

## FGK MODEL: AN EFFICIENT GRANULAR COMPUTING MODEL FOR PROTEIN SEQUENCE MOTIFS INFORMATION DISCOVERY

Bernard Chen<sup>1</sup>, Phang C. Tai<sup>2</sup>, Robert Harrison<sup>1,2</sup> and Yi Pan<sup>1</sup>  
Georgia State University, Computer Science<sup>1</sup> and Biology department<sup>2</sup>  
34 Peachtree Street Suit 1450  
Atlanta, GA, 30303  
bchen3@cs.gsu.edu

### ABSTRACT

Discovering protein sequence motif information is one of the most crucial tasks in bioinformatics research. In this paper, we try to obtain protein recurring patterns which are universally conserved across protein family boundaries. In order to achieve the goal, our dataset is extremely large. Therefore, an efficient technique is required. In this article, short recurring segments of proteins are explored by utilizing a granular computing strategy. First, Fuzzy C-Means clustering algorithm (FCM) is applied to separate the whole dataset into several smaller information granules and then followed by a novel greedy initialization OF K-means clustering algorithm on each granule to obtain the final results. A new evaluation method for sequence motif information, based on the function of the HSSP and the BLOSUM62 matrix, is also proposed. Compared with the existing IEEE Trans. research results, our method requires only one fifth of the execution time and shows better results in all three different quality measures.

### KEY WORDS

Fuzzy-Greedy-Kmeans model (FGK model), sequence motif, HSSP and BLOSUM 62.

## 1. Introduction

Proteins are a part of every cell in our body, and no other nutrient plays as many different roles in keeping us alive and healthy. The term “protein sequence motif” denotes amino-acid sequence pattern that is widespread and has biological significance. These motif patterns may be able to predict other proteins’ structural or functional areas, such as binding sites, conserved domains, prosthetic attachment site, etc.

There are some commonly used programs for protein sequence motif discover including MEME [11], Gibbs Sampling [12], and Block Maker [13]. Also, some of the latest algorithms include MITRA [14], ProfileBranching [15], and Gemoda[16]. Several protein sequences are required to be input by the user while using these tools. Since the size of input dataset is limited and discovered motifs are based on these input sequences, the obtained

information from above methods may carry little information about conserved sequence regions, which transcend protein families. Several popular motif databases, such as: PROSITE [8], PRINTS [9], BLOCKS [1, etc, also share the same weakness because most of the data are developed base on multiple alignments.

K-means clustering algorithm with random initial centroids is utilized by Han et al [2] to find recurring protein sequence motifs across the boundaries of a protein family. To overcome the innate problem of K-means clustering algorithm, Wei et al proposed an improved K-means clustering algorithm to obtain initial centroid locations more wisely [1]. Due to the fact that the performance of K-means clustering is very sensitive to initial point selection, the results published by Wei et al have been improved in their experiment. Because of the extremely large input dataset, both of the above papers use K-means instead of some other more advanced clustering technologies. Fast computation is always one of the advantages for K-means, other clustering methods with higher time and space costs may not be suitable for this task.

In order to overcome the high computational cost caused by a huge input dataset, we proposed a granular computing model in our most recent work called FIK model [20] which utilizes a Fuzzy C-means clustering algorithm to divide the whole data space into several smaller subsets and then applies a standard improved K-means algorithm to each subset to discover relevant information. In this paper, we develop a new greedy K-means algorithm to further improve secondary structural similarity of our discovered sequence motifs. Three evaluation methods are applied: Structural similarity, DBI measure, and a novel HSSP-BLOSUM62 evaluation method. Our sequence motif information is represented by frequency profiles.

