International Journal for Computational Biology (IJCB)
Vol.1, No.1, 2012, Page. 14-27
ISSN: 2278-8115

Protein Local Tertiary Structure Prediction by Super Granule Support Vector Machines with Chou-Fasman Parameter

Bernard Chen[1], MinwooKim[1],Matthew Johnson[1], Wooyoung Kim[2] and Yi Pan[2]*

[1]Department of Computer Science, University of Central Arkansas, Conway, AR 72035,
[2]Computer Science Department, Georgia State University, Atlanta, GA 30303

Article Info

Article history:

Received Dec 10th, 2011
Revised Jan 20th, 2012
Accepted Jan 30th, 2012

Keyword:

Protein Structure Prediction
Local Tertiary Structure
Sequence Motif
Chou-Fasman Parameter
Super Granule SVM

ABSTRACT

Prediction of a protein's tertiary structure from its sequence information alone is considered a major task in modern computational biology. In order to closer the gap between protein sequences to its tertiary structures, we discuss the correlation between protein sequence and local tertiary structure information in this paper. The strategy we used in this work is predict small portions (local) of protein tertiary structure with high confidence from conserved protein sequences, which are called "protein sequence motifs". 799 protein sequence motifs that transcend protein family boundaries were obtained from our previous work. The prediction accuracy generated from the best group of protein sequence motifs always keep higher than 90% while more than 8% of the independent testing data segments are predicted. Since the most meaningful result published in latest publication is merely 70.02% accuracy under the coverage of 4.45%, the research results achieved in this paper are obviously outperformed. Besides, we also set up a stricter evaluation to our prediction to further understand the relation between protein sequence motifs and tertiary structure predictions. The results suggest that the hidden sequence-to-structure relationship can be uncovered using the Super Granule SVM Model with the Chou-Fasman Parameter. With the high local tertiary structure prediction accuracy provided in this article, the hidden relation between protein primary sequences and their 3D structure are uncovered considerably.

Corresponding Author:

Yi Pan
Computer Science Department,
Georgia State University, Atlanta,
GA 30303.
Email: pan@cs.gsu.edu

How to Cite:

Bernard Chen *et. al.* Protein Local Tertiary Structure Prediction by Super Granule Support Vector Machines with Chou-Fasman Parameter. IJCB. 2012; Volume 1 (Issue 1): 14-27.

1. INTRODUCTION

Proteins are used by organisms for virtually every life function. Understanding the relationship between the amino acid sequence and the resulting protein structure is one of the most important research topics: First of all, based on many biochemical experiments, it is believed that a sequence is the sole determinate in a polypeptide's structural conformation. Most proteins have just one shape for their lifetime, but a handful -- in particular: proteins associated with viruses such as HIV, the influenza virus, and with alpha-synuclein (protein involved in Parkinson disease) -- have two dramatically different shapes; one before the disease attacks and one after.Second, the function of a protein is directly dependent on its three-dimensional structure. Last but not least, structural-based drug design in the medical field relies heavily on protein tertiary structural information which is usuallyobtained from expensive X-ray crystallography or NMR spectroscopy.

Sequence motifs are referred to as the conserved sequence patterns either functionally or structurally similar in a group of related proteins. The role of motifs is in predicting functional or structural portion of other proteins including prosthetic attachment sites, enzyme-binding sites and DNA /RNA binding sites, and so on.

Even though the discovery of new motifs requires tremendous time and effort, the modification of known motifs and the generalization of new motifs are major issues in academia. Protein sequence motifs are usually categories into families. The signatures can be derived as complex descriptors, or simple consensus patterns, such as blocks or profiles [1]. Some popular motifs databases include PROSITE [2], BLOACK [3], PRINTS [4], SBASE [5], and PFAM [6]. In terms of techniques, protein sequence motif discovery tools such as MEME [7], Gibbs Sampling [8], Block Maker [9], MITRA [10] and Profile Branching [11] are extensively adopted by the bioinformatics communities. These applications, however, suffer a common issue of limiting the size of input dataset. Consequently, little information that crosses family boundaries can be discovered by these databases and tools. In order to find out protein sequence motifs information that crosses family boundaries, the input dataset need to be big enough to cover all representative sequences for all known protein sequences. As a result, efficient techniques are demanded. Clustering is one of the most popular data mining techniques and has been studied extensively for protein sequence motifs discovering [13, 14, 16, 20-27]. Han et al produced high quality protein clusters from protein sequence frequency profiles [13, 21] using the K-means clustering algorithm. These recurring patterns were regarded as vocabularies to understand the whole sentence encoded in protein structure. Subsequently, they used the sequence clusters combined with Hidden Markov Model (HMM) [28] to predict local protein structures. However, these conventional clustering algorithms assumed that the distance between data points could be calculated. While the distance function was not well characterized, this approach might not reveal the true sequence-to-structure relationship [30].

