Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Super Granular SVM Feature Elimination (Super GSVM-FE) Model for Protein Sequence Motif Information Extraction

Bernard Chen[1], Stephen Pellicer[1], Phang C. Tai[2], Robert Harrison[12] and Yi Pan[1]
Georgia State University, Computer Science[1] and Biology[2] department
34 Peachtree Street Suit 1450 Atlanta, GA, 30303
bchen3@cs.gsu.edu

Abstract— **Protein sequence motifs are gathering more and more attention in the sequence analysis area. These recurring regions have the potential to determine protein's conformation, function and activities. In our previous work, we tried to obtain protein sequence motifs which are universally conserved across protein family boundaries. Therefore, unlike most popular motif discovering algorithms, our input dataset is extremely large. In order to deal with large input datasets, we provided two granular computing models (FIK and FGK model) to efficiently generate protein motifs information. In this article, we develop a new method which combines the concept of granular computing and the power of ranking SVM to further extract protein sequence motif information. There are two reasons to eliminate redundant data: First, the information we try to generate is about sequence motifs, but the original input data are derived from whole protein sequences by a sliding window technique; second, during fuzzy c-means clustering, it has the ability to assign one segment to more than one information granule. However, not all data segments have a direct relation to the granule they assigned. The quality of motif information increases dramatically in all three evaluation measures by applying this new feature elimination model. Compared with traditional methods which shrink cluster size to obtain a more compact one, our approach shows improved results.**

***Index Terms*—FIK Model, FGK Model, Ranking SVM, Feature Elimination, Protein Sequence Motif.**

I. INTRODUCTION

It is generally believe that proteins are the most varied macromolecules in life process and play an extremely important role in all biological activities. To understand the close relationship between protein sequences and structures is one of the most-valuable tasks in bioinformatics research. In genetics, a sequence motif is a pattern of amino acid sequence that is widespread and contains biological significance. These recurring patterns have the potential to predict other protein's structural or functional area.

Some popular sequence motifs databases [8-10] and some tools [11-16] designed for motifs discovering are developed from multiple alignments. However, due to the limitation of input data size, these sequence motifs only search conserved elements of sequence alignment from the same protein family and carry little information about conserved sequence regions, which transcend protein families [1].

Traditional K-means clustering algorithm which selects initial centroids by random is employed by Han and Baker [2] to find recurring protein sequence patterns. Wei et al [1] proposed an improved K-means clustering algorithm to obtain initial centroids location by greedy concept. Since selecting the initial centroids by random is one major drawback of K-means clustering algorithm, the quality of cluster published by Wei et al has been improved in their experiment. The extremely large input dataset is the main reason that both above papers select K-means instead of some other advanced clustering technology. Since K-means is famous for its efficiency, other clustering methods with higher time and space costs may not be suitable for this task. In our previous work [22,23], we proposed granular computing models that utilized Fuzzy C-means clustering algorithm to divide the whole data space into ten smaller subsets and then apply improved K-means algorithm to each subset to discover relevant information. The total execution time is only 20% of [1] and obtains even higher quality of sequence motif information.

The input dataset in our previous works [22,23] and related works [1,2] are generated from whole protein sequences by sliding window technique, however, the information we tried to obtain is sequence motif knowledge which is only several small parts of each sequence. Therefore, not all segments in the dataset can provide significant information. Besides, fuzzy C-means (FCM) is capable of assigning one data point to more than one information granule, but not all granules that FCM assigned need the information from the data point. In this paper, in order to obtain more precise motif information, we utilize our previous effort and combine it with ranking SVM to eliminate the redundant or less meaningful segments in our original dataset. In [18], a granular feature elimination model applied on Microarray data called GSVM-RFE is proposed. Although Microarray has the potential to deal with tens of thousands genes simultaneously, compared with our data size, we have much larger dataset: four clusters are divided in [18], however, 48 clusters are mined in this paper. In addition, we use greedy K-means cluster algorithm to fix the initial centroid location to optimize the effect of feature

1-4244-0710-9/07/$20.00 ©2007 IEEE

elimination. Therefore, we called our model super GSVM-FE to indicate that our model has the potential to apply on huge data space and obtain consistent results.

