International Journal of Computational Intelligence and Informatics, Vol. 4: No. 3, October - December 2014

An Empirical Analysis of Flame and Fuzzy C-Means Clustering for Protein Sequences

C Murugananthi
Assistant Professor
Department of Computer Science
Saratha College of Arts and Science
Erode, Tamil Nadu
murugananthiselvi3@gmail.com

D Ramyachitra
Assistant Professor
Department of Computer Science
Bharathiar University
Coimbatore, Tamil Nadu
jaichitra1@yahoo.co.in

***Abstract*- Biological data have to be analyzed, interpreted and processed to deal with the problems in life sciences. Bioinformatics addresses the biological problems using computational methods. Clustering is one of the computational techniques for analyzing biological data. Clustering protein sequences into families with similar patterns is important in Bioinformatics. Many clustering algorithms are available for rapid development of protein sequences. In this paper, we compare and evaluate the performance of two clustering algorithms, namely fuzzy c-means and flame for protein sequences. First, we describe each clustering method and compare them through the validity indices and execution time as well.**

Keywords- Bioinformatics, Protein sequence, Fuzzy c-means clustering, Flame clustering

I. INTRODUCTION

Bioinformatics is the use of computer technology for managing biological data and solving complex biological problems. Numerous genomics projects have caused a rapid growth of the protein databases. Most of the problems in Bioinformatics related to the analysis of DNA or protein sequences. Among biological sequences, protein sequences are important for protein is essential in all living organisms [1].

Proteins are composed of twenty amino acids and arranged in a sequential form. Each protein has unique structure and functions. Protein sequences are represented by a combination of alphabets, each representing different amino acids. These sequences are called the primary structure of the proteins. A protein sequence determines its structure and the structure determines function. These primary sequences of proteins are used for clustering proteins effectively. Determining the relationship between biological objects such as protein sequences and structures is important in Bioinformatics [2].

Clustering is partitioning of objects into different groups, so that the objects in each group share some common features [1, 3]. Data points within group are similar and different between groups [3]. Clustering techniques are used in various fields such as data mining, bioinformatics, web mining, biometrics, biomedical data analysis and document processing [4, 5, 6, 7].

Large amount of proteomic and genomic data has rapidly increased in bioinformatics. So there is a need of advanced computational tools to analyze and manage the data [8]. Clustering is one of the methods used in biological data analysis. Advantages of protein sequence clustering includes determining the protein structure/function, categorizing a new sequence, predicting the relationships between protein sequences, and extracting similar sequences for a given query sequence [9, 10]. Protein clustering also used in protein 3-dimensional structure discovery for understanding protein's function [1].

Many clustering algorithms and methods are available for clustering proteins. Most of the works uses the hierarchical, partitioning, graph based techniques and clustering the proteins by sequences [11, 12, 13, 14, 15, 16]. In this paper, we compare two clustering algorithms, fuzzy c-means and flame. The paper is organized as follows. Section 2 discusses the problem objective and presents the two algorithms used for comparison. Section 3 describes the performance evaluation of two algorithms and the conclusion is given in Section 4.

II. PROBLEM OBJECTIVE AND METHODOLOGY

Clustering proteins are used to identify the relationship between protein sequences and structures. Two techniques namely flame and fuzzy c-means clustering are used for clustering proteins and the performance of these algorithms are analyzed and compared for finding the efficient technique. The system architecture of this work is as follows:

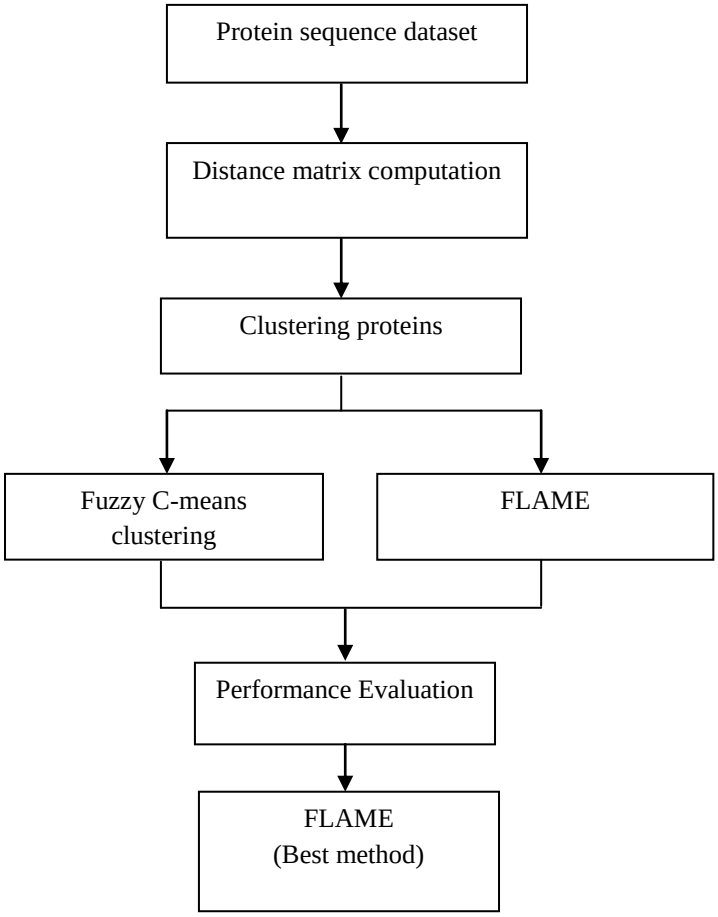

Figure 1. System architecture

2.1 Distance Matrix Computation

We used Smith-Waterman local alignment algorithm [17] for calculating alignment score. This method compares all sequences with each other and computes the alignment score. The distance matrix can be computed after finding the alignment score matrix. Distance between two protein sequences can be derived from its similarity score [18]. For a given set of protein sequences, distance between any two sequences is calculated as

$$D(G,H) = -\ln S_n(G,H) \quad (1)$$

where G and H are protein sequences, $D(G,H)$ is the distance between G and H, ln is natural logarithm, $S_n(G,H)$ is the normalized similarity score between G and H. Here $0 \leq S_n(G,H) \leq 1$ for any protein sequences G and H, and $S_n(G,H) = 1$ if sequences G and H are same. The normalized similarity score is obtained by using the below formula

$$S_n(G,H) \cong \frac{S(G,H)}{L.Q} \quad (2)$$

where $S(G,H)$ is the similarity score of G and H, L denote the length of the local alignment of G and H, and Q is normalization parameter. The normalization parameter Q is computed as a value when two residues are matched with each other. This value depends on the distribution of residues in the local alignment of G and H, and the scoring matrix between residues.

2.2 Clustering Algorithms

2.2.1 Fuzzy C-Means Clustering

In hard clustering or crisp clustering, each object belongs to exactly one cluster. In soft clustering (fuzzy clustering), objects can belong to more than one cluster [19]. Fuzzy clustering is a process of assigning membership and then using them to assign objects to one or more clusters. One of the widely used fuzzy clustering algorithms is Fuzzy C-Means (FCM) Algorithm [20, 21]. In this algorithm membership is assigned to each protein corresponding to each cluster based on the distance between protein and cluster center. Summation of membership of each protein should be equal to one. After each iteration, membership and cluster centers are updated.

Pseudo code

Let *D* is protein dataset, *D*= $\{x_i\}$, where i= 1, 2… n; n is the size of *D* and *k* is number of clusters

1) Randomly select *'c'* proteins as cluster centers.

Repeat

2) Calculate the fuzzy membership $'\mu_{ij}'$ using:

$$u_{ij} = 1/\sum_{l=1}^{c}\left(\frac{d_{ij}^2}{d_{il}^2}\right)^{\frac{1}{m-1}} \quad (3)$$

3) Update the fuzzy centers $'v_j'$ using:

$$V_j = \frac{\sum_{i=1}^{n}((u_{ij})^m x_i)}{\sum_{i=1}^{n}(u_{ij})^m},\ where\ j = 1,2,\dots,c \quad (4)$$

Until V_j estimate stabilize.

where $1 \le i \le n$ and $1 \le j \le C$. n is the number of proteins and c is the number of clusters, u_{ij} is the membership of ith protein in the jth cluster. d_{ij} is the distance between the ith protein and jth cluster center. m is the fuzzification parameter and it should be more than one. If m=1, then the problem is a crisp clustering. $m \in [1, \infty]$ and usually m is set to 2 [22]. V_j is the jth cluster center and c is the number of cluster. FCM minimize the objective function [20] in Eq. (5) and summation of membership of each protein should be equal to one in Eq. (6).

