Extraction of Motif Patterns from Protein Sequences Using K-Means with segment pruning methods

E.Elayaraja, K.Thangavel, P.Ashok, T.Chandrasekhar

Abstract— Bioinformatics is the application of information technology to the management of molecular biological data. Motif finding in protein sequence is one of the most crucial tasks in bioinformatics research. Motifs are identifying as overly recurring sub-patterns in segment of protein sequence biological data. Sequence motifs are verifying by their structural similarities or their functional roles in DNA or protein sequences. The generated sequence segments do not have classes or labels. Hence, unsupervised segment selection technique is adopted to select significant segments. In this work, K-Means clustering algorithm is applied for selecting significant segments. The K-Means clustering algorithm is improved by implementing initial centroids selection technique instead of random centroids selection. All clustering segments are not being important to produce good motif patterns. Therefore Average Distance between Object and Centroid (ADOC) and Outlier Removal Clustering (ORC) methods are applied to select significant segments. The experimental results show that the K-Means with segment pruning methods perform better than the traditional K-Means algorithm by producing a larger number of high-quality of motif patterns.

Index Terms— Proteins, Motif, Clustering, K-Means, ADOC, ORC, ICS.

1 INTRODUCTION

BIOINFORMATICS involves the use of several different techniques, including Computer Science, Data Mining, and Computational Intelligence, to solve the problems of Molecular Biology. Understanding the hidden knowledge between protein structures and their sequences is an important task in bioinformatics research. The biological term sequence motif denotes a relatively small number of functionally or structurally conserved sequence patterns that occurs repeatedly in a group of related proteins [3]. Proteins are vary in structures and as well as in functions. The term "protein sequence motif" denotes amino-acid sequence pattern that is widespread and has biological significance. These motif patterns may be able to predict other proteins' structural or functional areas, such as binding sites, conserved domains, and prosthetic attachment site [2]. There are several databases available for sequence motifs but the most popular ones are PROSITE [8], PRINTS [1] and BLOCKS [7].

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 [11]. 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. These sliding sequence segments are grouped by K-Means and segment pruning methods. The structural similarity of these groups is evaluated using the secondary structure information obtained from the DSSP (Dictionary of Secondary Structure of Proteins) file [15]. The results of the K-Means clustering algorithm compared to K-Means clustering with segment pruning methods. The comparative results shows, motif patterns obtained from the pruning methods are said to be efficient.

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

2 CLUSTERING TECHNIQUE AND PRUNING METHODS

2.1 K-Means Clustering Algorithm

K-Means [5], [6], [12] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group is done. At this point, it is need to re-calculate k new centroids as centres of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data points and the nearest new centroid. A loop has

- *E.Elayaraja is currently pursuing Ph.D., in Computer Science in Periyar University, Salem, India. E-mail: elayarajaphd.e@gmail.com*
- *K. Thangavel is currently working as Professor and Head, Department of Computer Science in Periyar University, Salem, India. E-mail: drktvelu@yahoo.com*

P.Ashok is currently pursuing Ph.D., in Computer Science in Bharthiyar University, India. E-mail: ashokcutee@gmail.com

- *T. Chandrasekhar is currently pursuing Ph.D., in Computer Science in Bharthiyar University, India. E-mail: ch_ansekh80@rediffmail.com*

been generated. As a result of this loop it may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

The K-Means algorithm is effective in producing clusters for many practical applications. But the computational complexity of the original K-Means algorithm is very high, especially for large Data sets. The K-Means clustering algorithm is a partitioning clustering method that separates data into K groups. For real life problems, the effective clusters centroids cannot be initialized. To overcome the above drawback the current research focused on developing the clustering algorithms instead of selecting the initial centroids randomly. The algorithm is composed in the following steps:

Algorithm 1: **K-Means Clustering Algorithm**

Input:

D = {d1, d2, d3... ... dn } // Set of n data points.

k =Number of desired clusters

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest centroids.

Methods

1. Arbitrarily choose *k* data points from *D* as initial centroids;
2. Assign each point d_i to the cluster which has the closest centroid
3. Calculate the new mean for each cluster;
4. **Repeat step 2** and **step 3 until convergence** criteria is met.

