New Seed Selection Technique for Protein Sequeunce Motif Identification

M. Chitralegha
Banker
Salem, India
chitra_legha04@yahoo.co.in

Dr. K Thangavel
Professor and Head
Department of Computer Science, Periyar University
Salem, India
drktvelu@yahoo.co.in

Abstract-**Bioinformatics is a field devoted to the interpretation and analysis of biological data using computational techniques. In recent years the study of bioinformatics has grown tremendously due to huge amount of biological information generated by the scientific community. Protein sequence motifs are short fragments of conserved amino acids often associated with specific function. Identifying such motifs is one of the challenging tasks in the area of bioinformatics. Data mining is one technique to explore sequence motif from protein sequences. In this proposed work, recurring sequence motifs are identified by adopting new seed initialization technique for K-Means clustering algorithm. This proposed work combine's local density approximation utilizes sorted pair wise distance calculation for identifying potential seeds for K-Means clustering. This new initialization technique enhances K-Means learning characteristics towards better cluster separation to identify the significant motif patterns.**

Keywords-Clustering, Data mining, Protein sequence, Motif, KKZ, DBI and HSSP-BLOSUM62.

I. INTRODUCTION

Proteins can be considered as one of the most important elements in the process of life. They regulate variety of activities in all known organisms, from replication of the genetic code to transporting oxygen, and are generally responsible for regulating the cellular machinery and determining the phenotype of an organism. The term 'Motif' refers to a region or portion of a protein sequence that has specific structure and is functionally significant. Detection of such motifs in proteins is an important problem in today's bioinformatics research. These motif patterns may able to predict other protein's structural or functional area, such as De-oxyribo Nucleic Acid (DNA) or Ribo Nucleic Acid (RNA) binding sites, conserved domains etc [6].

There are several popular motif databases. PROSITE [11], PRINTS [2] AND BLOCKS [10] are the three most popular motif databases. The most important motif finding tools are MITRA, Profile Branching, EMOTIF, CoSMos and Motif Scan [7]. But, these methods will generate motif patterns only for a single protein sequence. The patterns obtained by using above methods, may carry only a little information about conserved sequence regions which transcend protein families. Instead, in this paper, huge numbers of segments are generated from HSSP file for all protein sequences.

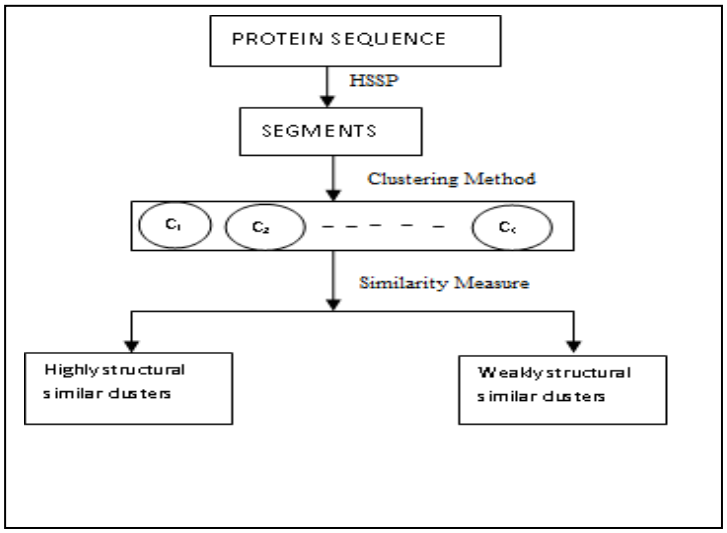

Figure 1. Flowchart of Motif Detection Method

The resulted sequence segments are then clustered using benchmark K-Means algorithm. Each generated cluster is assessed by using structural similarity measure [3]. Based on similarity measure clusters are classified into two types such as highly structural similar clusters and weakly structural similar clusters. Finally, highly structural similar clusters are considered for potential motif generation. The different steps are depicted in Figure 1.