## 2. Granular Computing Strategies

### 2.1 New Greedy K-means Clustering Algorithm

In order to overcome the potential problem of random initialization, Wei et al [1] developed a greedy

initialization method that tries to choose suitable initial points so that the final partitions can represent a more consistent and accurate result. In their experiment, the original random K-means clustering algorithm was performed five times. In each round, randomly generated initial points which have the potential to form clusters with high structural similarity are chosen for the improved K-means clustering algorithm. For each time a new potential initial center is chosen, its distance is checked against all points that are already selected in the initialization array. If the minimum distance of a new point is greater than the threshold distance, this point is included in the initialization array; otherwise, this point is discarded and another potential initial centroid is tried until the desired number of centroids is chosen.

Our method is similar to Wei’s method, but greedier. Instead of picking randomly generated initial seeds in each round of the original K-means, we collect all five K-means results and then select the initial centroids. Due to the fact that the centroids in higher quality clusters have the potential to generate better clusters in the sixth round, we divide our 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 minimum distance strategy mentioned in Wei’s approach also applies to this method. Results with different distance thresholds are given in section four. Compare with the first improved K-means algorithm, this method can gather more initial seeds. If we set minimum distance to 250 while gathering initial seeds for the sixth round, we can always obtain many more centroids than the number we need. Therefore, in this case, we only collect initial seeds until the amount is met and discard the rest. However, if the distance threshold is set to 350, sometimes the number of initial centroids acquired is not enough. In this case, we use a random method with minimum distance 800 to choose the rest of required centroids.

## 2.2 The FGK Model

Granular computing represents information in the form of aggregates, also called “information granules” [17] [18]. For a huge 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 [18].

A granular computing based model called “Fuzzy-Greedy-Kmeans model” (FGK model) is proposed in this work. This model works by building a set of information granules by FCM and then applying our new greedy K-means clustering algorithm to obtain the final information. Major advantages of the FGK model are similar to [20] which includes reduced time- and space-

complexity, filtered outliers, and higher quality granular information results. We will present comparative results with [1] in section 4 of this paper. Figure 1 shows the sketch of the model. At the first stage, all of the data segments are clustered by Fuzzy C-Means into several “functional granules” by a membership threshold cut. In each functional granule, the new greedy initialization K-means clustering is performed. At the final stage, we join the information generated by all granules and obtain the final sequence motif information.

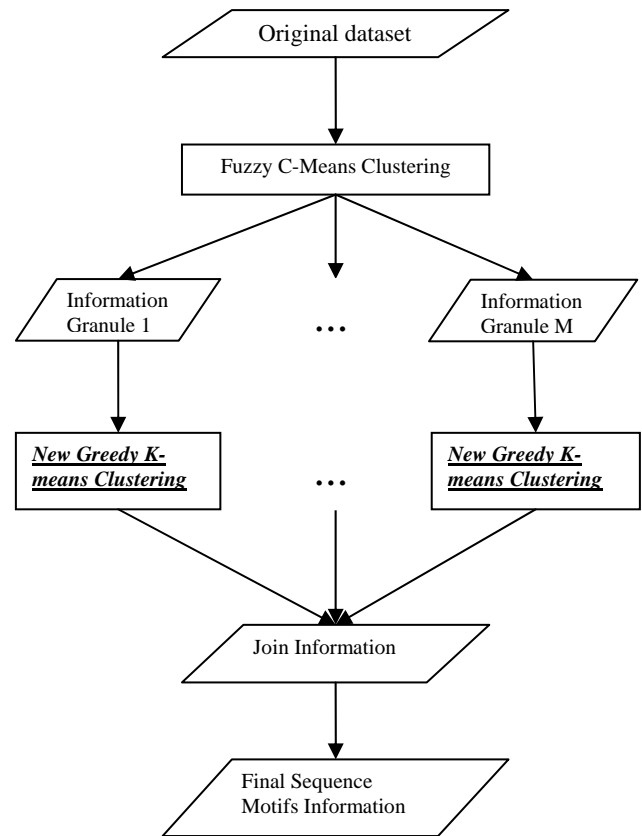


Figure 1. The FGK model

## 3. Experiment Setup

### 3.1 Dataset

2710 protein sequences obtained from Protein Sequence Culling Server (PISCES) [5] are included as the dataset of our work. No sequences in this database share more than 25% sequence identity. Sliding windows with 9 successive residues are generated from each protein sequence. Each window represents one sequence segment of nine continuous positions. More than 560,000 segments are generated by this method. 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 obtain secondary structure from DSSP [4], which is a database of

secondary structure assignments for all protein entries in the Protein Data Bank.