II. Granular Computing Models

A. FGK Model

Granular computing represents information in the form of aggregates, also called “information granules” [20, 21]. For a large and complicated problem, it uses the divide-and-conquer concept to split the original task into several smaller subtasks to save time and space complexity. Also, in the process of splitting the original task, it comprehends the problem without including meaningless information. As opposed to traditional data-oriented numeric computing, granular computing is knowledge-oriented [21].

A granular computing based model called “Fuzzy-Greedy-Kmeans model” (FGK model) is proposed in our previous work [23]. This model works by using FCM to build a set of information granules and then applying our new greedy K-means clustering algorithm to obtain the final information. The new greedy method collects five traditional K-means results and then selects the initial centroids based on those results. Due to the fact that the centroids in higher quality clusters have the potential to generate better clusters in the sixth round, we divided our initial centroid selection procedure into five steps: initially gathering centroid seeds belonging to clusters with structural similarity greater than 80% and then proceeding with 75%, 70%, 65% and 60%. The major advantages of the FGK model are reduced time- and space- complexity, filtered outliers, and higher quality granular information results.

B. Super Granular SVM Feature Elimination Model

Basically, this new model is the next generation of FGK model. It also use fuzzy concept to divide the original dataset into several smaller information granules. For each granule, after five iterations of traditional K-means clustering, the greedy k-means is applied. The next step is different from the FGK model: we adapt the ranking SVM [17] to rank all members in each cluster generated by greedy K-means clustering algorithm, and then we filter out lower ranked members. The number of segments being eliminated is decided by user defined filtration percentage. The results of different percentages are discussed in section four. After the feature elimination step, we collect all surviving data points in each information granule and then run greedy K-means with the same initial centroids we previously generated. Finally, we collect all results in all granules to create the final protein sequence motif information.

In order to compare the results, we present another similar feature elimination approach by modifying only one component of the model: we utilize a cluster shrinking instead of ranking SVM. The number of segments being eliminated is decided by a user defined distance threshold. If the distance between a member and the center of the cluster is greater than the threshold, the data point is filtered. The major advantage of this approach is that not all clusters get rid of the same amount members. If the cluster is compact at the beginning, fewer members should be eliminated. On the other hand, if the cluster is in a loose form, more data points should be eliminated. The results of different thresholds are also discussed and compared in section four.

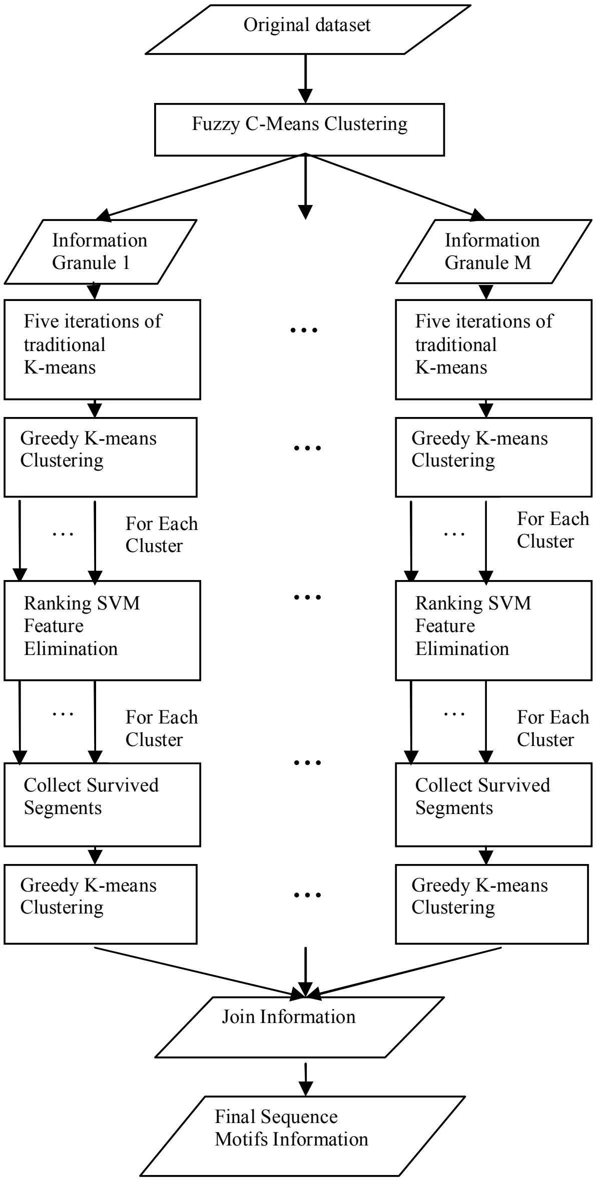
Original dataset
Fuzzy C-Means Clustering
Information Granule 1
Information Granule M
Five iterations of traditional K-means
Greedy K-means Clustering
For Each Cluster
Ranking SVM Feature Elimination
For Each Cluster
Collect Survived Segments
Greedy K-means Clustering
Join Information
Final Sequence Motifs Information