Clustering refers to the task of partitioning unlabelled data into meaningful groups. K-Means clustering is computationally efficient for large input datasets [14]. K-means algorithm selects K centroids in a random manner from whole input dataset. Then the algorithm iteratively updates the centers until no reassignment of data point to new cluster occurs. K-Means clustering algorithm has been used by many researchers for knowledge discovery in the area of bioinformatics research.

Cluster initialization methods have been broadly classified into three major categories, namely Random Sampling, Distance Optimization and Density Estimation Method. Random Sampling Method selects initial centroids in a random manner from input data set. Under this method, Forgy technique [8] adopts to select initial seeds in a uniform random manner. The other MacQueen technique [16] simply initializes the seeds to the cluster with the first K input samples. Secondly, the Distance Optimization Method tries to optimize the distance among seeds beforehand towards satisfactory of inter cluster separation. Simple cluster seeking technique [19] adopts FASTCLUS procedure for cluster initialization. Katsavoundiset. al [13] proposed a method that utilizes sorted pairwise distances for initialization.

Third, Density Estimation Method assumes that input samples uses Gaussian mixture distribution. Hence by identifying dense areas of input domain helps clustering technique to create compact clusters. Kaufman and Rousseuw [14] introduced a method that estimates density through pairwise distance comparison and initializes the seed clusters using the input samples from the area with high local density. Recently, Al-Daoud and Roberts [1] proposed a method that combines local density approximation and random initialization.

In this paper, seed selection to cluster granules is generated by local density approximation method. Careful seed selection helps us to locate hidden sequence motifs that transcend protein sequences. These identified motifs may have biological importance in day to day life.

The rest of the paper is organized as follows. Section II shows related work in this area of research. Clustering algorithm is explained in section III. In section IV the method adopted in the proposed work for seed selection process have been explained. In section V, experimental analysis and motif patterns are provided. Section VI concludes the paper with directions for further enhancement.

II. Related Work

Han and Baker [9] have first used K-Means clustering algorithm for finding protein sequence motif. They have chosen set of initial points for cluster centers in a random manner. Selecting initial points randomly leads to an unsatisfactory partition because some initial points may lie close to each other. In order to overcome the above mentioned problem, Wei Zhong [21] has proposed Improved K-Means clustering to explore sequence motifs. Improved K-Means algorithm tries to obtain initial seeds by using Greedy approach. In this approach, for each run, clustering algorithm will be executed for fixed number of iterations and then selects initial seeds which have capacity to form clusters with good structural similarity. The distance of chosen initial seeds 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 seeds using above greedy approach leads to high computational cost. Computational cost is a major problem to be faced when input data-set is very large.

Hence, Bernard Chen [3], [4] has proposed granular computing model using Fuzzy clustering technique. In his work of Fuzzy Improved K-Means algorithm, 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, 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.

III. Clustering Algorithm

This section explains the original K-Means clustering algorithm. The idea is to classify a set of input samples into K number of disjoint clusters, where the value of K is fixed in advance. The algorithm consists of two

separate phases: the first phase is to define K seeds, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the seeds are included in some cluster, first step is completed and initial grouping is done.

Next, we need to recalculate the new centroids by including new seeds which leads to a change in the cluster centroids. Once we find K new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, K centroids may change their position in a step by step manner. Finally, a situation will be reached where centroids do not move anymore. This signifies the convergence criterion for clustering. Pseudocode for the K-Means clustering algorithm is listed as Figure 2.

The K-Means algorithm is the most widely studied clustering algorithm and is generally effective in producing good results [14]. The major drawback of this algorithm is that it produces different clusters for different set initial centroids. Quality of the final clusters highly depends on the selection of the initial centroids. Therefore our proposed work tries to find potential seeds in a different manner.

