Extraction of Motif Patterns from Protein Sequences Using SVD with Rough K-Means Algorithm

E.Elayaraja[1], K.Thangavel[2], M.Chitralegha[3] and T.Chandrasekhar[4]

**[1,2,3,4] Department of Computer Science, Periyar University,
Salem, Tamilnadu – 636 011, India**

Abstract

Discovering protein sequence motif information is one of the most crucial tasks in bioinformatics research. In this work, we try to obtain protein recurring patterns which are universally conserved across protein family boundaries. In order to generate higher quality protein sequence motif information from Protein Sequence Culling Server (PISCES) dataset, we tried several different advanced clustering algorithms, such as hierarchical clustering, Self-Organizing Maps (SOM) etc. However, since the dataset itself contains more than 6, 60,000 segments where each segment contains 180 dimensions, any clustering algorithm required more than O(n) complexity is not applicable. Therefore, the very first step of our research is trying to reduce segments. The results suggest that the Singular Value Decomposition (SVD) computing technique is more suits for reducing segments. After that the reduced segments are followed by applying Rough K-Means clustering algorithm. Our experiments indicate that the Rough K-Means algorithm satisfactorily increases the percentage of sequence segments belonging to clusters with high structural similarity than K-Means. The experimental results suggest that the SVD with Rough K-Means algorithm may be applied to other areas of bioinformatics research in order to explore the underlying relationships between data samples more effectively.

Keywords: *Clustering, Motif, Protein Sequence, SVD, HSSP, DSSP, HSSP-BLOSUM62.*

1. Introduction

Clustering is an active research topic in pattern recognition, data mining, statistics, and machine learning with diverse emphasis. Clustering algorithms are probably the most commonly used methods in data mining. Data mining is the process of extracting unknown but useful information which from mass of data that is incomplete, ambiguous, noisy and random. Data mining technology is used to detect large-scale database and find an unknown model [7]. Bioinformatics is the application of computer technology to the management of biological information [6]. Discovering protein sequence motif information is one of the most popular problems in bioinformatics research [2]. In this work, we try to acquire protein recurring sequence patterns which are universally conserved across protein family boundaries. Such conserved sequence patterns are denoted as sequence motifs. Our input dataset is too large; hence an efficient technique is required.

The popular databases for sequence motifs are PROSITE [1], PRINTS [2], BLOCKS [3]. The commonly used tools for protein sequence motif discovery include MEME, Gibbs Sampling, and Block Maker.

In this paper Protein sequences are converted into sliding sequence segments by applying sliding window technique on HSSP (Homology-derived Secondary Structure of Proteins) file [4]. Each sequence segment is represented by the 10×20 matrix. Ten rows represent each position of the sliding window and twenty columns represent 20 amino acids. The total sliding sequence segments are trim by Singular Value Decomposition (SVD) [15]. These sliding sequence segments are classified into different groups with the K-Means and Rough K-Means clustering algorithms. The structural similarity of these groups is evaluated using the secondary structure information obtained from the DSSP (Dictionary of Secondary Structure of Proteins) file. The recurrent groups with high structural similarity will become the candidate to generate sequence motifs representing common structure. Identified sequence motifs are represented by frequency profiles.

This paper has been organized into five sections. In Section 2, various clustering approaches used so for are mentioned in brief. In Section 3, the experimental setup is explained. In Section 4, experimental results and discussion are presented. In section 5, conclusions and further research scope are presented.

2. Clustering Techniques

In this section, we review the K-Means and Rough K-Means clustering algorithms.

2.1 K-Means Clustering Algorithm

K-Means algorithm [11] is a prototype-based, partitional clustering technique that attempts to find user-specified number of clusters, which are represented by their

centroids. In most of the cases Euclidean distance measure is chosen as a common measure. A set of *n* objects x_i, i=1, 2… *n*, are to be partitioned into K groups. The cost function, based on the Euclidean distance between a vector x in group j and the corresponding cluster centroid c_j, can be defined by

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i - C_j \right\|^2$$

2.2 Rough K-Means Algorithm

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.

2.2.1 Rough K-Means 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.

This algorithm can also be interpreted as two layer interval clustering approach with lower and upper approximation. The figure 1 shows Rough K-Means algorithm [9, 10].

---

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≤ i, j ≤ k is used to determine the membership of v as follows:
   - If d (v,x_i) / d(v,x_j) ≤ thershold, then v ∈U(x_i) & v ∈ U(x_j). Furthermore, v will not be a part of any lower bound.
   - Otherwise, v∈L(x_i),such that d(v,x_i) is the minimum for 1≤ i ≤ k. In addition, v∈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.

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

   Where 1≤ j ≤ 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.