Support Vector Machines (SVMs) [31] have established their importance in various research fields. SVMs implement the soft margin concept to bear mislabeled examples for the purpose of maximizing the margin. Therefore, SVMs are capable of handling non-linear classification by implicitly mapping input samples into a higher dimension for maximum-margin hyperplane generation. Under this point of view, the SVM can be more efficient to discover the non-linear sequence-to-structure relationship than the K-means clustering algorithm [30]. However, applying the SVM to this problem is not feasible because of the high computational cost of the SVM algorithm [17]. It is almost infeasible to model a SVM over half a million data segments, which is the necessary requirement for generating protein sequence motifs that cross protein family boundaries. However, combining the SVM and the granule computing allows for uncovering the unknown behind the sequence-to-structure relationship.

Recently, Zhong et al [30] proposed the Clustering SVM for protein local tertiary structure prediction. With an aim to evaluate recurring pattern quality, 3D information including RMSD and Torsion Angle are integrated in the motifs evaluation process. Our research goal is to reveal the correlation between protein primary sequences and the structures; As a result, none of 3D information is included during the generation of protein sequence motifs. In this paper, we explain how to combine granule computing, the SVM and, the Chou-Fasman parameter to achieve our research goal. A detailed report on local protein structure prediction based on sequence information is also provided.

2. SUPER GRANULE COMPUTING MODELS

Super Granule Support Vector Machine (Super GSVM) with Chou-Fasman parameter is a new model specifically designed for protein local tertiary structure prediction. It is founded on the FGK model [16] and the Super GSVM-FE model [23]. In this section, we explain the FGK model and the Super GSVM-FE model, and then propose the Super Granule Support Vector Machine (Super GSVM) with Chou-Fasman parameter model.

2.1 The FGK Model for Protein Sequence Motifs Discovery

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 our previous work [16]. This model works by using FCM to building a set of information granules and then applying our new greedy K-means clustering algorithm to obtain the final information. The basic idea of FGK model is showed in Figure 1. The 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 selection initial centroids 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%. Major advantages of the FGK model are reduced time- and space- complexity, filtered outliers, and higher quality granular information results.

