SOFT GRANULAR COMPUTING MODEL FOR IDENTIFYING PROTEIN SEQUENCE MOTIF BASED ON SVD-ENTROPY METHOD

M Chitralegha and Dr K Thangavel

Abstract— Bioinformatics is a field devoted to the interpretation and analysis of biological data using computational techniques. In recent years the study of bioinformatics has grown tremendously due to huge amount of biological information generated by scientific community. Proteins are made up of chain of amino acids. Protein sequence motifs are small fragments of conserved amino acids often associated with specific function. These sequence motifs are identified from protein sequence segments generated from large number of protein sequences. All generated sequence segments may not yield potential motif patterns. Selecting subset of segments can improve clustering and lead to better understanding of the motif patterns. Protein sequence segments have no classes or labels, so one has to apply unsupervised segment selection method. Hence, Singular Value Decomposition (SVD) segment selection technique combined with Fuzzy C-Means (FCM) granular computing model has been proposed for the first time. Computational results demonstrate the efficiency and beneficial outcome of the proposed method. This approach shows the better performance at finding significant motif patterns that transcend different protein families.

Index Terms— Clustering, Singular Value Decomposition, Protein Sequence, Motif,DBI,HSSP-BLOSM62,Fuzzy c-Means.

1 INTRODUCTION

The word "protein" is derived from the Greek word "Protos", means "Primary" or "first rank of importance". Proteins form the very basis of life. They regulate variety of activities in all known organisms, from replication of the genetic code to transporting oxygen and are generally responsible for regulating cellular machinery and consequently the phenotype of organism. A protein is a long polypeptide chain. It is the chemical properties of each amino acid and its unique sequencing of peptide chain that gives a protein its distinct function and structure. Proteins have several different levels of organisation. The first level of protein structure is its primary structure. The primary structure of protein is the linear sequence of its constituent amino acids. The next level of organization is secondary structure of proteins. The most common secondary structure elements in proteins are the alpha helix and beta sheets. Polypeptide chains begin to interact with their respective side chains thus creating more complex level of folding called tertiary structure [7].

A sequence motif is a nucleotide or amino acid sequence pattern that has biological significance. Protein sequence motifs are signatures of protein families and can often be used as tools for the prediction of protein function. Detection of such motifs in proteins is a crucial problem in today's bioinformatics research. There are several databases available for sequence motifs but the most popular ones are PROSITE [11], PRINTS [2] and BLOCKS [10]. MEME, Gibbs Sampling and Motif Scan [8] are some of the tools to discover protein sequence motifs that transcend protein families. But these methods will generate motif patterns only for a single protein sequence. The patterns obtained by using above methods, may carry only a little information about conserved sequence regions which transcend protein families. Instead, in this paper, a huge number of segments are generated using sliding window technique and patterns are extracted from selected segments. Multiple protein sequences are represented by their corresponding HSSP file [16].

All generated sequence segments may not be significant and may also affect final motif patterns. In Super Granular SVM feature elimination technique segment selection process has been performed after clustering [5]. In this paper, SVD Entropy is used for segment selection process before clustering technique followed by FCM granular computing method. Finally information generated by all granules is combined to obtain the final sequence motif.

The rest of the paper is organized as follows. Section 2 presents related work in this area of research. Section 3 introduces SVD-Entropy based segment selection process. Section 4 discusses the K-Means clustering algorithm and it's variant. Section 5 focuses on the proposed K-Means granular with SVD Entropy and FCM granular with SVD Entropy algorithm. In section 6, experimental analysis and motif patterns are provided. Section 7 concludes the paper with directions for further enhancement.

II. RELATED WORK

Han and Baker [9] have first used K-Means clustering algorithm for finding protein sequence motif. They have chosen set of initial points for cluster centers in a random manner. Selecting initial points randomly leads to an unsatisfactory partition because some initial points may lie close to each other. In order to overcome the above mentioned problem, Wei Zhong [20] has proposed Improved K-Means clustering to explore sequence motifs. Improved K-Means algorithm tries to obtain initial points by using Greedy approach. In this approach, for each run, clustering algorithm will be executed for fixed number of iterations and then selects initial points which have capacity to form clusters with good structural similarity. The distance of chosen initial points will be checked against points already available in the initialization array. If minimum distance of newly selected points is greater than threshold value, these points will be added to the initialization array. In this research, data set is said to be huge and selecting initial points using above greedy approach leads to high computational cost. Computational cost is a major problem to be faced when input data-set is very large.

Bernard Chen [3, 4] has proposed granular computing model using Fuzzy C-Means clustering technique. In his work the segments first partitioned into small information granules using Fuzzy C-Means clustering method. Then, for each granule Improved K-Means algorithm has been executed. Finally, the clusters formed in each granule are combined to find final sequence motif information. In his another work, Fuzzy Greedy K-Means approach, granular computing technique is adopted and then initial points chosen greedier than Improved K-Means algorithm. In the Greedy K-Means, the best centriods are selected after five runs of K-Means and then K-Means algorithm is executed by considering those centriods. It consumes more time and complexity is also high.

