

# K-Harmonic Means Granular Computing Model for Protein Sequence Motif Identification

M Chitralegha  
Research Scholar  
Department of Computer Science  
Periyar University  
Salem

K Thangavel, Ph.D.  
Professor and Head  
Department of Computer Science  
Periyar University  
Salem

## ABSTRACT

Bioinformatics is concerned with creation and advancement of algorithms using techniques such as computational intelligence, applied mathematics and statistics to solve biological problems. Sequence analysis, protein structure alignment analysis and prediction, gene finding are said to be major research efforts done in the area of bioinformatics. Proteins are considered as one of the most important elements in the process of life. The activities and functions of proteins can be determined by protein sequence motifs. Identifying such motifs is one of the crucial tasks in the area of bioinformatics. In this study, Singular Value Decomposition (SVD) is adopted to select significant sequence segments and then K-Harmonic Means granular computing model is proposed to generate protein sequence motif information efficiently. Experimental result shows that K-Harmonic granular computing model outperforms K-Means granular technique.

## Keywords

Protein sequence, Motif, Clustering, HSSP-BLOSUM62, SVD

## 1. INTRODUCTION

“Bioinformatics” refers to the study of information processes in biotic systems. The field of bioinformatics involves in analysing and interpreting various types of data. These include nucleotide and amino acid sequence, protein domain and protein structure. The concept of protein domain has gained increasing interest from the biology research community because of its importance in protein classification, protein function assignment and protein engineering. Protein domains are generally considered as protein fragments of common structures which may independently fold or have their own functions. In protein sequences, motifs or patterns enclose significant homologous attributes since they correspond to conserved regions in protein families holding useful structural and functional biological information [6]. They are considered as islands of amino acids conserved in the same order of a given family. Therefore they can be seen as local feature characterizing the sequence.

The mining of sequence patterns also called ‘motifs’ is one of the most important tasks in protein sequence analysis and it continues to be an active topic of research in the area of bioinformatics. There are several databases available for sequence motifs but the most popular ones are PROSITE [9], PRINTS [2] and BLOCKS [8]. MEME, Gibbs Sampling and Block Maker 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 the 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 from HSSP file of each protein sequences using sliding window technique and the patterns are extracted from selected segments. Multiple protein sequences are represented by their corresponding HSSP file [12].

All the generated segments may not be important to yield potential motif patterns. Hence, SVD (Singular Value Decomposition) Entropy segment selection method has been adopted to select significant segments. These segments are then partitioned into small information granules using proposed granular computing method. Finally, information generated by all granules is combined to obtain final motif information.

In this paper, SVD-Entropy segment selection technique is combined with K-Harmonic granular computing technique. Benchmark K-Means algorithm is applied on the granules generated by K-Harmonic means algorithm. The main reason for using K-Means clustering instead of some other advanced clustering technology is due to extremely large input dataset. K-Means is said to be efficient than other clustering methods with time and space complexity. The objective of the proposed granular computing technique is to identify more number of hidden patterns that transcend in different protein families.

The rest of the paper is organized as follows. Section 2 shows related work in this area of research. Section 3 presents methodology of the technique. Section 4 presents segment selection process. Clustering technique is presented in section 5. Section 6 focuses on the proposed granular computing techniques. In section 7, experimental analysis and motif patterns are provided. Section 8 concludes the paper with directions for further enhancement.

## 2. RELATED WORK

An automated approach to identify local sequence motifs that transcend protein families was described by Han and Baker [7]. Benchmark K-Means clustering algorithm was used to identify sequence motifs. 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 has proposed Improved K-Means clustering to explore sequence motifs [16]. 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 area of 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 that occurs when input data-set is very large.

Hence, Bernard Chen has proposed granular computing model using Fuzzy clustering technique. In his work, of Fuzzy Improved K-Means algorithm [3], the segments are first partitioned into small information granules using fuzzy 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 [4], granular computing technique is adopted and then initial points chosen greedier than Improved K-Means algorithm. 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.

The main disadvantage of Fuzzy C-Means granular technique is that sum of membership values of data points in all clusters must be equal to one which directs to assign high membership values for the outlier points. Hence in this paper K-Harmonic Means granular computing technique is introduced. K-Harmonic Means has tendency to assign dynamic weights to a data point based on harmonic average. The harmonic average will assign a large weight to a data point that is not close to any centers and a small weight to data point closer to cluster centers. This concept avoids creating densely packed area of multiple centers. This principle is central to KHM being less sensitive to initialization than KM.