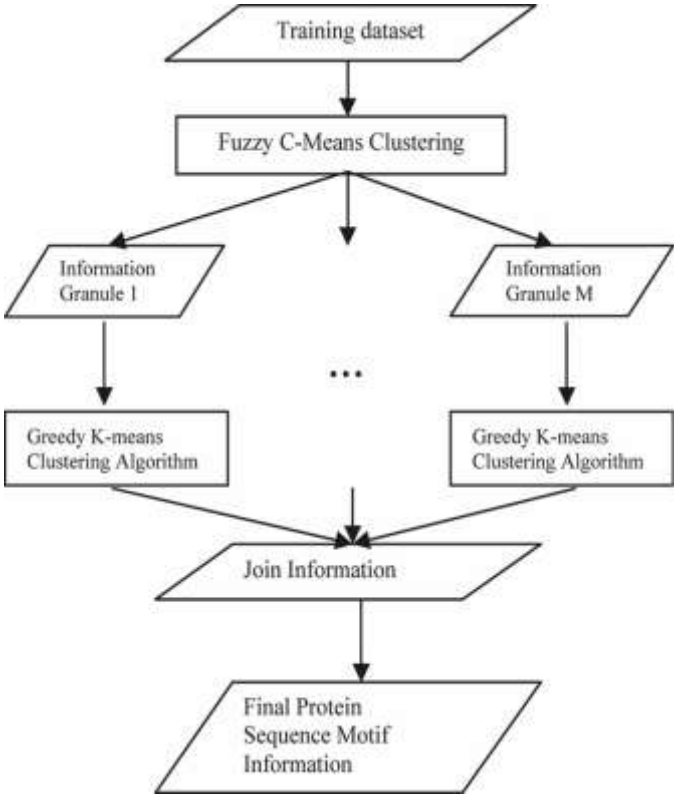

Fig. 1 The sketch of the Fuzzy Greedy K-means (FGK) Model

Table 1. Chou-Fasman Parameters [32]. The first column indicates the name of 20 amino acids. The next three columns represent the propensities of each amino acid for helices (P(a)), sheets (P(b)) or turns (P(t)).

Symbol and name of Amino Acid	P(a)	P(b)	P(t)
A : Alanine	142	83	66
R : Arginine	8	93	95
D : Aspartic Acid	101	54	146
N : Asparagine	67	89	156
C : Cysteine	70	119	119
E : Glutamic	151	37	74
Q : Glutamine	111	110	98
G : Glycine	57	75	156
H : Histidine	100	87	95
I : Isoleucine	108	160	47
L : Leucine	121	130	59
K : Lysine	114	74	101
M : Methionine	145	105	60
F : Phenylalanine	113	138	60
P : Proline	57	55	152
S : Serine	77	75	143
T : Threonine	83	119	96
W : Tryptophan	108	137	96
Y : Tyrosine	69	147	114
V : Valine	106	170	50

2.2 The Super GSVM-FE Model for Protein sequence Motifs Extraction

Basically, this new model is the next generation of the FGK model. It also uses the 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 ranking SVM (2002 et al, 2002) to rank all members in each cluster generated by the greedy K-means clustering algorithm, and then we filter out lower ranked members. The number of segments to eliminate is decided by a user defined filtrate percentage. The results of different percentage are discussed in [23] and "20%" provides the best tradeoff value. After the feature elimination step, we collect all surviving data points in

each information granule and then run greedy K-means with same initial centroids we previously generated. Finally, we collect all the results in all granules to create the final protein sequence motif information. Figure 2 demonstrates the Super GSVM-FE model.

2.3 The Super Granule Support Vector Machine (Super GSVM) with Chou-Fasman Parameter Model for Protein Location Tertiary Structure Prediction

The sketch of the proposed model has been shown in Figure 3 and 4. The whole model can be divided into two parts: 1. Generate and Extract protein sequence motifs generated mainly from primary sequence information (Figure 3); 2. Predict protein local tertiary structures through the obtained motifs. (Figure 4).

In order to discover protein sequence motif information which is universally conserved across protein family boundaries, our original input dataset is extremely large. Therefore, an efficient granule computing technique is applied: Fuzzy C-means clustering algorithm is utilized as the first step to softly divide the huge training dataset into 10 smaller information granules. For each information granule, we then carry out the Greedy K-means clustering algorithm [16], which performs the traditional K-means clustering five iterations and then brings together the good clusters' centroids as the starting centroids for the sixth round. 343 among 799 clusters are considered meaningful recurring patterns (for more information including parameter setup and detail results, lease reference [16]). After the quality evaluation, the Chou-Fasman parameter is calculated and appended to all data segments. Since the size of the clusters (the average size of the clusters is 905.75 members) is much smaller than the original training dataset (more than half million data segments), we are able to train the Ranking-SVM based on secondary structure for each cluster. Based on the trained Ranking-SVM models, we generate the rank of all members within the cluster. The research results in [23] have shown that eliminating 20% of the lower ranked members for each cluster generates the optimal protein sequence motifs information in the biological and biochemical perspective. Thus, we purge 20% of the lower ranked members from each cluster resulting in 536 out of 799 meaningful recurring patterns. To conclude the first part of the model, we collect all extracted recurring patterns for the next part of the model: local tertiary structure prediction. It is important to note that during the first part of the model, none of the 3D information is involved. After the sequence motifs are formed, for each cluster, we use all members' 3D structure to calculate the represented 3D structure of the cluster. 3D information is only appended after the cluster is generated and extracted. Our objective is to anticipate the similar 3D structure of discovered protein sequence recurring patterns and independent testing dataset on the basis of similarity shared in primary sequence.