1. Advantages of K-Means Algorithm

- It is easy to implement and works with any of the standard norms.
- It allows straightforward parallelization.
- It is incentive with respect to data ordering

2. Drawbacks of K-Means Algorithm

- The final clusters do not represent a global optimization result but only the local one, and complete different final clusters can arise from difference in the initial randomly chosen cluster centres.
- We have to know how many clusters we will have at the first

2.2 *Initial Centroid* Selection (ICS) Method

In the K-Means clustering algorithm, the initial centroids are selected randomly from the given data set. The initial centroids plays the main role in the clustering process, sometimes the algorithm produce bad clustering results due to selecting centroids randomly in the given data set, to avoid this problem we introduce algorithms to select the initial centroid [14] for the K-Means algorithm which is given below.

Algorithm 3: Initial Centroid Selection (ICS) method

Steps

1. Using Euclidean distance as a dissimilarity measure, compute the distance between every pair of all objects as follows:

$$\boldsymbol{d_{ij}} = \sqrt{\sum_{a=}^{p}(\boldsymbol{X_{ia}} - \boldsymbol{X_{ja}})^2} \quad \boldsymbol{i} \;\&\; j = 1 \ldots . n (1)$$

2. Calculate $\mathbf{M}_{ij}$ to make an initial guess at the centers of the clusters

$$\boldsymbol{M_{ij}} = \frac{\boldsymbol{d_{ij}}}{\sum_{i=1}^{n} \boldsymbol{d_{ij}}} \quad \boldsymbol{i = 1 \ldots n; j = 1 \ldots n} \qquad \textbf{(2)}$$

3. Calculate $\sum_{i=1}^{n} \boldsymbol{M}^2_{\boldsymbol{ij}_{(j=1\ldots.n)}}$ at each object and sort them in ascending order.
4. Select Kobjects having the minimum value as initial cluster medoids.

2.3 Pruning methods

The pruning is the method which is used to reduce the unnecessary data in the database. In our protein segment dataset is very huge so the execution time of the K-Means clustering algorithm is required too many hours and days so this is one of the problems to execute the methods with minimum time complexity. By reducing dataset with pruning methods the K-Means clustering methods can able to improve the performance and similarity structure of the protein sequence segment.

2.3.1 Pruning Method-1

The objective of the pruning [5] algorithm that we call Outlier Removal Clustering (ORC), is to produce a codebook as close as possible to the mean vector parameters that generated the original data. It consists of two consecutive stages, which are repeated several times. In the first stage, we perform K-Means algorithm until convergence, and in the second stage, we assign an outlyingness factor for each vector. Factor depends on its distance from the cluster centroid. Then algorithm iterations start, with first finding the vector with maximum distance to the partition centroid

$$\boldsymbol{d_{max}} = \mathbf{max}\{||\boldsymbol{x_i} - \boldsymbol{c_i}||\} \qquad (\mathbf{3})$$

Outlyingness factors for each vector are then calculated. We define the outlyingness of a vector *xi* as follows:

$$O_i = \frac{||x_i - C_i||}{d_{max}} \qquad (4)$$

We see that all outlyingness factors of the dataset are normalized to the scale [0, 1]. The greater the value, the more likely the vector is an outlier.

Algorithm2: ORC (I, T)

C←Run K-Means with multiple initial solutions
For j=1 to I do

$$d_{max} = \max\{||x_i - c_i||\}$$

For I =1 to n do

$$O_i = \frac{||x_i - C_i||}{d_{max}}$$

If O_i>Tthen
X=X/{x_i}
endif
endfor

(C,P)=Kmeans(X,C)
endfor

The vectors for which $\mathbf{o_i} > T$, are defined as outliers and removed from the dataset. At the end of each iteration, K-Means is run with previous the *C* as the initial codebook, so new solution will be a fine-tuned solution for the reduced dataset. By setting the threshold to $T < 1$, at least one vector is removed. Thus, increasing the number of iterations and decreasing the threshold

2.3.2 Pruning Method-2

The protein segments are also reduced by the pruning method-2 techniques [4][12][13].We use the alternate pruning method to detect the unwanted protein segments. In this method the distance between each objects and its centroid is calculated, after that calculates the Average Distance between the Object and Centroid (ADOC) and the obtained value is fixed as target value for detecting pruning protein segments, this process is implemented for all clusters. If the distance between the object and centroid value is greater than ADOC value than the object in dataset that is detected as pruning object. The ADOC value is calculated by using the expression which is given below.

$$ADOC = \left(\frac{1}{n_i}\sum_{x \sum p_i} ||x - V_i||^2\right) \quad i = 1, 2, \ldots, k. \qquad (5)$$

V_i is the centroid of the i^{th} Cluster
X is the object in the i^{th} Cluster

3 EXPERIMENTAL SETUP

In this section, we introduce experimental parameters, the dataset; represent the sequence segments, and distance measure. 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 800 to 1300 initial clusters are chosen arbitrarily for the K-Means and K-Means with pruning clustering algorithms. The each cluster interval is 100. The K-Means and K-Means with pruning 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) [16]. 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 [3]: 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 [18]. 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 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:

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

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

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.

