

Protein Structure Classification Based on Distance Feature

Dr. Sheshang Degadwala¹, Dhairya Vyas², Harsh S Dave³

¹Associate Professor, Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India ²Managing Director, Shree Drashti Infotech LLP, Vadodara, Gujarat, India ³Intern MBBS, Smt.B.K.Shah Medical Institute &Research Centre, Vadodara, Gujarat, India

ABSTRACT

Article Info Volume 6, Issue 4 Page Number: 263-269 Publication Issue : July-August-2020	In Bioinformatics field Protein Structure Classification is the hugest undertaking. The realized proteins have been requested subject to their level, feature, work, amino destructive and family and superfamily. Protein structure segregated into four sorts: all α protein structure, all β protein structure, $\alpha+\beta$ protein structure, and α/β protein structure. The use of a standard way to deal with perform plan is a very inconvenient and dreary task. The quantity of cutting edge Machine Intelligence enrolling strategies such Support Vector Machine, Random Forest, Artificial Neural Network, Decision Tree and Naïve Bayes Classifier had been proposed in the composition. Our objective right currently is to develop a system that performs better than anything past
Article History Accepted : 25 July 2020 Published : 30 July 2020	 markers for protein structure gathering by thinking about the separation among the distinctive Amino Acid buildups. To take a gander at the display of proposed work particular datasets are used. Keywords : Protein, Structure, Distance, Amino Acid, Sequence, SVM (Support Vector Machine), DT (Decision Tree), ANN (Artificial Neural Network), NB (Naive Bayer).

I. INTRODUCTION

Proteins are fundamental upgrades for the human body. They are one of the structure squares of body tissue and can in like way fill in as a fuel source. They take after machines that make each living thing, whether or not diseases, microorganisms, butterflies, jellyfish, plants or human limit. The human body includes around 100 trillion cells.

Proteins include hundreds or thousands of increasingly humble units called amino acids, which are added to each other in long chains. 20 stand-out

sorts of amino acids can be joined to make a protein, named as A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y.

Amino acids are organized in four sorts of structures, this is known as protein structure. The name of the structure is Primary Structure, Secondary Structure, Tertiary Structure, and Quaternary Structure. There are four degrees of Tertiary Structure: all α protein structure, all β protein structure, $\alpha+\beta$ protein structure, and α/β protein structure.

Copyright: © the author(s), publisher and licensee Technoscience Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited

All proteins limits are dependent upon their structure, which, in this manner, depends upon physical and compound parameters. There are other significant realities on examining these atoms; traditional organic, physical, synthetic, numerical and informatics sciences have been collaborating in another domain known as bioinformatics to allow another level of data about presence affiliation.

II. Related Work

Experts have been done different progression forms at this moment. They secured an unmistakable element extraction framework and different orders to get unrivaled forecast exactness. In Satpute et al. proposed separation highlight, creator download 600 protein grouping which is a gap in three classes, for example, class 1, class 2, and class 3. In each protein arrangement they compute the separation of all amino corrosive from the principal amino corrosive. By then take the normal of all separation of a particular Amino Acid buildup from the main buildup. For each grouping we get 20 such partitions. Subsequently the all out limit of the dataset is 600X20. They utilized credulous Bayes, Decision Tree, Support Vector Machine and Artificial neural system AI calculation for execution measures. Choice Tree gave better outcomes contrasted with others [1]. Wenzheng et al. had used three component extraction systems, one the sythesis of amino acids, the second is structure highlights and the connection coefficient of a polypeptide. Adaptable Neutral Tree is better than the Artificial Neural Network in part of precision and another measure parameters. The separation among various Amino Acid ought to be viewed as later on inspect. In explore ASTRAL, 640, 1189 dataset have been used as request resource [2].