Figure 1. Super GSVM-FE Model

III. Experiment Setup

A. Dataset

The dataset used in this work includes 2710 protein sequences obtained from Protein Sequence Culling Server (PISCES) [5]. No sequence in this database share more than 25% sequence identities. Sliding windows with 9 successive residues are generated from protein sequence. Each window represents one sequence segment of nine continuous positions. More than 560,000 segments are generated by this method and clustered into 800 clusters. The frequency profile from the HSSP [3] 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 sequence database. We also obtained secondary structure from DSSP [4], which is a database of secondary structure assignments for all protein entries in the Protein Data Bank.

Due to time limitation, we applied our new approach on information granule number eight in [23] which contains 43254 data segments and 48 clusters. After five iterations of traditional K-means clustering are executed, 45 initial centroids are decided for greedy K-means. All initial centers have at least 250 distance measure from existing centroids. Another three initial seeds are generated randomly with minimum distance threshold check.

B. Representation of Sequence Segment

The sliding windows with nine successive residues are generated from protein sequences. Each window corresponds to a sequence segment, which is represented by a 9 × 20 matrix plus additional nine corresponding secondary structure information obtained from DSSP. Twenty rows represent 20 amino acids and 9 columns represent each position of the sliding window. 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 paper, we convert those eight classes into three classes based on the following method: H, G and I to H (Helices); B and E to E (Sheets); all others to C (Coils).

C. Distance Measure

According to [1, 2], the city block metric is more suitable for this field of study since it will consider every position of the frequency profile equally. The following formula is used to calculate the distance between two sequence segments [2]:

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

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.

D. Structure Similarity Measure

Cluster's average structure is calculated using the following formula:

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

Where ws is the window size and $P_{i,H}$ shows the frequency of occurrence of helix among the segments for the cluster in position i. $P_{i,E}$ and $P_{i,C}$ are defined in a similar way. If the structural homology for a cluster exceeds 70%, the cluster can be considered structurally identical [7]. If the structural homology for the cluster exceeds 60% and bellows 70%, the cluster can be considered weakly structurally homologous [1].

E. Davis-Bouldin Index (DBI) Measure

Besides using secondary structure information as a biological evaluation criterion, we include an evaluation method used in computer science on this dataset in our previous work [22]. The DBI measure [6] is a function of the inter-cluster and intra-cluster distances. A good cluster result should reflect a relatively large inter-cluster distance and a relatively small intra-cluster distance. The DBI measure combines these two distance information into one function, which is defined as follows:

$$\text{DBI}=\frac{1}{k}\sum_{p=1}^{k}\underset{p\neq q}{MAX}\left\{\frac{d_{\text{int}ra}(C_p)+d_{\text{int}ra}(C_q)}{d_{\text{int}er}(C_p,C_q)}\right\},\ \text{where}$$

$$d_{\text{int}ra}(C_p)=\frac{\sum_{i=1}^{n_p}\left\|g_i-g_{pc}\right\|}{n_p}\quad\text{and}\quad d_{\text{int}er}(C_p,C_q)=\left\|g_{pc}-g_{qc}\right\|$$

k is the total number of clusters, $d_{\text{int}ra}$ and $d_{\text{int}er}$ 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 distance 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 higher quality of the cluster result.