Algorithm: K-MeansClustering

Input: D = {d_1, d_2,.....,d_n} // set of n data points

K // Number of desired clusters

Output: Partitions of the input data into K clusters

Steps:

Phase 1: Determine the initial centroids of the cluster by using Figure 4.

Phase 2: Assign each data point to the appropriate clusters by using Figure 5.

Figure 2. K-Means Clustering Algorithm

IV. PROPOSED WORK

Granular computing represents information in the form of "information granules". The strategy of divide and conquer can be used effectively to solve many types of large and complicated problem [3]. It can also be related in the sense that a large problem is decomposed into a smaller problems and solution of the large problem is obtained by combining the solutions of smaller problems. An information granule may be interpreted as one of the numerous small particles forming a larger unit. By considering a small group as a granule, we can obtain results from the theory of small groups

Cluster analysis is the unsupervised classification of patterns into similar grouping. It is useful in various applications. One of the most popular clustering algorithms is the K-Means algorithm. The performance of K-Means clustering depends on the initial guess of partition. The concept behind the proposed seed selection technique is that the whole input samples are evenly divided into M subspaces and number of clusters in each subspace is been calculated. For each subspace KKZ method of centroid initialization is adopted [1].

Finally combine all seeds obtained in the previous step based on different distance threshold value. As distance threshold increases the number of initial centroid's obtained may not be enough. In this case, we use a random method with minimum distance threshold to choose the rest of required seeds. K-Means clustering is performed on the seeds obtained by new initialization technique. Final, motif information is been extracted from highly structural similar clusters. Figure 3 depicts the flow of proposed seed selection technique and Figure 4 shows the proposed seed selection algorithm. Figure 5 shows the algorithm for assigning data points to each cluster.

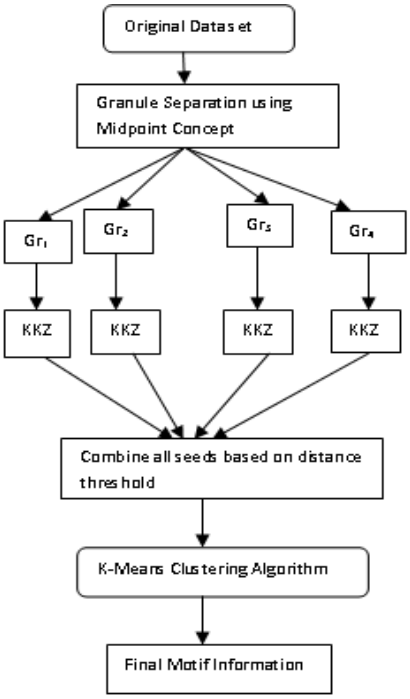

Figure 3. Flow of Proposed Seed selection Technique

Algorithm: New Seed Selection Algorithm

Input: D = {d_1, d_2,.....,d_n} // set of n data points

C // Number of desired seeds

Output: A set of C initial seeds.

Procedure:

Step 1: Divide whole input domain into x subspaces

S_j, j = 1....x. Let y_j be samples in each subspace.

Step 2: Decide number of seeds in each subspace by rounded integer $N.y_j/y$.

Let it be K.

Step 3: For each subspace x

Step 4: Initialize first seed cluster

$C_1 = y_{j1} \equiv \text{argmax}\{ \| y_j \| \}$

Step 5: For i = 2 to K

For each input sample $y_j$calculate its distance to the closest seed cluster

$d_j = \min \{ \| y_j - C_k \|$ for all existing $C_k \}$ and set

$C_i = y_{j1} \equiv \text{argmax} \{ d_j \}$

End For

End For

Step 6: For each subspace x

C = combine all seeds based on distance threshold.

End for

Figure 4. Proposed Seed Selection Algorithm

Algorithm: Assigning data points to clusters

Input: D = {d_1, d_2,.....,d_n} // set of n data points

C = {c_1, c_2,.....,c_k} // set of k centroids

Output:

Partitions of the input data into K clusters

Procedure

Step 1: Compute the distance of each data point d_i ($1 \le i \le n$) to all the centriodsc_j ($1 \le j \le k$) as$dis_{ij} = \|d_i - c_j\|$using City Block distance.