### 3.2 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  $9 \times 20$  matrix plus the additional nine corresponding secondary structure data obtained from DSSP. Twenty rows represent 20 amino acids and 9 columns represent each position of the sliding window. For the frequency profile (HSSP) representations of 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).

### 3.3 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]:

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

Where L is the window size and N is 20 which represent 20 different amino acids.  $F_k(i, j)$  is the value of the matrix at row i and column j used to represent the sequence segment.  $F_c(i, j)$  is the value of the matrix at row i and column j used to represent the centroid of a give sequence cluster.

### 3.4 Novel 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. Hence, the measure is defined as the following:

If  $k = 0$  or  $1$  Then HSSP-BLOSUM62 measure =  $0$

Else: HSSP-BLOSUM62 measure =

$$\frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^k \text{HSSP}_i \cdot \text{HSSP}_j \cdot \text{BLOSUM}_{62_{ij}}}{\sum_{i=1}^{k-1} \sum_{j=i+1}^k \text{HSSP}_i \cdot \text{HSSP}_j}$$

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).  $\text{HSSP}_i$  indicates the percent of amino acid i to be appeared.  $\text{BLOSUM}_{62_{ij}}$  denotes the value of BLOSUM62 on amino acid i and j. The higher HSSP-BLOSUM62 value indicates more significant motif information. To the best of our knowledge, it is the first time that HSSP and BLOSUM62 are combined and used as an evaluation method.

### 3.5 Secondary Structural Similarity Measure

A 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 is bellow 70%, the cluster can be considered weakly structurally homologous [1].

### 3.6 David-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 [20]. The DBI measure [6] is a function of the inter-cluster and intra-cluster distances. Good cluster results should reflect a relatively large inter-cluster distance and a relatively small intra-cluster distance. The DBI measure combines these two distance measurements into one function, which is defined as follows:

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

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

k is the total number of clusters,  $d_{\text{int ra}}$  and  $d_{\text{inter}}$  denote the intra- cluster and inter-cluster distances respectively.  $n_p$  is the number of members in the cluster  $C_p$ . The intra-cluster distance is 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.

### 3.7 Parameter Setup

For the Fuzzy C-means clustering, the fuzzification factor is set to 1.05 and the number of clusters is equal to 10. These settings yielded the best results in our specific dataset. For example, if we set the fuzzification factor as above and instead cluster the whole dataset into 20 groups, the membership function cannot perform as well and causing each segment to have almost the same membership to every cluster. If we further decrease the fuzzification factor, overflow may occur. In order to separate information granules from FCM results, the membership threshold is set to 12%. Using this value, we filter out around 15% of the dataset and assign the rest of the data to one or more clusters. 800 is the number we used for total number of clusters. The formula that decides how many clusters should be included in each information granule is given below:

$$C_k = \frac{n_k}{\sum_{i=1}^m n_i} \times \text{total number of clusters}$$

Where  $C_k$  denotes the number of clusters assigned to information granule k.  $n_k$  is the number of members belonging to information granule k.  $m$  is the number of clusters in FCM. Although the total data size increased from 413MB to 529MB and the total number of members increased from 562745 to 721390 by this method, we only deal with one information granule at a time. For example, the largest information granule in our work contains 136112 members, 151 clusters should be computed from those members and the data size is 99.9MB. Comparing with the original dataset, the biggest granule is only 25% in size. Therefore, the computation time for all information granules (231720 seconds) is merely around 20% of [1] (1285928 seconds). Detail execution time information and numerical data about number of members, number of clusters, and data size of each information granule can be found in [20]. Based on those results, reduced space- and time- complexities of our work are achieved.