F. New HSSP-BLOSUM62 Measure

BLOSUM62 [19] is a scoring matrix based on known alignments of diverse sequences. By using this matrix, we may tell 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. For example, if the rule indicates amino acid A1 and A2 are two elements frequently appear in some specific position; A1 and A2 should have similar biochemical property. Hence, the measure is defined as the following:

If k = 0: HSSP-BLOSUM62 measure = 0

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

Else: HSSP-BLOSUM62 measure =

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

k is the number of amino acids with frequency higher than a certain threshold in the same position (in this paper, 8% is the threshold). $HSSP_i$ indicates the percent that amino acid *i* appears. $BLOSUM62_{ij}$ denotes the value of BLOSUM62 on amino acid *i* and *j*. The higher HSSP-BLOSUM62 value indicates more significant motif information. When *k* equals zero, it indicates that there is no amino acid appearing in the specific position, so the value for this measure is assigned zero. While *k* equals one, it indicates that there is only one amino acid appearing in the position. Unlike the first time we proposed this measurement in [23] where we assign zero score to this situation; we believe it should assign some positive value to appreciate the clear information. Therefore, we assign the corresponding amino acid's diagonal value in BLOSUM62. To the best of our knowledge, it is the first evaluation method based on biochemical point of view that considers both HSSP and BLOSUM62 information.

G. *Ranking SVM Setup*

In ranking SVM [17], the target value is used to denote pair wise preference constraints. We set the target value for each member in the cluster by counting the number of matching secondary structures between a member's structure and the representative structure of the cluster. Since we use window size nine in our experiment, the highest target value is 9 and the lowest is 0.

IV. EXPERIMENTAL RESULTS

A. *Quality of Sequence Motifs Comparison*

In table 1, the number of clusters which contain higher than 60% and 70% structural similarity generated by different methods is given. Both the DBI measure and the average HSSP-BLOSUM62 value on high structural similarity (>60%) clusters are also available in the same table. The leftmost column indicates the percentage of whole dataset been filtered out. Figure 2 to 5 are interpreted from first table.

The results of Table 1 and figures from 2 to 4 reveal that the quality of clusters improved in all three measures steady by filtering out part of the original data. Comparing the Shrink approach and Ranking SVM method from the secondary structural similarity point of view, it is not hard to tell that ranking SVM generates much better results. The support vector machine approach produces more clusters with higher than 60% structural similarity almost all the time. If we compare the number of cluster that share over 70% structural similarity, ranking SVM unquestionably surpasses the shrink approach. It indicates that our proposed model has the higher potential to bring forth high quality motif information.

When it comes to the DBI measure which is a pure computer science aspect evaluation, shrink method always receives a lower (indicates better) value. This is mainly because the shrink method is based on simply narrow cluster size from outside, in other words, it focus on shorter intra-cluster distance. Therefore, it can always generate a better DBI value. Although the SVM approach has a larger DBI value, the difference between the compared methods is small. More importantly, the ranking SVM shows the same tendency of decreasing the DBI measurement with shrink approach. In our HSSP-BLOSUM62 measure aspect, ranking support vector machine gives higher value throughout most of the time. It implies that our model generates more biochemical meaningful motif information by ruling out some less meaningful data points.

	Shrink Approach				Ranking SVM Approach			
	>60%	>70%	DBI	H-B	>60%	>70%	DBI	H-B
0%	12	0	6.387	.749	12	0	6.387	.749
≈10%	15	0	6.135	.634	14	1	6.352	.773
≈15%	15	0	6.060	.731	15	2	6.213	.908
≈20%	14	0	5.988	.818	18	2	6.058	.802
≈25%	15	1	5.793	.760	18	5	6.069	.884
≈30%	17	1	5.794	.655	25	6	5.892	.916
≈35%	18	1	5.719	.597	26	6	5.919	.720
≈40%	19	1	5.639	.694	26	7	5.856	.821
≈45%	19	1	5.604	.595	28	9	5.678	.736

Table 1 Comparison of all measures

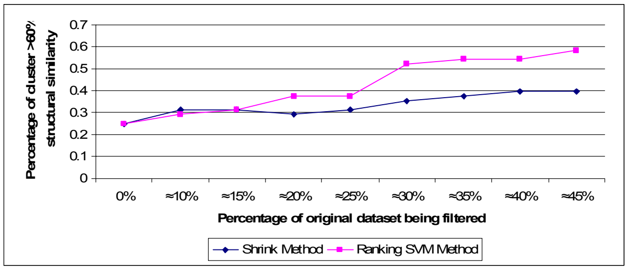

Figure 2 Comparison of percentage of sequence segments belonging to cluster with structure similarity higher than 60%

