International Journal of Computational Intelligence and Informatics, Vol. 5: No. 2, September 2015

Exploring Highly Structure Similar Protein Sequence Motifs using SVD with Soft Granular Computing Models

E Elayaraja
Periyar University
Salem, Tamilnadu, India
elayarajaphd.e@gmail.com

K Thangavel
Periyar University
Salem, Tamilnadu, India
drktvelu@yahoo.com

***Abstract*- Vital areas in Bioinformatics research is one of the Protein sequence analysis. Protein sequence motifs are determining the structure, function, and activities of the particular protein. The main objective of this paper is to obtain protein sequence motifs which are universally conserved across protein family boundaries. In this research, the input dataset is extremely large. Hence, an efficient technique is demanded. A Rough Granular computing model is created to efficiently extracting protein motif data that transcends protein families. Before apply this model, the very first step of this research is trying to reduce segments. The literature suggests that the Singular Value Decomposition (SVD) computing technique is more suited for reducing segments. After that the reduced segments are followed by applying Rough Granular computing model. The effectiveness of final results effectiveness is tested by several measures. The experimental results suggest that the SVD with Rough Granular computing model generates more number of highly structured motif patterns.**

Keywords- Protein Sequence Motifs, DBI, DI, HSSP-BLOSUM62, Granular Computing, K-Means, Adaptive Fuzzy C-Means, Rough K-Means.

I. INTRODUCTION

A thick relationship between protein sequence and its structure plays a vital role in current bioinformatics research. The biological term 'sequence motif' denotes a relatively, functionally or structurally conserved sequence patterns that occur repeatedly in a group of related proteins [12]. These motif patterns may be able to predict the structural or functional area of other proteins, such as enzyme-binding sites, DNA or RNA binding sites, prosthetic attachment sites, or regions involved in binding other small molecules.

PROSITE [1], PRINTS [2], BLOCKS [3], SBASE [4], and PFAM [5] are five popular databases for sequence motifs. There are some commonly used softwares for protein sequence motif discover including MEME [6], Gibbs Sampling [7, 8], Block Maker [9] and some of the latest algorithms include MITRA [10], and Gemoda [11]. These applications, endure a common issue of limiting the size of input dataset. Several protein sequences are required to be input by the user while using these tools.

In this research, protein sequences are converted into segments using sliding window concepts and patterns are extracted from the selected segments. The total sliding sequence segments are trim by Singular Value Decomposition (SVD) [13].These sliding sequence segments are separated into different groups with granular computing models that utilized Fuzzy C-Means, Adaptive Fuzzy C-Means and Rough K-Means clustering algorithms to divide the set of segments into several smaller subsets and then apply K-Means and Rough K-Means algorithm to each subset to discover relevant information. Finally, we merge the information generated by all granules and obtain the final sequence motif information. Various evaluation methods are applied in this study such as structural similarity, Dunn Index (DI) measure, Davis-Bouldin Index (DBI) measure, and HSSP-BLOSUM62 evaluation method. The hybridization of the SVD with Rough Granular computing model generates more number of highly structured motif patterns.

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. In section 4, the description of granular computing techniques and clustering algorithms are explained. Experimental setup is explained in section 5. In section 6, experimental results are explained. Section 7 concludes the paper with directions for further enhancement.

ISSN: 2349-6363

II. RELATED WORKS

K-Means clustering algorithm with random initial centroids is utilized by Han et al. [14] to find recurring protein sequence motifs across the boundaries of a protein family. To overcome the inherent problem of K-Means clustering algorithm, Wei et al. proposed an improved K-Means clustering algorithm to obtain initial centroid locations more wisely [15] and the results published by Wei et al. have been improved in their experiment.

Bernard Chen et al. proposed a granular computing model work called FIK model [16, 17] for overcome the high computational cost, which utilizes a Fuzzy C-Means clustering algorithm to divide the whole data space into several smaller subsets and then applies a standard improved K-Means algorithm to each subset to discover relevant information. In FGK model [16, 17] Bernard Chen et al. develop a new greedy K-Means algorithm to further improve secondary structural similarity sequence motifs. In the Greedy K-Means, the best centroids are selected after five runs of K-Means and then K-Means algorithm is executed by considering those centroids. 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 [18] 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 addresses the problem of selecting the significant segments in the area of protein sequence motif identification [13, 32]. The city block metric is used for calculating the difference between a sequence segment and the centroid of a given sequence cluster. The formula for calculating entropy each sequence segment is given here under.

segment is given here under.

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

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

The resulting SVD- Entropy is as follows

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

1) E < m + n, features with high contribution.
2) m + n > E > m - n, features with average contribution.
3) E < m - n, features with negative contribution.

The segments obtained in the first group are said to relevant to our problem. The segments in the second group are said to be neutral and the third group segments will reduce total SVD entropy. In this work, we have selected only those segments which fall under the first category.

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)
    Let K is the number of non zero SVD entries along with window size
    For j varies from 1 to K
      Apply SVD Entropy using (2)
    End For
End For
```

Step2: Selection of Sequence segments