Fatima et al. introduced a worldwide structure roused by the data extraction process from organic data reliant on the affiliation rules. This structure has three principal propels: (1) the pre-handling stage removing the descriptors, utilizing the N-Gram method, (2) separating the affiliation governs between the proteins parts, utilizing the apriori calculation, (3) chose the applicable principles to classified the obscure protein. Moreover, they applied this classifier on five classes of protein, removed from the Uniprot data bank differentiated among a five methodologies for classification in WEKA stag, other order strategy perspective, their classifiers are given better outcomes

In [4] the creator gave another AI framework that relies upon solidifying a couple of protein descriptors isolated from different protein depictions, for example, Position-Specific Scoring Matrix (PSSM), the amino corrosive arrangement, and auxiliary basic groupings. The forecast motor framework is worked by a troupe of help vector machines, where each SVM is set up on a substitute mark. FC699, 1189, 640, 25PDB are utilized as preparing dataset.

There are two types of vector representation. One is n-gram and the second is Keras embedding. These vectors pass in different deep learning layers such as DNN, RNN, CNN, and LSTM. The Deep learning method with Keras embedding has performed much better than n-gram with deep neural networks [5].

III. Proposed Approach

A. Classification Based on Distance Approach

Right now the existing framework is clarified which integrate protein family classification based on distance feature include after that proposed framework is clarified. In the proposed framework we characterize protein structure dependent on ngram feature. So here the principle distinction is the feature extraction system.

Download the protein grouping from UniProtKB to around 600 arrangements. Concentrate highlights from those groupings. Ascertain the separation of all amino corrosive from the principal amino corrosive. By then take the normal of all separation of a particular amino corrosive from the main amino corrosive.

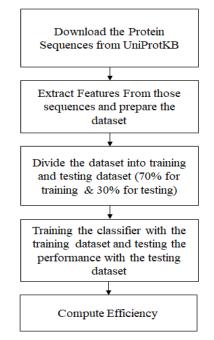


Figure 1. Existing System Architecture

For example, Amino Acid Sequence C D A G H A C A D C A G M D H A M A E A A is repeated 14 times. TABLE 1 shows the distance vector of Amino Acid 'A'. Then take the mean of all those 14 distances.

TABLE I. DISTANCE VECTOR OF 'A'

3	6	8	1	1	1 8	2	2	2	3	3	4	4	4
			1	6	8	0	6	9	1	4	0	5	7

Count of An Amino Acid same as others C, D, G, H, L, M for all sequences and average them. Table 2 shows the average distance of all the occurrence of Amino Acids in 600 sequences. For 600 sequences, we get 20 sizes of distances vector. 600 X 20 is the size of the dataset.

TABLE II. SAMPLE DATA SET

AA Sequence	A	С	D	G	Н	L	Μ
Seq 1	18	22	16	8	10	24	12
Seq 2	9	8	11	14	8	11	13
Seq 3	25	18	11	17	22	14	12
Seq 4	11	7	19	9	12	14	26

The above system is used for protein classification but when it is used for sequence classification it does not give batter performance as we can see that in the next section results and analysis.

B. Classification Based on N-gram Approach

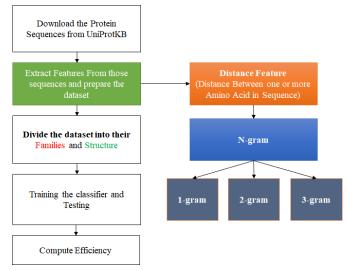


Figure 2. The Flow of the Proposed System

Step: 1 Protein sequence data

Download dataset of all $\alpha,$ all $\beta,$ $\alpha{+}\beta,$ and $\alpha{/}\beta$ classes.

Step: 2 Pre-process data

We will extract sequences from the table and make an equal sample.

Step: 3 Extract N-gram

A feature like 1-1, 2-2, 3-3 pair amino acid repetition pattern count.

Step: 4 Labelling

Labeling using four class all α , all β , $\alpha+\beta$, α/β .