Motif detection from a huge amount of sequences is a challenging task and not all the segments generated are so important. Therefore, Bernard Chen [5] has proposed Super Granular SVM Feature Elimination. In this approach the original dataset is first partitioned using Fuzzy C-Means clustering and then for each partition Greedy K-Means clustering algorithm is been implemented. Then ranking SVM based segment selection is done on each cluster to collect survived sequence segments. The survived segments are then clustered once again using Greedy K-Means to generate motif information.

The Super Granular SVM segment selection technique requires more computational time for segment selection process. Here, the computational time includes time taken for Fuzzy Clustering plus time taken for Greedy K-Means clustering before segment selection. In this paper, SVD Entropy segment selection technique is applied before clustering, which helps us to reduce computational time. Here, all sequence segments generated by sliding window technique may not yield highly structural similar clusters. Therefore, removing such noisy segments using entropy segment selection [19] helps us to produce clusters with good structural similarity.

III. Segment Selection Technique

A. SVD- Entropy Based Segment Selection Technique

SVD based entropy is proposed for the first time to address the problem of selecting the significant segments in the area of protein sequence motifs identification. The formula for calculating singular value decomposition for each sequence segment is given here under [1].

$$V_j = S_j^2 / \sum_w S_w^2 \qquad (1)$$

where S_j denotes singular values of the segment, S_w^2 denotes eigen values of the segment, w denotes window size.

The resulting SVD- Entropy is as follows

$$\text{E} = -\frac{1}{\log(w)} \sum_{j=1}^{w} V_j \log(V_j) \qquad (2)$$

```
Algorithm: SVD Entropy Based Segment Selection
Input: Sequence segments of N numbers.
Output: Significant protein sequence segments.
Procedure:
Step1: Computation of SVD - Entropy
    For i = 1 to N
        Calculate singular value decomposition for each sequence segment using (1)
        S = Number of non zero SVD entries along window size
        For j varies from 1 to S
            Apply SVD Entropy using (2)
        End For
    End For
Step2: Selection of Sequence segments
If (entropy of each sequence segment) < (threshold value) then
    Select the sequence segments.
End If
```

Fig. 1 SVD Entropy Segment Selection Algorithm

Fig. 1 shows SVD Entropy Selection algorithm applied in proposed K-Means granular with SVD Entropy and FCM granular with SVD Entropy technique.

IV Clustering Algorithms

Clustering is a process of grouping a set of data objects into clusters based on the information found in the data objects, in such a way that the objects in the same cluster are similar where as objects in different clusters are different. The clustering plays an important role in various data analysis fields including statistics, pattern recognition, machine learning, data mining, information retrieval and bioinformatics. Generally, clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. An excellent survey of clustering techniques can be found in [12]. In this research, K-Means clustering algorithm and its variant are cored to cluster the segments.

A. K-Means Clustering

K-Means clustering is one of the simplest unsupervised learning algorithm in Data Mining to identify patterns. This section explains the original K-Means clustering algorithm. The idea is to classify a set of input samples into K number of disjoint clusters, where the value of K is fixed in advance. The algorithm consists of two separate phases [9, 15]:

The first phase is to define K seeds, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some cluster, the first step is completed and initial grouping is done. Next we need to recalculate the new centroids, including new points may lead to a change in the cluster centroids. Once we find K new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, K centroids may change their position in a step by step manner. Finally, a situation will be reached where centriods do not move anymore. This signifies the convergence criterion for clustering.

In this proposed work, the original data set is said to be very large. To overcome high computational cost caused due to large input dataset a granular computing model [14] that utilized Fuzzy C-Means clustering algorithm to divide the whole data space into several small subsets and then apply K-Means clustering algorithm to each subset to find motif information.

B. Fuzzy C-Means Clustering

Fuzzy C-Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method developed by Dunn in 1973 and improved by Bezdek in 1981 is frequently used in pattern recognition. It is based on minimization of the following objective function [3]:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2$$

where $1 \le m < \infty$, m is any real number greater than 1, uij is the degree of membership of xi in the cluster j, xi is the ith of d-dimensional measured data, cj is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership uij and the cluster centers cj by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}$$

This iteration will stop when $\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$ where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of Jm.

The algorithm is composed of the following steps:

1. *Initialize U=[u_{ij}] matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}$=[c_j] with $U^{(k)}$*

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update $U^{(k)}$, $U^{(k+1)}$*

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}$$

4. *If $|| U^{(k+1)} - U^{(k)} || < \varepsilon$ then STOP; otherwise return to step 2.*

Fig. 2 Fuzzy C-Means clustering

V. Proposed Work

A. K-Means Granular with SVD Entropy

K-Means Granular with SVD technique comprised of three stages. Stage one selects significant protein sequence segments using SVD-Entropy method. In stage two, the survived segments are then clustered into small information granules using traditional K-Means algorithm. In this proposed work number of granules has been set to ten. Finally in stage three, for each granule once again SVD segment selection method is applied which removes if there are still more uncertain segments available in input dataset.