Fig. 1 Rough K-Means algorithm

3. Experiment Setup

In this section, we introduce experimental parameters, the dataset; represent the sequence segments, distance measure and SVD. Finally we preserve Davis-Bouldin Index (DBI) and HSSP_BLOSUM62 measures in order to evaluate the performance of clustering algorithms.

3.1 Experimental Parameters

In this research, there are 1500 to 2000 initial clusters are chosen arbitrarily for the K-Means and Rough K-Means clustering algorithms. The each cluster interval is 100. The K-Means and Rough K-Means clustering algorithms are estimated to five times with different random starting points in each cluster interval. The result obtained by using city-block distance metric for calculating distance between segments and the centroid.

3.2 Dataset

Since the major purpose of this work is to obtain protein sequence motif information across protein family boundaries, the dataset of our work is supposed to collect all known protein sequences. However, without a systematic approach, it is very difficult to extract useful knowledge from an extremely large volume of data. The original dataset used in this work includes 4000 protein sequences obtained from Protein Sequence Culling Server (PISCES) [12]. No sequence in this database shares more than 25% sequence identity. The frequency profile from the HSSP is constructed based on the alignment of each protein sequence from the Protein Data Bank (PDB) where 3000 sequences are considered homologous in the sequence database.

3.3 Representation of Sequence Segment

The sliding windows with ten successive residues are generated from protein sequences. Each window corresponds to a sequence segment, which is represented by a 10 × 20 matrix plus additional ten corresponding secondary structure information obtained from DSSP. Ten rows represent each position of the sliding window and twenty columns represent 20 amino acids. For the frequency profiles (HSSP) representation for sequence segments, each position of the matrix represents the frequency for a specified amino acid residue in a sequence position for the multiple sequence alignment. DSSP originally assigns the secondary structure to eight different classes. In this work, we convert those eight classes into three classes based on the following method [14] : H, G and I to H (Helices); B and E to E (Sheets); all others to C (Coils).

3.4 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 [8]. The following formula is used to calculate the distance between two sequence segments:

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

Where L 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.

3.5 SVD Entropy Based Segment Selection Technique

In [15] SVD based entropy has been proposed for the first time to address the problem of selecting the significant segments in the area of protein sequence motif identification. 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.

$$V_j = S_j^2 / \sum_w S_w^2$$

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

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

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. These meaningful segments are then clustered by using traditional K-Means [9] and Rough K-Means clustering algorithms. 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.

3.6 Davis-Bouldin Index (DBI) Measure

The DBI measure [13] 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:

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

$$d_{intra}(C_p) = \frac{\sum_{i=1}^{n_p} \|g_i - g_{pc}\|}{n_p} \quad and \quad d_{inter}(C_p, C_q) = \|g_{pc} - g_{qc}\|$$

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.

3.7 HSSP-BLOSUM62 Measure

BLOSUM62 [5] (Fig. 2.) 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

Fig. 2 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

If k = 0: HSSP-BLOSUM62 measure = 0

Else If k = 1:

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

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

Else: HSSP-BLOSUM62 measure = $\frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} HSSP_i \cdot HSSP_j \cdot BLOSUM\ 62_{ij}}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} HSSP_i \cdot HSSP_j}$

4. Experimental Results

In this work, 3000 protein sequences are extracted from the Protein Sequence Culling Server (PISCES) as the dataset. In this protein database, the percentage identity cutoff is 25%, the resolution cutoff is 2.2, and the R-factor cutoff is 1.0. With these protein sequences, sliding windows with ten consecutive residues are obtained. Each window contains one sequence segment of ten continuous positions. This sliding window approach generates 6, 60,364 segments. K-Means and Rough K-Means algorithms were applied to these segments and they are clustered between 1500 and 2000 clusters. The threshold value is set as 1, w_{lower}= 0.7, w_{upper}= 0.3 for Rough K-Means algorithm. The secondary structure information is used as biological evaluation criteria. The higher HSSPBLOSUM62 value indicates more significant motif information. We also use DBI measure to identify the best cluster. The lower DBI value indicates the high quality of the cluster result.

Table 1: Comparison of HSSP-BLOSUM62 measure and DBI measure belonging to K-Means clusters with high structural similarity.

Number of Clusters	Number of Iterations 5				
	K-Means				
	>60	>70	Without SVD DBI Measure	SVD Applied DBI Measure	BLOSUM 62 Measure
1500	332	154	5.3377	5.1808	0.6822
1600	349	166	5.2791	5.0809	0.6780
1700	380	185	5.2369	5.1353	0.7328
1800	403	190	5.1837	5.1055	0.6776
1900	415	204	5.1464	5.1014	0.6585
2000	441	214	5.1110	5.0359	0.7055

Table 2: Comparison of HSSP-BLOSUM62 measure and DBI measure belonging to Rough K-Means clusters with high structural similarity.