Step 2: For each data point d_i, find the closest centroid c_j and assign d_i to cluster j.

Step 3: Compute new cluster centroids $c_j$by taking mean for each cluster.

Step 4*:* Repeat steps 1 to 3 until convergence.

Figure 5. Assigning Data Points to Clusters

V. EXPERIMENTAL SETUP

A. Data Set

The latest dataset obtained from Protein Culling Server (PISCES) [20] 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. Homology Secondary Structure Prediction (HSSP) frequency profiles are used to represent each segment [14]. Database of Secondary Structure Prediction (DSSP) assigns secondary structure to eight different classes [12], [18]. In this paper, we convert those eight classes to three different classes based on the CASP experiment as follows [3]: H, G and I to H (Helices); B and E to E (Sheets); all others to C (Coils)

B. Structural Similarity Measure

Average structural similarity of a cluster is calculated using the following formula:

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

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

C. Distance Measure

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

$$\text{Distance} = \sum_{i=1}^{w}\sum_{j=1}^{N} |D_s(i,j) - D_c(i,j)| \quad (2)$$

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

D. David-Bouldin Index (DBI) Measure

Davis-Bouldin Index, measures how compact and well separated the clusters are. To obtain clusters with these characteristics, the dispersion measure for each cluster needs to be small and dissimilarity measure between clusters need to be large [5].

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} R_i \quad (3)$$

Where $R_i = \max_{j=1\ldots k, j\neq i} R_{ij}$, i=1...k (4)

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

$$\text{dinter}(c_i, c_j) = \sum_{k}^{l} || \bar{x}_{ik} - \bar{x}_{jk} || \quad (5)$$

and dispersion of a cluster c_i is defined as

$$\text{dintra}(c_i) = \sum_{i=1}^{Np} || x - \bar{x}_i || \quad (6)$$

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

E.HSSP-BLOSUM62 Measure

HSSP stands for Homology-Derived Secondary Structure of Proteins [17]. 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 our method. HSSP frequency profile and BLOSUM62 matrix has been combined to obtain significance of motif information. Hence, the measure is defined as the following [3].

If m = 0: HSSP-BLOSUM62 measure = 0

Else If m = 1:HSSP-BLOSUM62 measure = BLOSUM62_{ii}

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

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

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

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

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

F. Experimental Results

The proposed new seed selection technique is applied for K-Means clustering. In this work, the number of clusters has been set to 900. Cluster quality and significance of motif information are measured using two metrics such as DBI measure and HSSP-Blossum62 measure. In table I, number of clusters with high structural similarity and average percentage of sequence segments belonging to clusters with high structural similarity are shown. The first column refers different parameters and second column shows results of standard K-Means algorithm applied on whole dataset. Third column "KKZ 1200" indicates dataset is clustered by using new seed selection technique and distance threshold is set as 1200. "KKZ 1300", "KKZ 1400", "KKZ 1500" are defined in a similar manner.

From Table I, it is observed that as distance between initial seeds are high we can able to find increased number of highly and weakly structural similar clusters. Number of sequence segments of clusters does get increased. It is noted from table I that the results obtained after new seed selection technique generates more biochemical meaningful motif information more than random initialization of K-Means algorithm.