## 4. Experiment Results

### 4.1 Sequence Motifs Quality Comparison

In table 1, the DBI measure, the novel HSSP-BLOSUM62 measure and average percentage of sequence segments belonging to clusters with high structural similarity for different methods is given. The first column shows the different methods with different parameters. “Traditional” refers to the original K-means algorithm applied to the whole dataset. “FCM-K-means” indicates the original K-means clustering method applied to information granules generated by FCM. “FGK model 250” indicates that the dataset is clustered by the FGK model with the new greedy initialization K-means clustering algorithm, and the distance threshold is set as

250. “FGK model 300”, “FGK model 350”, and “FGK model 400” are defined similarly. The second column of Table 1 gives the average percentage of sequence segments belonging to the cluster with structural similarity greater than 60%. Similarly, the third column contains the average percentage of the clusters with the structural similarity higher than 70%. The third column denotes the average DBI measure (The lower DBI value indicates the higher quality of the cluster result) and the last column indicates the average value of HSSP-BLOSUM62 evaluation. Figures 2 to 4 are interpreted from table 1.

Different Methods	> 60%	> 70%	DBI Measure	HSSP-BLOSUM62
Traditional	25.82%	10.44%	6.098	0.254299
FCM-K-means	37.14%	12.99%	4.359	0.358886
FGK model 250	42.93%	14.39%	4.627	0.346093
FGK model 300	41.80%	13.89%	4.619	0.325683
FGK model 350	37.67%	13.64%	4.577	0.357170
FGK model 400	36.17%	13.01%	3.981	0.375084

Table 1 Comparison of all measures

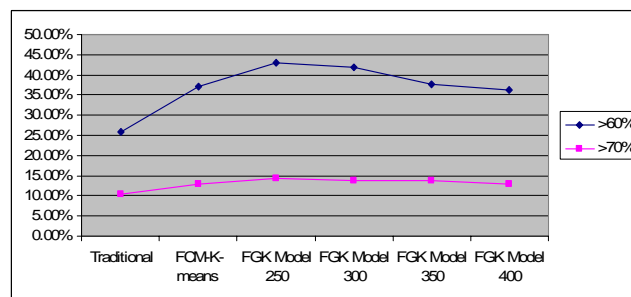


Figure 2 Comparison of percentage of sequence segments belonging to cluster with high structure similarity.

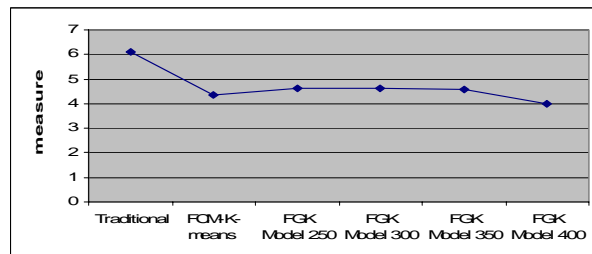


Figure 3 Comparison of the DBI measure.

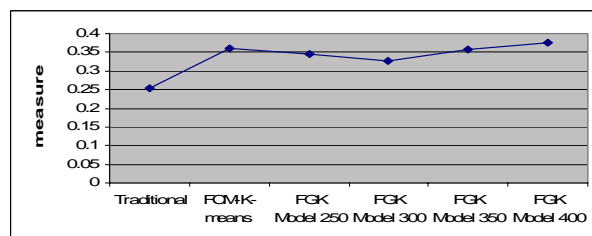


Figure 4 Comparison of the HSSP-BLOSUM62 measure

The results of Table 1 and all figure from 2 to 4 reveal that the quality of clusters improved dramatically by applying granular computing techniques which utilizes