```
If (entropy of each sequence segment < threshold value) then
        Select those sequence segments for clustering process
Else
        Eliminate the segments from clustering process
End If
```

Figure 1. SVD Entropy Segment Selection Algorithm

Fig. 1 shows SVD Entropy Selection algorithm applied in Fuzzy Granular Model (FGM), Adaptive Fuzzy Granular Model (AFGM) and Rough Granular Model (RGM). The motif information obtained after the segment selection process is said to be more meaningful as well as DBI value considerably decreased after the feature selection process.

IV. GRANULAR COMPUTING TECHNIQUES

A. Fuzzy Granular Model with SVD Entropy

This computation work consists of two phases. Phase one selects significant segments using SVD-Entropy method. Phase two adopts FGM computing technique. The SVD-Entropy has been combined with FGM to identify hidden motif patterns that are available in different protein families. As the dataset is very large, hence the work focuses on segment selection technique to be applied before granular computing which helps us to reduce computational cost. Traditional K-Means [20] and Rough K-Means Clustering algorithms are 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. The Figures 2 and 3 show the structure of FGM using K-Means and FGM using Rough K-Means respectively.

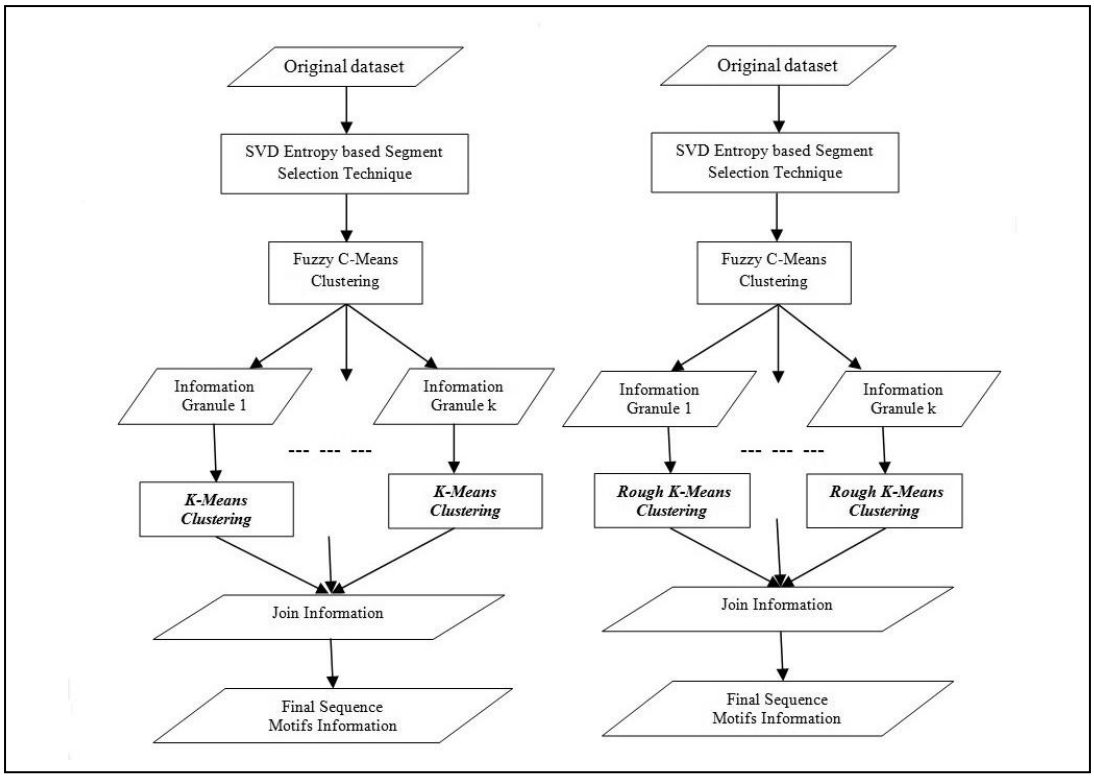


Figure 2. Sketch of FGM using K-Means Computing Model with SVD Entropy

Figure 3. Sketch of FGM using K-Means Computing Model with SVD Entropy

B. Fuzzy C-Means

Fuzzy C-Means (FCM) is a clustering algorithm which allows one segment of data is belongs to one or more clusters. This algorithm is to minimize the following objective function [16]:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\|x_i - c_j\right\|^2, 1 \le m \infty \quad (3)$$

where m, the fuzzification factor, is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x is the i [th] of d-dimensional measured data, c 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 u_{ij} and the cluster centers c_j by:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (4)$$

where

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \quad (5)$$

This iteration will stop when $\max_{ij}\left\{\left|U^{(k+1)} - U^{(k)}\right|\right\} < \delta$ where δ is a termination criterion between 0 and 1, whereas k is the iteration step. This procedure converges to a local minimum or a saddle point of J_m.

The Fuzzy C-Means Clustering algorithm is described as following:

---

i. Initialize membership function matrix U = [u_{ij}], and U (0).

ii. at k step: Calculate the centroid point by the equation (4)

iii. Update $U^{(k)}$and $U^{(k+1)}$by using equation (5).

iv. if $\left|U^{(k+1)} - U^{k}\right| < \mathcal{E}$ then stop; otherwise return to step 2.

C. Adaptive Fuzzy Granular Model with SVD Entropy

The SVD-Entropy has been combined with AFGM to identify more hidden motif patterns. Traditional K-Means and Rough K-Means Clustering algorithms are performed on each information granule generated by AFCM. At the final stage, we combine information generated by all granules and obtain final sequence motif information. The Figures 4 and 5 show the structure of AFGM with SVD Entropy [21, 22].

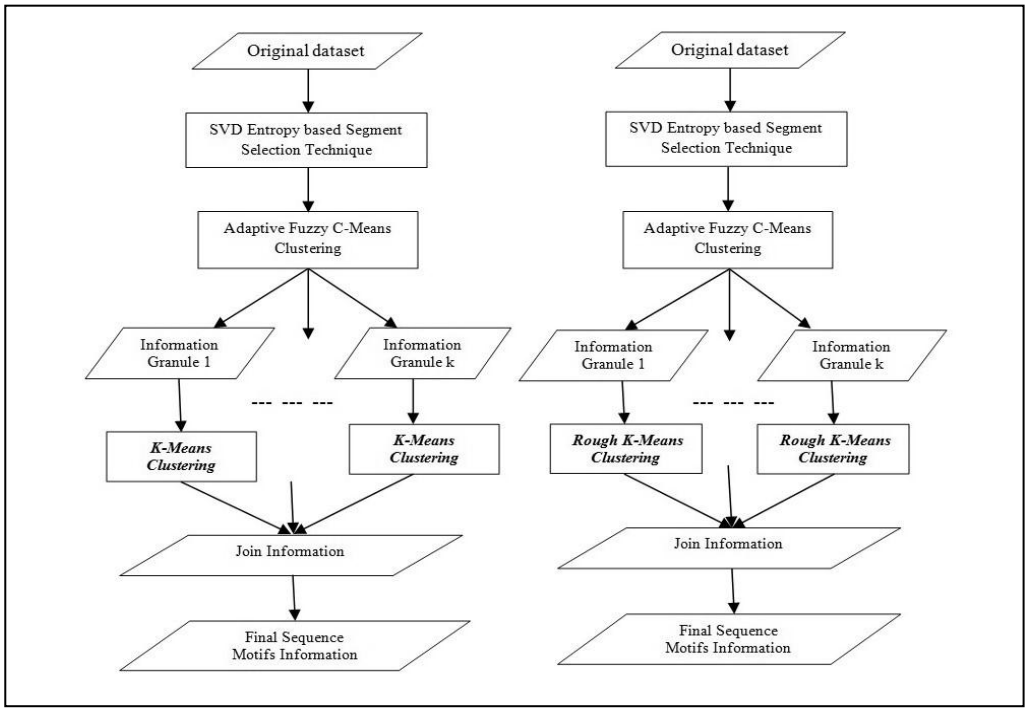


Figure 4. Sketch of AFGM using K-Means Computing Model with SVD Entropy

Figure 5. Sketch of AFGM using K-Means Computing Model with SVD Entropy

Many of the behavioural problems with standard Fuzzy C-Means algorithm are eliminated when we relax probabilistic constraint imposed on membership function. Further Krishnapuram R and Keller JM [21, 33] have modified the approach for calculating membership values. Equation (6) shows membership calculation.

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\mu_{j\,(x_i)=n} \tag{6}$$

Here,

μ_j (x_i) is the membership of x_i in j^{th} cluster

k is the specified number of clusters

n is the number of data points

In Adaptive Fuzzy C-Means (AFCM), the total membership quantifiers for all sample points are equal to n. This flexible approach leads to clustering optimization problem, provides a way to improve cluster robustness. Here the algorithm is adaptive; that is membership is based on sample size rather than fixed to upper limit as one in Fuzzy C-Means clustering. The membership values in this method are calculated using Equation (7)

$$\mu_j(x_i) = \frac{n\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\sum_{z=1}^{n}\left(\frac{1}{d_{kz}}\right)^{\frac{1}{m-1}}} \quad (7)$$