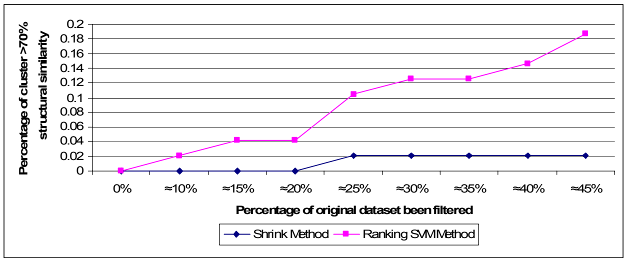

Figure 3 Comparison of percentage of sequence segments belonging to cluster with structure similarity higher than 70%

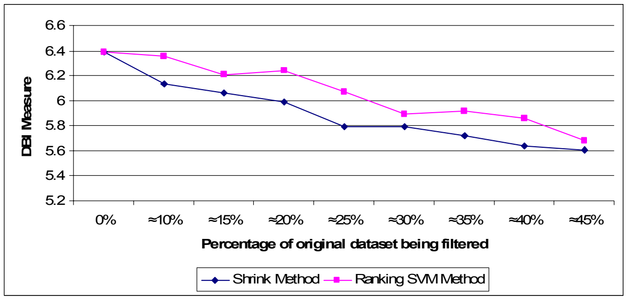

Figure 4 Comparison of the DBI measure (lower indicates better).

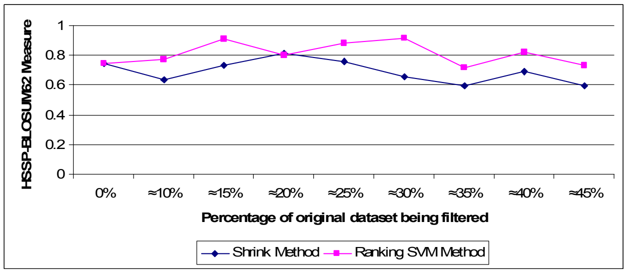

Figure 5 Comparison of the HSSP-BLOSUM62 measure

B. *Sequence Motifs*

Based on the results shown above, we select some motif information generated from filtering 30% of the whole data size and compare it with the original information. The reason we choose 30% as representative is that it matches the criteria of filtering part of data point and creates higher structural similarity results. Also, the lower DBI measure (almost equal to shrink method) and the highest HSSP-BLOSUM62 measure are considered. Tables 4 through 7 illustrate four different sequence motifs before and after feature elimination. The following format is used for representation of each motif table.

1. The first row represents number of members belonging to this motif and the secondary structural similarity.
2. The first column stands for the position of amino acid profiles in each motif with window size nine.
3. The second column expresses the type of amino acid frequently appeared in the given position. If the amino acids appearing with the frequency higher than 10%, they are indicated by upper case; If the amino acids appearing with the frequency between 8% and 10%, they are indicated by lower case.
4. 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.
5. The fourth column gives the HSSP-BLOSUM62 value.
6. The last column indicates the representative secondary structure to the position.

Before Feature Elimination						**After Feature Elimination**				
Number of segments: 830 **Structure homology:** 67.36% **Avg. HSSP-BLOSUM62:** 1.167					» » »	**Number of segments:** 624 **Structure homology:**74.02% **Avg. HSSP-BLOSUM62:** 1.562				
#	**Noticeable Amino Acid**	**H**	**B**	**S**		**#**	**Noticeabl Amino Acid**	**H**	**B**	**S**
1	V L I	.74	1.9	E	»	**1**	V L I	.74	2.0	E
2	V L I	.61	1.9	E	»	**2**	V L I	.62	2.0	E
3	s t n D	.22	.18	C	»	**3**	s t n D	.22	.18	C
4	g A p K	.27	-.76	C	»	**4**	g A pskEd	.27	-.59	C
5	Ed	.13	.38	C	»	**5**	s E N D	.11	.84	C
6	g e N D	.09	6.0	C	»	**6**	G	.07	6.0	C
7	G	.24	1.0	C	»	**7**	t K e	.22	-.24	C
8	K e	.51	-2.0	E	»	**8**	V l i	.55	2.0	E
9	V p V L I	.58	1.9	E	»	**9**	V L I	.53	1.9	E

