Journal of Engineering Science and Technology
Vol. 12, No. 3 (2017) 820 - 833

CLUSTERING ON STRUCTURED PROTEINS WITH FILTERING INSTANCES ON BIOWEKA

VIGNESH U.*, PARVATHI R.

School of Computing Science and Engineering, VIT University, Chennai Campus,
Vandalur – Kelambakkam Road, Chennai – 600 127, India
*Corresponding Author: vignesh.u2014phd1182@vit.ac.in

Abstract

This paper presents a synthesis on the analysis of structured proteins through data mining techniques. The protein structure is predicted by using target sequences from BLAST. The predicted protein structure of macropus rufus and structured proteolyzed lysozyme protein are interacted together using mergesets filter for clustering. The selection of a clustering technique for protein clustering is a problem. In order to estimate, performance on three clustering algorithms that are SimpleKMeans, ExpectationMaximization and MakeDensityBasedClusterer are analysed. Carried the simulation experiment on open source data mining tool Bioweka. Comparative analysis with various clustering algorithms illustrates the efficiency of the better clustering algorithm. This method can be applied for the large class of applications such as drug discovery, protein-protein analysis, docking, etc.

Keywords: Clustering, Protein structure, BLAST, Mergesets.

1. Introduction

The process of extracting unknown information from considered large database is said to be data mining. The extracted data has meaningful information. The research in this area has various advantages to the society, which led to accurate and efficient techniques for the process of mining useful patterns from a given database. For example, consider our government sectors where the database is maintained with details of each and every citizen in the country. To retrieve an information from that database, a KDD process is needed, i.e., extracting potential patterns in considered data from the database. Proteins are taken as input; it is a collection of molecules that used to form the mass of living beings.

Nomenclatures

D	Defines the distance between data points
d(M,Y)	Squared error distortion
d(m,y)	Euclidean distance
m	Given data point
Y	Set of K clusters

Abbreviations

ARFF	Attribute Relation File Format
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOck SUbstitution Matrix
CST	Correlation Search Technique
DNA	Deoxyribo Nucleic Acid
FPTAS	Flight Path Threat Analysis Simulation
NCBI	National Center for Biotechnology Information
PDB	Protein Data Bank
PPI	Protein-Protein Interaction
SOPMA	Self-Optimized Prediction Method with Alignment
TLBO	Teaching Learning Based Optimization

The information extracted from proteins that are taken as input. The extracted information is useful to apply in the aspect of drug discovery especially. Data mining techniques play a major role to retrieve the useful information. Allowing a technician to easily search and retrieve the information from the database is known as retrieval. Clustering technique is applied on two structured proteins to retrieve the useful information.

Bioinformatics is the biological data, where the data mining techniques and machine learning algorithms are applied on those data to retrieve the information. The biological data consists of DNA and proteins, which differs from each other in their own individual characteristics. DNA is made up of genes, nucleotides, whole together known as the genome. Proteins have its own structure defined with the amino acids. The unstructured proteins are not powerful for the data mining techniques to perform clustering and hence converting the unstructured proteins to structured proteins by using the modeller 9.15 software. Whereas, the other methods to convert these proteins into structured one doesn't satisfy the time complexity aspects. Since these formats to change are very difficult and to identify the correct predictions are challenging.

Figure 1 shows the conversion of unstructured proteins to structured proteins takes place in a modeller software. Then, the input crab dataset is taken from NCBI public database, i.e., unstructured proteins. It is given as input in a BLAST algorithm. BLAST used to compare a given input sequence against a database. It can also be done in FASTA, but still it has disadvantages in their efficiency of matching the protein structure in time. BLAST uses Altschal-Dembo-Kerlin statistical methods for calculating the statistical significances of identified matches. Based on the identity of BLAST, target sequences are selected. These unstructured target sequences are noted down for identification in a public database PDB for identifying the structured protein sequences. The structured

protein sequences of target sequences are identified and based on it query sequence structure predicted on modeller 9.15.

In this paper, apart from modeller 9.15, with the predicted structure, FARFF concept are applied on it.

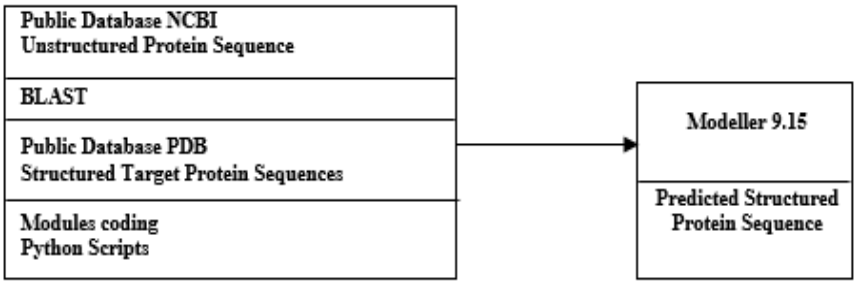


Fig. 1. Predicting structured protein sequence.