The Adaptive fuzzy clustering algorithm is efficient in handling data with outlier points. It gives very low membership values for outliers since the sum of distances of points in all the clusters involves in membership calculation.

D. Rough Granular Model with SVD Entropy

A set of information granules is built using the Rough Granular Model (RGM) with SVD entropy and then applying K-Means and Rough K-Means Clustering algorithms to obtain the final information. The RGM process is given below in Fig. 6 and Fig. 7 [21, 22].

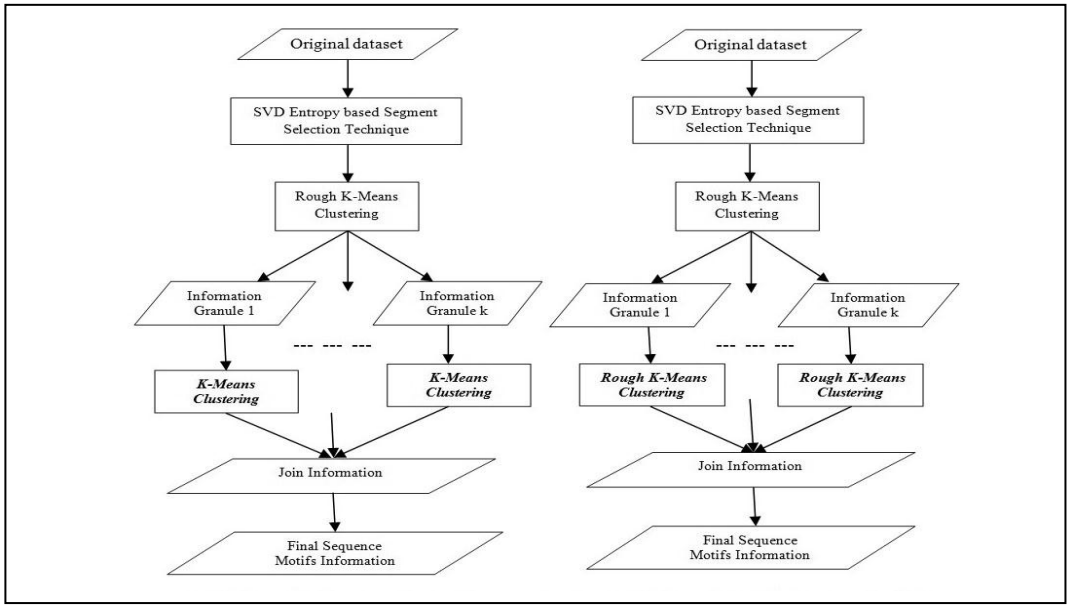


Figure 6. Sketch of RGM using K-Means Computing Model with SVD Entropy

Figure 7. Sketch of RGM using K-Means Computing Model with SVD Entropy

E. Rough Clustering

In rough clustering each cluster has two approximations, a lower and an upper approximation. The lower approximation is a subset of the upper approximation. The members of the lower approximation belong certainly to the cluster; therefore they cannot belong to any other cluster. The data objects in an upper approximation may belong to the cluster. Since their membership is uncertain they must be a member of an upper approximation of at least another cluster.

F. Properties for the Rough Clustering Algorithm

Property 1: a data object can be a member of one lower approximation at most.

Property 2: a data object that is a member of the lower approximation of a cluster is also member of the upper approximation of the same cluster.

Property 3: a data object that does not belong to any lower approximation is member of at least two upper approximations [23].

The Rough K-Means algorithm provides a rough set theoretic flavour to the conventional K-Means algorithm to deal with uncertainty involved in cluster analysis. The Rough K-Means algorithm [24, 25] described as follows:

---

1. Select initial clusters of n objects into K clusters.
2. Assign each object to the Lower bound (L(x)) or upper bound (U(x)) of cluster/ clusters respectively as: For each object v, let d (v,x_i) be the distance between itself and the centroid of cluster x_i. The difference between d (v,x_i) / d(v,x_j), $1 \leq i, j \leq k$ is used to determine the membership of v as follows:
   - If d (v,x_i) / d$(v,x_j) \leq$ thershold, then $v \in U(x_i)$ & $v \in U(x_j)$. Furthermore, v will not be a part of any

lower bound.

- Otherwise, $v \in L(x_i)$, such that $d(v,x_i)$ is the minimum for $1 \leq i \leq k$. In addition, $v \in U(x_i)$.

3. For each cluster x_i re-compute center according to the following equations the weighted combination of the data points in its lower_bound and upper_bound.

$$x_i = \begin{cases} w_{lower} \times \frac{\sum_{v \in L(x)} v_j}{|L(x)|} + w_{upper} \times \frac{\sum_{v \in U(x)-L(x)} v_j}{|U(x)-L(x)|} & \text{if } |U(x)-L(x) \neq \phi| \\ w_{lower} \times \frac{\sum_{v \in L(x)} v_j}{|L(x)|} & \text{otherwise} \end{cases}$$

where $1 \leq j \leq k$. The parameters w_{lower} and w_{upper} correspond to the relative importance of lower and upper bounds. If convergence criterion is met, i.e. cluster centers are same to those in previous iteration, then stop; else go to step2.

G. K-Means Clustering Algorithm

Among all clustering algorithms, K-Means clustering algorithm has the advantages of easy interpretation and implementation, high scalability, and low computation complexity. The K-Means clustering take the user input parameter K, and partitions a set of n objects into K clusters then iteratively updates the centers until no reassignment of patterns to new cluster centers occurs. In every step, each sample is allocated to its closest cluster center and cluster centers are reevaluated based on current cluster memberships [26].

V. EXPERIMENTAL SETUP

A. Data Set

The dataset obtained from Protein Sequence Culling Server (PISCES) includes 4946 protein sequences [27]. 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. Homology Derived Secondary Structure of Proteins (HSSP) frequency profiles is used to represent each segment [4, 5]. 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. Fig. 8 shows sliding window technique. In this sliding window technique we can generate n number of sequence segments (10 X 20 matrices).

Dictionary of Secondary Structure of Proteins (DSSP) assigns secondary structure to eight different classes [28]. These eight structural classes can be reduced to three using reduction method as follows: H, G and I to H (Helices); B and E to E (Sheets); all others to C (Coils) [29].

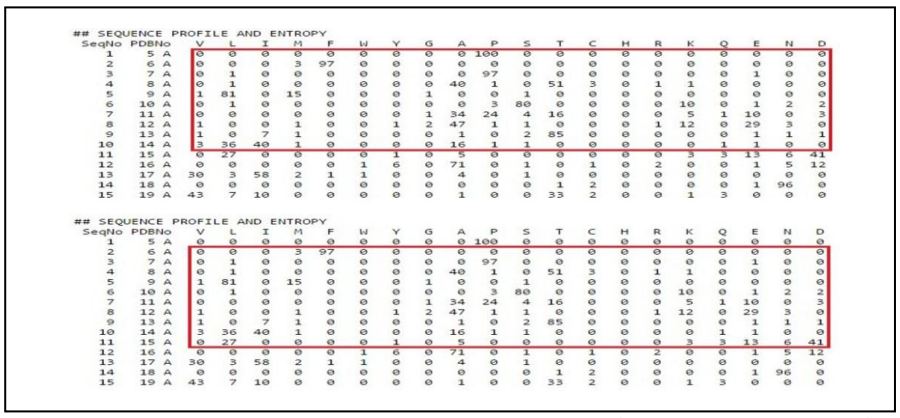


Figure 8. Sliding Window techniques with a window size of 10 applied on 3CA8 HSSP file

B. Structural Similarity Measures

A cluster's average structure is calculated using the following formula:

$$\frac{\sum_{i=1}^{WS} \max(P_{i,H}, P_{i,E}, P_{i,C})}{WS}$$

where WS is the window size and $(P_{i,H})$shows the frequency of occurrence of helix among the segments for the cluster in position i. $(P_{i,E})$ and $(P_{i,C})$ are defined in a similar way. If the structural homology for a cluster exceeds 70%, the cluster can be considered structurally identical [16]. If the structural homology for the cluster exceeds 60% and is below 70%, the cluster can be considered weakly structurally homologous.

C. Distance Measure

The city block metric is more suitable for this field of study since it will consider every position of the frequency profile equally. The city block metric is used for calculating the difference between a sequence segment and the centroid of a given sequence cluster. Han and Baker also chose the city block metric because of complications associated with the use of Euclidean metric for clustering algorithms [14]. The following formula is used to calculate the distance between two sequence segments:

$$\text{Distance} = \sum_{i=1}^{WS} \sum_{j=1}^{N} |F_k(i,j) - F_c(i,j)|$$

where WS is the window size and N is 20 which represent 20 different amino acids. F_k (i, j) is the value of the matrix at row i and column j used to represent the sequence segment. F_c (i, j) is the value of the matrix at row i and column j used to represent the centroid of a give sequence cluster.

D. Dunn Index Measure

The Dunn Index (DI) also favours clustering with low intra-cluster and high inter-cluster distances, although the compactness of the clusters is assessed in a different way [42]. This index is originally proposed to use at the identification of "compact and well separated clusters". So the result of the clustering has to be recalculated as it was a hard partition algorithm. Dunn Index was proposed by J. C. Dunn in 1974. Similar to the DBI index, the DI index measures the quality of clustering result. The goal of DI index is the same with other cluster validity indexes which tries to find a good intra cluster and inter cluster relationships. It is used to measure the goodness of a clustering structure without respect to external information. The Dunn Index has a value between zero and ∞, it should be maximized.

The DI index is defined as follows [43]:

$$D_{nc} = \min_{i=1,\ldots,nc} \left\{ \min_{j=i+1,\ldots,nc} \left(\frac{d(c_i,c_j)}{\max_{k=1,\ldots,nc} diam(c_k)} \right) \right\}$$

where $d(c_i,c_j)$ the dissimilarity function between two clusters is $c_i, and\ c_j$ defined as

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_j} d(x,y)$$

$diam(c)$ is the diameter of a cluster, which may be considered as a measure of dispersion of the clusters. The diameter of a cluster c can be defined as follows:

$$diam\ (c) = \max_{x,y \in C} d(x,y)$$

It is clear that if the dataset contains compact and well-separated clusters, the distance between the clusters is expected to be large and the diameter of the cluster is expected to be small. Thus, based on the Dunn's index definition, we may conclude that large values of the index indicate the presence of compact and well-separated clusters.

The index D_{nc} does not exhibit any trend with respect to number of clusters. Thus, the maximum in the plot of D_{nc} versus the number of clusters can be an indication of the number of clusters that fits the data.

The implications of the Dunn index are:

- The considerable amount of time required for its computation.
- The sensitive to the presence of noise in datasets.

Since these are likely to increase the values of $diam\ (c)$.

E. Davis-Bouldin Index (DBI) Measure

The DBI measure [17] is a function of the inter-cluster and intra-cluster distance. A good cluster result should reflect a relatively large inter-cluster distance and a relatively small intra-cluster distance. The DBI measure combines both distance information into one function, which is defined as follows:

$$DBI = \frac{1}{k}\sum_{p=1}^{k}\max_{p\neq q}\left\{\frac{d_{intra}\left(C_p\right)+d_{intra}\left(C_q\right)}{d_{inter}\left(C_p,C_q\right)}\right\}, where$$

$$d_{intra}\left(C_p\right)=\frac{\sum_{i=1}^{n_p}\left\|g_i-g_{pc}\right\|}{n_p} \quad and$$

$$d_{inter}\left(C_p,C_q\right)=\left\|g_{pc}-g_{qc}\right\|$$

K is the total number of clusters, d_{intra} and d_{inter} denote the intra- cluster and inter-cluster distances respectively. n_p is the number of members in the cluster C_p. The intra-cluster distance defined as the average of all pair wise distances between the members in cluster P and cluster P's centroid g_{pc}. The inter-cluster distance of two clusters is computed by the distance between two clusters' centroids. The lower DBI value indicates the high quality of the cluster result.