Step: 5 Train/ Test

Data are train and test using K-Nearest Neighbor, Support Vector Machine, Artificial Neural Network, Random Forest this technique.

Step: 6 Result

Structure Type	Amino Sequence	1-1	2-2	3-3	1-gram	2-gram	3-gram
Alpha-All	1-gram ACAMCDAMCACDAMCAC 2-gram ACAMCDAMCACDAMCAC 3-gram	1,3,2,2	8,4	1,3	2	6	2
	ACAMCDAMCACDAMCAC						

TABLE III. N-GRAM FEATURE

Results and Analysis

Experiment on UPF family dataset and protein tertiary structure dataset results are shown in figure 3 to figure 14 and TABLE IV.

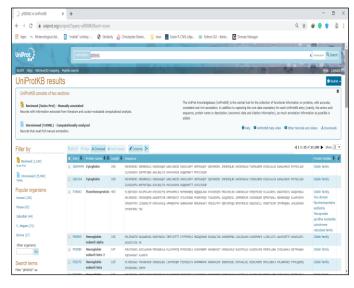
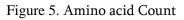


Figure 3. Dataset Downloading

	A	В	С	D	E	F	G	н	1.1	J	K	L	М	N	0	P	Q	R	S	T	U
1	Entry	Entry nam L	ength	Sequenc	e Protein	families															
2	095696	BRD1_HUM	105	MRRKGR	CHRGSAA	RHPSSPCS	VKHSPTRET	LTYAQAQA	RMVEIEIEG	RUHRISIFE	PLEIILEDDI	TAQEMSE	CNSNKENSE	RPPVCLRTK	RHKNNRVI	KKKNEALP	SAHGTPAS	SASALPEPK	WRIVEYSPI	PSAPRRPPV	YYKFI
3	Q9NPI1	BRD7_HU!	65:	MGKKHR	KHKSDKH	LYEEYVEKP	UKLVLKVGO	GNEVTELST	GSSGHDS	SLFEDKND	HDKHKDRK	RKKRKKG	EKQIPGEEKG	RKRRRVKED	KKKRDRDF	RVENEAEK	DLQCHAP	VRLDLPPE	PLTSSLAK	QEEVEQTPU	QEAU
4	060885	BRD4_HU1	1363	MSAESG	PGTRLRNI	PVMGDGL	ETSQMSTTO	DADADAD	PANAASTI	VPPPPETS	NPNKPKRO	πναιαγμ	RVVLKTLWA	HQFAWPF	QPVDAV	KUNUPOYYA	UIKTPMDN	AGTIKKRLE	NNYYWNA	QECIQDENT	MFT
5	Q58F21	BRDT_HU!	943	MSLPSRO	QTAIIVNP	PPEYINTK	KNGRLTNO	LQYLQKV	UKDLWKH	SFSWPFQ	RPVDAVKU	QUPDYYTII	KNPMDLNTI	KRLENKYY	AKASECIED	FNTMFSN	CYLYNKPG	DDIVLMA	QALEKLFM	OKLSOMPO	EEQV
6	Q15059	BRD3_HU1	72	MSTATT	/APAGIPA	TPGPVNP	PPPEVSNPS	KPGRIKTNI	alayman	VVVKTLW	KHQFAWP	FYQPVDAI	KLNLPDYHKI	IKNPMDMG	TIKKRLEN	WYWSASE	CMQDFNT	IMFINCYI	NKPTODIV	UMAQALEK	IFLQ)
7	Q9UIF8	BAZ28_HL	216	MESGER	LF WAL far	nily															
8	P25440	BRD2_HUP	80	MLQNVT	PHNKLPG	EGNAGLLG	LGPEAAAP	GKRIRKPS.	LYEGFESP	TMASVPA	LOLTPANP	PPEVSNP	KKPGRVTNO	LQYLHKVVI	MKALWKH	QFAWPER	QPVDAVK	LGLPDYHKI	IKQPMDM	GTIKRRLENI	WYW
9	Q9NRL2	BAZ1A_HL	155	MPLLHR	GWAL far	nily															
10	Q9UIF9	BAZZA_HL	1903	MEMEAN	KE WAL far	nily															
11	Q12830	BPTF_HUN	304	MRGRRG	R PBTF fa	nily															
12	Q9NSI6	BRWD1_H	232	MAEPSS	ARRPVPU	SELYFUAR	YLSAGPCR	RAAQVLVO	ELEQYOLL	PKRLDWE	GNEHNRSY	EELVLSNK	HVAPDHLLQ	CORIGPMU	DKEIPPSISE	RVTSLLGA	ROSLIRTA	KDCRHTV	WKGSAFA	ALHRGRPPE	MPV