Figure 2 shows the clear view of the concept of FARRF, i.e., FASTA to ARFF file conversion, which is done by the technique of ordering process. With this processes, the resulting ARFF file is given as input sequence1 into bioweka software. Taking another structured protein sequence and applying the same processes results in input sequence2. Clustering is done on these two structured input sequences, which results in a group of clusters on two protein sequences.

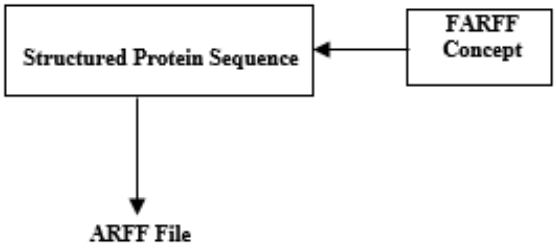


Fig. 2. Format conversion of protein sequence.

2. Related Work

Protein clustering is involved in the aspects of drug discovery, biomedical applications. Song et al. [1] used the parameterized BLOSUM matrices for their molecular biological research on protein alignment. They have used a simulation calculation on BLOSUM matrices and find linear relationships between values of matrices to prove their effectiveness. Protein secondary structure prediction was done by using the deep learning network approach and the resultant known as Ab initio protein secondary structure for further prediction. To identify the functions Spencer et al. [2] uses the scoring matrix generated by PSI-BLAST. This method was failed to improve the mentioned field as suggested by its alluring complexity. One-Prototype-Take-One-Cluster competitive learning paradigm was mentioned in Wu et al. [3]. They simulated gene expression data for cluster analysis, which can require more time to complete the process.

Gonzalez-Alvarez et al. [4] proposed hybrid multiobjective teaching learning based optimization algorithm for finding patterns in a protein sequences. Here, they combined the TLBO algorithm with a local search tool to predict the common patterns. This method proposed their concept based on inspiration by the relationship between student and teacher, taken as instances but not solved for the complex instances. The longest common segments in the protein sequences are found by using FPTAS, which finds a root mean squared deviation for calculation as mentioned in Ng et al. [5]. The structure is implemented by using C++ language. It is done instead of local/global alignment. They calculated LCS score to find the similarities between two 3D structures. It does not define the stable structures in protein complexes automatically.

Birlutiu et al. [6] researches protein-protein interactions prediction by a Bayesian framework. Which combines network topology information and protein. Here, they proposed an unsupervised learning method. They have set hyperparameters roughly match PPI network characteristics and not framed the empirical Bayesian framework. The powerful heuristic method of calculating distance using pearson correlation measure for accelerating K-means clustering in life sciences data are mentioned in Ichikawa et al. [7] known as boostKCP software. It is not proven for other clustering methods. Tseng et al. [8] proposed a correlation search technique (CST) for extracting the data from gene expression efficiently. It needs large memory requirements to store the similarity of the structures they identified. Their identity view also requires large memory allotments compared to other search algorithms in data mining techniques. It is not applicable for all kinds of real microarray datasets to estimate the validity of clustering methods.

Roy et al. [9] proposed a genetic algorithm for clustering, which paves the way for efficient result only for numerical and categorical data and not for highly sensitive large biological applications. Dong et al. [10] describes the overview of data mining concept with relation to the biological sequence. Fayyad [11] in detail explains how to extract the useful biological information from database and make sense of the extracted data. Guralnik et al. [12] proposed a scalable algorithm that deals with the less complexity measures compares to the other sequential clustering algorithms. Zhou et al. [13] proposed a mMBioPM algorithm for pattern mining with the advantage of high efficiency by using hashing technique in it. Pipenbacher et al. [14] explains the graph based approach for clustering protein sequences only which has their functions identified. Yang et al. [15] deals with a new CLUSEQ model for clustering the sequence based on the statistical analysis of matching measure and behaviour of the given biological sequence.

3. Proposed Work

The interaction between protein and protein are said to be docking, where the combination of amino acids complexes matches identified and their ancestors to which they match also identified known as homology.