Table 4 Sheet-Coil-Sheet motif

Number of segments: 966 **Structure homology:** 61.02% **Avg. HSSP-BLOSUM62:** 0.775					» » »	**Number of segments:**730 **Structure homology:** 65.51% **Avg. HSSP-BLOSUM62:** 0.890				
#	**Noticeable Amino Acid**	**H**	**B**	**S**		**#**	**Noticeabl Amino Acid**	**H**	**B**	**S**
1	V L I	.85	2.3	E	»	**1**	V L I	.85	2.3	E
2	V L I	.87	2.2	E	»	**2**	V L I	.83	2.2	E
3	A S D	.31	-.33	E	»	**3**	g a s t d	.33	-.39	E
4	S t n	.28	.70	C	»	**4**	s r n	.28	.04	C
5	A p S t	.33	-.13	C	»	**5**	a P S t	.32	-.27	C
6	a p D	.32	-.64	C	»	**6**	a p D	.33	-1.3	C
7	G d	.30	-1.0	C	»	**7**	s e d	.30	.74	C
8	a E D	.32	-.15	C	»	**8**	k E D	.24	.78	C
9	A	.38	4.0	C	»	**9**	a	.41	4.0	C

Table 5 Sheet-Coil motif

Number of segments: 703 **Structure homology:** 51.75% **Avg. HSSP-BLOSUM62:** 1.051					» » »	**Number of segments:**510 **Structure homology:** 60.11% **Avg. HSSP-BLOSUM62:** 1.262				
#	**Noticeable Amino Acid**	**H**	**B**	**S**		**#**	**Noticeabl Amino Acid**	**H**	**B**	**S**
1	V L I	.68	1.9	H	»	**1**	V L I f	.66	1.1	H
2	F y	.87	3.0	H	»	**2**	F y	.85	3.0	H
3	a S d	.30	-.17	C	»	**3**	G a p S e	.27	-.69	C
4	G a s e n d	.23	-.25	C	»	**4**	G a s e n d	.23	-.24	C
5	s d	.30	0.0	C	»	**5**	g s e d	.28	-.12	C
6	I p	.48	-3.0	C	»	**6**	I P s	.43	-2.0	C
7	p s	.35	-1.0	C	»	**7**	p s d	.35	-.70	C
8	e	.32	5.0	C	»	**8**	e	.31	5.0	C
9	a	.38	4.0	C	»	**9**	d	.35	6.0	C

Table 6 Helix-Coil motif

Number of segments: 1453 **Structure homology:** 57.77% **Avg. HSSP-BLOSUM62:** 1.712					» » »	**Number of segments:**1120 **Structure homology:** 71.44% **Avg. HSSP-BLOSUM62:** 0.858				
#	**Noticeable Amino Acid**	**H**	**B**	**S**		**#**	**Noticeabl Amino Acid**	**H**	**B**	**S**
1	V L I	.66	2.0	H	»	**1**	V L I	.67	1.9	H
2	V L I	.74	2.2	H	»	**2**	V L I	.82	1.9	H
3	A s	.35	2.0	H	»	**3**	g A s e	.34	-.22	H
4	a s E n	.22	1.1	H	»	**4**	a s r k qE	.20	.19	H
5	A	.41	1.9	H	»	**5**	I A	.47	-1.0	H
6	L	.94	-.29	H	»	**6**	L i	.94	2.0	H
7	g a k	.34	.67	H	»	**7**	I A r k	.37	-.83	H
8	A k E d	.30	-.64	H	»	**8**	A r k E d	.28	-.21	H
9	a	.39	-.28	H	»	**9**	A	.44	4.0	H

Table 7 Helices motif

V. Conclusion

A novel granular feature elimination model called Super GSVM-FE which combines Fuzzy C-means, Greedy K-means clustering algorithm and Ranking SVM has been proposed to extract protein sequence motif information. In this model, we utilize fuzzy clustering to split the whole dataset into several information granules and analyze each granule by Greedy K-means clustering algorithm. After that, we rate all members in all clusters by ranking SVM, and then filter out less meaningful segments to obtain higher quality motif knowledge. Analysis of sequence motifs also shows that by filtering some portion of original dataset may reveal some subtle motif information hidden behind some ordinary data points. Additionally, the latest version of HSSP-BLOSUM62 for motif information biochemical aspect measurement is also proposed. We believe some other research with large input data size may adapt our model to generate high quality purified results.