$$J(U,V) = \sum_{j=1}^{C}\sum_{i=1}^{N}(u_{ij})^m\,(d_{ij})^2 \quad (5)$$

$$\sum_{j=1}^{C} u_{ij} = 1 \quad (6)$$

2.2.2 Flame Clustering

Fuzzy clustering by Local Approximation of MEmberships (FLAME) [23] defines clusters in the dense parts of a dataset and performs cluster assignment based on the neighborhood relationships among objects. The FLAME constructs k-Nearest Neighbors graph to identify the cluster centers and outliers. Proteins with the highest local density called Cluster Supporting Objects (CSO) and proteins with a local density lower than a threshold are called outliers. CSOs are assigned with full membership to represent itself as cluster centers. Outliers are assigned with full membership to the outlier group. Fuzzy memberships are then assigned to remaining proteins with varying degrees of memberships to the cluster supporting objects. There is no need to specify the predefined number of clusters. It automatically determines the numbers of cluster and outliers. FLAME requires the number of k-Nearest Neighbors and threshold value for outliers as initial parameters.

Pseudo code

1. Extraction of the structure information from the dataset
    1. Construct a neighborhood graph to connect each object to its K-Nearest Neighbors (KNN)
    2. Estimate a density for each object based on its proximities to its KNN
    3. Objects are classified into 3 types
        1. Cluster Supporting Object (CSO): object with density higher than all its neighbors
        2. Cluster Outliers: object with density lower than all its neighbors and lower than a predefined threshold
        3. The rest.
2. Local/Neighborhood approximation of fuzzy memberships
    1. Initialization of fuzzy membership
        1. Each CSO is assigned with fixed and full membership to itself to represent one cluster
        2. All outliers are assigned with fixed and full membership to the outlier group
        3. The rest are assigned with equal memberships to all clusters and the outlier group
    2. Then the fuzzy memberships of all type 3 objects are updated by a converging iterative procedure called Local/Neighborhood Approximation of Fuzzy Memberships, in which the fuzzy membership of each object is updated by a linear combination of the fuzzy memberships of its nearest neighbors.
3. Cluster construction from fuzzy memberships in two possible ways
    1. One-to-one object-cluster assignment, to assign each object to the cluster in which it has the highest membership
    2. One-to-multiple object-clusters assignment, to assign each object to the cluster in which it has a membership higher than a threshold.

III. PERFORMANCE EVALUATION

3.1 Protein Sequence Datasets

The experiment was conducted on four different protein data sets: Dengue virus proteins, Human Leukocyte Antigen (HLA) proteins, Globins proteins and Saccharomyces cerevisiae (Yeast) proteins. Dengue virus protein sequences are extracted from Protein Data Bank [24] and named as DS1. Sequences of Globins protein family and Human Leukocyte Antigen (HLA) proteins were collected from European Bioinformatics Institute (EMBL-EBI) database [25] and named as DS2, DS3 respectively. Yeast proteins are collected from Saccharomyces Genome database [26] and named as DS4.

3.2 Validity Indices

To assess the performance of clustering algorithms, we used two validity indices silhouette index and partition index.

3.2.1 Silhouette Index

The silhouette index [27] is a cluster validity index used to assess the quality of any clustering. The silhouette index of a protein defines its closeness to its own cluster relative to its closeness to other clusters. The silhouette width s(x) of the protein x is defined as

$$S(x) = \frac{b(x)-a(x)}{\max[b(x),a(x)]} \quad (7)$$

where a(x) is the average distance between protein x and all other proteins in its cluster and b(x) is the minimum of the average distances between protein x and the proteins in the other clusters. The silhouette index s(c) of cluster c is defined as the average silhouette width of its all proteins. Finally, silhouette index of the whole clustering is the average silhouette width of all clusters. It reflects the compactness and separation of clusters. The value of the silhouette index varies from -1 to 1 and higher values indicate a better clustering result.

3.2.2 Partition Index

The partition index p(c) [23] is defined as the ratio between the overall within-cluster variability and the overall between-cluster distance. Based on this validation index, a good data clustering results in low intra cluster variation and high inter cluster variation. To find the overall within-cluster variation, the variation within each cluster is calculated as the average distance between each pair of proteins in the cluster and then averaged for all clusters. The between-cluster variation is obtained by averaging the distance between each pair of clusters. Each single between-cluster distance is calculated by averaging the distance between each pair of protein from the two clusters. The lower partition index value indicates the better clustering result.