FCM to separate the whole dataset into several information granules. In the FCM-K-means approach, the average percentage of clusters with structural similarity greater than 60% increased more 11%, which translates to more than 90 meaningful sequence motifs that cannot be disclosed by traditional methods but are discovered by our approach. The DBI measure also successfully decreased from 6.098 to 4.359, implying that our model not only generates more biologically meaningful results but that these results are supported by statistical/computer-science techniques. Also, the HSSP-BLOSUM62 measurement increasing from 0.254 to 0.358 proves that the motif information is more consistent and meaningful under the granular computing strategy.

For FGK model 250, although the measurement of DBI and HSSP-BLOSUM62 decreased slightly, the result achieves the highest percentage (42.93%) of clusters with high structure similarity among all methods. It indicates that greedy initialization method can reveal some hidden motif information that the traditional one can not. Since we have a larger dataset and different window sizes from Wei et al [1], we cannot directly compare the results. However, more than 17% of high structural similarity is increased by our model, while their best work increased 4.51%; our achievement is stronger.

## 4.2 Sequence Motifs

Tables 2 to 7 illustrate six different sequence motif examples are generated by our method. Due to space limitation, we only present part of our recurring pattern information in this paper. However, all of the clusters with 60% secondary structural similarity obtained by different parameters are available under <http://www.cs.gsu.edu/~cscbecx/Bioinformatics%20Information.htm>. The following format is used for representation of each motif table. The first row represents the number of members belonging to this motif and the secondary structural similarity. The first column stands for the position of amino acid profiles in each motif with window size nine. The second column expresses the type of amino acid frequently appearing in the given position. If the amino acids appear with a frequency higher than 10%, they are indicated by upper case; if the amino acids appear with the frequency between 8% and 10%, they are indicated by lower case. The third column corresponds to the hydrophobicity value, which is the summation of the frequencies of occurrence of Leu, Pro, Met, Trp, Ala, Val, Phe, and Ile. The fourth column indicates the value of the HSSP-BLOSUM62 measure. The last column shows the representative secondary structure of the position.

## 5. Conclusion

A new greedy initialization K-means clustering algorithm and Fuzzy C-means clustering algorithm have been combined as a new granular computing model and

proposed in this paper. In this model, we utilize fuzzy clustering to split the whole dataset into several information granules and analyze each granule by the K-means clustering algorithm with a greedy method for choosing centroids. Analysis of sequence motifs also shows that the granular computing technology may detect some subtle sequence information overlooked by the K-means clustering algorithm alone. Besides, a biochemical measurement called HSSP-BLOSUM62 for discovered motifs is proposed. Since our FGK model is capable of decreasing time and space complexity, filtering outliers, and capturing better results, we believe some other bioinformatics research with large database may also adapt this granular computing strategy to perform satisfied results.

Number of segments: 1369 Structure homology: 80.87% Avg. HSSP-BLOSUM62: 0.114				
#	Noticeable Amino Acid	H	B	S
1	A k E D	.25	-.13	H
2	V L I A	.70	.52	H
3	V L I a	.67	1.0	H
4	A r K E d	.23	-.16	H
5	L A r k e	.44	-.93	H
6	A	.83	.00	H
7	A R K E	.25	.21	H
8	A R K E	.24	.00	H
9	v L i A	.56	.48	H

Table 2 Helices motif with conserved A

Number of segments: 565 Structure homology: 60.26% Avg. HSSP-BLOSUM62: 0.781				
#	Noticeabl Amino Acid	H	B	S
1	v L I	.86	1.7	H
2	V L I	.83	1.7	H
3	a r K E	.28	-.01	H
4	a r K E	.23	.11	H
5	V L I	.60	1.9	H
6	v L I	.54	1.8	H
7	A s k E d	.27	-.10	C
8	a E D	.27	-.03	C
9	a e d	.34	-.25	C

