International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206464

Protein Structure Classification Based on Distance Feature

Dr. Sheshang Degadwala[1], Dhairya Vyas[2], Harsh S Dave[3]
[1]Associate Professor, Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India
[2]Managing Director, Shree Drashti Infotech LLP, Vadodara, Gujarat, India
[3]Intern MBBS, Smt.B.K.Shah Medical Institute &Research Centre, Vadodara, Gujarat, India

Article Info
Volume 6, Issue 4
Page Number: 263-269
Publication Issue :
July-August-2020

Article History
Accepted : 25 July 2020
Published : 30 July 2020

ABSTRACT

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, α+β 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 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, α+β protein structure, and α/β protein structure.

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 n-gram 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.

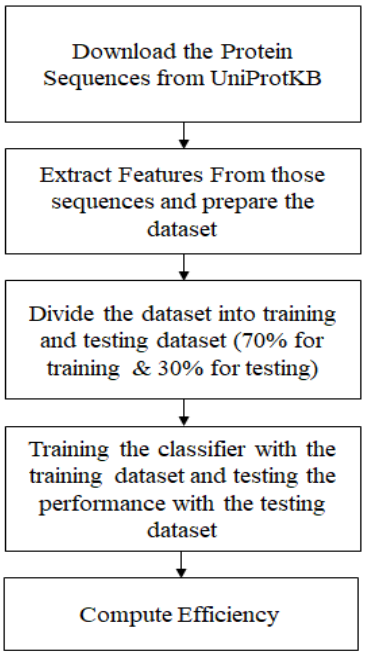


Figure 1. Existing System Architecture

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	11	16	18	20	26	29	31	34	40	45	47

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	C	D	G	H	L	M
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

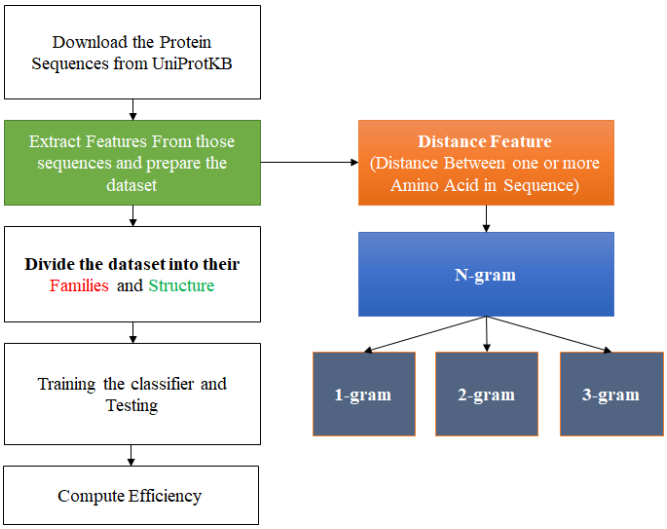


Figure 2. The Flow of the Proposed System

Step: 1 Protein sequence data

Download dataset of all α, all β, α+β, and α/β 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, α+β, 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

TABLE III. N-GRAM FEATURE

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 ACAMCDAMCACDAMCAC	1,3,2,2	8,4	1,3	2	6	2

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.

