A Novel Entropy Based Segment Selection Technique for Extraction of Protein Sequence Motifs

M Chitralegha[1], Dr K Thangavel[2]

**[1] Research Scholar, Department of Computer Science, Periyar University,
Salem, Tamil Nadu, India- 636 011.**

**[2] Professor and Head, Department of Computer Science, Periyar University,
Salem, Tamil Nadu, India- 636 011.**

Abstract

Bioinformatics is the combination of Biology, Mathematics and Information Technology. It is a study of management and analysis of De-oxyribo Nucleic Acid, Ribo Nucleic Acid and protein sequence data. In Bioinformatics, motif finding is one of the most popular problems which have got lot of applications in diagnosing the diseases, drug designing and protein classification. It is essential to have an efficient technique to explore sequence motif from protein sequences. Data mining is one such technique. Bioinformatics dataset frequently contains large volume of segments generated from protein sequences. However, all the generated protein segments may not yield potential motif patterns. The segments have no labels or classes. Hence, one has to apply unsupervised segment selection method to select the potential segments. In this paper, two novel unsupervised segment selection methods are proposed for first time based on Shannon Entropy and Singular Value Decomposition (SVD) based - Entropy. The proposed methods are evaluated using the benchmark K-Means clustering method. It is found that the proposed SVD-Entropy based segment selection produces more number of highly structurally similar clusters, through which we are able to generate significant motif patterns.

Keywords: *Clustering, Data mining, protein sequence, Motif, SVD – Entropy.*

1. Introduction

Proteins can be regarded as one of the most important elements in the process of life; they are more flexible than nucleic acids in structure because of larger number of type of residues, the increased flexibility and lower charge density of the polypeptide backbone. It can serve many roles in the cell as enzymes, structural components, and membrane components etc. Some of the most important functions of proteins are to regulate the expression of other proteins. Higher order structures such as motifs and domains are said to be some of Components of proteins. The proteins are end products of the genes which encode them [7].

The term “motif” refers to a region or portion of a protein sequence that has specific structure and is functionally significant. Protein families are often characterized by one or more such motifs. 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, prosthetic attachment site etc. [7]

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 [8]. 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, a huge number of segments are generated using sliding window technique [17] and patterns are extracted from the selected segments. Multiple protein sequences are represented by their corresponding HSSP file [14].

All the generated sequence segments may not be significant and may also affect final motif patterns. In [5], the segment selection process has been performed after clustering. In this paper Shannon and SVD Entropy are proposed to eliminate some of the segments generated from HSSP file using sliding window concept before clustering. The resulted sequence segments are then Clustered using 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 step of the procedure is depicted in figure 1.

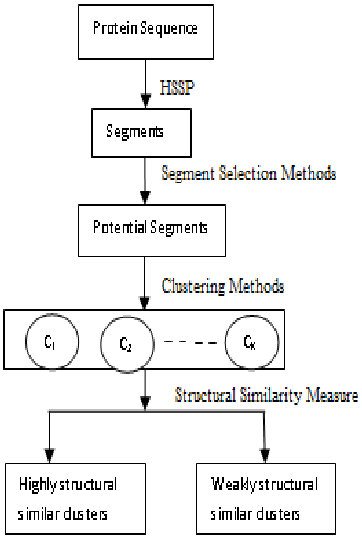

Protein Sequence
HSSP
Segments
Segment Selection Methods
Potential Segments
Clustering Methods
C_1
C_2
- - - -
C_k
Structural Similarity Measure
Highly structural similar clusters
Weakly structural similar clusters

Figure 1: Flowchart of the Method

The rest of the paper is organized as follows. Section 2 shows related work in this area of research. Section 3, presents the research motivation. In section 4, the methods adopted in the proposed work for segment selection process have been explained. Clustering algorithm is explained in section 5. In section 6, experimental analysis and motif patterns are provided. Section 7 concludes the paper with directions for further enhancement.

2. 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 [19] has proposed Improved K-Means clustering to explore sequence motifs. 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 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 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 centriods are selected after five runs of K-Means and then K-Means algorithm is executed by considering those centriods. It consumes more time and complexity is also high.