Table 3 Helices-Coil motif

Number of segments: 1163 Structure homology: 62.71% Avg. HSSP-BLOSUM62: 0.209				
#	Noticeable Amino Acid	H	B	S
1	G A S t	.34	.04	C
2	A S t	.39	.68	C
3	g A S t	.40	.10	C
4	G A S t	.32	.05	C
5	g A S	.37	.38	C
6	G A S t	.32	.00	C
7	G A S t	.33	-.06	C
8	g A S t	.37	.06	C
9	A S t	.40	.64	C

Table 4 Hydrophilic Coli motif with conserved G, A, S and T

Number of segments: 1008 Structure homology: 69.57% Avg. HSSP-BLOSUM62: 0.342				
#	Noticeabl Amino Acid	H	B	S
1	v l A t	.48	-.20	H
2	v L A	.51	-.15	H
3	v l A	.51	-.18	H
4	A s	.76	1.0	H
5	V L I a	.76	.93	H
6	V L i A	.59	.34	H
7	l A s	.47	-.42	H
8	A s	.54	1.0	H
9	V L I a	.65	.77	H

Table5 Hydrophobic Helices motif

Number of segments: 695 Structure homology: 70.49% Avg. HSSP-BLOSUM62: 0.973				
#	Noticeable Amino Acid	H	B	S
1	R K	.29	2.0	E
2	V L I	.83	2.2	E
3	V L I	.74	2.0	E
4	V L I f	.76	1.1	E
5	V L I	.78	1.9	E
6	G a s t	.29	-.29	E
7	a s t	.40	.67	C
8	G a p S d	.32	-.64	C
9	G a S d	.26	-.28	C

Table 6 Sheet-Coil motif

Number of segments: 1328 Structure homology: 72.16% Avg. HSSP-BLOSUM62: 0.307				
#	Noticeabl Amino Acid	H	B	S
1	L A	.78	-1.0	H
2	A R K E	.22	.22	H
3	A r K E	.25	-.05	H
4	L A	.54	-1.0	H
5	G	.06	0	H
6	V L I A	.64	.71	H
7	P K E D	.27	-.19	H
8	V L I	.69	1.9	H
9	V L I	.65	2.1	H

Table 7 Helices-coil-sheet motif

## Acknowledgements

The authors would like to thank Dr. Wei Zhong for sharing information and helping with this work. 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] Karen F. Han and David 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. Mabeay, 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," *Bioinformatics*, vol. 15, no. 6, pp. 417-479, 1999.
- [11] Bailey,T.L. and Elkan,C. Fitting a mixture model by expectation Maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol., 2, 28-36. 1994.
- [12] Lawrence,C.E. et al. "Detecting subtle sequence signals: a Gibbs Sampling strategy for multiple alignment." *Science*, 262, 208-214. 1993
- [13] Henikoff,S. et al. "Automated construction and graphical presentation of Protein blocks from unaligned sequences." *Gene*, 163, GC17-GC26, 1995
- [14] Eskin,E. and Pevzner,P.A. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18 ((Suppl. 1)), 354-363, 2002
- [15] Price,A et al. "Finding subtle motifs by branching from sample strings." *Bioinformatics*, 19 (Suppl. 2), II149-II155, 2003
- [16] Kyle L. Jensen et al, "A Generic motif discovery algorithm for sequential data", *Bioinformatics*, vol 22, no.1, pp. 21-28, 2006.
- [17] 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.
- [18] Y.Y. Yao, "On Modeling data mining with granular computing," *Proceedings of COMPSAC 2001*, pp.638-643, 2001.
- [19] Henikoff, S. and Henikoff, J. G. (1992). "Amino acid substitution matrices from protein blocks." *Proc. Natl. Acad. Sci. USA* 89: 10915-10919.
- [20] 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", *Proceedings of BIBE 2006*, full paper accepted.