This technique adopts double refinement, which helps us to remove noisy sequence segments which may affect final motif patterns. Finally, we collect all survived segments which are then clustered using benchmark K-Means algorithm. Experimental results show that K-Means granular with double refinement SVD is better than single refinement SVD with K-Means. Fig. 3 depicts the structure of proposed K-Means granular with SVD Entropy.

B. FCM Granular with SVD Entropy

This proposed work consists of two phases. Phase one selects significant segments using SVD-Entropy method. Phase two adopts FCM of granular computing technique. This is for the first time SVD-Entropy has been combined with FCM granular to identify hidden motif patterns that are available in different protein families. As the dataset is very large, hence the proposed work focuses on segment selection technique to be applied before granular computing which helps us to reduce computational cost.

Traditional K-Means Clustering is performed on each information granule generated by FCM. At the final stage, we combine information generated by all granules and obtain final sequence motif information. It is noted from the results that Fuzzy C-Means granular with SVD Entropy technique able to identify more number of hidden motif patterns compared with that of K-Means granular with SVD-Entropy. Figure 4 shows the structure of FCM granular with SVD Entropy.

Comparing the above two proposed techniques the advantage of FCM granular with SVD Entropy is that number of stages to generate motif information has been decreased. Since the stages is been decreased from three to two in FCM granular with SVD – Entropy technique computational cost also decreases to find motif information. The quality of clusters and motif information obtained in our proposed work is said to be more significant compared to K-Means Granular with SVD – Entropy.

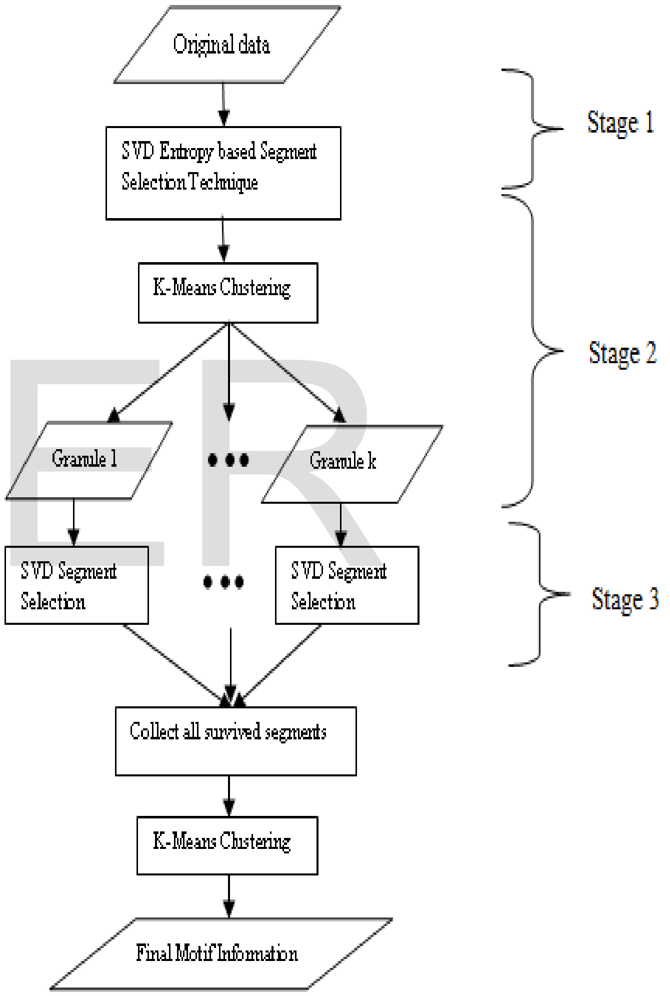


Fig. 3 Sketch of K-Means Granular with SVD Entropy

VI. Experimental Setup

A. Data Set

The latest dataset obtained from Protein Culling Server (PISCES) [18] includes 4946 protein sequences. In this work, we have considered 3000 protein sequences to extract sequence motifs that transcend in protein sequences. The threshold for percentage identity cut-off is set as less than or equal to 25%, resolution cut-off is 0.0 to 2.2, R-factor cut-off is 1.0 and length of each sequence varies from 40 to 10,000.