Number of Clusters	Number of Iterations 5				
	Rough K-Means				
	>60	>70	Without SVD DBI Measure	SVD Applied DBI Measure	BLOSUM 62 Measure
1500	337	152	4.9431	4.6565	0.6627
1600	342	166	4.8771	4.7321	0.6461
1700	382	200	4.8738	4.7099	0.6701
1800	422	200	4.8220	4.5988	0.6008
1900	455	198	4.7885	4.5916	0.6922
2000	442	217	4.7436	4.6719	0.6153

The following Figures 3 and 4 are interpreted from table 1 and 2.

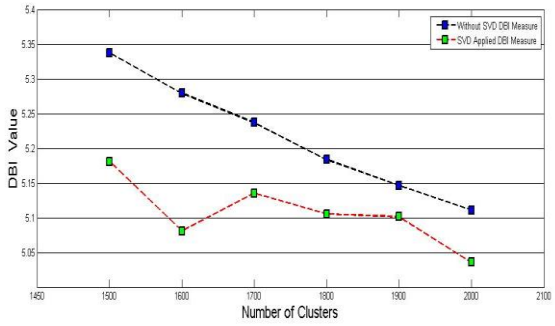

Fig. 3 Comparison of K-Means DBI values of sequence segments belonging to cluster with high structure similarity

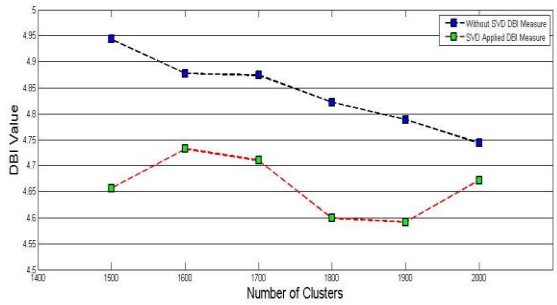

Fig. 4 Comparison of Rough K-Means DBI values of sequence segments belonging to cluster with high structure similarity

The results of Table 1 and 2 with the figure 3 and 4 reveal that the quality of clusters improved dramatically by applying the SVD computing technique which utilizes K-Means and Rough K-Means. In the Rough K-Means approach, the average percentage of clusters with structural similarity increased more. The DBI measure also successfully decreased, implying that our model not only generates more biologically meaningful results but that these results are supported by statistical/computer-science techniques. Also, the HSSP-BLOSUM62 measurement increasing proves that the motif information is more consistent and meaningful under the SVD computing strategy.

4.1 Representation of Motif Patterns

The table 3 to 8 illustrates six different sequence motifs generated by our method. The following format is used for the representation of each motif table.

- The first row represents the number of members belonging to this motif, the secondary structural similarity and the average HSSP-BLOSUM62 value.
- The first column stands for the position of amino acid profiles in each motif with window size ten.
- The second column expresses the type of amino acid frequently appearing in the given Position. If the amino acids are appearing with the frequency higher than 10%, they are indicated by upper case; if the amino acids are appearing with the frequency between 8% and 10%, they are indicated by lower case.
- The third column corresponds to the hydrophobicity value, which is the summation of the Frequencies of occurrence of Leu, Pro, Met, Trp, Ala, Val, Phe, and Ile.
- The fourth column indicates the value of the HSSP-BLOSUM62 measure.
- The last column indicates the representative secondary structure to the position.

Table 3: Hydrophobic Helixes motif

Number of segments: 785 **Structure homology**: 78.203822% **Avg. HSSP-BLOSUM62**: 0.598				
#	**Noticeable Amino Acid**	**H**	**B**	**S**
1	aSt	0.38	0.72	H
2	Ap	0.46	-1	H
3	AskED	0.28	-0.12	H
4	AEd	0.38	-0.39	H
5	VLI	0.90	2.38	H
6	aRK	0.36	0.22	H
7	AKqE	0.23	0.16	H
8	vA	0.55	0.00	H
9	L	0.96	4	H
10	arKE	0.26	0.01	H

Table 4: Helices motif with conserved A

Number of segments: 765 **Structure homology**: 75.921569% **Avg. HSSP-BLOSUM62**: 1.738				
#	**Noticeable Amino Acid**	**H**	**B**	**S**
1	Ae	0.36	-1	H
2	A	0.73	4	H
3	A	0.71	4	H
4	vLia	0.57	0.64	H
5	Ad	0.40	-2	H
6	A	0.77	4	H
7	vlA	0.52	-0.12	H
8	Ark	0.37	-0.14	H
9	A	0.45	4	H
10	A	0.48	4	H