The second part of the Super GSVM with Chou-Fasman parameter model is straightforward: for each independent testing sequence segment, we first append its Chou-Fasman value and then calculate the total distance (including the difference of primary sequence and the Chou-Fasman value) by formula (2) with all sequence clusters. Due to the fact that the protein sequence motifs we discovered are transcend protein family boundaries, we can directly search for a match without pre-processing the testing dataset into protein categories or families. If we find a closest cluster within a given distance threshold, we can say that the testing segment is close enough to our discovered sequence motif and it should have a similar tertiary structure to the representative 3D structure of the discovered sequence motif. Needless to say, how to setup this threshold is a research problem. The stricter threshold we set, the higher prediction accuracy should be achieved. However, the stricter threshold we set, the fewer testing segments can be predicted. Detail results related to the threshold, the prediction coverage and the prediction accuracy are showed in section 4. Due to the fact that sequence motifs, by definition, only occur in a limited number of positions within a proteins sequence, we emphasize "local" tertiary structure prediction [29] instead of complete tertiary structure prediction. Detailed experimental results are provided in the results section.

3. EXPERIMENTAL SETUP

3.1 Training dataset

Since the major purpose of this work is to obtain protein sequence motif information across protein family boundaries, the dataset of our work is supposed to represent all known protein sequences. However, without a systematic approach, it is very difficult to extract useful knowledge from an extremely large volume of data. The basic principle we use is to choose representative protein files from the whole PDB database, and then use the profile in HSSP to expand each file.

The dataset used in this work includes 2710 PDB protein sequences obtained from Protein Sequence Culling Server (PISCES) [19]. Among these 2710 protein sequences, no sequence in this database shares more than 25% sequence identity. **HSSP** is a derived database merging structural (3-D) and sequence (1-D) information. For each protein of known 3-D structure from the Protein Data Bank (PDB), the database has a multiple sequence alignment of all available homologues and a sequence profile characteristic of the family [35]. In the end of

each HSSP file, it calculates the occurrence percentage of every amino acid on each sequence position. An example of the 1b25 HSSP file is given in Figure 5.

The sliding window technique with nine successive residues is generated from protein sequences. Each window represents one sequence segment of nine contiguous positions. More than 560,000 segments are generated by this method. Figure 6 shows how we apply the sliding window technique on the 1b25 HSSP file. Each window corresponds to a sequence segment, which is represented by a 9 × 20 matrix. 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.

3.2 Independent testing dataset

The latest release of PISCES includes 4345 PDB files. Compared with the dataset in our experiment (obtained in 2005), 2419 PDB files are excluded. In this experiment, we use the protein sequence motifs information generated from our old dataset (2710 protein sequence files) to predict the tertiary structure of these 2419 protein files. Therefore, we regard our 2710 protein files as the training dataset and 2419 protein files as the independent testing dataset, which generates around 486,234 segments by the sliding window approach.

3.3 The source of secondary and tertiary structure information

We also obtained secondary structure from DSSP [34], which is a database of secondary structure assignments for all protein entries in the Protein Data Bank, for each sequence segment. The main uses of secondary structure information are to evaluate sequence clusters and train the ranking SVM. Originally, DSSP allocates the secondary structure to eight different classes. However, in this study, those eight classes are reclassified into three categories according to the following conversion model: assigning H, G, and I to H (Helices), assigning B and E to E (Sheets), and assigning all others to C (Coils). The tertiary structure of protein sequence segments in the training set and testing set are available from Protein Data Bank (PDB).