13	P55201	BRPF1_HU	1214	MGVDFD	WKTECHN	LRATKPPY	ECPVETCRK	VYKSYSGIE	YHLYHYDH	IDNPPPPO	QQTPLRKHK	KKGRQSR	PANKQSPSP	SEVSQSPGF	EVMSYAQ	AQRMIVEV	DLHGRVH	RISIFONLD	VVSEDEEA	PEEAPENGS	NKE
14	Q9UIGD	BAZ1B_HL	1483	MAPLLG	RI WAL far	nily, BAZ18	8 subfamily														
15	Q9H0E9	BRD8_HUI	123	MATGTG	KHKLLSTG	PTEPWSIR	EKLCLASSV	MRSGDQN	WVSVSRA	IKPFAEPG	RPPDWFSC	KHCASQY	SELLETTETP	RKRGEKGE	VVETVEDV	IVRKLTAE	RVEELKKVI	KETQERYR	RLKRDAEL	QAGHMDSF	ILDEI
16	Q9H8M2	BRD9_HUT	593	MGKKHR	KHKAEWI	ISSYEDYAD	KPLEKPLKL	VLKVGGSE	VTELSGSG	HDSSYYDO	RSDHERER	HKEKKKOO	KKKSEKEKH	LDDEERRKR	KEEKKRKRE	EREHCOTE	GEADOFDF	GKKVEVER	PPPDRPVR	ACRTOPAEN	IEST
17	Q6RI45	BRWD3_H	1803	мааарт	QIEAELYY	LIARFLOSG	PCNKSAQV	LVQELEEH	QUPRRLD	WEGKEHR	RSFEDLVAJ	NAHIPPO	YLLKICERIGP	LLDKEIPQSV	PGVQTLLO	SVGRQSLU	RDAKDCKS	TLWNGSA	FAALHRGR	PPELPVNYV	KPP
18	Q14140	SRTD2_HU	314	MLGKGG	KRKFDEH	DGLEGKIV	SPCDGPSK	VSYTLORO	TIFNISUMK	UNHRPL	TEPSLOKTV	UNNMURR	IQEELKQEGS	URPMFTPSS	QPTTEPSD	SYREAPPA	FSHLASPS	SHPCDLGS	TTPLEACLT	PASLLEDOO	DTFO
19	Q9ULD4	BRPF3_HU	1203	MRKPRR	KSRQNAE	GRRSPSPYS	SUKCSPTRET	ITYAQAQ	RIVEVDIDG	GRUHRISIM	DPLKIITEDE	LTAQDITE	CNSNKENSE	LPOFPGKS	KPSSKGKK	KESCSKHA	SGTSFHLP	QPSFRMVI	DSGIQPEA	PPLPAAYYR	ЛЕКР
20	Q9UHV2	SRTD1_HU	23	MLSKGU	RKREEEE	KEPLAVDS	WWLDPGH	TAVAQAP	PAVASSSL	FDLSVLKU	HHSLQQSEF	POLRHLVU	/VNTURRIQA	SMAPAAAL	PPVPSPPA	APSVADN	LLASSDAA	LSASMASL	LEDUSHIEG	LSQAPQPLA	DEG
21	Q98ZH6	WDR11_H	122	MUPYTVI	VEKVSART	LTGALNAH	INKAAVDW	IGWQGLIA	YGCHSLVV	VIDSITAC	TLQVLEKH	ADVVKV	WARENYHH	NIGSPYCLR	LASADVNG	SKIIVWDV/	AGVAQC	еіденакрі	IQDVQWLI	VNQDASRD	LLLAI
22	Q92831	KAT28_HL	83	MSEAGG	A Acetylti	ansferase	family, GCI	V5 subfam	ily												
23	P51532	SMCA4_H	164	MSTPDP	PISNF2/R	AD54 helic	ase family														
24	Q92793	CBP_HUM	2443	MAENUU	DGPPNPK	RAKLSSPGR	SANDSTDF	GSLFDLEN	DLPDEUPN	GGELGLU	VSGNLVPD	ASKHKQ	SELLRGGSGS	SINPGIGN	SASSPVQ	QGLGGQA	QGQPNSAI	NMASLSAN	MGKSPLSQ	GOSSAPSUP	(QAJ
25	Q92830	KAT2A_HL	83	MAEPSO	A Acetyltr	ansferase	family, GCI	VS subfam	ily												
26	Q95U85	PB1_HUM	168	MGSKRR	RATSPSSS	VSGDFDDO	HHSVSTPG	PSRKRRL	SNLPTVDP	AVCHELY	NTIROYKOB	QGRLLCEL	FIRAPKRRN	OPDIMEVVS	QPIDLMKI	QQKLKME	EYDDVNU.	TADFQLLFI	NNAKSYYK	POSPEYKAA	CKLV
27	015164	TIF1A_HU	1050	MEVAVE	KAVAAAJ	AASAAAS	GGPSAAPS	GENEAESR	QGPDSERG	GGEAARLN	LLDTCAVC	IQNIQSRA	PKLLPCLHSF	CORCLPAPI	DRYLMLPA	PMLGSAET	PPPVPAP	GSPVSGSS	PFATQVGV	IRCPVCSQE	CAE
28	P21675	TAF1 HUN	187	MGPGCE	L TAF1 fa	nilv															
		Sheet0																			