3.1. System architecture

The unstructured protein sequence of crab taken and converted into a structured protein sequence using modeller9.15 by identifying a target protein sequence

from PDB database. These structured protein sequence of crab are in the format of pdb. The target sequences referred pdb id are, viz. 2KLR, 3L1G, 2WJ7, 2Y1Z. Then, the resultant protein structure can be viewed in the 3dimensional model by using the software rasmol2.7.5. Then, use the FARFF concept for format conversion in protein. At first, the pdb structure is converted into an FASTA format using online pdb to FASTA converter. With this FASTA file, the secondary structure predicted. It is done by using the SOPMA tool, which results in FASTA file with secondary structure prediction. Then, this FARFF concept works for conversion to ARFF file. Thus, Fig. 3 shows the conversion occurs in the protein sequence format. Thus, this ARFF file is named as input sequence1. Another structured protein sequence from pdb database downloaded and converted into ARFF file using the same method, the resultant known as input sequence2. These two input sequences are pre-processed in bioweka software and chosen the filter named mergesets from the root of bioweka. Applying mergesets filter on this two protein datasets, single dataset results with the class relation_name. Then, on this interacted dataset, do cluster analysis by using the different clustering algorithm and results in a group of clusters based on the algorithm as shown in Fig. 4. The detailed description of the system architecture with the images of python scripts execution are presented in Appendix (A).

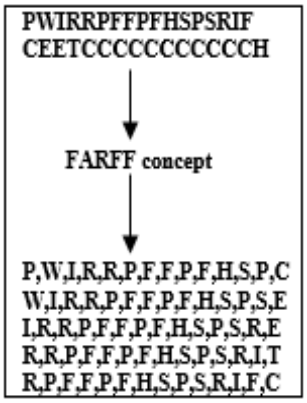


Fig. 3. FARFF converter.

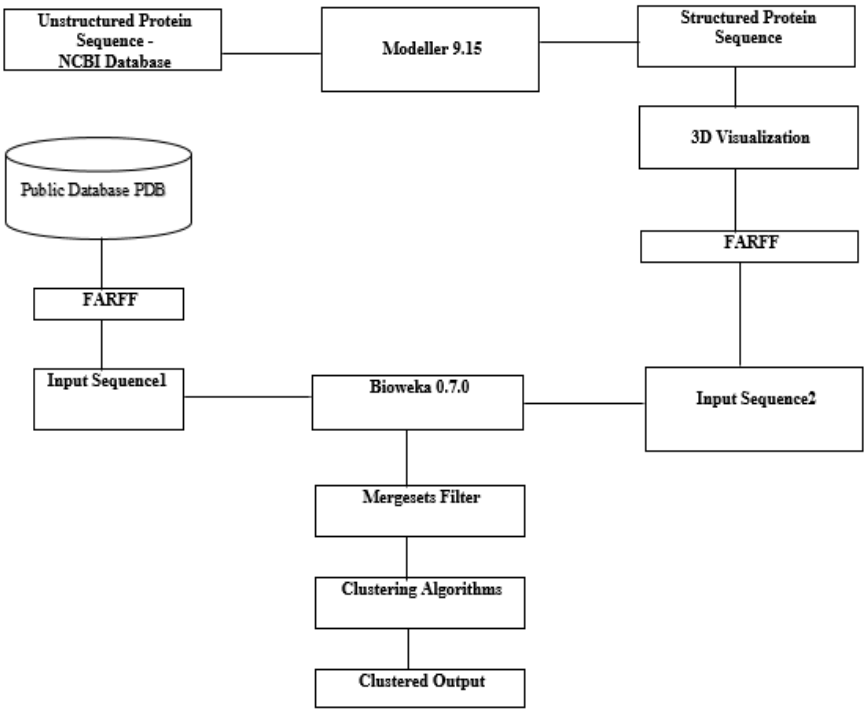


Fig. 4. System architecture.

3.2. K-means clustering

In this algorithm, give a set of J data points in L dimensional space and an integer K. the objective is to evaluate the set of K-points that reduce squared error distortion as in Eq. (2) shown below. The distance between the data points calculated by using the Euclidean distances. The time complexity of k-means clustering algorithm takes O (lknm).

$$d(m, y) = \min_{1<i<k} d(m, y_i) \tag{1}$$

$$d(M, Y) = \frac{\sum d(m_i, Y)^2}{J} \tag{2}$$

where $M = \{m_1, \ldots m_J\}$ are the set of K centres, $Y = \{y_1, \ldots y_k\}$ are mean squared distance from data points to center, l = iteration it performs, k = number of clusters, n = number of objects, and m = dimensional vectors.

It is very difficult here to compare the quality of clusters produced. It also proves its inability in predicting K value, when fixed number of clusters are already assigned. It produces very strong tighter cluster than hierarchical clustering. Obviously, it does not prove its work in non-globular clusters. In K-means clustering, we follow these two steps, viz.