F. HSSP-BLOSUM62 Measure

BLOSUM62 [30] (Fig. 9.) is a scoring matrix based on known alignments of diverse Sequences.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Figure 9. BLOSUM62 Matrix

By using this matrix, we may access the consistency of the amino acids appearing in the same position of the motif information generated by our method. Because different amino acids appearing in the same position should be close to each other, the corresponding value in the BLOSUM62 matrix will give a positive value. Hence, the measure is defined as the following [31]:

If k=0: HSSP-BLOSUM62 measure = 0

Else If k=1:

If $HSSP_i > 10\%$: HSSP-BLOSUM62 measure = $BLOSUM62_{ii}$

If $8\% \leq HSSP_i < 10\%$: HSSP-BLOSUM62 measure = $\frac{1}{2}BLOSUM62_{ii}$

Else: HSSP-BLOSUM62 measure = $\dfrac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} HSSP_i.HSSP_j.BLOSUM62_{ij}}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} HSSP_i.HSSP_j}$

G. 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, the fuzzification factor is 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 FGM results, the membership threshold is set to 18% [32]. The function that decides how many numbers of clusters should be in each information granule is given below:

$$C_k = \frac{n_k}{\sum_{i=1}^{m} n_i} \times m$$

where C_k denotes the number of clusters assigned to information granule k, n_k is the number of members belonging to information granule k, m is the number of clusters in Fuzzy C-Means. In this technique we are able to indentify 900 clusters.

For Adaptive Fuzzy C-Means, the fuzzification factor is considered as 1.15 and membership threshold is set to 13% [32]. The number of clusters in each granule is decided by the function given below:

$$C_k = \frac{n_k}{\sum_{i=1}^{m} n_i} \times m$$

where C_k denotes the number of clusters assigned to information granule k, n_k is the number of members belonging to information granule k, m is the number of clusters in Adaptive Fuzzy C-Means. In this technique we are able to indentify 901 clusters.

For Rough K-Means, the epsilon value is considered as 1.001 and the number of clusters in each granule is been decided by the function given below:

$$C_k = \frac{n_k}{\sum_{i=1}^{m} n_i} \times m$$

where C_k denotes the number of clusters assigned to information granule k, n_k is the number of members belonging to information granule k, m is the number of clusters in Rough K-Means. In this technique we are able to indentify 900 clusters.

VI. EXPERIMENTAL RESULTS

TABLE I. SUMMARY OF THE RESULTS OBTAINED BY THE FCM

Granules	Number of Members	Number of Clusters	Data Size (in MB)
Granule 1	76090	85	5.06
Granule 2	39915	45	2.48
Granule 3	60151	45	3.58
Granule 4	265960	297	11.10
Granule 5	120024	134	7.44
Granule 6	23348	26	1.36
Granule 7	9612	11	0.49
Granule 8	151631	169	8.39
Granule 9	45472	51	3.03
Granule 10	13666	15	0.64
Total	805869	900	43.57
Original Data Set	660364	900	19.20

The summary of the results obtained from FCM granular method is shown in Table I. Although the total segment increased from 660364 to 805869, we achieved the goal of reduced data size is to deal with one information granule at a time [22].

The summary of the results obtained from FCM granular method with SVD entropy is shown in Table II. The total number of segments are slight increased, but we achieved the goal of reduced data size is to deal with one information granule at a time.

TABLE II. SUMMARY OF THE RESULTS OBTAINED BY THE SVD-FCM

Granules	Number of Members	Number of Clusters	Data Size (in MB)
Granule 1	24412	32	1.68
Granule 2	100385	131	6.27
Granule 3	44428	58	3.30
Granule 4	98815	130	6.52
Granule 5	41557	54	2.58
Granule 6	33376	44	2.32
Granule 7	67448	88	4.45
Granule 8	133945	176	7.47
Granule 9	42674	56	3.00
Granule 10	99339	130	6.24
Total	686379	899	43.83
Original Data Set	565314	900	17.70

TABLE III. SUMMARY OF THE RESULTS OBTAINED BY THE AFCM

Granules	Number of Members	Number of Clusters	Data Size (in MB)
Granule 1	20675	28	1.74
Granule 2	35324	48	2.65
Granule 3	215674	292	8.98
Granule 4	62388	85	3.78
Granule 5	4376	6	0.38
Granule 6	125769	170	6.34
Granule 7	2409	3	0.23
Granule 8	65409	89	4.14
Granule 9	2824	4	0.22
Granule 10	129761	176	6.47
Total	664609	901	34.93
Original Data Set	660364	900	19.20

The summary of the results obtained from AFCM granular method is shown in Table III. Although the total number of members increased from 562745 to 721390, we only deal with one information granule at a time. Therefore, we achieved the goal of reduced space-complexity [22].