3.6 HSSP-BLOSUM62 Measure

BLOSUM62 [19] (Fig. 1.) 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. 1. 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

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 K-Means with pruning algorithms were applied to these segments and they are clustered between 800 and 1300 clusters. The secondary structure information is used as biological evaluation criteria. The higher HSSP-BLOSOM62 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. The results are obtained which are depicted in the below Table 1.

TABLE1
COMPARISON OF HSSP-BLOSOM62 MEASURE AND DBI MEASURE BELONGING TO K-MEANS CLUSTERS WITHOUT APPLYING PRUNING METHODS

Clusters	Number of Iterations 5			
	K-Means			
	≤70 & >60	>70	Without Pruning Method	
			DB Index Measure	HSSP BLOSOM 62 Measure
800	183	78	4.2431	0.7612
900	203	89	4.2771	0.7589
1000	212	105	4.7754	0.7400
1100	239	119	4.3451	0.6989
1200	274	132	4.5045	0.7008
1300	290	142	4.4502	0.7321

The K-Means clustering method is executed with two Pruning methods are Pruning method1 and pruning method2 for 5 iteration and varying cluster values from 800 to 1300. The DB index value and HSSP-BLOSOM62 Measure are obtained from the result which are depicted in the below Table 2 and Table 3.

TABLE 2
COMPARISON OF HSSP-BLOSOM62 MEASURE AND DBI MEASURE BELONGING TO K-MEANS CLUSTERS WITH APPLYING PRUNING METHOD1

Clusters	Number of Iterations 5			
	K-Means			
	≤70& >60	>70	Pruning Method1	
			DB Index Measure	HSSP-BLOSOM62 Measure
800	180	81	4.4431	0.7734
900	206	96	3.8451	0.7991
1000	213	112	3.8547	0.7664
1100	247	108	4.3547	0.7146
1200	270	121	3.6481	0.7132
1300	303	146	4.0085	0.7045

TABLE3
COMPARISON OF HSSP-BLOSOM62 MEASURE AND DBI MEASURE BELONGING TO K-MEANS CLUSTERS WITH APPLYING PRUNING METHOD2

Clusters	Number of Iterations 5			
	K-Means			
	≤70 & >60	>70	Pruning Method2	
			DB Index Measure	Blosum Measure
800	187	85	4.3217	0.7692
900	213	90	4.0045	0.7679
1000	225	109	4.6784	0.7400
1100	264	121	4.2859	0.7004
1200	274	132	4.1916	0.7108
1300	287	135	4.4650	0.7358

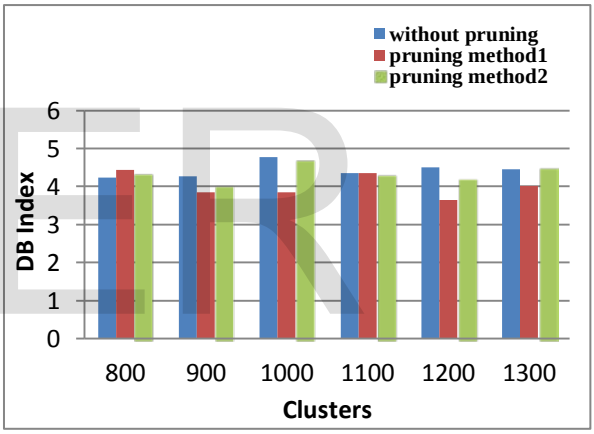

Fig.2. Davis Bouldin index measure chart K-Means method

The results in the Table 1, 2 and 3 with the Fig. 2 reveal that the quality of clusters improved dramatically by applying the Pruning technique which utilizes K-Means In the K-Means approach, the percentage of clusters with structural similarity increased by applying two pruning methods. The DBI measure also successfully decreased by applying K-Means with Pruning methods, hence the pruning method2 improves the results and quality of the clusters than pruning method1 implying that our model not only generates more biologically meaningful results but, these results are supported by statistical/computer-science techniques. Also, the HSSP-BLOSOM62 measurement increasing proves that the motif information is more consistent and meaningful.