In the Super GSVM with Chou-Fasman parameter model, Chou-Fasman parameter is encoded right after the protein recurring patterns (clusters) are generated and the testing data are read-in. The encoded value is computed as follows. For each location within one window size, we calculate the propensity value for helices, sheets and turns. Since the window size we select in the paper is 9 and 3 and different secondary structures are considered, an additional 9 × 3 information segment is added after the encoding procedure of Chou-Fasman parameter. As we previously mentioned, for each location within a window size, HSSP provides the probability of each amino acid to be appeared. Since the Chou-Fasman parameter (Table 1) provides the relative value for secondary structure determination, if we sum up the twenty cross value of the probability of each amino acid and its corresponding helices value in Chou-Fasman parameter, we can determine the total helices value. Sheets and turns (or coil) share the same trends. For example, if a sequence with A (10%), R (2%), D (20%)…, and the total helices value equals to 10%*142 + 2%*8 + 20%*101 + … and so on. Sheets and turns (or coil) share the same trends.

3.5 Distance Measure

Since the Manhattan distance is featured by every position of the frequency profile equally, this distance measure is the most suitable measurement for this research [13]. The following formulation is adopted to obtain the distance between two sequence segments [13].

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

Where L is the window size and N is 20 representing 20 different amino acids. $Fk(i,j)$ is the value of the matrix at row i and column j andrepresents the sequence segment. $Fc(i,j)$ is the value of the matrix at row i and column j andrepresents the centroids of a give sequence cluster. The lower the dissimilarity value, the higher similarity the two segments have.

3.6 Distance Measure together with Chou-Fasman Parameter

City block distance measure is still valid after the Chou-Fasman parameter is encoded in each sequence segment. The following formula is used to calculate the similarity of two sequence segments:

$$\text{Total\_Dissimilarity} = w_1\sum_{i=1}^{L}\sum_{j=1}^{N}\left|F_k(i,j) - F_c(i,j)\right| + w_2\sum_{i=1}^{L}\sum_{k=1}^{M}\left|F_k(i,k) - F_c(i,k)\right|$$

Where w_1 and w_2 indicate the weight of the sequence dissimilarity and Chou-Fasman value. In this paper, both weights are equal to 1. L is the window size and M is 3 for the 3 different secondary structures (H, E and C) score values. The lower total dissimilarity value, the higher similarity the two segments have.

3.7 Secondary Structural Similarity Measure

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

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

Where ws is the window size and Pi,H shows the frequency of occurrence of helix among the segments for the cluster in position i. Pi,E and Pi,C are defined in a similar way. If the structural homology for a cluster exceeds 70%, the cluster can be considered structurally identical [15]. If the structural homology for the cluster is between 60% and 70%, the cluster can be viewed weakly structurally homologous [20].

3.8 Tertiary Structure Distance (dmRMSD)

In this research, we use "Distance Matrix", which is the mutual distance among Cα carbans, to represent the real 3D structure as well as predicted 3D structure. The distance matrix to represent the structural segment stores the distance from the first atom's Cα carban to the second atom's Cα carban, the distance from the first atom's Cα carban to the third atom's Cα carban, …, the distance from the first atom's Cα carban to the ninth atom's Cα carban and then the distance from the second atom's Cα carban to the third atom's Cα carban… and so on. In our example, since the window size equals to nine, the distance matrix stores 36 distances in total.

In order to describe the representative 3D structure of a cluster, we introduce Average Distance Matrix (ADM), which records the average for the distance matrices of all the sequence segments in one cluster, using the following formula:

$$\alpha_{i\to j}^{ADM} = \frac{\sum_{k=1}^{N} \alpha_{i\to j}^{k}}{N}$$

Where $\alpha_{i\to j}^{k}$ is referred to the distance between α-carbon atom i and α-carbon atom j in the sequence segment k of the length L. N is the total number of sequences in the cluster.

To calculate the structure distance between the real one and the predicted one, we use dmRMSD [36, 37] described as follows:

$$\text{dmRMSD} = \sqrt{\frac{\sum_{i=1}^{L}\sum_{j=i+1}^{L}\left(\alpha_{i\to j}^{s1} - \alpha_{i\to j}^{ADM}\right)^2}{M}}$$

$$M = \frac{L \times L - L}{2}$$

Where $\alpha_{i\to j}^{ADM}$ is used to represent the predicted sequence cluster's 3D structure and $\alpha_{i\to j}^{s1}$ is the structure information to be predicted. M is the number of distances in the distance matrix. Since the window size we assumed is nine (L=9), M = 36.