Figure 4. 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



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								
Posit	ion d	of Ar	nino Ac:	ids:							
(10, 3	12, 3	30, 3	38, 43,	54,	55,	61,	62,	71,	91,	94,	103)

Figure 6. Mean feature

Confusion Matr [[71 0 5] [0 38 2] [8 0 56]] Accuracy Score Report:		66666666	i.	
	precision	recall	f1-score	support
FUPF0060 FUPF0061	0.90 1.00		0.92 0.97	
FUPF0102	0.89	0.88	0.88	64
accuracy macro avg weighted avg	0.93 0.92		0.92 0.92 0.92	180

Figure 7. UPF family Classification DT

Using UPF Family dataset Decision Tree classifier gives confusion matrix accuracy is 91.66%

Confusion Matrix [[55 11 25 3] [21 87 45 5] [18 61 75 14] [5 3 11 10]] Accuracy Score : Report:		287305123		
	ecision	recall	f1-score	support
all alpha	0.56	0.59	0.57	94
all beta	0.54	0.55	0.54	158
alpha+beta	0.48	0.45	0.46	168
alphab-beta	0.31	0.34	0.33	29
accuracy			0.51	449
macro avg	0.47	0.48	0.48	449
weighted avg	0.51	0.51	0.51	449

Figure 8. Protein Structure Classification DT

Using protein structure dataset Decision Tree classifier gives confusion matrix accuracy is 50.55%

[[72 0 4] [0 39 1] [22 6 36]] Accuracy Scor Report:	e : 81.66666	666666666	,	
hepor er	precision	recall	f1-score	support
FUPF0060	0.77	0.95	0.85	76
FUPF0061				
FUPF0102	0.88	0.56	0.69	64
accuracy			0.82	180
macro avg	0.84	0.83	0.82	180
weighted avg	0.83	0.82	0.81	180