In this paper, K-Harmonic Means granular computing technique is proposed to obtain hidden protein sequence motifs present across different protein families. The proposed method is then compared with K-Means granular technique and experimental results shows that KHM able to identify more number of hidden protein sequence motifs that has common structural similarity.

### 3. METHODOLOGY

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. Each protein sequence is represented by their corresponding HSSP file and sequence segments are generated by their HSSP file using sliding window technique. All sequence segments generated by sliding window method may not participate to yield significant motif patterns. Hence, we adopted SVD Entropy based segment selection technique to select segments that yields important motif patterns that has common structural similarity. Figure 1 shows sketch of the proposed method.

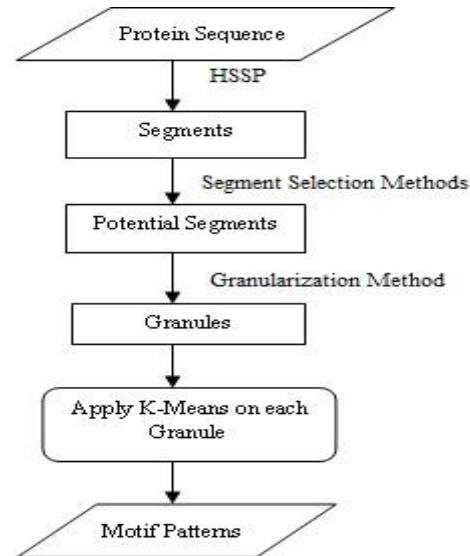


Figure 1: Sketch of the Proposed Method

### 4. SVD- ENTROPY BASED SEGMENT SELECTION TECHNIQUE

SVD Entropy based method is a revival 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].

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

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

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

Figure 2 shows SVD Entropy Selection algorithm adopted in this proposed work

**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)

Select the sequence segments for clustering process

End If

**Figure 2: SVD Entropy Segment Selection Algorithm**

## 5. GRANULAR COMPUTING METHOD: A REVIEW

Granular computing represents information in the form of small granules. Large complicated problems can be solved using divide and conquer strategy. Solution to a large problem can be easily obtained by adopting divide and conquer method.

### 5.2 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. Figure 3 depicts the structure of K-Means granular with SVD Entropy.

## 6. PROPOSED GRANULAR COMPUTING TECHNIQUE

### 6.1 K-Harmonic Means

K-Harmonic clustering is a center-based clustering method proposed by Zhang in 1999 [17]. The following notations are used for K-Harmonic Means algorithm

$x_i$   $i$ th data point,  $i=1,2,...,N$

$c_j$   $j$ th cluster center,  $j=1,2,...,K$

$u_{ij}$  membership value of  $i$ th data point to  $j$ th cluster

$d_{(i,j)}$  distance of  $i$ th data point to  $j$ th cluster

$w_{(x_i)}$  weight of  $x_i$   $i$ th data point

Figure 4 shows steps adopted for K-Harmonic Means clustering algorithm.

**Algorithm :** K-Harmonic Means clustering

**Input :**  $N$  number of data points,  $K$  number of seeds.

**Output:**  $K$  number of partitions

1. Initially, partition  $N$  data points into  $K$  cluster by selecting initial centers randomly. Calculate objective function value according to

$$KHM(X, C) = \sum_{i=1}^N \frac{k}{\sum_{j=1}^k \frac{1}{\|x_i - c_j\|^p}}$$

2. Obtain membership matrix according to

$$u_{ij} = \frac{\|x_i - c_j\|^{-p-2}}{\sum_{j=1}^k \|x_i - c_j\|^{-p-2}}$$

$$u_{ij} \in [0, 1]$$

3. Calculate weight for each point according to

$$w(x_i) = \frac{\sum_{j=1}^k \|x_i - c_j\|^{-p-2}}{\left(\sum_{j=1}^k \|x_i - c_j\|^{-p}\right)^2}$$