Motif detection from a huge amount of sequences is a challenging task and not all the segments generated are so important. Therefore, Bernard Chen [5] has proposed Super Granular SVM Feature Elimination. In this approach the original dataset is first partitioned using Fuzzy C-Means clustering and then for each partition Greedy K-Means clustering algorithm is been implemented. Then ranking SVM based segment selection is done on each cluster to collect survived sequence segments. The survived segments are then clustered once again using Greedy K-Means to generate motif information.

The Super Granular SVM feature elimination technique requires more computational time for segment selection process. Here the computational time includes time taken for Fuzzy clustering plus time taken for Greedy K-Means clustering before segment selection. In this paper, our goal is to reduce computational time for segment selection process before applying the clustering method.

3. Research Motivation

Super Granular SVM Feature Elimination Model for protein sequence Motif Information Extraction has been proposed in [4]. In their work, granular computing technique is applied on whole dataset, and then adopted segment selection method to identify significant sequence segments. These selected sequence segments are clustered using various clustering algorithms. Computational time taken in Super Granular SVM feature elimination technique includes O(nkl). This motivated us to use segment selection technique without applying granular computing model which can reduce computational time for segment selection. Hence, in this proposed work, two unsupervised segment selection techniques such as Shannon – Entropy measure and SVD – Entropy measure have been proposed to identify significant sequence segments. Computational time for Shannon and SVD Entropy for sequence segments said to be comparatively less than granular computing. In the proposed work computational time for protein sequence segment selection is reduced to O(n).

4. Proposed Work

4.1 Segment Selection Techniques

A biological study needs an efficient tool for feature selection process. Feature selection is a problem that has to be addressed in many areas especially in Bioinformatics, Text Analysis and Object Recognition etc. Bioinformatics dataset frequently contain thousands or even hundreds of thousands of segments. All segments may not be important for some problems but sometimes only small subset of segments is usually relevant. Pre-processing is one of the step for searching small subset of segment to reduce computational constraints, allows to handle high throughput biological experiments and to separate good segments from noisy data [18].

Basically there are two types of feature (segments) selection methods. One is wrapper method and the other is filter method. Wrapper method has well specified objective function which is optimized through the selection. Segment filtering is a process of selecting segments without referring any other objective function. Therefore filtering method is more suitable process that can be applied in unsupervised manner.

In this proposed work, all sequence segments generated by sliding window technique may not yield highly structural similar clusters. Therefore, removing such noisy segments using entropy segment selection helped us to produce clusters with good structural similarity.

4.2 Shannon Entropy-Based Segment Selection Algorithm

One commonly used sequence conservation measure is the entropy score. Therefore Entropy-based segment selection method is proposed to address the problem of selecting the significant segments [16]. The Shannon entropy is defined as

$$H(x) = -\sum_{i}^{n_{aa}} P_i log_2 P_i$$

where n_{aa} is the number of residue types in the column representing an alignment position, and p_i represents the observed frequency of residue type i in the aligned column.

```
Algorithm: Shannon Entropy Based Segment Selection

Input: Sequence segments of N numbers.

Output: Significant protein sequence segments.

Procedure:
Step1: Calculation of entropy
    For i = 1 to Number of sequence segments
    For j varies from 1 to window size of sequence segment
        Select number of non zero entries of amino acids
        For k varies from 1 to length of non zero entries
            Apply Shannon entropy formula
    End For
    End For
        Sum of entropies of each window will represent entropy of a sequence segment.
    End For
Step2: Selection of Sequence segments
    If entropy of each sequence segment is less
        than threshold value then Select those
        segments for clustering process.
    Else
        Eliminate the segments from clustering
        process.
    End If
```

Figure 2: Shannon Entropy based Segment Selection Algorithm

The entropy for each protein sequence segment is computed and selected only those segments that satisfy threshold criteria. These meaningful segments are then clustered by traditional K-Means [13].

4.3 SVD- Entropy Based Segment Selection Technique

SVD based entropy is proposed for the first time to address the problem of selecting the significant segments in the area of protein sequence motifs identification. The formula for calculating singular value decomposition of 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 window size.