Table 5: Helixes-Coil motif

Number of segments: 424 **Structure homology**: 72.429245% **Avg. HSSP-BLOSUM62**: 0.804				
#	**Noticeable Amino Acid**	**H**	**B**	**S**
1	VL	0.56	1	C
2	vL	0.46	1	C
3	GA	0.40	0	C
4	VLi	0.54	1.87	C
5	STD	0.18	0.29	C
6	vlApE	0.51	-1.33	H
7	AED	0.19	0.35	H
8	qED	0.12	1.82	H
9	A	0.76	4	H
10	vLArke	0.46	-0.96	H

Table 6: Helices-coil-sheet motif

Number of segments: 844 **Structure homology**: 73.388626% **Avg. HSSP-BLOSUM62**: 0.209				
#	**Noticeable Amino Acid**	**H**	**B**	**S**
1	ArKEd	0.26	-0.24	H
2	lAr	0.42	-1.18	C
3	G	0.03	6.00	C
4	VLIA	0.65	0.67	C
5	RKE	0.25	1.09	E
6	VlI	0.77	2.19	E
7	VLI	0.68	2.18	E
8	VLia	0.53	0.81	E
9	VLI	0.69	1.90	E
10	STD	0.27	-0.01	C

Table 7: Helices motif with conserved A

Number of segments: 785 **Structure homology**: 79.388535% **Avg. HSSP-BLOSUM62**: 1.501				
#	**Noticeable Amino Acid**	**H**	**B**	**S**
1	Lar	0.40	-1.31	H
2	Ae	0.37	-1.0	H
3	A	0.71	4.0	H
4	A	0.70	4.0	H
5	vLia	0.52	0.65	H
6	A	0.38	4.0	H
7	A	0.84	4.0	H
8	VLiA	0.54	0.40	H
9	Are	0.37	-0.73	H
10	As	0.43	1.0	H

Table 8: Coils Sheets motif with conserved V L and I

Number of segments: 381 **Structure homology**: 79.317585% **Avg. HSSP-BLOSUM62**: 0.7340				
#	**Noticeable Amino Acid**	**H**	**B**	**S**
1	Vl	0.51	1.0	E
2	VLi	0.55	1.90	E
3	VLI	0.77	2.05	E
4	vlE	0.42	-1.47	E
5	VlI	0.63	2.27	E
6	gEND	0.08	0.46	C
7	Gd	0.06	-1.0	C
8	RKqE	0.18	1.15	E
9	vLp	0.54	-1.19	E
10	VLI	0.71	2.17	E

5. Conclusion

Proteins are involved in every body functions including nutrient transportation, muscle building, metabolism regulation, etc. Understanding the functions and structures of proteins encourages cellular process discovery. In this work we have obtained the data set from the Protein Sequence Culling Server (PISCES). The sliding windows with ten successive residues were generated from protein sequences. These sequence segments of ten continuous positions were clustered into different groups with K-Means and Rough K-Means algorithms. Before clustering we try to reduce unwanted segments using SVD. The SVD resultant segments are then grouped using K-Means and Rough K-Means clustering with respect to similarity of secondary structure. Clusters with similarity higher than a pre-determined threshold are taken to obtain sequence motifs. The Rough K-Means clustering followed SVD technique is capable of decreasing time and space complexity, filtering outliers, and capturing better results. We believe some other bioinformatics research with large database may also adapt this SVD computing strategy to perform well.

Acknowledgement

The first and second author would like to thank UGC, New Delhi for the financial support received under UGC Major Research Project No. F-34-105/2008.

The First Author extends his gratitude to UGC as this research work was supported by Basic Scientist Research (BSR) Non-SAP Scheme, under grant reference number, F-41/2006(BSR)/11-142/2010(BSR) UGC XI Plan.

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 PROCITE database”, Nucleic Acids Research, vol. 32, no. Database, pp. D134-137, 2004.

[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] 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.

[5] Henikoff, S. and Henikoff, J. G. (1992), Amino Acid Substitution Matrices from Protein Blocks, Proceedings of the National Academy of Sciences of the United States of America. 89, 10915-10919.

[6] G. Karp, Cell and Molecular Biology (Concepts and Experiments), 3rd Ed. New York: Wiley, 2002, pp. 52–65.

[7] Zhong, W., Altun, G., Harrison, R., Tai, P. C. & Pan, Y. (2005) Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property, NanoBioscience, IEEE Transactions on. 4, 255-265.

[8] K. F. Han and D. Baker, "Recurring local sequence motifs in proteins", J. Mol. Biol., vol 251, no. 1, pp. 176–187, 1995.

[9] P. Lingras, C. West, Interval set clustering of web users with rough k-means, J. Intell. Inform. Syst. 23 (2004) 5–16.

[10] P. Lingras, R. Yan, C. West, Comparison of conventional and rough k-means clustering, in: International conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Lecture Notes in Artificial Intelligence, vol. 2639, Springer, Berlin, 2003, pp. 130–137.

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

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

[13] 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.

[14] 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.

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