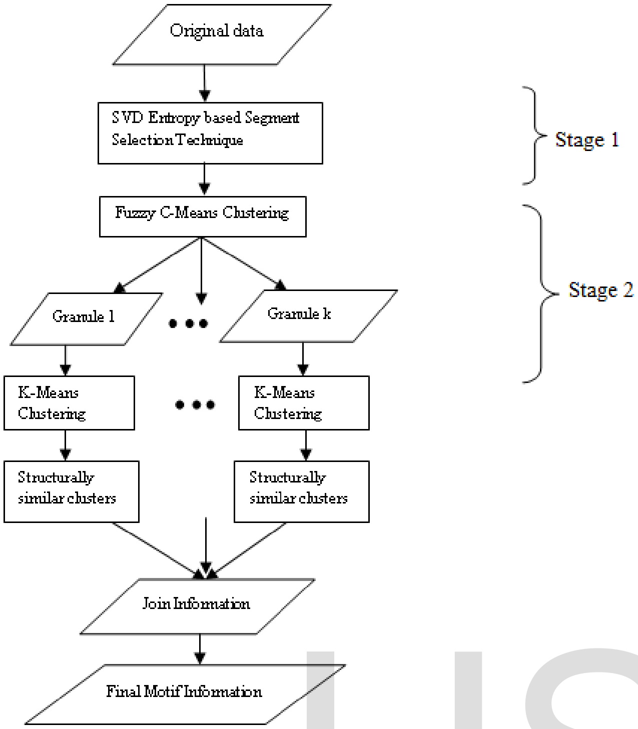
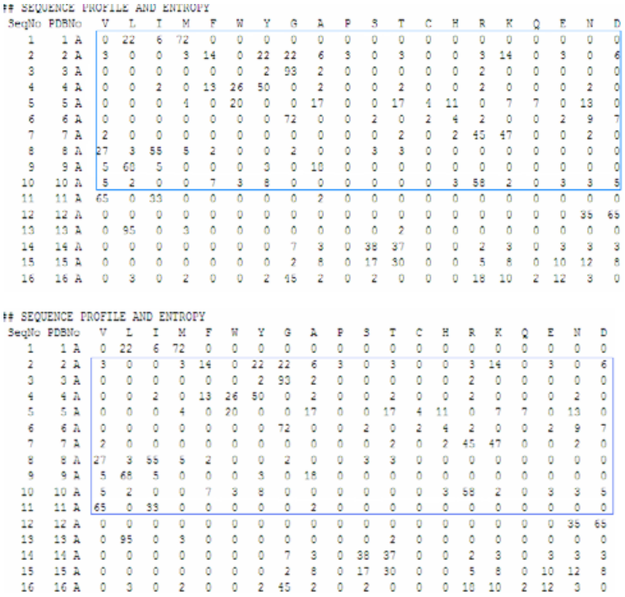


Fig. 4 Sketch of FCM Granular with SVD Entropy

Each protein sequence is represented by their corresponding frequency profile from HSSP [16]. The sliding windows with ten successive residues are generated from protein sequences. Each window represents one sequence segment of ten continuous positions. Around 6, 60,364 sequence segments are generated by sliding window method, from 3000 protein sequences. Each sequence segment is represented by 10 X 20 matrix, where ten rows represent each position of sliding window and 20 columns represent 20 amino acids.

Homology Secondary Structure Prediction (HSSP) frequency profiles are used to represent each segment [16]. Database of Secondary Structure Prediction (DSSP) [17] assigns secondary structure to eight different classes [13]. In this paper, we convert those eight classes to three different classes based on the CASP experiment as follows [3]: H, G and I to H (Helices); B and E to E (Sheets); all others to C (Coils)



```
## SEQUENCE PROFILE AND ENTROPY
SeqNo PDBNo   V   L   I   M   F   W   Y   G   A   P   S   T   C   H   R   K   Q   E   N   D
    1     1 A   0  22   6  72   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    2     2 A   3   0   0   3  14   0  22  22   6   3   0   3   0   0   3  14   0   3   0   6
    3     3 A   0   0   0   0   0   0   2  93   2   0   0   0   0   0   2   0   0   0   0   0
    4     4 A   0   0   2   0  13  26  50   0   2   0   0   2   0   0   2   0   0   0   2   0
    5     5 A   0   0   0   4   0  20   0   0  17   0   0  17   4  11   0   7   7   0  13   0
    6     6 A   0   0   0   0   0   0   0  72   0   0   2   0   2   4   2   0   0   2   9   7
    7     7 A   2   0   0   0   0   0   0   0   0   0   0   2   0   2  45  47   0   0   2   0
    8     8 A  27   3  55   5   2   0   0   2   0   0   3   3   0   0   0   0   0   0   0   0
    9     9 A   5  68   5   0   0   0   3   0  18   0   0   0   0   0   0   0   0   0   0   0
   10    10 A   5   2   0   0   7   3   8   0   0   0   0   0   0   3  58   2   0   3   3   5
   11    11 A  65   0  33   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0
   12    12 A   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  35  65
   13    13 A   0  95   0   3   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0
   14    14 A   0   0   0   0   0   0   0   7   3   0  38  37   0   0   2   3   0   3   3   3
   15    15 A   0   0   0   0   0   0   0   2   8   0  17  30   0   0   5   8   0  10  12   8
   16    16 A   0   3   0   2   0   0   2  45   2   0   2   0   0   0  18  10   2  12   3   0
```

```
## SEQUENCE PROFILE AND ENTROPY
SeqNo PDBNo   V   L   I   M   F   W   Y   G   A   P   S   T   C   H   R   K   Q   E   N   D
    1     1 A   0  22   6  72   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    2     2 A   3   0   0   3  14   0  22  22   6   3   0   3   0   0   3  14   0   3   0   6
    3     3 A   0   0   0   0   0   0   2  93   2   0   0   0   0   0   2   0   0   0   0   0
    4     4 A   0   0   2   0  13  26  50   0   2   0   0   2   0   0   2   0   0   0   2   0
    5     5 A   0   0   0   4   0  20   0   0  17   0   0  17   4  11   0   7   7   0  13   0
    6     6 A   0   0   0   0   0   0   0  72   0   0   2   0   2   4   2   0   0   2   9   7
    7     7 A   2   0   0   0   0   0   0   0   0   0   0   2   0   2  45  47   0   0   2   0
    8     8 A  27   3  55   5   2   0   0   2   0   0   3   3   0   0   0   0   0   0   0   0
    9     9 A   5  68   5   0   0   0   3   0  18   0   0   0   0   0   0   0   0   0   0   0
   10    10 A   5   2   0   0   7   3   8   0   0   0   0   0   0   3  58   2   0   3   3   5
   11    11 A  65   0  33   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0
   12    12 A   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  35  65
   13    13 A   0  95   0   3   0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0
   14    14 A   0   0   0   0   0   0   0   7   3   0  38  37   0   0   2   3   0   3   3   3
   15    15 A   0   0   0   0   0   0   0   2   8   0  17  30   0   0   5   8   0  10  12   8
   16    16 A   0   3   0   2   0   0   2  45   2   0   2   0   0   0  18  10   2  12   3   0
```