The resulting SVD- Entropy is as follows [1]

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

```
Algorithm:  SVD Entropy Based Segment Selection

Input:    Sequence segments of N numbers.

Output:  Significant protein sequence segments.

Procedure:
Step1: Calculation of entropy

      For i = 1 to Number of sequence segments
          Calculate singular value decomposition
          for each sequence segment
      Let K = Number of non zero SVD entries
              along window size
      For j varies from 1 to K
              Apply SVD Entropy formula
      End For
      End For
Step2: Selection of Sequence segments
        If entropy of each sequence segment is less
           than threshold value then Select those
           segments for clustering process.
        Else
              Eliminate the segments from
              clustering process.
        End If
```

Figure 3: SVD – Entropy based Segment Selection Algorithm

The segments are categorized into three categories [1]. Let m to be average of all SVD entropy and their standard deviation to be n.

i. E < m + n, segments with high contribution.

ii. m + n > E > m - n, segments with average contribution.

iii. E < m - n, segments with negative contribution.

The segments obtained in first group are said to relevant to the identification of sequence motif 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 clustering algorithm [13]. The motif information obtained after segment selection process is said to be more meaningful as well as DBI value is also considerably decreased after segment selection process.

5. Clustering Algorithms

5.1 K-Means Clustering

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 o 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 points are included in some cluster, the first step is completed and initial grouping is done.

Next we need to recalculate the new centroids, including new points may lead 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 centriods do not move anymore. This signifies the convergence criterion for clustering [13].The step by step procedure of K-Means algorithm is detailed in figure 4.

```
Algorithm: K-Means

Input : Dataset $X$ of $n$ objects with $d$
        features and number of clusters $K$, $(K < n)$

Output: Partitions of the input data into K
        clusters

Procedure

Step-1: Declare a membership matrix $U$ of
        size n x K

Step-2: Generate $K$ cluster centroids
        randomly within the range of the
        data or select K objects randomly as
        initial cluster centroids. Let the
        centroids of each cluster be $c_1, c_2, \dots, c_K$

Step-3: Calculate the distance measure
        $d_{ij} = \| x_i - c_j \|$ using City Block
        distance, for all cluster centroids $c_j$,
        $j = 1, 2, \dots, K$

Step-4: Compute the $U$ membership matrix
        $\forall_{1 \le i \le n,\ 1 \le j \le K} \ U_{ij} = \begin{cases} 1; & d_{ij} \le d_{il},\ j \ne l \\ 0; & otherwise \end{cases}$

Step-5: Compute new cluster centroids $c_j$
        with each data object
        $\forall_{1 \le j K} \ c_j = \frac{\sum_{i=1}^{n} (U_{ij})\, x_i}{\sum_{i=1}^{n} (U_{ij})}$

Step-6: Repeat steps 3 to 5 until
        convergence.
```

Figure 4: K-Means Algorithm

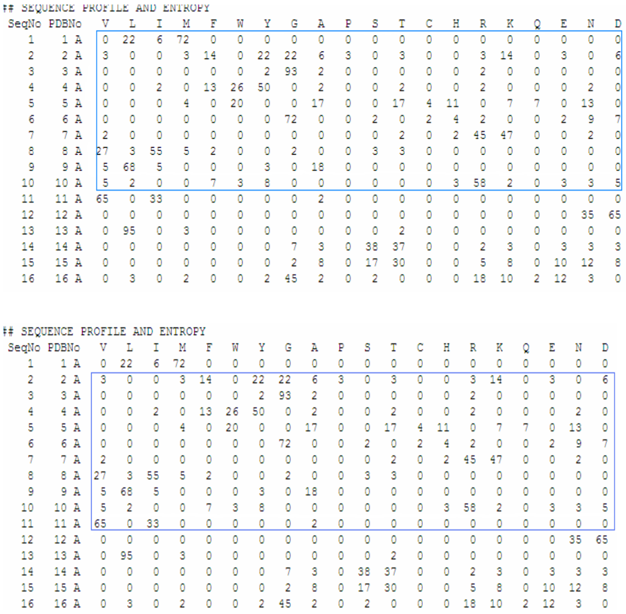
6. Experimental Setup

6.1 Data Set

The latest dataset obtained from Protein Culling Server (PISCES) [17] which includes 4946 protein sequences. In our 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 660,364 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.