VI. Acknowledgement

This research was supported in part by the U.S. National Institutes of Health (NIH) under grants R01 GM34766-17S1 and P20 GM065762-01A1, and the U.S. National Science Foundation (NSF) under grants CCF-0514750 and ECS-0334813. This work was also supported by the Georgia Cancer Coalition and used computer hardware supplied by the Georgia Research Alliance.

References

[1] 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, vol4, no.3, pp. 255-265. 2005

[2] K. F. Han and D. Baker, "Recurring Local Sequence Motifs in Proteins ", J. Mol. Biol, vol. 251 pp. 176-187

[3] C. Sander and R. Schneider, "Database of similarity derived protein structures and the structure meaning of sequence alignment," *Proteins: Struct. Funct.* Genet., vol.9 no. 1, pp. 56-68, 1991.

[4] 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.

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

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

[7] C. Sander and R. Schneider, "Database of similarity derived protein structures and the structural meaning of sequence alignment," Proteins: Struct. Funct. Genet., vol. 9, no.1, pp. 56-68, 1991

[8] N. Hulo, C. J. A. Sigrist, V. Le Saux, P. S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, and A. Bairoch, "Recent improvements to the PROSITE database," *Nucleic Acids Res*., vol. 32, Database issue: D134-137, 2004

[9] 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, pp. 239-241, 2002

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

[11] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation Maximization to discover motifs in biopolymers.", Proc. Int. Conf. Intell. Syst. Mol. Biol., 2, 28–36. 1994.

[12] C. E. Lawrence, S. F. Altschul, M. S. Boquski, J. S. Liu, A. F. Neuwald, J. C. Wootton "Detecting subtle sequence signals: a Gibbs Sampling strategy for multiple alignment. "Science, 262, 208–214, 1993.

[13] S. Henikoff, J. G. Henikoff, W. J. Alford, S. Pietrokovski, "Automate construction and graphical presentation of Protein blocks from unaligned sequences." Gene, 163, GC17–GC26, 1995.

[14] E. Eskin and P. A. Pevzner, "Finding composite regulatory patterns in DNA sequences." Bioinformatics, 18 (Suppl. 1), 354–363, 2002.

[15] A. Price, S. Ramabhadran, P. A. Pevzner "Finding subtle motifs by branching from sample strings." Bioinformatics, 19 (Suppl. 2), II149–II155, 2003

[16] K. L. Jensen, M. P. Styczynski, I. Rigoutsos, G. N. Stephanopoulos "A Generic motif discovery algorithm for sequential data", Bioinformatics, Vol 22, no.1, pp. 21-28, 2006.

[17] T. Joachims, "*Optimizing Search Engines Using Clickthrough data"* Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002.

[18] Y. Tang, Y. Zhang, Z. Huang, X. Hu, "Granular SVM-RFE gene selection algorithm for reliable prostate cancer classification on microarray expression data" , in *IEEE proc. 5th symposium on Bioinformatics and Bioengineering (BIBE),* Minneapolis, 2005, pp. 290 – 293.

[19] S Henikoff and J G Henikoff "Amino acid substitution matrices from protein blocks." Proc. Natl. Acad. Sci. USA 89: 10915-10919, 1992.

[20] T. Y. Lin, "Data Mining and Machine Oriented Modeling: A Granular Computing Approach," Journal of Applied Intelligence, Kluwer, Vol 13, No 2, 113-124, 2002.

[21] Y.Y. Yao, "On Modeling data mining with granular computing," Proceedings of COMPSAC 2001, pp.638-643, 2001.

[22] B. Chen, P. C. Tai, R. Harrison and Y. Pan, "FIK Model: Novel Efficient Granular Computing Model for Protein Sequence Motifs and Structure Information Discovery", in *IEEE proc. 6th symposium On Bioinformatics and Bioengineering (BIBE),* Washington DC, 2006, pp. 20-26, 2006.

[23] B. Chen, P. C. Tai, R. Harrison and Y. Pan, "FGK Model: An Efficient Granular Computing Model for Protein Sequence Motifs Information Discovery", in I*ASTED proc. International conference on Computational and Systems Biology (CASB)*, Dallas 2006, pp56-61