4. For each data point calculate new center locations according to

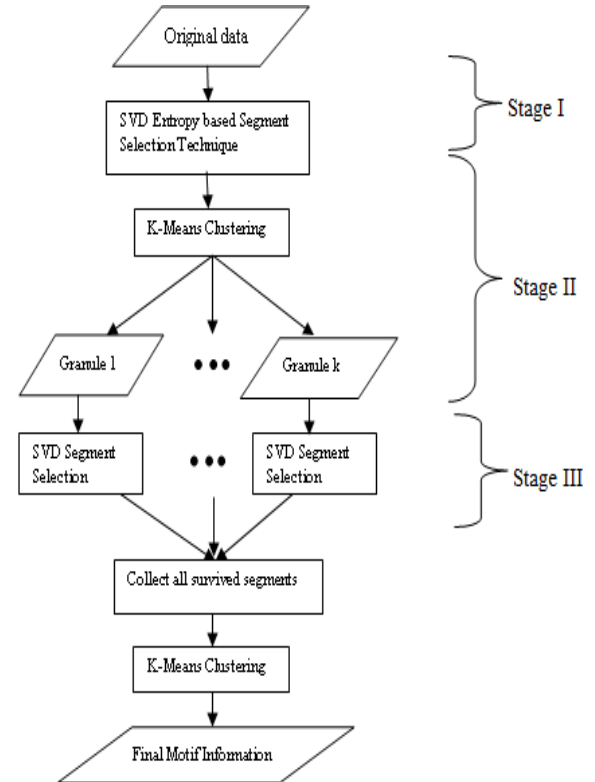
$$c_k = \frac{\sum_{i=1}^N \frac{1}{d_{i,k}^{p+2}} \left( \sum_{l=1}^k \frac{1}{d_{i,l}^p} \right)^2 x_i}{\sum_{i=1}^N \frac{1}{d_{i,k}^{p+2}} \left( \sum_{l=1}^k \frac{1}{d_{i,l}^p} \right)^2}$$

5. Calculate objective function value with new center values.

6. Repeat step 2-5 for predefined number of iterations or until  $KHM(X,C)$  does not change.

7. Assign point  $i$  to cluster  $j$  with biggest  $u_{ij}$  value.

**Figure 4: K-Harmonic Means Clustering Algorithm**



**Figure 3: Sketch of K-Means Granular with SVD Entropy**

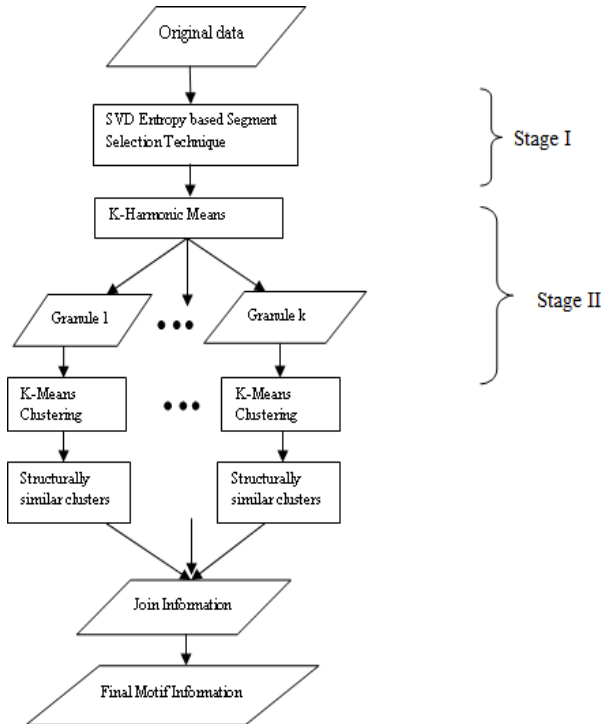


Figure 5: Sketch of KHM Granular with SVD Entropy

Steps of the proposed technique are shown in figure 5. Comparing the above two proposed techniques the advantage of KHM granular with SVD Entropy is that number of stages to generate motif information has been decreased. The quality of clusters and motif information obtained in this proposed work is said to be more significant compared to K-Means Granular with SVD – Entropy.

## 7. EXPERIMENTAL SETUP

### 7.1 Data Set

The latest dataset obtained from Protein Culling Server (PISCES) [14] 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.

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. Figure 6 shows structure of sliding window technique.

## 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	35	65	0
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	18	10	2	12	3	0	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	35	65	0
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	18	10	2	12	3	0	0

Figure 6: Sliding Window techniques with a window size of 10 applied on 1b25 HSSP file.

Database of Secondary Structure Prediction (DSSP) assigns secondary structure to eight different classes [13]. Eight different classes are converted to three 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)

### 7.2 Structural Similarity Measure

Each cluster can be identified as structurally similar by using the following formula [3]:

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

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 structural similarity of a cluster lies between 60% and 70% then the cluster is said to weakly structurally homologous. If the structural homology for a cluster is more than 70% the cluster can be considered more structurally similar [3].

### 7.3 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.

### 7.4 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 [5].

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

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

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

## 7.5 HSSP-BLOSUM62 Measure

HSSP stands for Homology-Derived Secondary Structure of Proteins [12]. 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.