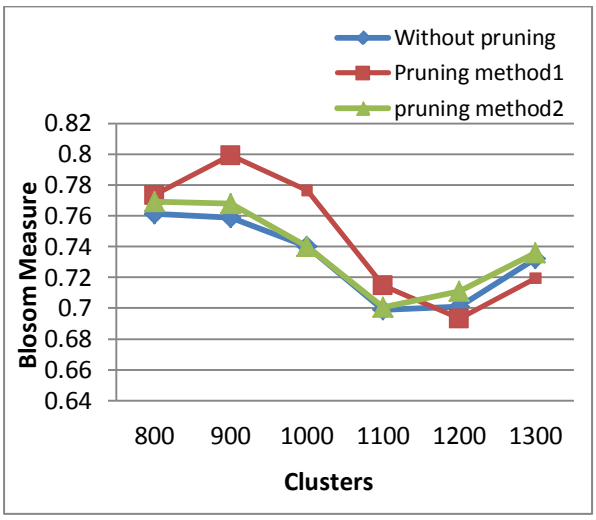

Fig.3. HSSP-BLOSOM62 measure comparison chart for K-Means algorithm

From the above Fig. 3, it clearly shows that the K-Means approach, the percentage of clusters with structural similarity increased by applying two pruning methods. The pruning method2 perform well then pruning method1 which improve structural similarity, the HSSP-BLOSOM62 measure and also successfully increased. Hence the pruning method2 improves the results and quality of the clusters than pruning method1. The HSSP-BLOSOM62 measurement increasing proves that the motif information is more consistent and meaningful.

3.6.1 Initial centroid selection method

The K-Means clustering method is improved by initialize the cluster centroids by Initial Centroid Selection (ICS) method instead of random centroid selection method, the centroids selection method is explain in the section II, the K-Means clustering method is executed with initial centroid selection method varying from 800 to 1300 with pruning method. The obtained results and DB index values are depicted in the below Table 4.

TABLE4
PERFORMANCE ANALYSIS TABLE FOR INITIAL CENTROID SELECTION METHOD OF K-MEANS

Clusters	Number of Iterations 5			
	K-Means			
	≤70 & >60	>70	Centroid Selection Method	
			Random method (DB Index Measure)	Proposed Method (DB index Measure)
800	178	84	4.4431	3.5481
900	210	99	3.8451	3.7812
1000	216	118	3.8547	3.6157
1100	240	120	4.3547	3.9154
1200	276	129	3.6481	3.7548
1300	309	140	4.0085	4.1254

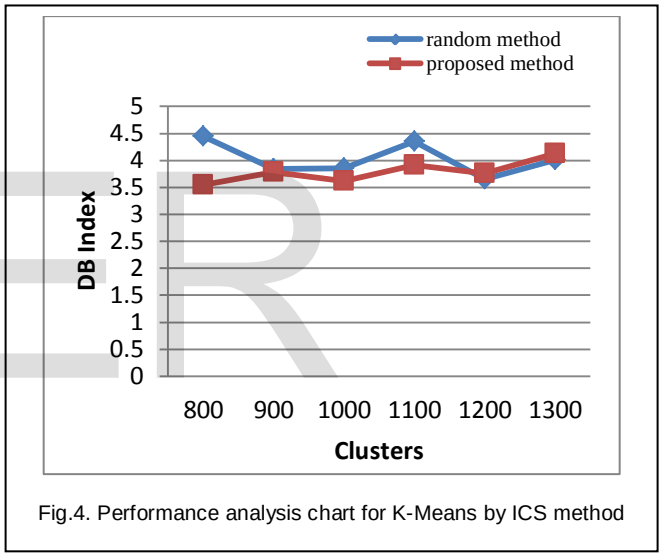

Fig.4. Performance analysis chart for K-Means by ICS method

From the above Fig. 4, it clearly shows that the K-Means approach and initial centroid selection method perform well then random centroids selection method which improves structural similarity and successfully decreasing the DB index measure. Hence the initial centroid selection method improves the results and quality of the clusters than random centroid selection method. The DB index measurement decreasing proves that the motif information is more consistent and meaningful.

3.6.2 Representation of Motif Patterns

The Table 5 to 10 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 averageHSSP-BLOSOM62 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-BLOSOM62 measure.
- The last column indicates the representative secondary structure to the position.