TABLE I. COMPARATIVE ANALYSIS OF DIFFERENT MEASURES

	K-Means	**KKZ with 1200**	**KKZ with 1300**	**KKZ with 1400**	**KKZ with 1500**
Number of clusters with structural similarity > 70%	110	105	111	112	116
Number of clusters with structural similarity > 60%<70%	148	151	155	157	159
Number of sequence segments > 70% structural similarity	0.1537	0.1562	0.1604	0.1666	0.1684
Number of sequence segments > 60%<70%	0.1629	0.1650	0.1679	0.1782	0.1807

TABLE II. COMPARISON OF DBI AND HSSP-BLOSUM62 MEASURE

	DBI Measure	**HSSP-BLOSUM62 Measure**
Random Initialization	6.1515	0.5432
KKZ with 1200	6.1570	0.7796
KKZ with 1300	6.1511	0.7846
KKZ with 1400	6.1472	0.7868
KKZ with 1500	6.1055	0.7969

Table II, shows DBI measure values decreases on increase in distance threshold which shows quality of clusters gets increased. Motif information obtained in new cluster initialization technique has also been increased by looking towards HSSP Blosum62 Measure.

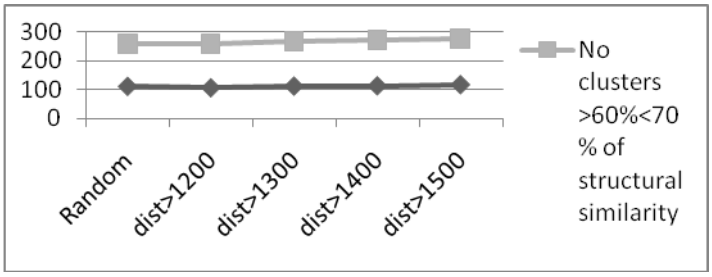

Figure 7. Comparison of clusters with high and weak structural similarity

Figure 7 is interpreted from table I which shows graphical representation of highly structural similar clusters and weakly structural similar clusters. Figure 8 is interpreted from table I which shows percentage of sequence segments belonging to clusters with high and weak structural similarity. From figure 7 and figure 8 we can able to understand that as distance between initial seed gets increased we can able to identify more number of clusters with high and weak structural similarity.

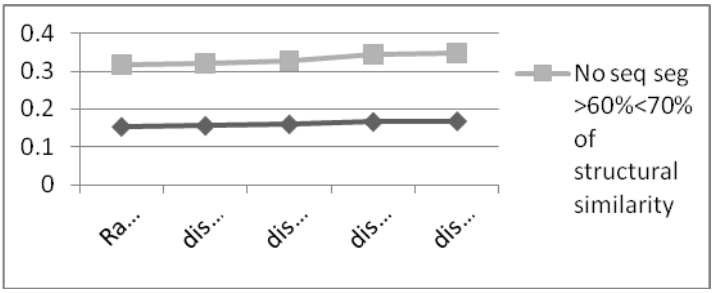

Figure 8. Comparison of Percentage of sequence segments belonging to cluster with high and weak structural similarity.

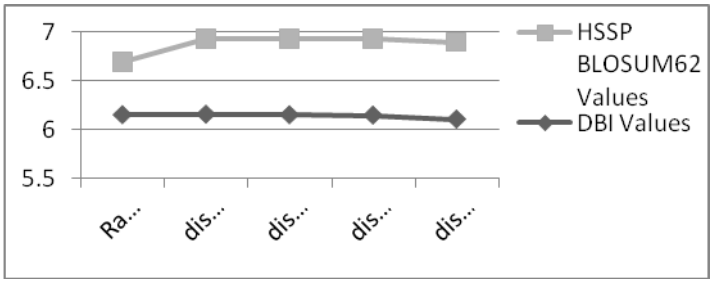

Figure 9. Comparison of DBI and HSSP BLOSUm62 Measure Values

Figure 9 shows comparative analysis of cluster quality and quality of motif information. Decreased DBI value and increased HSSP-BLOSUM62 values shows the performance of clustering and significance of motif information gets increased when initial seeds are more apart.

G. Sequence Motifs

Different motif patterns are shown in figure 10 and 11. The following format is used for representation of each sequence motif table. In this proposed work protein logo representation has been used to represent sequence motifs generated by new seed selection technique.