1- Assignment, 2- Update

This algorithm works on above criteria, as pick the seeds, reassigning the clusters, computing the centroids again vice versa until the convergence occurs, i.e., the state where the clusters do not change.

3.3. EM clustering

EM stands for Expectation and Maximization. Real-world datasets in EM algorithm gives an extremely useful result. It is highly recommended when K-means algorithm result does not satisfy the dataset evaluation and also for clustering a very small kind of scene or a specific region of interest. It is highly complex in nature to find the probability and log likelihood statistical calculations. Equations (3) and (4) show the clear view of the likelihood function evaluation to find the best model for data generation. In log likelihood function data is a constant one and discovers the best model over probability. It is an iteration method for finding log likelihood estimates in discrete model. Here, the class labels are sorted in order with respective to the probability. Equations (5) and (6) shows, how the expectation and maximization are calculated by maximizing F with respect to theta (observed data) and vice versa. Here, the expectation step is the assignment and the maximization step is the update of centers as compared with the K-means algorithm. Expectation maximization algorithm takes the time complexity of O (nkl).

$$P\left(Model\middle|Data\right)=\frac{P\left(Data\middle|Model\right)P(Model)}{P(Data)} \tag{3}$$

$$l(\theta)=\log_2 p(D|\theta)=\log_2\sum_H p(D,H|\theta) \tag{4}$$

$$Q^{k+1}=\arg\max_Q F(Q^k,\theta^k) \tag{5}$$

$$\theta^{k+1}=\arg\max_\theta F(Q^{k+1},\theta^k) \tag{6}$$

where $D = \{x(1),x(2),\ldots. x(n)\}$ are the set of n observations, $H = \{z(1),z(2),\ldots. z(n)\}$ with $z(i)$ corresponds to $x(i)$, and z = discrete value.

3.4. Density based clustering

In this algorithm, the number of clusters decided based on the density distribution function of the respective nodes. It has the ability to identify the noise in the database if the dataset does not undergo pre-processing analysis in an efficient manner. Here, mentioning the number of clusters are not necessary. It fails, when chosen for datasets with higher differences in their densities. This density based algorithm also uses the Euclidean distance for calculating its distance matrix. The most appreciated function is that the algorithm can be able to find the clusters although it is surrounded by different clusters. Consider a graph with the nodes assigned as a point, which has to be clustered. For each considered point T, create an edge from T to every point S in the neighbourhood of T. Set P to the assigned nodes in the graph. Pick a set point T in P. Let Y be the set of nodes that can be reached from T by moving forward through following steps, viz.

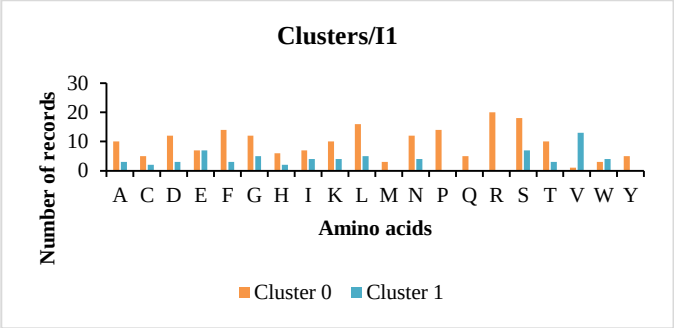
- Step 1: Create a cluster that contains $Y \cup \{T\}$
- $P = P \Big/ \left(Y \cup \{T\}\right)$ (7)

Repeat the process until the convergence occurs. Density based clustering algorithm also calculates log likelihood function to estimate their discrete model in proper manner. Time complexity of density based clustering algorithm takes O (n^2). Complexity for each point has to be determined if it is a core point, it can be reduced to O (n log (n)) in lower dimensional spaces by using efficient data structures (n is the number of objects to be clustered).

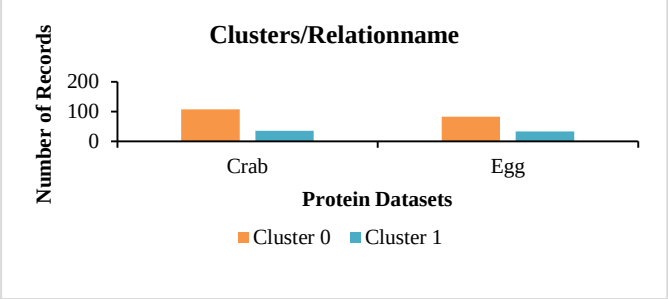
4. Results and Discussion

In this section, unstructured protein sequence selected from NCBI protein database. The length of the protein sequence in the database are nearly same, since they are homologous. This unstructured protein sequence is converted into a structured sequence and taken as input sequence1 in ARFF format. Another protein structure from PDB database taken as input sequence 2. These 2 sequences are merged and clustered by using the three clustering algorithms, viz. MakeDensityBasedClusterer, ExpectationMaximization and SimpleKMeans to find the best similarity matches in combined protein dataset. This data mining technique is performed in bioweka software.