TABLE IV. SUMMARY OF THE RESULTS OBTAINED BY THE SVD-AFCM

Granules	Number of Members	Number of Clusters	Data Size (in MB)
Granule 1	221515	290	9.40
Granule 2	3478	5	0.28
Granule 3	42792	56	3.11
Granule 4	129507	170	6.85
Granule 5	97204	127	5.81
Granule 6	25899	34	2.15
Granule 7	49103	64	3.44
Granule 8	102852	135	5.95
Granule 9	4615	6	0.37
Granule 10	9587	13	0.82
Total	686552	900	38.18
Original Data Set	660364	900	17.70

The summary of the results obtained from AFCM granular method with SVD entropy is shown in Table IV. Although the total number of members increased at 686552, we only deal with one information granule at a time. Hence, we achieved the goal of reduced space-complexity with more number of highly structure motif patterns.

The summary of the results obtained from RKM granular method is shown in Table V. The total number of members is exactly same as original data set but identifies more number of hidden highly structure motif patterns.

TABLE V. SUMMARY OF THE RESULTS OBTAINED BY THE RKM

Granules	Number of Members	Number of Clusters	Data Size (in MB)
Granule 1	122260	167	7.37
Granule 2	11112	15	0.967
Granule 3	6794	9	0.591
Granule 4	7552	10	0.675
Granule 5	167789	229	9.00
Granule 6	3369	5	0.319
Granule 7	44961	61	2.56
Granule 8	143504	196	7.95
Granule 9	37645	51	2.19
Granule 10	115378	157	6.77
Total	660364	900	38.392
Original Data Set	660364	900	19.20

TABLE VI. SUMMARY OF THE RESULTS OBTAINED BY THE SVD-RKM

Granules	Number of Members	Number of Clusters	Data Size (in MB)
Granule 1	80341	128	5.55
Granule 2	21425	34	1.48
Granule 3	77671	124	5.28
Granule 4	54727	87	3.89
Granule 5	43451	69	2.60
Granule 6	53482	85	3.77
Granule 7	60673	97	4.13
Granule 8	45012	72	2.96
Granule 9	66865	107	4.72
Granule 10	61623	98	4.17
Total	565270	901	38.55
Original Data Set	660364	900	17.70

The summary of the results obtained from RKM granular method with SVD entropy is shown in Table VI. The total number of members is smaller than original data set but identifies more number of hidden highly structure motif patterns.

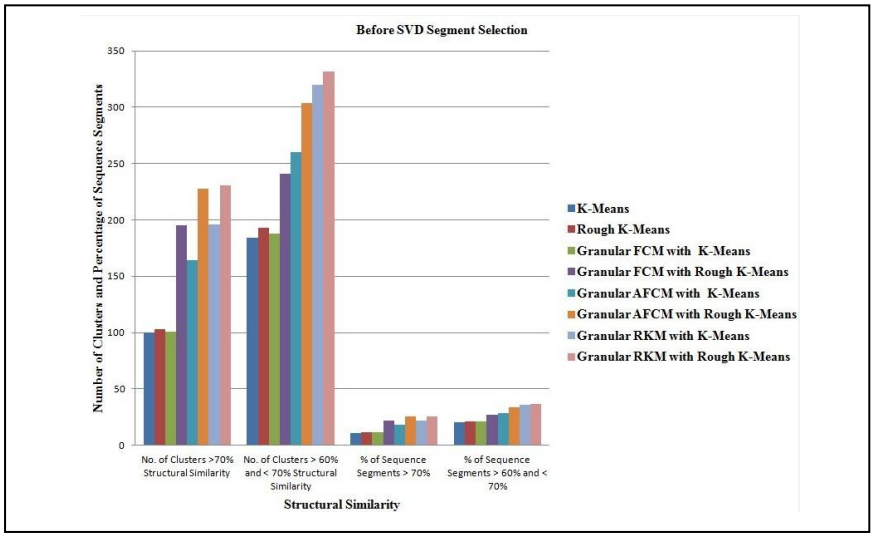


Figure 10. BLOSUM62 Matrix

Fig. 10 has been interpreted from table VII. From the Fig. 9 we state that the number of strong and weak clusters have been increased in Granular RKM with Rough K-Means technique as well as percentage of sequence segments have also been increased considerably.

TABLE VII. COMPARISON RESULTS OF DIFFERENT ALGORITHMS

	Before SVD Segment Selection							
	K-Means	Rough K-Means	Granular FCM with K-Means	Granular FCM with Rough K-Means	Granular AFCM with K-Means	Granular AFCM with Rough K-Means	Granular RKM with K-Means	**Granular RKM with Rough K-Means**
No. of Clusters >70% Structural Similarity	100	103	101	195	164	228	196	**231**
No. of Clusters > 60% and < 70% Structural Similarity	184	193	188	241	260	304	320	**332**
% of Sequence Segments > 70%	11.11	11.44	11.22	21.67	18.20	25.31	21.78	**25.67**
% of Sequence Segments > 60% and < 70%	20.44	21.44	20.89	26.78	28.86	33.74	35.56	**36.89**
DI Measure	0.1978	0.2104	0.2553	0.2652	0.2285	0.2753	0.2251	**0.3142**
DBI Measure	6.2409	6.1985	4.2163	3.7339	3.9268	3.6186	3.8721	**3.6005**
Avg. HSSP-BLOSUM62	0.5268	0.6010	0.6125	0.6617	0.7325	0.7901	0.8125	**0.8227**

Table VII shows the comparative results obtained from different algorithms and granularization methods. From the table VII, we can infer that RGM with Rough K-Means method able to identify more number of hidden motif patterns.

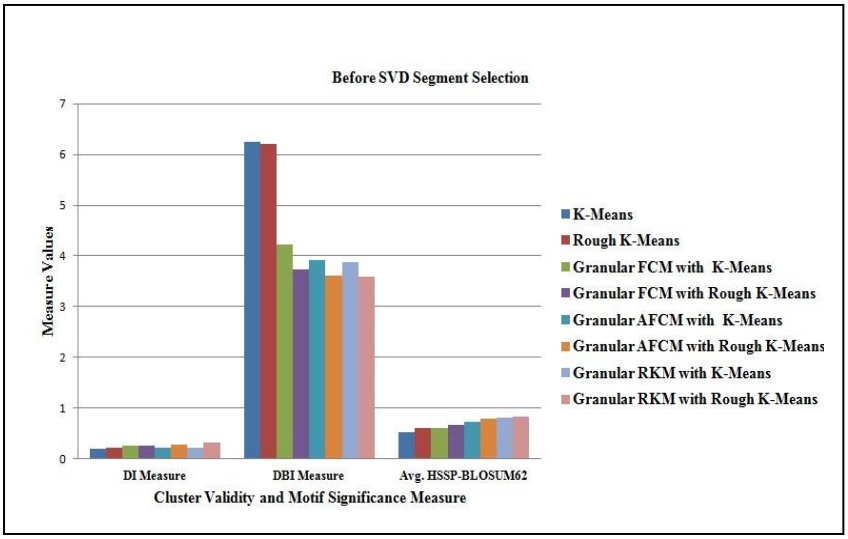


Figure 11. Comparison of DBI, DI and BLOSUM62 measure values

Fig. 11 shows DBI, DI and HSSP-BLOSUM62 measure values obtained from different methods and different granular computing techniques.

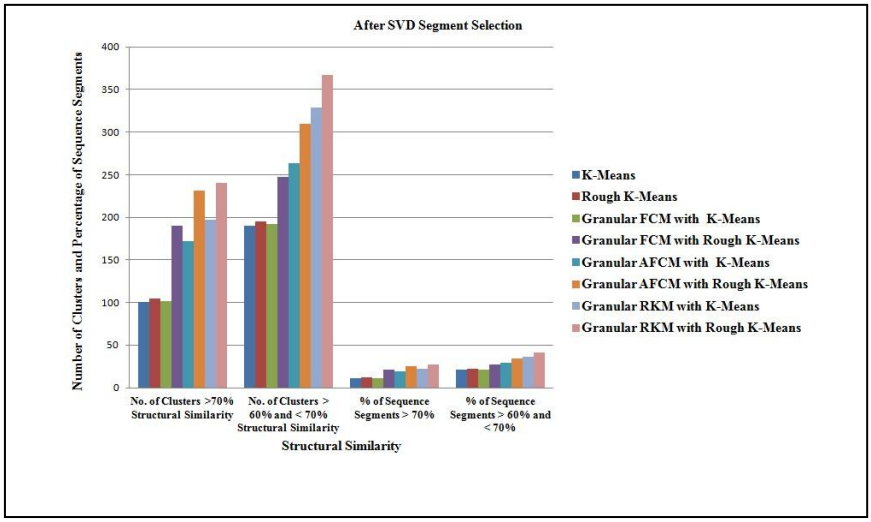


Figure 12. Comparison of Structural Similarity Values

Fig. 12 shows percentage of structural similarity belonging to clusters obtained from different methods and different granular computing techniques. Fig. 11 has been interpreted from table VIII. From the Fig. 11, we state that the number of strong and weak clusters have been increased in RGM with SVD entropy along with Rough K-Means.

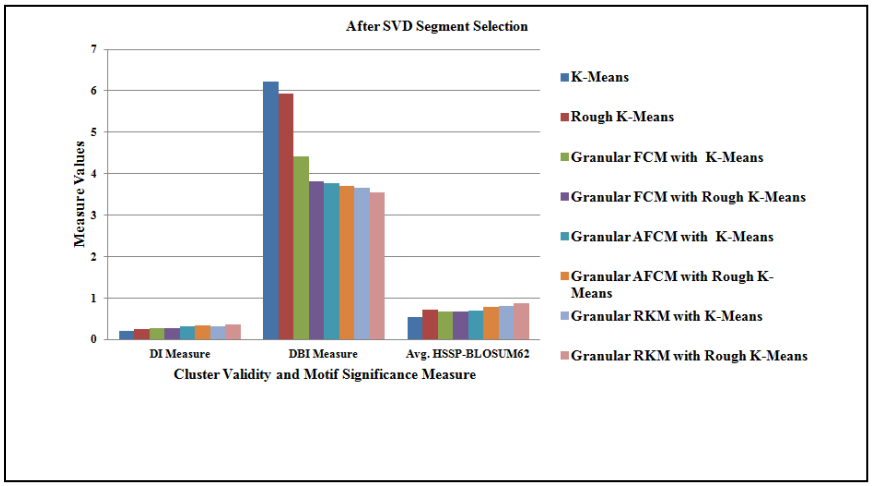


Figure 13. Comparison of DBI, DI and BLOSUM62 measure values

Fig. 13 shows DBI and HSSP-BLOSUM62 measure values obtained from different methods and different granular computing techniques. The low DBI measure and high HSSP-BLOSUM62 values indicate the improvement of the quality of clusters achieved by RGM with SVD entropy along with Rough K-Means technique.

TABLE VIII. COMPARISON RESULTS OF DIFFERENT ALGORITHMS

	After SVD Segment Selection							