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 this method. HSSP frequency profile and BLOSUM62 matrix has been combined to obtain significance of motif information [3]. Hence, the measure is defined as the following.

If  $m = 0$ : HSSP-BLOSUM62 measure = 0

Else If  $m = 1$ : HSSP-BLOSUM62 measure = BLOSUM62<sub>ij</sub>

$$\text{Else: } \text{HSSP-BLOSUM62 measure} = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^m \text{HSSP}_i \cdot \text{HSSP}_j \cdot \text{BLOSUM62}_{ij}}{\sum_{i=1}^{m-1} \sum_{j=i+1}^m \text{HSSP}_i \cdot \text{HSSP}_j}$$

where

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

$\text{HSSP}_i$  indicates the percent of amino acid  $i$  to be appeared.

$\text{BLOSUM62}_{ij}$  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.

## 7.6 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 KHM 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 order to separate information granules from KHM results, the membership threshold is set to 0.1255. The function that decides how many numbers of clusters should be in each information granule is given below [3]:

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

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 KHM. In this technique, we are able to identify 899 clusters instead of 900 clusters applied in benchmark K-Means clustering.

## 7.7 Experimental Results

Table 1 is the summary of the results from KHM granular with SVD Entropy. Although data size increased from 565314 to 735280, it is to be dealt with one information granule at a time.

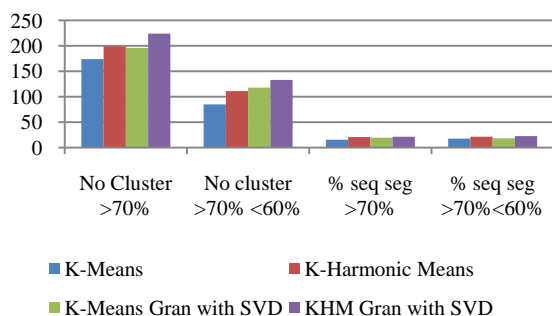
Table 1: Results obtained by SVD-KHM

	Number of Members	Number of Clusters
Granule 1	833650	102
Granule 2	88846	109
Granule 3	108858	133
Granule 4	59938	73
Granule 5	120922	148
Granule 6	83462	102
Granule 7	18482	23
Granule 8	110709	136
Granule 9	30508	37
Granule 10	30205	37
Total	735280	899
Original dataset	565314	900

Table 2 shows the comparative results obtained from different clustering algorithms. It is observed that KHM granular technique with SVD able to identify more number of hidden motif patterns by looking towards structural similarity values. Low DBI value shows that cluster quality is increased in KHM- based K-Means algorithm and in KHM granular technique with SVD. The motif information obtained in KHM granular with SVD Entropy technique is said to be more significant compared to all other algorithms.

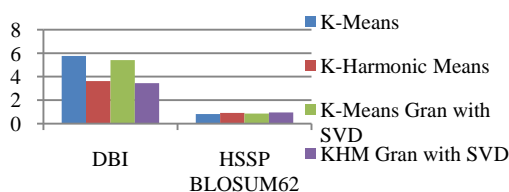
Table 2. Comparison of Different Algorithms

	K-Means Gran	KHM Gran	K-Means Gran. Tech with SVD	KHM Gran. Tech with SVD
No of clusters >60% and < 70%	136	198	196	204
No of clusters > 70%	81	110	118	123
% of Seq Segments > 70%	11.30	18.26	19.44	18.97
% of Seq Segments > 60% <70%	15.42	19.19	18.32	20.10
DBI Measure	6.20	3.62	5.41	3.48
Average HSSP-BLOSUM62	0.49	0.71	0.77	0.79



**Figure 7: Comparison of Structural Similarity values**

Figure 7 is interpreted from results tabulated in table 2. From the above figure 7, it is stated that the number of strong and weak clusters have been increased in KHM granular SVD technique as well as percentage of sequence segments have also been increased considerably.



**Figure 8: Comparison of DBI measure and HSSP-BLOSUM62 values**

Figure 8 shows comparative analysis of cluster quality and quality of motif information. Decreased DBI value and increased HSSP-BLOSUM62 values show the performance of clustering and significance of motif information obtained in KHM granular technique with SVD Entropy segment selection process is good.

From the above table 2, it is inferred that the results obtained in proposed KHM granular with SVD Entropy segment selection technique generates more biochemical meaningful information by eliminating some less meaningful data points. Figures 7 and 8 are interpreted from the results tabulated in table 2.

