAGHMDSRLDELC
16	Q9H8M2	BRD9_HUI	597	MGKKHKKHKAEWRSSYEDYADKPLEKPLKLVLKVGGSEVTELSGSGHDSSYYDDRSDHERERHKEKKKKKKKKSEKEKHLDDEERRKRKEEKKRKRREREHCDTEGEADDFDPGKKVEVEPPPDRPVRACRTQPAENESTPI
17	Q6RI45	BRWD3_H	1802	MAAAPTQIEAELYYLIARFLQSGPCNKSAQVLVQELEEHQLIPRRLDWEGKEHRRSFEDLVAANAHIPPDYLLKICERIGPLLDKEIPQSVPGVQTLLGVGRQSLLRDAKDCKSTLWNGSAFAALHRGRPPELPVNYVKPPNY
18	Q14140	SRTD2_HU	314	MLGKGGKRKFDEHEDGLEGKIVSPCDGPSKVSYTLQRQTIFNISLMKLYNHRPLTEPSLQKTVLINNMLRRIQEELKQEGSLRPMFTPSSQPTTEPSDSYREAPPAFSHLASPSSHPCDLGSTTPLEACLTPASLLEDDDDTFCT
19	Q9ULD4	BRPF3_HU	1205	MRKPRRKSRQNAEGRRSPSPYSLKCSPTRETLTYAQAQRIVEVDIDGRLHRISIYDPLKIITEDELTAQDITECNSNKENSEQPQFPGKSKKPSSKGKKKESCSKHASGTSFHLPQPSFRMVDSGIQPEAPPLPAAYYRYIEKPP
20	Q9UHV2	SRTD1_HU	236	MLSKGLKRKREEEEEKEPLAVDSWWLDPGHTAVAQAPPAVASSSLFDLSVLKLHHSLQQSEPDLRHLVLVVNTLRRIQASMAPAAALPPVPSPPAAPSVADNLLASSDAALSASMASLLEDLSHIEGLSQAPQPLADEGP
21	Q9BZH6	WDR11_H	1224	MLPYTVNFKVSARTLTGALNAHNKAAVDWGWQGLIAYGCHSLVVVIDSITAQTLQVLEKHKADVVKVKWARENYHHNIGSPYCLRLASADVNGKIIVWDVAAGVAQCEIQEHAKPIQDVQWLWNQDASRDLLLAIH
22	Q92831	KAT2B_HU	832	MSEAGGA Acetyltransferase family, GCN5 subfamily
23	P51532	SMCA4_HI	1647	MSTPDPP SNF2/RAD54 helicase family
24	Q92793	CBP_HUM	2442	MAENLLDGPPNPKRAKLSSPGFSANDSTDFGSLFDLENDLPDELIPNGGELGLLNSGNLVPDAASKHKQLSELLRGGSGSSINPGIGNVSASSPVQQGLGGQAQGQPNSANMASLSAMGKSPLSQGDSSAPSLPKQAAS
25	Q92830	KAT2A_HU	837	MAEPSQA Acetyltransferase family, GCN5 subfamily
26	Q86U86	PB1_HUM	1689	MGSKRRRATSPSSSVSGDFDDGHHSVSTPGPSRKRRRLSNLPTVDPIAVCHELYNTIRDYKDEQGRLLCELFIRAPKRRNQPDYYEVVSQPIDLMKIQQKLKMEEYDDVNLLTADFQLLFNNAKSYYKPDSPEYKAACKLWD
27	O15164	TIF1A_HUI	1050	MEVAVEKAVAAAAAAASAAASGGPSAAPSGENEAESRQGPDSERGGEAARLNLLDTCAVCHQNIQSRAPKLLPCLHSFCQRCLPAPQRYLMLPAPMLGSAETPPPVPAPGSPVSGSSPFATQVGVIRCPVCSQECAERH
28	P21675	TAF1_HUM	1872	MGPGCDL TAF1 family

Figure. 7. Dataset

```
Sequence:
MIKTTLLFFATALCEIIGCFLPWLWLKRNASIWLLLPAGISLALFVWLLTLHPAASGRVYAAYGGVYVCTALMWLRVVDGVKLTLYDWTG
ALIALCGMLIIVAGWGRT
Length of Sequence: 108
Count of A: 13
Count of R: 4
Count of N: 1
Count of D: 2
Count of C: 4
Count of E: 1
Count of Q: 0
Count of G: 10
Count of H: 1
Count of I: 8
Count of L: 21
Count of K: 3
Count of M: 3
Count of F: 4
Count of P: 3
Count of S: 3
Count of T: 8
Count of W: 7
Count of Y: 4
Count of V: 8
```