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

Motif Table 1
Helices Motif with conserved G
Number of Sequence Segments: 1300
Structural Similarity: 85.06%
HSSP-BLOSUM62: 0.3578

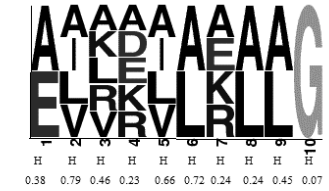

Figure. 10.Helices Motif with Conserved Glycine

Figure 10 shows helices motif with conserved with Glycine amino acid. Glycine can be substituted by all other small amino acids. It is a unique residue as it contains hydrogen as its side chain meaning more conformational flexibility. Clusters with conserved glycine residues are particularly common because of its conformational flexibility. These patterns may be advantageous in local structures with unusual backbone torsion angles.

Motif Table 2
Helices Motif with conserved L,A
Number of Sequence Segments: 1438
Structural Similarity: 77.09%
HSSP-BLOSUM62: 0.2546

Figure 11. Helices Motif with Conserved A

Figure 11 shows Helices Motif conserved with Analine amino acid. Alanine is a non-polar amino acid. The side chain of alanine is very non-reactive. It plays a role in substrate recognition or in specificity.

VI. CONCLUSION

The K-Means algorithm is commonly used for large datasets. But, the benchmark algorithm do not always guarantee good results as the accuracy of final clusters depends on the selection of initial centroids. In this paper, an attempt has been made to select seeds for K-Means Clustering algorithm. Furthermore, our seeding technique is fast and simple which makes it attractive and practice in the field of bioinformatics for indentifying protein sequence motifs from large volume of segments. Our work highlights that granular computing along with KKZ tend to help K-Means in producing clustering solution with significantly better cluster separation. The new seeding method helps us to identify more number of sequence motifs that are hidden inside proteins in an efficient manner. Cluster compactness of our proposed seeding technique outputs is satisfactory as well. Our future work aims by extending this frame work with methods for refining the computation of initial centroids.

REFERENCES

[1] M. Al-Daoud and S. Roberts. “New methods for the initialization of clusters”, Technical Report 94.34, School of Computer Studies, University of Leeds, 1994.

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

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

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

[5] D.L Davies, and D.WBuldin, “A cluster separation measure”,IEEE Trans. Pattern Recogn. Machine Intell.,vol. 1, pp. 224-227,1979.

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

[7] E.Eskin and P.APevzner, “Finding composite regulatory pattern in DNA sequences”,Bioinformatics, 18(Suppl.1)354-363, 2002.

[8] E. Forgy, “Cluster analysis of multivariate data: efficiency vs interpretability of classifications”, In WNAR meetings,Univ of Calif Riverside, pp. 768, 1965.

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

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

[11] N.Hullo, C.J.ASigrist, V.LeSaux, P.SLangendijk-Genevaux, L.Bordoli, A.Gattiker, E. De Castro, P. Bucher, and A. Bairoch, “Recent improvements to the PROSITE database”, Nucleic Acids Res, vol. 32, Database issue: D134-137, 2004.

[12] 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.

[13] I. Katsavounidis, C.Kuo, and Z.Zhang, “A new initialization technique for generalized Lloyd iteration”, IEEE Signal Proceesing Letters, 1(10):144-146, 1994.

[14] L.Kaufman and Rousseeuw, “Finding groups in data: an introduction to cluster analysis”, Wiley, New York, 1990.

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

[16] J.BMacQueen, “Some methods for classification and analysis of multivariate observations”,In proceedings of the 5thBerkely symposium in mathematics and probability, pp. 281-297, 1967.

[17] 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.

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

[19] J.TTou and R.C. Gonzalez, “Pattern Recognition Principals”, Addison-wesley, Massachusetts, 1974.

[20] G.Wang and R.LDunbrack,Jr., “PISCES:a protein sequence culling server”, Bioinformatics,vol,19, No.12,pp.1589-1591,2003.

[21] 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.