Fig. 5 Sliding Window techniques with a window size of 10 applied on 1b25 HSSP file. Thus by applying the sliding window technique we can generate n number of sequence segments (10 X 20 matrices).

B. Structural Similarity Measure

Average structure of a cluster is calculated using the following formula:

$$\frac{\sum_{i=1}^{w} \max\left(P_{i,H}\,, P_{i,E}, P_{i,C}\right)}{W}$$

where w is the window size and $P_{i,H}$,$P_{i,E}$ and $P_{i,C}$ shows frequency of Helices, Sheets and Coils among the segments for the cluster in position i. If the structural homology for a cluster exceeds 70% the cluster can be considered more structurally similar [3] and if it is between 60% and 70% then the cluster is said to weakly structurally homologous.

C. Distance Measure

Dissimilarity between each sequence segment is calculated using city block metric. In this field of research city block metric is more suitable than Euclidean metric because it considers every position of the frequency profile equally. The following formula is used for distance calculation [3]:

$$\text{Distance} = \sum_{i=1}^{w} \sum_{j=1}^{N} |D_s\,(i,j) - D_c(i,j)\,|$$

where w is the window size and N is 20 amino acids. D_s(i, j) is the value of the matrix at row i and column j which represents sequence segment. D_c (i, j) is the value of the matrix at row i and column j which represents the centroid of a given cluster.

D. David-Bouldin Index (DBI) Measure

Davis-Bouldin Index, measures how compact and well separated the clusters are. To obtain clusters with these characteristics, the dispersion measure for each cluster needs to be small and dissimilarity measure between clusters need to be large [6].

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} R_i$$

Where $R_i = \max_{j=1....k_{j \neq i}} R_{ij}$, i=1...k

The dissimilarity between cluster c_i and c_j in l dimensional space is defined as

$$\text{dinter}\,(c_i\,,c_j\,) = \sum_{k}^{l} ||\,\bar{x}_{ik} - \bar{x}_{jk}\,||$$

and dispersion of a cluster c_i is defined as

$$\text{dintra}\,(\,c_i\,) = \sum_{i=1}^{Np} ||\,x - \bar{x}_{i}\,||$$

where Np is number of members in cluster c_i .Small values of DB are indicative of the presence of compact and well separated clusters.

E. HSSP-BLOSUM Measure

HSSP stands for Homology-Derived Secondary Structure of Proteins. It is a database that combines information from three dimensional protein structures and one dimensional sequence of proteins. BLOSUM stands for Block Substitution Matrix. It is a scoring matrix based on alignment of diverse sequence. A threshold of 62% identity or less resulted in the target frequencies for BLOSUM62 matrix. BLOSUM62 has become a defacto standard for many protein alignment programs [3].

This matrix lists the substitution score of every single amino acid. A score for an aligned amino acid pair is found at the intersection of the corresponding column and row. By using this matrix, we may tell the consistency of the amino acid appearing in the same position of motif information generated by our method. HSSP frequency profile and BLOSUM62 matrix has been combined to obtain significance of motif information. Hence, the measure is defined as the following [3].

If N = 0: HSSP-BLOSUM62 measure = 0

Else If N = 1: HSSP-BLOSUM62 measure = BLOSUM62ii

$$\text{Else:} \quad \text{HSSP-BLOSUM62 measure} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} HSSP_i\, HSSP_j . BLOSUM62_{ij}}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} HSSP_i\, HSSP_j}$$

where m is the number of amino acids with frequency higher than certain threshold in the same position.

$HSSP_i$ indicates the percent of amino acid i to be appeared.

BLOSUM62 $_{ij}$ denotes the value of BLOSUM62 on amino acid i and j.