	K-Means	Rough K-Means	Granular FCM with K-Means	Granular FCM with Rough K-Means	Granular AFCM with K-Means	Granular AFCM with Rough K-Means	Granular RKM with K-Means	**Granular RKM with Rough K-Means**
No. of Clusters >70% Structural Similarity	101	105	102	190	172	231	197	**240**
No. of Clusters > 60% and < 70% Structural Similarity	190	195	192	247	263	310	329	**367**
% of Sequence Segments > 70%	11.22	11.67	11.33	21.11	19.09	25.64	21.89	**26.67**
% of Sequence Segments > 60% and < 70%	21.11	21.67	21.33	27.44	29.19	34.41	36.56	**40.78**
DI Measure	0.2206	0.2486	0.2669	0.2854	0.3183	0.3513	0.326	**0.3591**
DBI Measure	6.2134	5.9241	4.4162	3.8115	3.7625	3.7132	3.6513	**3.5425**
Avg. HSSP-BLOSUM62	0.5361	0.7243	0.6723	0.6824	0.6913	0.7835	0.8072	**0.8725**

A. Sequence Motifs

Four different motif patterns obtained from RGM with SVD entropy along with Rough K-Means process are shown in tables IX to XII. The following format is used for representation of each sequence motif table. Instead of using the traditional format, in this paper protein logo representation has been used [18].

TABLE IX. SHEETS-COILS MOTIF

Number of Sequence Segments:171

Structural Similarity: 73.1

	1	2	3	4	5	6	7	8	9	10
S	E	E	E	E	C	C	C	C	E	E
HP	0.86	0.65	0.72	0.33	0.26	0.24	0.20	0.22	0.26	0.64
Var	5	6	5	9	9	9	8	9	11	5

TABLE X. COILS MOTIF

Number of Sequence Segments:34

Structural Similarity: 72.65

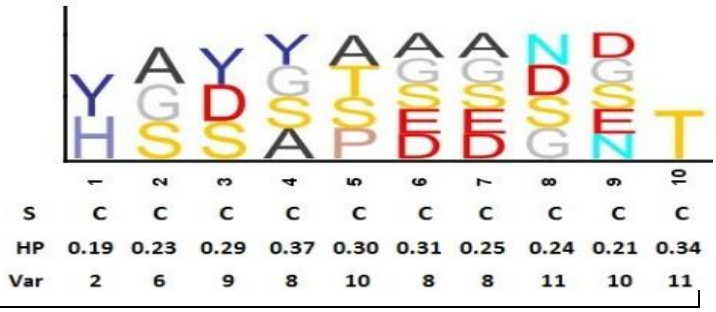


TABLE XI. HELICES MOTIF

Number of Sequence Segments:120

Structural Similarity: 73

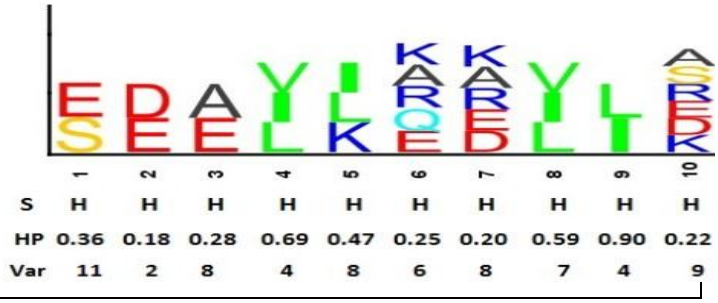
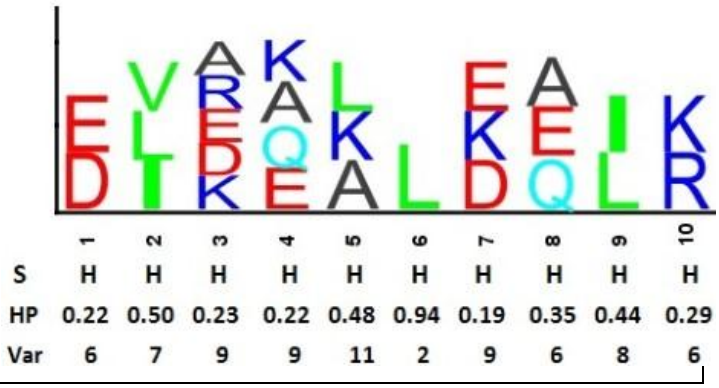


TABLE XII. HELICES MOTIF

Number of Sequence Segments:125

Structural Similarity: 70.96



- The above tables IX-XII show the number of sequence segments belonging to this motif, percentage of structural similarity. 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. The variability indicates the number of amino acids with the frequency greater than 5%.

VII. CONCLUSION

In this study, the granular computing models such as FGM, AFGM, RGM and combined these methods with SVD entropy have studied and implemented. The SVD with Rough Granular computing model generates more number of highly structured motif patterns in each granule. Further, the granules obtained in each of the above methods are clustered using K-Means and Rough K-Means. The highly structured clusters are used to construct the motif patterns. The main objective of generating more motif patterns has been achieved with the proposed SVD with rough granular approach and Rough K-Means clustering. It is believed that this SVD entropy with granular strategy make innovative ideas in bioinformatics research.

ACKNOWLEDGMENT

The second author would like to thank the presented work supported by Special Assistance Programme of University Grants Commission, New Delhi, India (Grant No. F.3-50/2011 (SAP II)).

REFERENCES

[1] N. Hulo, 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 Research*, vol. 32, pp. D134-D137, 2004. Database Issue doi:10.1093/nar/gkh044.

[2] T. K. Attwood, M. Blythe, D. R. Flower, A. Gaulton, J. E. Mabey, N. Naudling, L. McGregor, A. Mitchell, G. Moulton, K. Paine, and P. Scordis, "PRINTS and PRINTS-S shed light on protein ancestry", *Nucleic Acids Research*, vol. 30, no. 1, pp. 239-241, 2002.

[3] 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.

[4] Murval J., Gabrielian A., Fabian P., Hatsagi Z., Degtyarenko K., Hegyi H., Pongor S., "The SBASE protein domain library, release 4.0: a collection of annotated protein sequence segments", *Nucleic Acids Research*, vol. 24, no. 1, pp. 210-213, 1996.

[5] Bateman, A. et al., "The Pfam protein families database", *Nucleic Acids Research*, vol. 32, Issue D1, pp. D138-D141, 2004. Database Issue doi: 10.1093/nar/gkh121.

[6] Bailey T.L., Boden M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W. and Noble W.S., "MEME SUITE: Tools for motif discovery and searching", Nucleic Acids Research, vol. 37, Suppl. 2, pp. W202-W208, 2009. doi:10.1093/nar/gkp335.

[7] Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C., "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment", *Science*, vol. 262, Issue 5131, pp. 208-214, 1993.

[8] Bhattacharya, S., "Gibbs sampling based Bayesian analysis of mixtures with unknown number of components", Sankhya: The Indian Journal of Statistics, Series B, vol. 70-B, part 1, pp. 133-155, 2008.

[9] Henikoff, S., Henikoff, J.G., Alford, W.J. & Pietrokovski, S., "Automated construction and graphical presentation of protein blocks from unaligned sequences", *Gene*, vol. 163, Issue 2, pp. GC17-GC26, 1995.