The visualization aspect gives the comparison of clustered results with respect to their performance analysed. Figures 5, 6 and 7 shows the clustering results of K-Means, Expectation Maximization and MakeDensityBasedClusterer algorithm with respect to attributes struct, relationname and I1. Since the results of K-Means algorithm not satisfied for the protein datasets, process undergo the Expectation Maximization algorithm, this algorithm clusters the result based on two different datasets and calculates log likelihood function. MakeDensityBasedClusterer algorithm gives the efficient results with respect to log likelihood function, time, etc. When compared the result with the two existing algorithms previously analysed, ExpectationMaximization is proved to be a better clustering algorithm. The merged protein dataset in bioweka shown in Fig. 8 then the results are given as input to the Tableau software. The input protein sequence images are presented in Appendix (*B*).



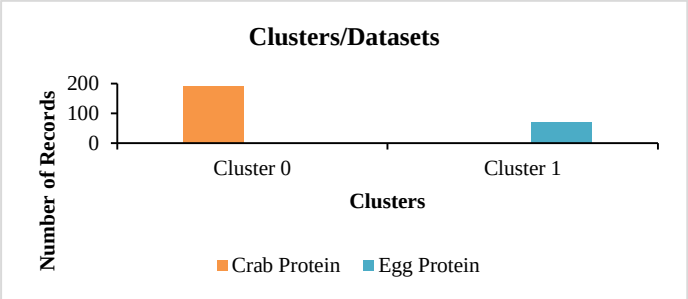
(a) Attribute I1 clusters on merged dataset.



(b) Merged dataset clusters.

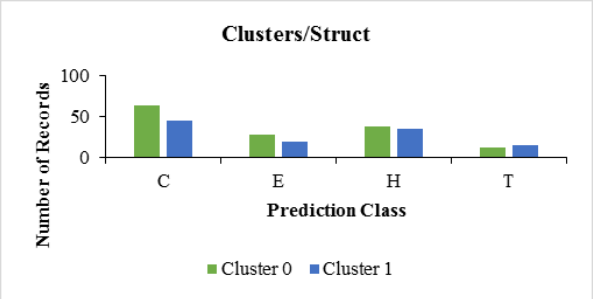
Fig. 5. Simple K-means clustering.

(a) Secondary structure predicted clusters.

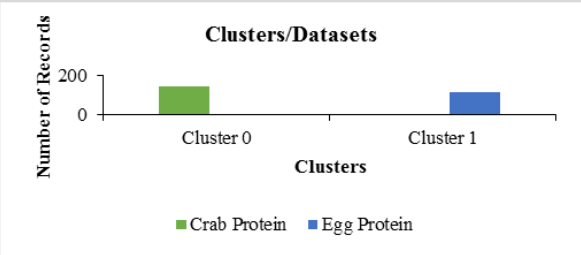


(b) Data clustered based on datasets.

Fig. 6. Density based clustering.



(a) Secondary structure predicted clusters.



(b) Data clustered based on datasets.

Fig. 7. Expectation maximization clustering.

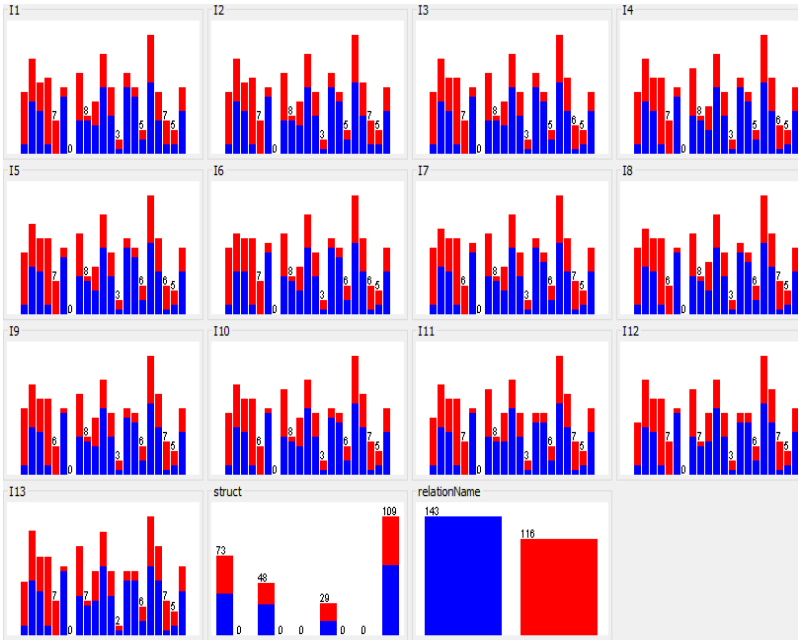


Fig. 8. Visualisation of merged dataset.