The higher HSSP-BLOSUM62 value indicates more significant motif information. Here, we adopted DBI measure and HSSP-BLOSUM62 measure to evaluate the performance of clustering algorithms and significance of motif information

.

F. Parameter setup

In this work, SVD - Entropy based segment selection is applied and selected around 85% of sequence segments from original data set. Number of clusters has been set to 900. For FCM granular with SVD – Entropy technique, fuzzification factor is been set to 1.15 and number of clusters is equal to ten. This setting produced better results in our specific dataset. In order to separate information granules from FCM results, the membership threshold is set to 18%. The function that decides how many numbers of clusters should be in each information granule is given below:

$$\text{C}_\text{k} = \frac{n_k}{\sum_{i=1}^{m} n_i} * \text{ total number of clusters}$$

where C_k denotes the number of clusters assigned to information granule , n_k is the number of members belonging to information granule k, m is the number of clusters in FCM. In this technique we are able to identify 899 clusters instead of 900 clusters applied in benchmark K-Means clustering.

TABLE I

Results obtained by SVD-FCM

	Number of Members	Number of Clusters
Granule 1	24412	32
Granule 2	100385	131

Granule 3	44428	58
Granule 4	98815	130
Granule 5	41557	54
Granule 6	33376	44
Granule 7	67448	88
Granule 8	133945	176
Granule 9	42674	56
Granule 10	99339	130
Total	686379	899
Original dataset	565314	900

Table I is the summary of the results from FCM granular with SVD Entropy. Although data size increased from 565314 to 686379, we are going to deal with one information granule at a time.

COMPARISON OF DIFFERENT ALGORITHMS

	K-Means	FCM –K-Means	K-Means with SVD Entropy	K-Means Granular Technique with SVD Entropy	FCM Granular Technique with SVD Entropy
No of clusters >60% and < 70%	174	178	185	196	209
No of clusters > 70%	85	101	113	118	127
% of Seq Segments > 70%	15.5329	19.6012	19.3188	19.4162	20.4862
% of Seq Segments > 60% <70%	17.563	21.2342	21.6265	18.3218	22.183
DBI Measure	5.7694	4.2163	5.4637	5.4072	4.0759
AverageHSSP-BLOSUM62	0.8165	0.8886	0.8567	0.8578	1.1525
Execution Time (in secs)	35831.82	21333.71	30674.34	28800.43	18263.02

Table II shows the comparative results obtained from different clustering algorithms. From above table II, we can interpret that FCM granular technique with SVD able to identify more number of hidden motif patterns by looking towards structural similarity values. From Table II, decreased DBI value shows that cluster quality is increased in FCM- based K-Means algorithm and in FCM granular technique with SVD. The motif information obtained in FCM granular with SVD Entropy technique is said to be more significant compared to all other algorithms. Execution time for the proposed FCM granular with SVD Entropy is said to be comparatively less than all other algorithms.

Fig. 7 shows comparative analysis of cluster quality and quality of motif information. Decreased DBI value and increased HSSP-BLOSUM62 values shows the performance of clustering and significance of motif information obtained in FCM granular technique with SVD Entropy segment selection process is good.

From the above table II, it is inferred that the results obtained in proposed FCM granular with SVD Entropy segment selection technique generates more biochemical meaningful information by eliminating some less meaningful data points. Fig. 6 and 7 are interpreted from the results given in table II.

Fig. 8 is interpreted from table II. From fig. 8 it is observed that execution time for FCM granular with SVD is said to be less than all other algorithms which shows that the goal of computational time has been decreased in proposed FCM granular with SVD Entropy algorithm.

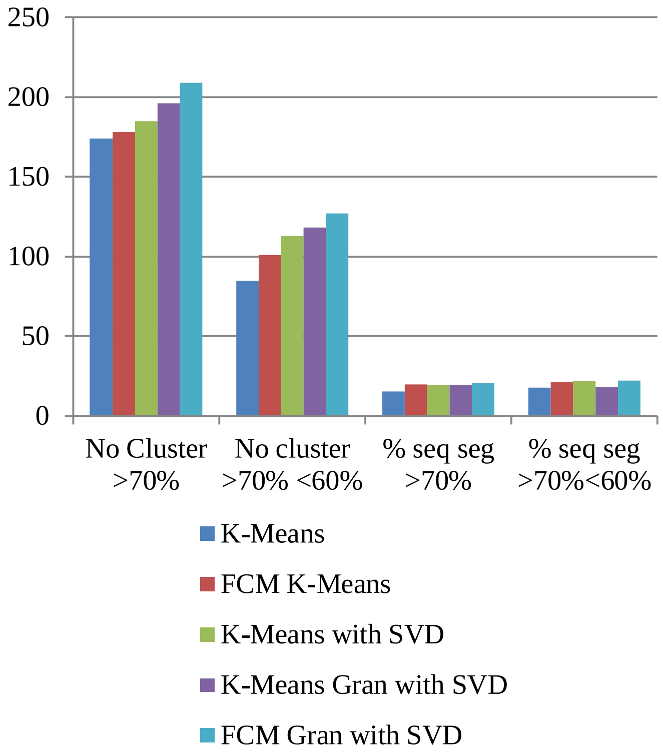


Fig. 6 Comparison of Structural Similarity values

Fig. 6 has been interpreted from table II. From the above fig. 6 we state that the number of strong and weak clusters have been increased in FCM granular SVD technique as well as percentage of sequence segments have also been increased considerably.