The frequency profile from HSSP [14] is constructed based on the alignment of each protein sequence from the protein data bank (PDB) where all the sequences are considered homologous in the sequences database. Figure 5 shows sliding window technique applied on 1b25 HSSP file.



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

Figure 5: Sliding Window techniques with a window size of 10 applied on 1b25 HSSP file. Thus by applying the sliding window technique we can generate n number of sequence segments (10 X 20 matrices).

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, 15]. 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).

6.2 Structural Similarity Measure

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

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

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.

6.3 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.

6.4 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 [6].

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

where $R_i = \max_{j=1....k_{j\neq i}} R_{ij}$, i=1...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 Np is number of members in cluster c_i .Small values of DB are indicative of the presence of compact and well separated clusters.

6.5 HSSP-BLOSUM Measure

HSSP stands for Homology-Derived Secondary Structure of Proteins. 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, one could tell the consistency of the amino acid appearing in the same position of motif information generated by the proposed 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.

6.6 Experimental Results

The proposed SVD - Entropy based feature selection is applied and selected around 85% of the segments. Then K-Means algorithm is applied to cluster sequence segments. 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. Table 1 shows comparative values of three distinguished features.

From table 1, we infer that the type-3 threshold criterion is able to produce large number of strong and weak clusters as well as DBI value has also been reduced considerably.

Table 1: Comparison of three threshold values

Type	Threshold criteria	Number of strong clusters (nstr)	Number of weak clusters (nwek)	DBI Value
Type-1	E < (m - n)	73	161	5.6551
Type-2	(m + n) > E > (m - n)	84	169	5.5612
Type-3	E < (m + n)	113	185	5.4647

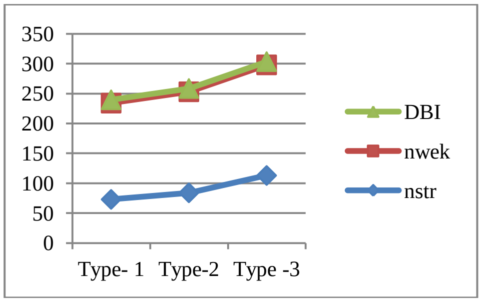


Figure 6: Comparison of three types of threshold values

Hence we have chosen type-3 criteria as our threshold value being relevant to our problem. Figure 6 is interpreted for the results given in table 1.

Table 2 – Comparison of Values Before and After Segment Selection process

	K-Means	K-Means with Shannon Entropy	K-Means with SVD Entropy
No of clusters >60% and < 70%	174	135	185
No of clusters > 70%	85	76	113
% of Seq Segments > 70%	15.5329	12.6515	19.3188
% of Seq Segments > 60% <70%	17.563	16.8132	21.2342
DBI Measure	5.7694	5.6576	5.4637
AverageHSSP-BLOSUM62	0.8165	0.8025	0.8567

Table 2 shows comparison between K-Means, Shannon Entropy and SVD-Entropy segment selection techniques. From above table 2, we can notice that structural similarity values have been increased in SVD Entropy segment selection technique.

It also seen that cluster quality has increased by looking towards DBI measure value and that motif information obtained in SVD Entropy segment selection technique is more significant compared to that of Shannon Entropy technique.

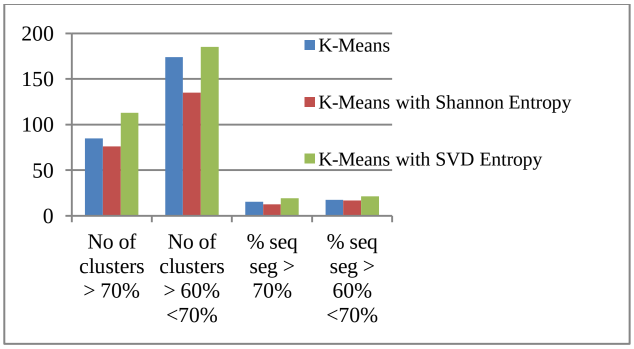


Figure 7: Comparison of Structural Similarity values

Figure 7 has been interpreted from Table 2. From the above Figure 7, we state that the number of strong and weak clusters has been increased after SVD- Entropy based segment selection technique as well as percentage of sequence segments have also been increased considerably.