[10] Eskin E, Pevzner P., "Finding composite regulatory patterns in DNA sequences", *Bioinformatics*, vol. 18, no. 1, pp. S354–S363, 2002.

[11] Kyle L. Jensen et al., "A Generic motif discovery algorithm for sequential data", *Bioinformatics*, vol. 22, no.1, pp. 21-28, 2006.

[12] Chen, B., Pellicer, S., Tai, P.C., Harrison, R. and Pan, Y., "Novel efficient granular computing models for protein sequence motifs and structure information discovery", *International Journal of Computational Biology and Drug Design*, vol. 2, Issue 2, pp. 168-186, 2009.

[13] M. Chitralegha, Dr K. Thangavel, "A Novel Entropy Based Segment Selection Technique for Extraction of Protein Sequence Motifs", *International Journal of Computer Science Issues (IJCSI)*, vol. 9, Issue 4, no. 3, July-2012. ISSN (Online): 1694-0814.

[14] Han K. F. and Bake D., "Recurring Local Sequence Motifs in Proteins", *Journal of Molecular Biology*, vol. 251, no. 1, pp. 176-187, 1995.

[15] Zhong, W., Altun, G., Harrison, R., Tai, P. C. & Pan, Y., "Improved K- Means clustering algorithm for exploring local protein sequence motifs representing common structural property", *NanoBioscience, IEEE Transactions*, vol. 4, no. 3, pp. 255-265, September-2005.

[16] Bernard Chen, Phang C. Tai, Robert Harrison, and Yi Pan, "FIK model: A Novel Efficient Granular Computing Model for Protein Sequence Motifs and Structure Information Discovery", *IEEE BIBE 2006*, Washington D.C., proceeding, pp. 20-26, 2006.

[17] Bernard Chen, Phang C. Tai, Robert Harrison, and Yi Pan, "FGK model: A Efficient Granular Computing Model for Protein Sequence Motifs Information Discovery", *IASTED CASB 2006*, Dallas, proceeding pp. 56-61, 2006.

[18] B.Chen, P.C Tai, R.Harrison and Y.Pan, "Super GSVM-FE model for protein Sequence Motif Information Extraction", Proceedings of the 2007 IEEE Symposium on *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Published in *IEEE xplore*, pp. 317-322, 2007.
[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*, vol. 19, pp. 764–771, 2003.
[20] Margaret H. Dunham, "Data Mining: Introductory and Advanced Concepts", *Pearson Education*, 2006.
[21] M. Chitralegha and K. Thangavel "Protein sequence motif patterns using adaptive Fuzzy C-Means granular computing model", Proceedings of the IEEE International Conference on *Pattern Recognition, Informatics and Medical Engineering (PRIME)*, Published in *IEEE xplore*, pp. 96 – 103, 2013. Print ISBN: 978-1-4673-5843-9.
[22] E. Elayaraja, K. Thangavel, M. Chitralegha, T. Chandrasekhar, "Exploring Highly Structure Similar Protein Sequence Motifs using Granular Computing Model Based on Adaptive FCM", *International Journal of Scientific and Engineering Research (IJSER)*, vol. 5, Issue 7, pp. 95-104, July-2014. ISSN (Online): 2229-5518.
[23] Peters G., "Some refinements of rough k-means clustering", *Pattern Recognition Letters*, vol. 39, no. 8, pp. 1481-1491, 2006.
[24] P. Lingras, C. West, "Interval set clustering of web users with rough K- Means", *Journal of Intelligent Information Systems(JIIS)*, vol. 23, Issue 1, pp. 5–16, 2004.
[25] P. Lingras, R. Yan, C. West, "Comparison of conventional and rough K-Means clustering", Proceedings of the 9th international conference on *Rough sets, fuzzy sets, data mining, and granular computing (RSFDGrC)*, Published in *Springer, Berlin*, vol. 2639, pp. 130–137, 2003.
[26] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review", *ACM Computing Surv*eys, vol. 31, no. 3, pp. 264–323, 1999.
[27] G. Wang and R. L. Dunbrack, Jr., "PISCES: a protein sequence-culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589-1591, 2003.
[28] 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.
[29] Cuff JA, Barton GJ., "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction", *Proteins*, vol. 34, Issue 4, pp. 508-519, 1999.
[30] Henikoff, S. and Henikoff, J. G., "Amino Acid Substitution Matrices from Protein Blocks", Proceedings of the *National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10915-10919, 1992.
[31] E. Elayaraja, K. Thangavel, M. Chitralegha, T. Chandrasekhar, "Extraction of Motif Patterns from Protein Sequences using SVD with Rough K-Means Algorithm", *International Journal of Computer Science Issues (IJCSI)*, vol. 9, Issue 6, no. 2, pp. 350-356, 2012. ISSN (Online): 1694-0814.
[32] M. Chitralegha, K. Thangavel, "Soft Granular Computing Model For Identifying Protein Sequence Motif Based On SVD-Entropy Method", *International Journal of Scientific & Engineering Research (IJSER)*, vol. 4, Issue 7, July-2013.
[33] E. Cox, Fuzzy Modelling and Genetic Algorithms for Data Mining and Exploration, San Francisco, CA: *Elsevier*, 2005. ISBN: 978-0-12-194275-5.