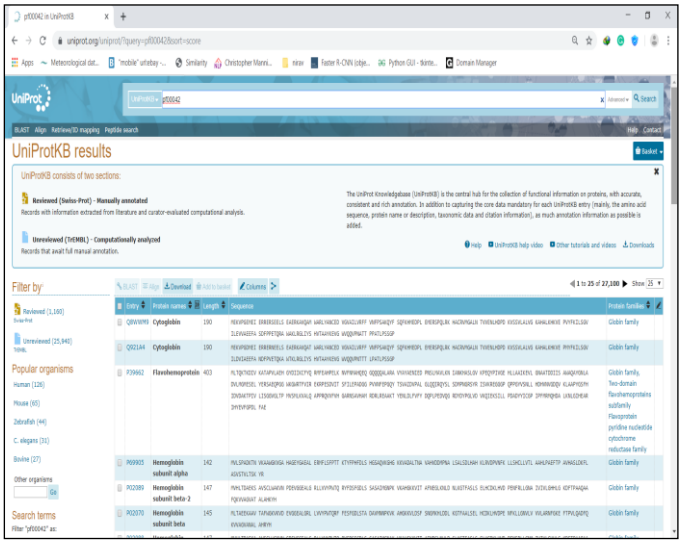


Figure 3. Dataset Downloading

Figure 4. Dataset

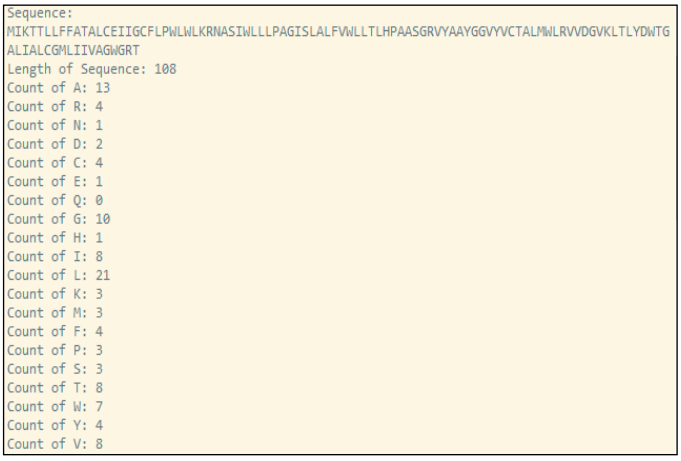


Figure 5. Amino acid Count

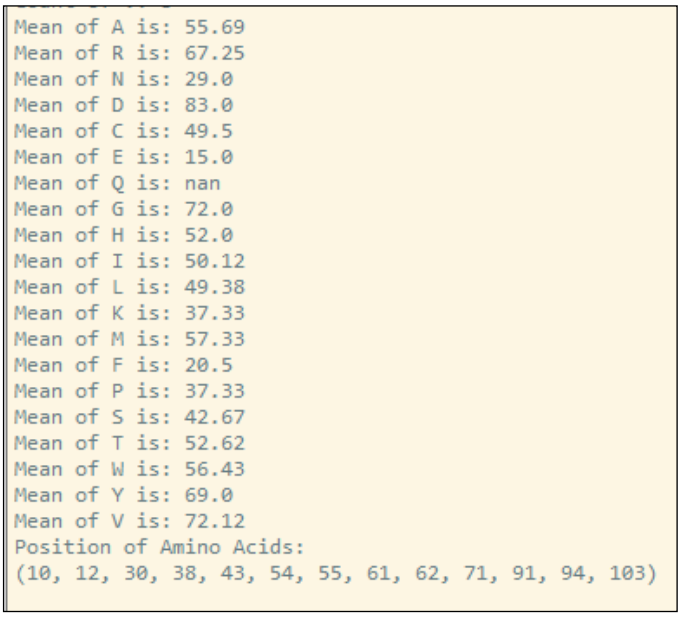


Figure 6. Mean feature

```
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 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 : 50.55679287305123
Report:
              precision    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 Score : 81.66666666666667
Report:
              precision    recall  f1-score   support

    FUPF0060       0.77      0.95      0.85        76
    FUPF0061       0.87      0.97      0.92        40
    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 Matrix :
[[ 28   0  19  47]
 [ 31   0  32  95]
 [ 26   0  35 107]
 [  0   0   2  27]]
Accuracy Score : 20.044543429844097
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 Matrix :
[[71  0  5]
 [ 0 39  1]
 [ 5  1 58]]
Accuracy Score : 93.33333333333333
Report:
              precision    recall  f1-score   support

    FUPF0060       0.93      0.93      0.93        76
    FUPF0061       0.97      0.97      0.97        40
    FUPF0102       0.91      0.91      0.91        64

    accuracy                           0.93       180
   macro avg       0.94      0.94      0.94       180
weighted avg       0.93      0.93      0.93       180
```

Figure 11. UPF family Classification SVM

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

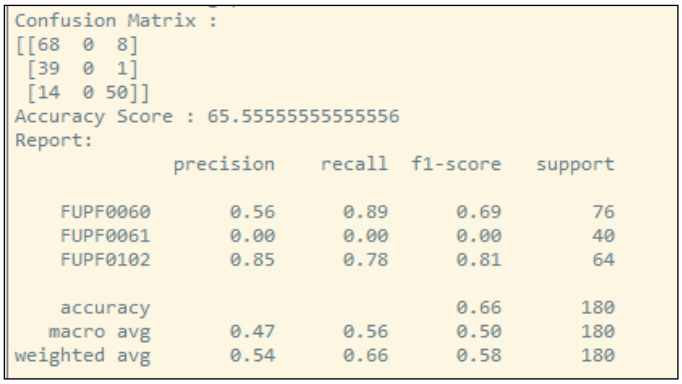
```
Confusion Matrix :
[[ 29  20  45   0]
 [ 16  77  65   0]
 [ 14  53 101   0]
 [  1   7  21   0]]
Accuracy Score : 46.10244988864143
Report:
              precision    recall  f1-score   support

   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%



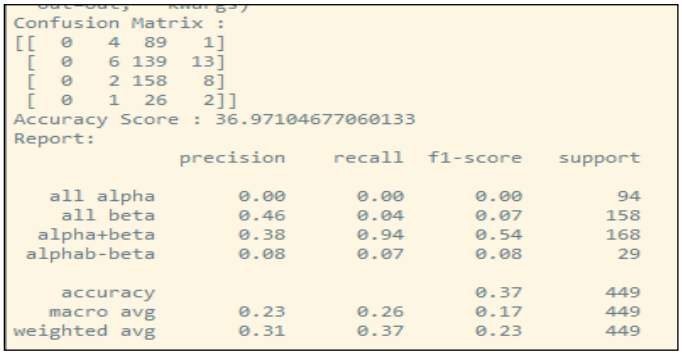
```
Confusion Matrix :
[[68  0  8]
 [39  0  1]
 [14  0 50]]
Accuracy Score : 65.55555555555556
Report:
              precision    recall  f1-score   support

    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%



```
Confusion Matrix :
[[  0   4  89   1]
 [  0   6 139  13]
 [  0   2 158   8]
 [  0   1  26   2]]
Accuracy Score : 36.97104677060133
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
```

Figure 14. Protein Structure Classification ANN

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

TABLE IV. COMPARATIVE STUDY

Classifiers	Accuracy base Paper PUF Classification [1]	Accuracy PUF family Classification	Accuracy Proposed Protein Structure Classification
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%

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, α+β, 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

[1] 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

[3] 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