Figure 8 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 obtained after SVD Entropy segment selection process is good. From the above Table 2, it is inferred that the results obtained after SVD Entropy segment selection process generates more biochemical meaningful motif information by eliminating some less meaningful data points. Figures 7 and 8 are interpreted for the results given in table 2.

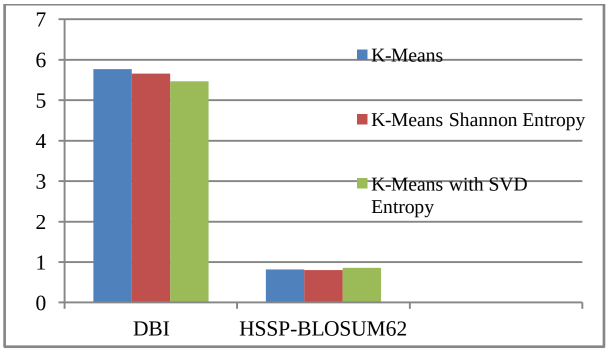


Figure 8: Comparison of DBI measure and HSSP-BLOSUM62 values

6.7 Sequence Motif Representation

Tables 3 to 5 show different sequence motif obtained before segment selection and after segment selection process of K-Means clustering. The following format is used for representation of sequence motif in each table [2]. The first row represents number of segments belonging to that particular motif, secondary structural similarity and average HSSP-BLOSUM62 value.

- First column represents window position.
- The second column represents noticeable amino acids in each position. If amino acids appearing with frequency greater than 10% are represented by upper case character and amino acids appearing with frequency between 8% and 10% are represented by lower case characters.
- The third column represents corresponding hydrophobicity value, which is the sum of frequencies of occurrences of Leu, Pro, Met, Trp, Ala, Val, Phe, Ile.
- The fourth column represents HSSP-BLOSUM62 measure value.
- The fifth column represents secondary structure of the position.

K-Means Clustering

Before Segment Selection Number of Segments 1770 Structural Homology : 51.08 Avg HSSP BLOSUM62 : 0.6762					After Segment Selection Number of Segments 823 Structural Homology : 63.73% Avg HSSP BLOSUM62 : 0.769				
pos	Noticeable Amino Acids	H	B	S	pos	Noticeable Amino Acids	H	B	S
1	lA	0.46	-1	H	1	AkE	0.34	-0.41	H
2	AkEd	0.39	-0.39	H	2	VA	0.55	0	H
3	lAe	0.4	-1.52	H	3	VA	0.52	0	H
4	VLI	0.88	2.19	H	4	ARKED	0.22	-0.18	H
5	ArKe	0.35	0.02	H	5	ARKe	0.32	0.04	H
6	AkqED	0.23	0.06	H	6	VLI	0.92	2.33	H
7	A	0.48	4	H	7	vlA	0.50	-0.13	H
8	L	0.96	4	H	8	AkED	0.20	-0.04	H
9	arkEd	0.28	-0.2	H	9	A	0.50	4.0	H
10	AkE	0.27	-0.35	H	10	VLI	0.84	2.09	H

Table 3: Conserved Helices motif

Table 3 shows the motif obtained are conserved with amino acid 'A' and 'L' which have secondary structure of Helices.

Table 4 shows the motif obtained are highly conserved with amino acid 'VLI' which has secondary structure turn of sheets and coils.

Before Segment Selection Number of Segments 1113 Structural Homology : 57.60% Avg HSSP BLOSUM62 : 1.0669					After Segment Selection Number of Segments 779 Structural Homology : 65.63% Avg HSSP BLOSUM62 1.4410				
pos	Noticeable Amino Acids	H	B	S	pos	Noticeable Amino Acids	H	B	S
1	t	0.47	5.0	E	1	V	0.41	4.0	E
2	VLI	0.35	-1.99	E	2	vlGa	0.46	-1.26	E
3	VlAt	0.85	-0.11	E	3	Vl	0.48	1.0	E
4	VLI	0.29	1.87	E	4	VLI	0.64	1.81	E
5	VLi	0.38	1.92	E	5	VLit	0.53	0.89	E
6	sTD	0.21	-0.19	C	6	nD	0.09	1.0	C
7	aped	0.43	-0.15	C	7	AE	0.35	-1.0	C
8	SeND	0.81	-0.80	C	8	eND	0.15	1.13	C
9	G	0.27	6	C	9	G	0.05	6.0	C
10	trKe	0.29	0.02	C	10	RKn	0.24	0.88	C