Figure. 8. Amino acid Count

```
Mean of A is: 55.69
Mean of R is: 67.25
Mean of N is: 29.0
Mean of D is: 83.0
Mean of C is: 49.5
Mean of E is: 15.0
Mean of Q is: nan
Mean of G is: 72.0
Mean of H is: 52.0
Mean of I is: 50.12
Mean of L is: 49.38
Mean of K is: 37.33
Mean of M is: 57.33
Mean of F is: 20.5
Mean of P is: 37.33
Mean of S is: 42.67
Mean of T is: 52.62
Mean of W is: 56.43
Mean of Y is: 69.0
Mean of V is: 72.12
Position of Amino Acids:
(10, 12, 30, 38, 43, 54, 55, 61, 62, 71, 91, 94, 103)
```

Figure. 9. Mean feature

```
Confusion Matrix :
[[71  0  5]
 [ 0 38  2]
 [ 8  0 56]]
Accuracy Score : 91.66666666666666
Report:
              precision    recall  f1-score   support

    FUPF0060       0.90      0.93      0.92        76
    FUPF0061       1.00      0.95      0.97        40
    FUPF0102       0.89      0.88      0.88        64

    accuracy                           0.92       180
   macro avg       0.93      0.92      0.92       180
weighted avg       0.92      0.92      0.92       180
```

Figure. 10. UPF family Classification DT

Figure. 11. Comparative Study

Classifiers	Accuracy Proposed Protein Structure Classification
SVM	46.00%
DT	50.00%
NB	20.00%
Extra tree	91.66%

VI. Conclusion

The inference of a protein's solid construction as of its amino acid sequence is known as protein structure prediction. The diverse structures of proteins FUPF0060, FUPF0061, and FUPF00102, as well as their characteristics extraction techniques, were examined in this study. Amino acid characteristics factor scale, association, Apriorist principles, and other factors are used in research. They employ SVM, DT, NB, and Extra Tree classification techniques for classification. The proposed study on distance combination feature with Extra Tree classier yields a classification result of 91.66 percent.

References

[1] S. Bankapur and N. Patil, "Enhanced Protein Structural Class Prediction using Effective Feature Modeling and Ensemble of Classifiers," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2020.2979430.

[2] Siddhant College of Engineering, Institute of Electrical and Electronics Engineers. Bombay Section., and Institute of Electrical and Electronics Engineers, 2018 3rd International Conference for Convergence in Technology (I2CT) : The Gateway Hotel, XION Complex, Wakad Road, Pune, India. Apr 06-08, 2018.

[3] W. Bao, D. Wang, and Y. Chen, "Classification of Protein Structure Classes on Flexible Neutral Tree," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 14, no. 5, pp. 1122–1133, 2017

[4] Hamou, R. M., Kabli, F. and Amine, A. (2017) "New classification system for protein sequences" 2017 First International Conference on Embedded & Distributed Systems (EDiS).

[5] M. R. Harun Babu and N. K. S, "Protein Family Classification using Deep Learning." bioRxiv preprint first posted online Sep. 11, 201

[6] S. Brahnam, L. Nanni, and A. Lumini "Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition," J. Theor. Biol., vol. 360, pp. 109–116, Nov. 2014.

[7] Wang D., "A novel protein structure classification model," no. September, 2015.

[8] A. Charuvaka and H. Rangwala, "Classifying protein sequences using regularized multi-task learning," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 11, no. 6, pp.1087–1098, 2014.

[9] K. M. Shawkat Zamil and J. Rahman, "Prediction of Protein-Protein Interaction from Amino Acid Sequence Using Ensemble Classifier," Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018, pp. 1–4, 2018.

[10] D. Zhang and M. R. Kabuka, "Protein Family Classification with Multi-Layer Graph Convolutional Networks," Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018, pp. 2390–2393, 2019.

[11] I. Wohlers, M. Le Boudic-jamin, and H. Djidjev, "LNBI 8542 - Exact Protein Structure Classification Using the Maximum Contact Map Overlap Metric," pp. 262– 273.

[12] S. Ji et al., "Deep CDpred: Inter-residue distance and contact prediction for improved prediction of protein structure," PLoS One, vol. 14, no. 1, pp. 1–15, 2019.

[13] A. Ghosh and B. Parai, "Protein secondary structure prediction using distance based classifiers," Int. J. Approx. Reason., vol. 47, no. 1, pp. 37–44, 2008.

[14] L. Zhu, S. P. Deng, and D. S. Huang, "A Two-Stage Geometric Method for Pruning Unreliable Links in Protein-Protein Networks," IEEE Trans. Nanobioscience, vol. 14, no. 5, pp. 528–534, 2015.

[15] S. Shatabda, M. A. H. Newton, M. A. Rashid, D. N. Pham, and A. Sattar, "How good are simplified models for protein structure prediction?," Adv. Bioinformatics, vol. 2014, 2014.

[16] D. S. Huang and H. J. Yu, "Normalized feature vectors: A novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 10, no. 2, pp. 457–467, 2013.

[17] K. M. Shawkat Zamil and J. Rahman, "Prediction of Protein-Protein Interaction from Amino Acid Sequence Using Ensemble Classifier," Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018, pp. 1–4, 2018.