## 7.8 Sequence Motifs

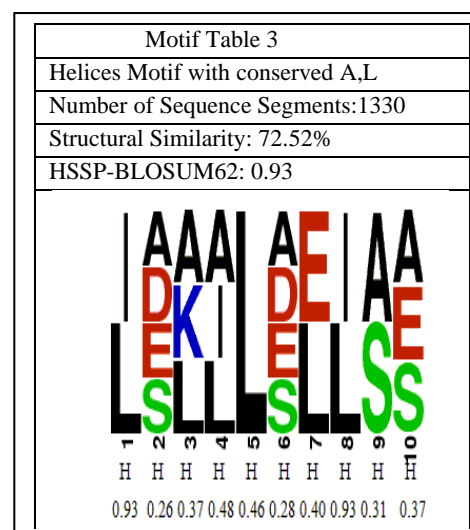
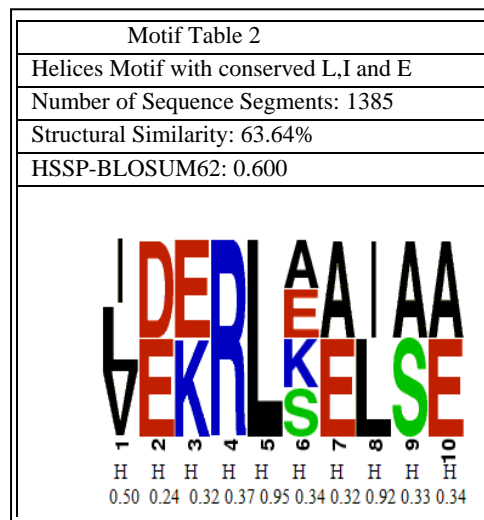
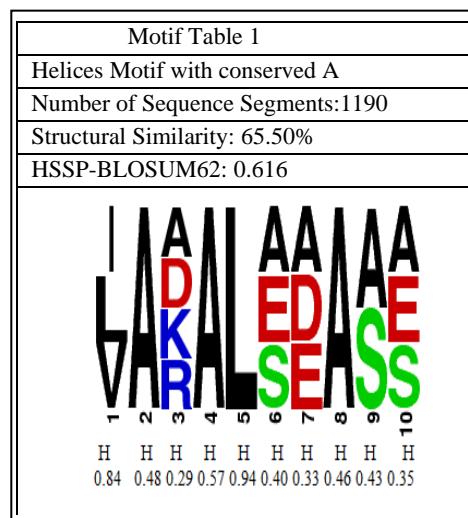
Four different motif patterns are shown in motif tables 1-4. The following format is used for representation of each sequence motif table. Instead of using existing format, in this paper protein logo representation has been used.

The top box shows the number of sequence segments belonging to this motif, percentage of structural similarity, and average HSSP-BLOSUM62 value.


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 of 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 4									
Helices-coils Motif with conserved A,K,E									
Number of Sequence Segments:613									
Structural Similarity: 70.20									
HSSP-BLOSUM62: 0.63									
									
H H H H H C C C C C 0.39 0.18 0.56 0.33 0.23 0.44 0.80 0.18 0.32 0.62									

## 8. 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, proposed work SVD Entropy segment selection technique is been combined with KHM 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 K-Harmonic 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 KHM 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.

## ACKNOWLEDGEMENT

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] O. Alter, P.O Brown, D. Botstein, “Singular value decomposition for genome-wide expression data preprocessing and modeling”, 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. Harrison and Y. Pan, “FIK Model: Novel Efficient Granular Computing Model for Protein Sequence Motifs and Structure Information Discovery”, in IEEE proc, 6<sup>th</sup> symposium on Bioinformatics and Bio Engineering (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] D. L Davies, and D. W Buldin, “A cluster separation measure”, IEEE Trans. Pattern Recogn. Machine Intell., 1,224-227, 1979.
- [6] David W. Mount, Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, New York, 2001.
- [7] K. F Han and D. Baker, “Recurring local sequence motifs in proteins”, J. Mol. Bio, Vol. 251, No. 1, pp. 176-187, 1995.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006.
- [12] 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.
- [13] C.Sander and R.Schneider, “Database of similarity derived protein structures and the structural meaning of sequence alignment”, Proteins: Struct. Funct. Genet. Vol. 9, No.1, pp.56-68, 1991.
- [14] G.Wang and R.L Dunbrack,Jr., “PISCES: a protein sequence culling server”, Bioinformatics, Vol.19, No.12, pp.1589-1591,2003.
- [15] 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.
- [16] B. Zhang, M. Hsu, U. Dayal, K-Harmonic means- a data clustering algorithm, Technical report HPL-1999-124. Hewlett Packard laboratories, 1999.