3.3 Results and Discussions

The experiments were conducted on Intel pentium-4 processor with 2GB RAM. Alignment scoring matrix for dataset given in section 3.1 was obtained by Smith-Waterman algorithm [17]. Then, the normalized similarity scores are calculated by Eq. (2). Distance matrix of protein sequences are calculated using similarity scores. After completing these processes, clustering algorithms are initialized, and run with the datasets and above predicted distance matrix. We calculate validity indices given in Section 3.2 for clustering algorithms on four datasets. Figure2 shows silhouette index on four datasets. Figure 3 shows partition index on four datasets.

TABLE I. SILHOUETTE INDEX OF ALGORITHMS ON FOUR DATASETS

Algorithms	**Datasets**			
	DS1	***DS2***	***DS3***	***DS4***
FCM	0.4837	0.4568	0.4488	0.4231
FLAME	0.5097	0.4798	0.4781	0.4797

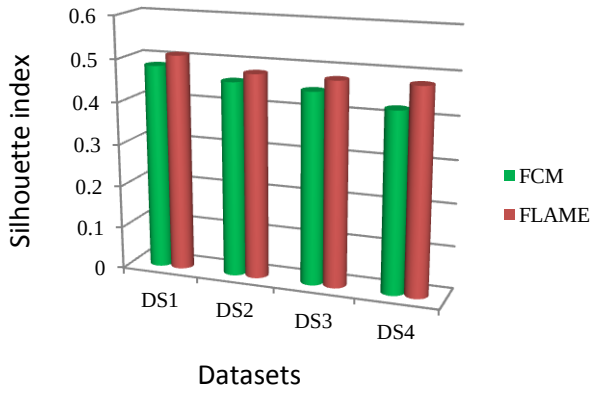

Figure 2. Clustering validation and comparison by silhouette index

TABLE II. PARTITION INDEX OF ALGORITHMS ON FOUR DATASETS

Algorithms	**Datasets**			
	DS1	***DS2***	***DS3***	***DS4***
FCM	0.3087	0.3022	0.3532	0.3279
FLAME	0.2595	0.2813	0.2931	0.2754

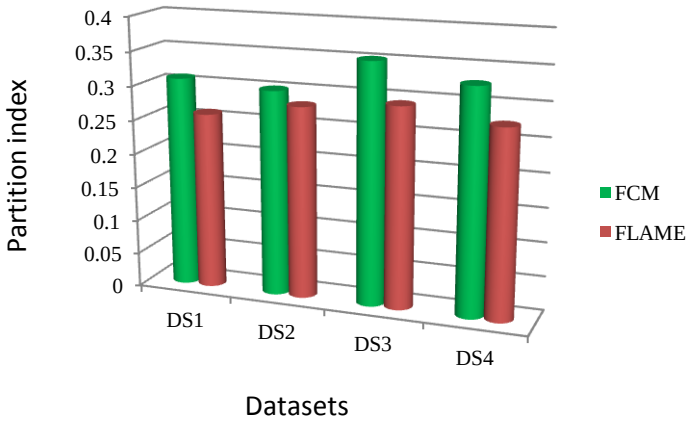

Figure 3. Clustering validation and comparison by partition index

TABLE III. EXECUTION TIME OF ALGORITHMS ON FOUR DATASETS

Algorithms	**Datasets**			
	DS1 (Sec.)	***DS2 (Sec.)***	***DS3 (Sec.)***	***DS4 (Sec.)***
FCM	74.7738	86.6731	80.0534	87.2015
FLAME	72.1925	80.9234	75.7239	84.8062

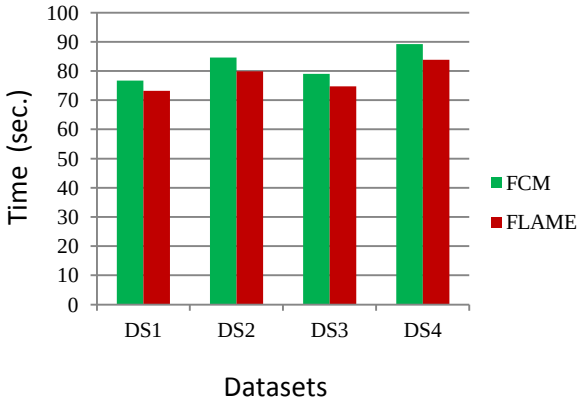

Figure 4. Execution time of algorithms on four datasets

According to both of the validity index analysis, flame is the best algorithm on four datasets. Figure 4 shows the execution time of clustering methods on four datasets. Execution time of flame is lower than fuzzy c-means clustering. From the results, it is inferred that flame performs better in terms of validity indices and execution time as well.