Figure 9. UPF family Classification NB

Using UPF Family dataset Naïve Bayes classifier gives confusion matrix accuracy is 81.66%

Confusion Mat	riv .			
[[28 0 19				
[31 0 32	-			
26 0 35	-			
0 0 2	27]]			
Accuracy Scor	e : 20.04454	342984409	7	
Report:				
	precision	recall	f1-score	support
all alpha	0.33	0.30	0.31	94
all beta	0.00	0.00	0.00	158
alpha+beta	0.40	0.21	0.27	168
alphab-beta	0.10	0.93	0.18	29
accuracy			0.20	449
macro avg	0.21	0.36	0.19	449
weighted avg	0.22	0.20	0.18	449

Figure 10. Protein Structure Classification NB

Using protein structure dataset Naïve Bayes classifier gives confusion matrix accuracy is 20.044%

Confusion Mat [[71 0 5] [0 39 1] [5 1 58]] Accuracy Scor		3333333333		
Report:	precision	recall	f1-score	support
FUPF0060 FUPF0061	0.93 0.97	0.93	0.93	76 40
FUPF0102	0.91	0.97	0.91	64
accuracy			0.93	180
macro avg weighted avg	0.94 0.93	0.94 0.93	0.94 0.93	180 180

Figure 11. UPF family Classification SVM

Using UPF family dataset Support Vector Machine classifier gives confusion matrix accuracy is 93.33%

Confusion Mat	rix :			
[[29 20 45	0]			
[16 77 65	0]			
[14 53 101	0]			
[1 7 21	0]]			
Accuracy Scor	e : 46.1024	4988864143		
Report:				
	precision	recall	f1-score	support
all alpha	0.48	0.31	0.38	94
all beta	0.49	0.49	0.49	158
alpha+beta	0.44	0.60	0.51	168
alphab-beta	0.00	0.00	0.00	29
accuracy			0.46	449
macro avg	0.35	0.35	0.34	449
weighted avg	0.44	0.46	0.44	449

Figure 12. Protein Structure Classification SVM

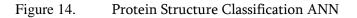
Using protein structure dataset Support Vector Machine classifier gives confusion matrix accuracy is 46.10%

Confusion Matr [[68 0 8] [39 0 1] [14 0 50]]	rix :			
Accuracy Score	. 65 55555	555555556		
Report:				
Report:	precision	recall	f1-score	support
FUPF0060	0.56	0.89	0.69	76
FUPF0061	0.00	0.00	0.00	40
FUPF0102	0.85	0.78	0.81	64
accuracy			0.66	180
macro avg	0.47	0.56	0.50	180
weighted avg	0.54	0.66	0.58	180

Figure 13. UPF family Classification ANN

Using UPF family dataset Artificial Nueral Network classifier gives confusion matrix accuracy is 65.55%

ouc-ouc,				
Confusion Mat	rix :			
[[0 4 89	1]			
[0 6 139	13]			
0 2 158	8]			
0 1 26	2]]			
Accuracy Score	e : 36.97104	677060133		
Report:				
	precision	recall	f1-score	support
all alpha	0.00	0.00	0.00	94
all beta	0.46	0.04	0.07	158
alpha+beta	0.38	0.94	0.54	168
alphab-beta	0.08	0.07	0.08	29
accuracy			0.37	449
macro avg	0.23	0.26	0.17	449
weighted avg	0.31	0.37	0.23	449



Using protein structure dataset Artificial Nueral Network classifier gives confusion matrix accuracy is 36.97%