The results of visualization are compared with literature survey for selecting the better clustering algorithm for protein clustering. The comparison of these clustering algorithm on the basis of performance analysis are mentioned in Table 1 and the comparisons of clustering algorithms mentioned in Fig. 9 gives the result based on time.

Table 1. Clustering algorithm comparisons.

Clustering Algorithms	Log Likelihood / Sum of Squared Errors	Time (Secs)	Number of Data's Clustered	
			Cluster 0	Cluster 1
MakeDensityBasedClusterer	-39.12727	0.01	189	70
ExpectationMaximization	-38.21832	0.02	143	116
SimpleKMeans	3183	0.01	190	69

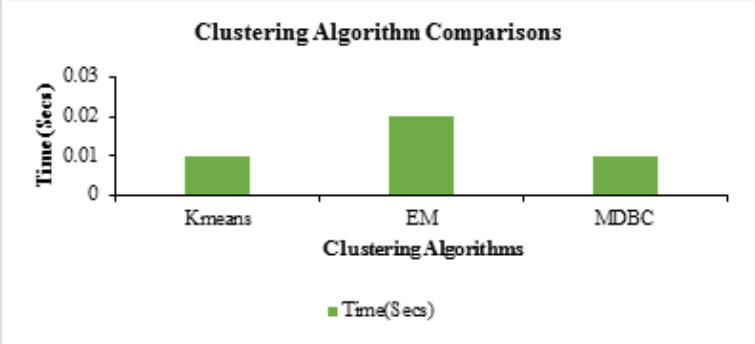


Fig. 9. Time related graph.

5. Conclusions

This paper proposed a protein data to protein data interaction analysis by applying clustering algorithms and performed a clustering on the interacted dataset of structured proteins with size less than 20GB with three different clustering algorithms and found a best clustering algorithm ExpectationMaximization for protein clustering. The sequences analysed and interaction features are discussed. The result predicted shows that mergesets are not only the concentrated field of the process. Interaction aspect in individual can't reach the performance to fulfil the system. Thus, the results obtained show the support of sequence analysis in addition to mergesets filter information. Finally, this procedure may be applied to predict the right drug discovered for respective disease or may produce allergic symptoms. As future work, it would be a challenge to study other data mining technique and apply to the proposed procedure.

References

1. Song, D.; Chen, J.; Chen, G.; Li, N.; Li, J.; Fan, J.; Bu, D.; and Li, S.C. (2015). Parameterized BLOSUM matrices for protein alignment. *IEEE Transactions on Computational Biology and Bioinformatics*, 12(3), 686-694.
2. Spencer, M.; Eickholt, J.; and Cheng, J. (2015). A deep learning network approach to ab initio protein secondary structure prediction. *IEEE Transactions on Computational Biology and Bioinformatics*, 12(1), 103-112.
3. Wu, S.; Liew, A.W.; Yan, H.; and Yang, M. (2004). Cluster analysis of gene expression data based on self-splitting and merging competitive learning. *IEEE Transactions on Information Technology in Biomedicine*, 8(1), 5-15.
4. Gonzalez-Alvarez, D.L.; Vega-Rodriguez, M.A.; and Rubio-Largo, A. (2015). Finding patterns in protein sequences by using a hybrid Multiobjective teaching learning based optimization algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(3), 656-666.
5. Ng, Y.K.; Yin, L.; Ono, H.; and Li, S.C. (2015). Finding all longest common segments in protein structures efficiently. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(3), 644-655.
6. Birlutiu, A.; d'Alche-Buc, F.; and Heskes, T. (2015). A Bayesian framework for combining protein and network Topology information for predicting protein-protein interactions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(3), 538-550.
7. Ichikawa, K.; and Morishita, S. (2014). A simple but powerful Heuristic method for accelerating k-means clustering of large-scale data in life science. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(4), 681-692.
8. Tseng, V.S.; and Kao, C-P. (2005). Efficiently mining gene expression data via a novel Parameterless clustering method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(4), 355-365.