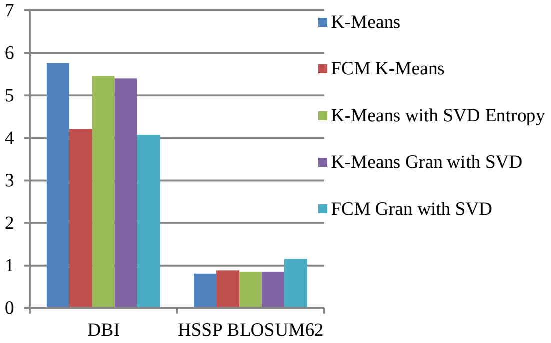


Fig. 7 Comparison of DBI measure and HSSP-BLOSUM62 values

F. Sequence Motifs

Five different motif patterns are shown in motif table 1-5. The following format is used for representation of each sequence motif table. Instead of using existing format in this paper protein logo representation has been used.

- The top box shows the number of sequence segments belonging to this motif, percentage of structural similarity, and average HSSP-BLOSUm62 value.

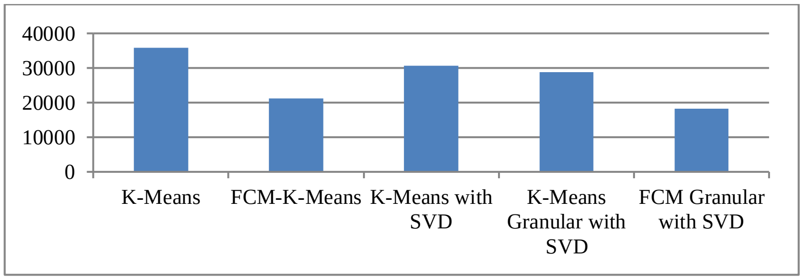


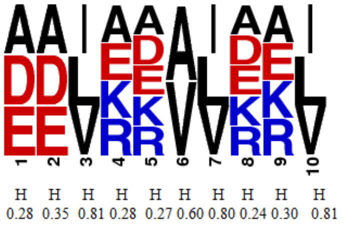
Fig. 8 Comparison of Execution time for different algorithms

- The graph demonstrates the type of amino acid frequently appearing in the given position by amino acid logo. It only shows the amino acid appearing with a frequency higher than 8%. The height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position.

- The x-axis label indicates the representative secondary structure (S), the hydrophobicity value (Hyd.) of the position. The hydrophobicity value is calculated from the summation of the frequencies of occurrence of Leu, Pro, Met, Trp, Ala, Val, Phe, and Ile.

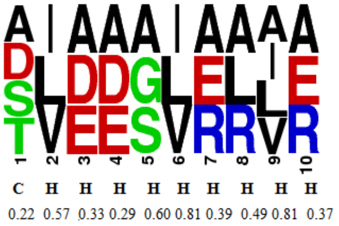
Motif Table 1

Helices Motif with conserved A,E,D
Number of Sequence Segments:1363
Structural Similarity: 83.37%
HSSP-BLOSUM62: 0.2242



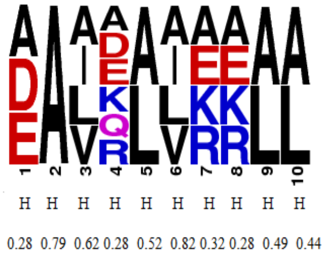
Motif Table 2

Helices Motif with conserved A,L
Number of Sequence Segments: 724
Structural Similarity: 74.20%
HSSP-BLOSUM62: 0.3739



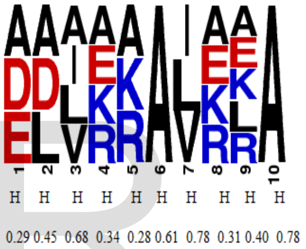
Motif Table 3

Helices Motif with conserved A,K
Number of Sequence Segments:1330
Structural Similarity: 77.60%
HSSP-BLOSUM62: 0.3167



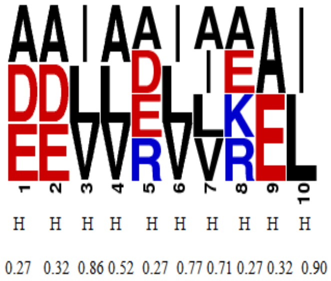
Motif Table 4

Helices Motif with conserved A,K,E
Number of Sequence Segments:1283
Structural Similarity: 79.53%
HSSP-BLOSUM62: 0.3361



Motif Table 5

Helices Motif with conserved V,L,I
Number of Sequence Segments:845
Structural Similarity: 75.51%
HSSP-BLOSUM62: 0.3220