Classifier	Accuracy	Accuracy	Accuracy
S	base Paper	PUF family	Proposed
	PUF	Classificatio	Protein
	Classificatio	n	Structure
	n [1]		Classificatio
			n
SVM	65.00%	93.00%	46.00%
DT	82.00%	91.00%	50.00%
NB	73.00%	81.00%	20.00%
ANN	69.00%	65.00%	36.00%

TABLE IV. COMPARATIVE STUDY

IV. CONCLUSION

Protein structure forecast is the surmising of the third dimension structure of a protein from its amino acid succession. In this research study about different structures of protein all α , all β , $\alpha+\beta$, and α/β , and their features extraction methods based on amino acid features factor scale, association, rules, etc. For classification, research uses SVM, NB, DT, and ANN classification approaches and analyze features capability of classifying correct protein structure and family class. From the comparative table, it can be said that the count feature gives low accuracy for protein structure classification. So, it can be said that count does not work for Protein Structure classification. So, in the future, if we work on distance combination feature (n-gram) with Decision Tree classier give batter output for classification.

V. REFERENCES

- Siddhant College of Engineering, Institute of Electrical and Electronics Engineers. Bombay Section., and Institute of Electrical and Electronics Engineers, Apr 06-08, 2018.
- [2] D. Wang, W. Bao and Y. Chen, "Arrangement of Protein Structure Classes on Flexible Neutral Tree," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 14, no. 5, pp. 1122–1133, 2017
- Kabli, F., Hamou, R. M., and Amine, A. (2017). New arrangement framework for protein successions. 2017 First International Conference on Embedded and Distributed Systems (EDiS).
- [4] N. K. S and M. R. Harun Babu, "Protein Family Classification utilizing Deep Learning." bioRxiv preprint first posted online Sep. 11, 201
- [5] S. Brahnam, L. Nanni, and A. Lumini, "Forecast of protein structure classes by joining diverse protein descriptors into general Chou's pseudo amino corrosive sythesis," J. Theor. Biol., vol. 360, pp. 109–116, Nov. 2014.

- [6] D. Wang, "An epic protein structure grouping model," no. September, 2015.
- [7] H. Rangwala, and A. Charuvaka "Ordering protein arrangements utilizing regularized perform various tasks learning," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 11, no. 6, pp.1087–1098, 2014.
- [8] J. Rahman, and K. M. Shawkat Zamil, "Expectation 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.
- [9] M. R. Kabuka and D. Zhang, "Protein Family Classification with Multi-Layer Graph Convolutional Networks," Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018, pp. 2390–2393, 2019.
- [10] 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.
- [11] S. Ji et al., "Profound CDpred: Inter-buildup separation and contact forecast for improved expectation of protein structure," PLoS One, vol. 14, no. 1, pp. 1–15, 2019.
- [12] B. Parai and A. Ghosh, "Protein auxiliary structure forecast utilizing separation based classifiers," Int. J. Approx. Reason., vol. 47, no. 1, pp. 37–44, 2008.
- [13] S. P. Deng, D. S. Huang, and L. Zhu, "A Two-Stage Geometric Method for Pruning Unreliable Links in Protein-Protein Networks," IEEE Trans. Nanobioscience, vol. 14, no. 5, pp. 528–534, 2015.
- [14] S. Shatabda, A. H. Newton, D. N. Pham, M. A. Rashid, and A. Sattar, "How great are disentangled models for protein structure forecast?," Adv. Bioinformatics, vol. 2014, 2014.

- [15] D. S. Huang and H. J. Yu, "Standardized element vectors: A tale arrangement free grouping correlation technique dependent on the quantities of neighboring amino acids," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 10, no. 2, pp. 457–467, 2013.
- [16] J. Rahman, and K. M. Shawkat Zamil, "Expectation 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.
- [17] https://www.uniprot.org

Cite this article as :

Dr. Sheshang Degadwala, Dhairya Vyas, Harsh S Dave "Protein Structure Classification Based on Distance Feature", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 263-269, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206464 Journal URL : http://ijsrcseit.com/CSEIT206464