IV. CONCLUSION

Many problems in bioinformatics are related to gene or protein sequence analysis. Clustering protein sequences is important problem in bioinformatics. Clustering is used to identify the relationship between proteins. In this paper, we compare and evaluate the performance of two clustering algorithms fuzzy c-means and flame. The experimental result shows that flame clustering performs better than fuzzy c-means clustering in terms of validity indices and execution time.

REFERENCES

[1] Yonghui Chen, Kevin D Reilly, Alan P Sprague, and Zhijie Guan, "SEQOPTICS: a protein sequence clustering system", BMC Bioinformatics, 7(Suppl 4), S10, 2006.

[2] Wooyoung Kim, Bernard Chen, Jingu Kim, Yi Pan, and Haesun Park, "Sparse nonnegative matrix factorization for protein sequence motif discovery", Expert Systems with Applications, 38, 13198–13207, 2011.

[3] Efendi Nasibov, and Cagin Kandemir-Cavas, "OWA-based linkage method in hierarchical clustering: Application on phylogenetic trees", Expert Systems with Applications, 38, 12684–12690, 2011.

[4] L. Baldacci, M. Golfarelli, A. Lumini, and S. Rizzi, "Clustering techniques for protein surfaces", Pattern Recognition, 39(12), 2370–2382, 2006.

[5] B.C.H. Chang and S.K. Halgamuge, "Protein motif extraction with neuro-fuzzy optimization", Bioinformatics, 18, 1084–1090, 2002.

[6] Z.S.H. Chan, L. Collins, and N. Kasabov, "An efficient greedy k-means algorithm for global gene trajectory clustering", Expert Systems with applications, 30(1), 137–141, 2006

[7] K.S. Lin and C.F. Chien, "Cluster analysis of genome-wide expression data for feature extraction", Expert Systems with Applications, 36(2), 3327–3335, 2009.

[8] Piotr Lukasiak, Jacek Blazewicz, and Maciej Milostan, "Some Operations Research Methods for Analyzing Protein Sequences and Structures", Annals of Operations Research, 175, 9-35, 2010.

[9] D.W. Mount, "Bioinformatics—Sequence and Genome Analysis", Cold Spring Harbor Lab Press, New York,2002.

[10] C. Peter and B. Rolf, "Computational Molecular Biology—An Introduction", Wiley, New York, 2000.

[11] Sondes Fayech, Nadia Essoussi, and Mohamed Limam, "Partitioning clustering algorithms for protein sequence data sets", BioData Mining, 2:3, 2009.

[12] S. Mitra and T. Acharya, "Data Mining: Multimedia, Soft Computing and Bioinformatics", Wiley, New York, 2003.

[13] A.K. Jain and R.C. Dubes, "Algorithms for Clustering Data", Prentice-Hall, Upper Saddle River, NJ, 1988.

[14] Jain, A.K., Murty, M.N., Flynn, P.J., (1999). Data clustering: a review. ACM Comput. Surv., 31(3), 264–323.

[15] A.K. Pujari, "Data Mining Techniques", Universities Press, Private Limited, India, 2000.

[16] K. Cios, W. Pedrycz, and R. Swiniarski, "Data Mining Techniques", Kluwer Academic Publishers, Dordrecht, MA, 1988.

[17] T.F. Smith and M.S. Waterman, "Identification of common molecular subsequences", Journal of Molecular Biology, 147, 195-197, 1981.

[18] H. Matsuda, T. Ishihara, and A. Hashimoto, "Classifying molecular sequences using a linkage graph with their pairwise similarities",. Theoretical Computer Science, 210, 305–325, 1999.

[19] L. Zadeh, "Fuzzy sets. Information and Control", 8, 338–353, 1965.

[20] J.C. Bezdek, " Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, 1981.

[21] F. Hoppner, F. Klawonn, and R. Kruse, "Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition", New York: Wiley, 1999.

[22] R. Hathaway and J.C. Bezdek, "Fuzzy c-means clustering of incomplete data", IEEE Transactions. Systems, Man, Cyberetics, vol. 31(5), pp. 735–744, 2001.

[23] Limin Fu and Enzo Medico, "FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data", BMC Bioinformatics, 8:3, 2007.

[24] Protein Data Bank, 2013. (http://www.rcsb.org) Last access 03.05.2013.

[25] The European Bioinformatics Institute (EMBL-EBI) database, 2013. (http://srs.ebi.ac.uk) Last access 10.06.2013.

[26] Saccharomyces Genome database, 2013. (http://www.yeastgenome.org) Last access 25.05.2013.

[27] P.J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, 1987.