9. Roy, D.K.; and Sharma, L.K. (2010). Genetic k-means clustering algorithm for mixed numeric and categorical data sets. *International Journal of Artificial Intelligence & Applications*, 1(2), 23–28.
10. Dong, G.; and Pei, J. (2007). *Sequence data mining.* Springer US, 47-65.
11. Fayyad, U.M. (1996). Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, 11(5), 20-25.
12. Guralnik, V.; and Karypis, G. (2001). A scalable algorithm for clustering sequential data. *SIGKDD Workshop on Bioinformatics, BIOKDD*, 179-186.
13. Zhou, Q.; Jiang, Q.; Li, S.; Xie, X.; and Lin, L. (2010). An efficient algorithm for protein sequence pattern mining. *The 5th International Conference on Computer Science & Education.* Hefei, 1876-1881.
14. Pipenbacher. P.; Schliep, A.; Schneckener, S.; Schonhuth, A.; Schomburg, D.; Schrader, R. (2002). ProClust: Improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics*, 18(2), S182–S191.
15. Yang, J.; and Wang, W. (2003). CLUSEQ: Efficient and effective sequence clustering. *19th International Conference on Data Engineering*. Los Alamitos, 101-112.

Appendix A

Modeller 9.15 & Bioweka 2.7.5

A program for the prediction of structured protein (Crab - alpha B crystallin, partial [Macropus rufus]**)** are coded in a python language and this scripts is done for the purpose of designing modules in the structure. The main flowchart of the proposed process is shown in Figs A-1 and A-2 with the scripts execution.

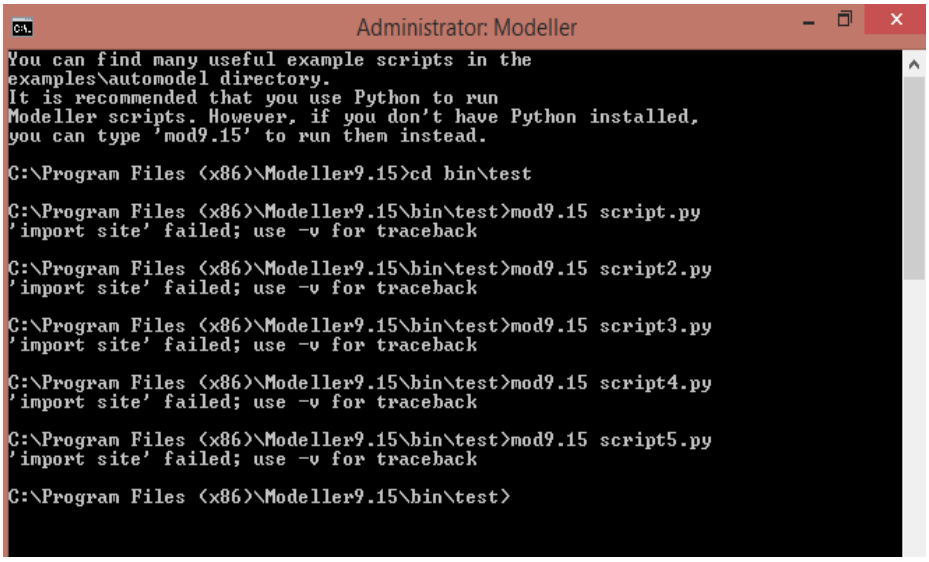


Fig. A-1. Modules execution.

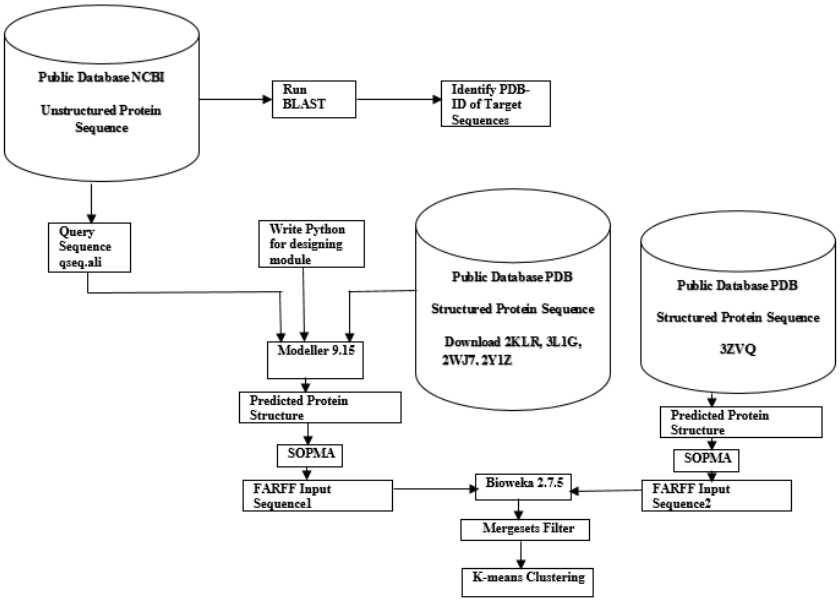


Fig. A-2. Detailed system architecture.