Table 4: Sheet-Coils Motif

Before Segment Selection Number of Segments: 1018 Structural Homology : 59.49% Avg HSSP BLOSUM62 : 0.2978					After Segment Selection Number of Segments: 551 Structural Homology : 63.47% Avg HSSP BLOSUM62 0.8078				
pos	Noticeable Amino Acids	H	B	S	pos	Noticeable Amino Acids	H	B	S
1	aED	0.27	0.08	H	1	AEd	0.35	-0.40	H
2	A	0.74	4.00	H	2	as	0.38	1.00	H
3	laRkE	0.40	-0.75	H	3	la	0.46	-1.00	H
4	ArKE	0.30	-0.27	H	4	AkED	0.27	-0.17	H
5	vLA	0.60	-0.22	H	5	AED	0.32	-0.41	H
6	vlA	0.68	-0.33	H	6	VLI	0.82	2.20	H
7	ARKED	0.23	-0.16	H	7	lAr	0.47	-1.25	H
8	ArkE	0.34	-0.20	H	8	AKqED	0.22	0.11	H
9	VLI	0.77	1.83	H	9	A	0.45	4.00	H
10	lA	0.46	-1.00	H	10	L	0.97	4.00	H

Table 5: Conserved Helices motif

Table 5 shows the motif obtained are conserved with amino acid 'V' and 'L' which has secondary structure of Helices.

7. Conclusion

Dataset is said to be very large and not all sequence segments generated by sliding window technique is useful for motif generation. Hence, it is necessary to select significant segments for motif generation. In this research, two unsupervised segment selection techniques have been proposed such as Shannon and SVD - Entropy based technique. The experimental results of two segment selection techniques are compared and it is observed that SVD – Entropy method produces better result than Shannon entropy technique. The Bench mark K-Means algorithm is executed to cluster the selected sequence segments. Finally we compared the results of clustering technique before and after segment selection process. The motif patterns obtained in the proposed method is as good as the existing methods. Instead of applying K-Means algorithm with random method of selecting centriods, centroid selection algorithm may be incorporated in clustering algorithm. This is the direction for future research.

References

[1]. O. Alter, P.O Brown and D. Boststein, “Singular value decomposition for genome-wide expression data preprocessing and modelling”, PNAS, Vol. 97, No.18, 2000, pp. 10101-10106.

[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 Acid Res. Vol. 30, No. 1, 2002, pp. 239-241.

[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), 2006, pp. 56-61.

[5]. B. Chen, P.C Tai, R. Harrison and Y. Pan, “Super GSVM-FE model for protein Sequence Motif Information Extraction”, in proc. IEEE symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2007, pp. 317-322.

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

[7]. David W. Mount, Sequence and Genome Analysis, New York: Cold Spring Harbor Laboratory Press, 2001.

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

[9]. K.F Han and D. Baker, “Recurring local sequence motifs in proteins”, J. Mol. Bio, Vol. 251, No. 1, 2005, pp. 176-187.

[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, 1999, pp. 417-479.

[11]. 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, 2004, Database issue: D134-137.

[12]. W. Kabsch and C. Sander, “Dictionary of protein secondary structure pattern recognition of hydrogen-bonded and geometrical features”, Biopolymers, Vol. 22, 1983, pp. 2577-2637.

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

[14]. 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, 1991, pp. 56-68.

[15]. 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, 1991, pp. 56-68.

[16]. Shannon, C.E and Weaver W. The Mathematical Theory of Communication, United States: University of Illinois press, 1949.

[17]. G. Wang and R. L. Dunbrack, Jr., “PISCES: a protein sequence-culling server,” Bioinformatics, Vol. 19, No. 12, 2003, pp.1589-1591.

[18]. J Weston, F Pérez-Cruz, O Bousquet, O Chapelle, A. Elisseeff, and B. Schölkopf: “Feature Selection and Transduction for Prediction of Molecular Bioactivity for Drug Design”, Bioinformatics 2002, **1:**1-8.

[19]. 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, 2005, pp. 255-265.