VII. Conclusion

The domain of bioinformatics deals with large voluminous of data. To deal with large input dimensionality dataset a wealth of feature selection technique has been designed by researchers in bioinformatics. The input dataset used in proposed work is said to be very large and all sequences generated by sliding window technique will not be able to produce significant motifs. Hence, for the first time in this proposed work SVD Entropy segment selection technique is been combined with FCM granular method to select important motifs that transcend in different protein families. In this proposed work, SVD Entropy segment selection method helps us to eliminate insignificant segments from large input dataset. Selecting significant segments before applying any clustering technique will help in reduction of computational time to generate significant motifs. The survived segments are clustered using Fuzzy C-Means clustering and on each granule benchmark K-Means clustering is performed. Finally we collect information from all granules to generate final motifs. Comparative results of different clustering technique shows that the proposed FCM granular with SVD – Entropy technique able to identify more number of hidden motifs in protein families. Future work aims to apply different types of clustering algorithm.

REFERENCES

[1]. O.Alter, P.O Brown, D.Boststein, "Singular value decomposition for genome-wide expression data preprocessing and modelling", PNAS, vol. 97, No.18, pp. 10101-10106, 2000.

[2]. T K Attwood, M E Beck, A J Bleasby, K Degtyarenko , DJP Smityh : Progress with the PRINTS protein fingerprint database. Nucleic Acids Res 1996, 24:182-183.

[3]. B.Chen, P.C Tai, R.Harrision and Y.Pan, "FIK Model: Novel Efficient Granular Computing Model for Protein Sequence Motifs and Structure Information Discovery", in IEEE proc, 6th symposium on Bioinformatics and BioEngineering (BIBE), Washington DC, 2006, pp. 20-26.

[4]. B.Chen, P.C Tai, R.Harrison and Y.Pan, "FGK Model: An Efficient Granular Computing Model for Protein Sequence Motifs Information Discovery", in IASTED proc. International conference on Computational and Systems Biology(CASB), Dallas 2006,pp 56-61.

[5]. B.Chen, P.C Tai, R.Harrison and Y.Pan, "Super GSVM-FE model for protein Sequence Motif Information Extraction", in proc.IEEE symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2007, pp. 317-322.

[6]. D.L Davies, and D.W Buldin, "A cluster separation measure",IEEE Trans. Pattern Recogn. Machine Intell., 1,224-227,1979.

[7]. David W.Mount, Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, New York, 2001.

[8]. E.Eskin and P.A Pevzner, "Finding composite regulatory pattern in DNA sequences", Bioinformatics, 18(Suppl.1)354-363, 2002.

[9]. K.F Han and D.Baker, "Recurring local sequence motifs in proteins", J.Mol.Bio, vol. 251, No. 1, pp. 176-187, 1995.

[10]. S.Henikoff, J.G.Henikoff and S.Pietrokovski, "Blocks+: a non redundant database of protein Alignment blocks derived from multiple compilation", Bioinformatics, vol.15, no.6, pp.417-479, 1999.

[11]. N.Hullo, C.J.A Sigrist, V.Le Saux, P.S Langendijk-Genevaux, L.Bordoli, A.Gattiker, E. De Castro, P. Bucher, and A. Bairoch, "Recent improvements to the PROSITE database", Nucleic Acids Res, vol. 32, Database issue: D134-137, 2004.

[12]. Jain A. K. Murthy M. N. Flynn P. J., "Data clustering: Areview", ACM Computing Surveys, pp. 265-323, Vol. 31(3), 1999.

[13]. W.Kabsch and C.Sander, "Dictionary of protein secondary structure pattern recognition of hydrogen-bonded and geometrical features", Biopolymers,vol 22,pp.2577-2637,1983.

[14]. Lin, T.Y. 'Data mining and machine oriented modeling: a granular computing approach', Journal of Applied Intelligence, Kluwer, Vol. 13, No. 2, pp.113–124, 2002.

[15]. Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006.

[16]. C.Sander and R.Schneider, "Database of Homology-derived protein structures and the structural meaning of sequence alignment", Proteins: Struct.Funct. Genet., vol. 9, No. 1, pp. 56-68, 1991.

[17]. C.Sander and R.Schneider, "Database of similarity derived protein structures and the structural meaning of sequence alignment, "Proteins: Struct. Funct. Gent., vol. 9,no.1,pp.56-68,1991.

[18]. G.Wang and R.L Dunbrack,Jr., "PISCES: a protein sequence culling server", Bioinformatics,vol,19,no.12,pp.1589-1591,2003.

[19]. J Weston , F Pérez-Cruz, O Bousquet, O Chapelle, A Elisseeff, B Schölkopf : Feature Selection and Transduction for Prediction of Molecular Bioactivity for Drug Design. Bioinformatics 2002, **1:**1-8.

[20]. W.Zhong, G.Altun, R.Harrison, P.C Tai and Yi Pan, "Improved K-Means Clustering algorithm for Exploring Local Protein Sequence motifs Representing Common Structural Property", IEEE transactions on Nanobioscience, Vol. 4, No.3, pp. 255-265, 2005.