Appendix B

Representation of Protein Structures

In the present work a unstructured protein sequence (Crab - alpha B crystallin, partial [Macropus rufus]) named as query sequence and the target sequences (2KLR - Solid-state NMR structure of the alpha-crystallin domain in alphaB-crystallin oligomers, 3LIG - Human AlphaB crystallin, 2WJ7 - Human AlphaB crystallin, 2Y1Z - Human Alphab crystallin ACD R120G) as shown in Figs. B-1 to B-4 are given as input to predict the protein structure as shown in Fig. B-5 was converted to ARFF file format known as input sequence1. Another structured protein (Egg - Crystal Structure of proteolyzed lysozyme) as shown in Fig. B-6 was converted into ARFF file known as input sequence2. This PPI method done through mergesets filter in Bioweka software and finds the best clustering algorithm for the real world applications.

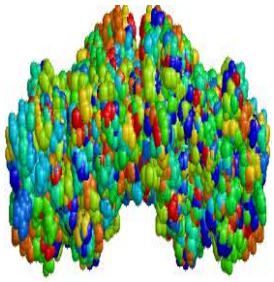


Fig. B-1. 2KLR - Solid-state NMR structure of the alpha-crystallin domain in alpha-crystallin oligomers.

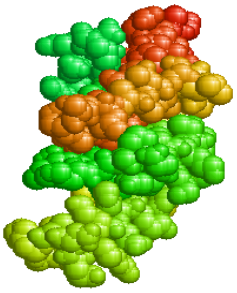


Fig. B-2. 3LIG - Human alphab crystalline.

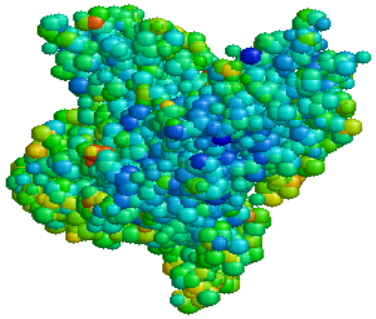


Fig. B-3. 2WJ7 - Human alphab crystalline.

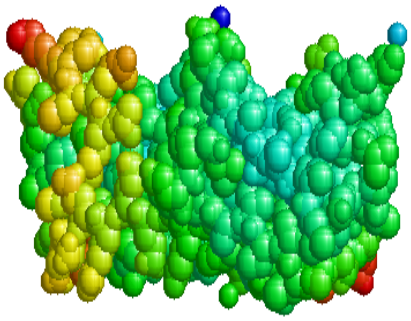


Fig. B-4. 2Y1Z - Human alphab crystalline ACD R120G.

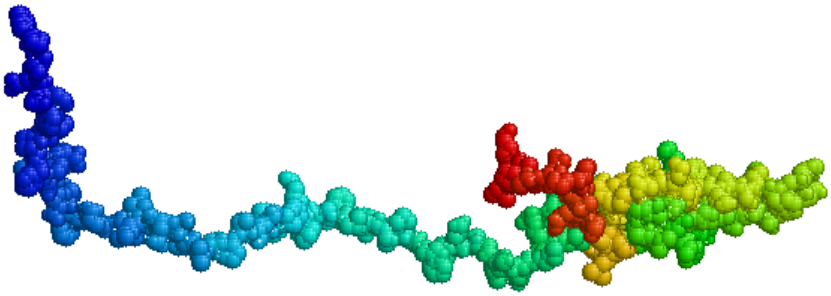


Fig. B-5. Predicted protein structure.

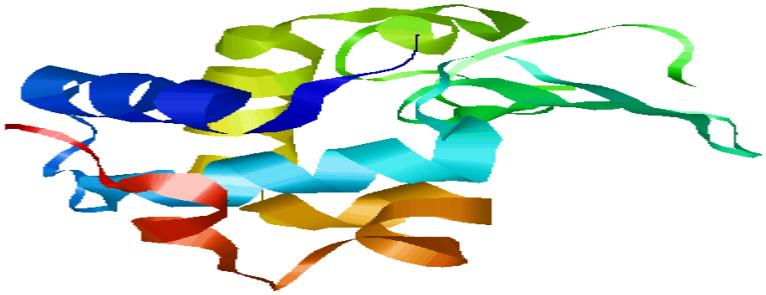


Fig. B-6. Crystal structure of proteolyzed lysozyme.