TABLE5 HYDROPHOBIC HELIXES MOTIF

Number of segments: 785 **Structure homology**: 78.2038% **Avg. HSSP-BLOSOM62**: 0.628				
#	**Noticeable Amino Acid**	**H**	**B**	**S**
1	AaSt	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

TABLE6 HELICES MOTIF WITH CONSERVED A

Number of segments: 765 **Structure homology**: 76.9215% **Avg. HSSP-BLOSOM62**: 1.531				
#	**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

TABLE7 HELIXES-COIL MOTIF

Number of segments: 400 **Structure homology**: 71.4292% **Avg. HSSP-BLOSOM62**: 0.904				
#	**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

TABLE8 HELICES-COIL-SHEET MOTIF

Number of segments: 840 **Structure homology**: 73.3886% **Avg. HSSP-BLOSOM62**: 0.509				
#	**Noticeable Amino Acid**	**H**	**B**	**S**
1	ArKEd	0.26	-0.24	H
2	lAr	0.42	-1.28	H
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

TABLE9 HELICES MOTIF WITH CONSERVED A

Number of segments: 785 **Structure homology**: 80.3885% **Avg.HSSP-BLOSOM62**: 1.502				
#	**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

TABLE10 COILS SHEETS MOTIF WITH CONSERVED V L AND I

Number of segments: 381 **Structure homology**: 79.3175% **Avg. HSSP-BLOSOM62**: 0.7440				
#	**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	ENDI	0.08	0.46	C
7	GdVL	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

6 CONCLUSION

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. We try to reduce unwanted segments using two effective Pruning methods. After pruning the sequence segments then the resultant segments are grouped using K-Means clustering with respect to similarity of secondary structure. The K-Means clustering followed with Initial Centroid Selection method is capable of decreasing DB index value and also increasing HSSP-BLOSOM62 Measure by filtering outliers, and capturing better results.

Acknowledgment

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.

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

REFERENCES

[1] C. Attwood T. K., 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.

[2] 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.

[3] 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.

[4] Chiu, A. and A. Fu, 2003. "Enhancement on Local Outlier Detection." 7th International Database Engineering and Application Symposium (IDEAS03)", pp. 298-307.

[5] HARTIGAN, J. and WONG, M. 1979. Algorithm AS136: "A K-Means clustering algorithm". Applied Statistics, 28, pp. 100-108.

[6] HEER, J. and CHI, E.2001. "Identification of Web user traffic composition using multimodal clustering and information scent.", 1st SIAM ICDM, Workshop on Web Mining, pp51-58.

[7] Henikoff S. Henikoff J. G. and S.Pietrokovski, "Blocks+: a non re-

dundant database of protein Alignment blocks derived from multiple compilation", Bioinformatics, vol. 15, no. 6, pp. 417-479, 1999.

[8] Hulo N. Sigrist., C. J. a, Le Saux V., 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.

[9] Knorr, E. and R. Ng, Algorithms for Mining Distance-based Outliers in Large Data Sets, 1998. Proc. the 24th International Conference on Very Large Databases (VLDB), pp. 392-403.

[10] Loureiro,A., L. Torgo and C. Soares, 2004. Outlier Detection using Clustering Methods: a Data Cleaning Application, in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

[11] Sander C., 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.

[12] Sauravjoyti Sarmah and Dhruba K.Bhattacharyya. May 2010 "An Effective Technique for Clustering Incremental Gene Expression data", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No. 3.

[13] Ville Hautam aki, Svetlana Cherednichenko,Ismo Karkkainen Tomi Kinnunen, and Pasi Fr anti, "Improving K-Means by Outlier removal", SCIA, LNCS 3540, 2005, pp.978–987.

[14] Yuan F, Z. H. Meng, H. X. Zhangz, C. R. Dong, August 2004 " A New Algorithm to Get the Initial Centroids", proceedings of the 3rdInternational Conference on Machine Learning and Cybernetics, pp. 26-29.

[15] E. Elayaraja, K. Thanagavel, M.Chitralegha and T. Chandrasekhar, "Extraction of Motif Pattern from Protein Sequences Using SVD with Rough K-Means Algorithm",*International Journal of Computer Science Issues (IJCSI)*, Volume 9, Issue 6, November 2012 edition ISSN (Online):1694-0814, pp. 350-356.

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

[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.

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

[19] 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.

[20] T. Chandrasekhar, K.Thangavel and E. Elayaraja "Performance Analysis of Enhanced Clustering Algorithm for Gene Expression Data", IJCSI International Journal of Computer Science Issues, Vol.8, Issue 6, No 3, November 2011. ISSN (online):1694-0814.