4. RESULTS AND ANALYSIS

4.1 Quality of Protein Sequence Motifs Information

Due to the fact that our main research idea is based on using protein sequence patterns generated from only sequence (1D) information to predict the protein tertiary (3D) structure, the quality of protein sequence recurring patterns dominates the success level of our experiment. As the result, improving the quality of our protein sequence pattern (motifs) information is our first priority. Intra-cluster secondary structure similarity within the protein sequence clusters is the major evaluation criteria. According to [15, 20], if the structural homology for a cluster exceeds 70%, the cluster can be considered structurally identical; If the structural homology for the cluster exceeds 60% and lower than 70%, the cluster can be viewed weakly structurally

homologous. Therefore, we separate the cluster quality into four classes based on the secondary structure similarity of clusters: Bad (<60%), Fair (60%~70%), Good (70%~80%) and Excellent (>80%).

In our previous work, we have successfully obtained 343 out of 799 clusters (detailed cluster quality information is available in Table 2). Next, we further extract all clusters by training a Ranking-SVM for each cluster and discard the lower 20% ranked data segments. Extra 200 high quality (secondary structural similarity > 70%) protein sequence patterns are produced [29] (detailed cluster quality information is also available in Table 2). Since our focus is on finding protein sequence motifs that crosses family boundaries, we are able to use our protein sequence patterns to predict protein local tertiary structures on all unknown protein sequences without being limited to a specific protein family.

Table 2. The comparison of number of clusters belongs to different group

Secondary Structure Quality	< 60% (Bad)	60%~70% (Fair)	70%~80% (Good)	> 80% (Excellent)
FGK250 [16]	456	231	88	24
Super GSVM [5]	256	287	156	100
Super GSVM with Chou-Fasman	274	267	160	99

The key difference between this experiment and the latest study [29] is the inclusion of the Chou-Fasman parameter [32, 33] on each data segment before the clusters are trained by the ranking SVM. The same training data set and independent testing dataset were used in this experiment as were used in our previous work [29]. Table2 demonstrates the number of sequence clusters belonging to different quality categories generated by different approaches. The first row of Table2 indicates the secondary structure quality category. According to [15, 20], if the structural homology for a cluster exceeds 70%, the cluster can be considered structurally identical; If the structural homology for the cluster exceeds 60% and lower than 70%, the cluster can be viewed weakly structurally homologous. Therefore, we separate the clusters into four classes based on the secondary structure similarity of clusters: Bad (<60%), Fair (60%~70%), Good (70%~80%) and Excellent (>80%).

As shown from Table2, both Super GSVM models outperform the original FGK model. Comparing the two Super GSVM models, we find that the one with Chou-Fasman does produce one more high quality clusters (secondary structural similarity > 70%). This suggests that the addition of the Chou-Fasman parameter enabled the ranking SVM to rank the belongingness of each cluster member to its particular cluster more intelligently, resulting in higher quality clusters. Although the total number of clusters greater than 60% is reduced and the difference of number of high quality clusters is not huge, the prediction accuracy for protein local tertiary structure is increased dramatically as reported in the next section.

4.2 Prediction Accuracy Comparison

In Figure 7, we provide a visual description to explain the use of protein local tertiary structure prediction.The colorful portion of figure is depicted from the local tertiary structure prediction. Researchers can use the predicted portion as anchors to expend consecutive fractions and form global tertiary structure prediction. Undoubtedly, the prediction accuracy and the prediction coverage (how many colorful portions are formed) play the key role to the success of global prediction.

Table 3. Prediction Accuracy with 1.5 Å criteria and coverage on three clustering quality groups under different distance threshold

	Excellent Group			Good Group			Fair Group		
Distance Threshold	Prediction accuracy (%)	#segment Predicted	Coverage(%)	Prediction accuracy (%)	#segment Predicted	Coverage(%)	Prediction accuracy (%)	#segment Predicted	Coverage(%)
600	100.00%	20	≈0%	57.14%	14	≈0%	12.90%	31	0.01%
700	99.21%	254	0.05%	63.87%	155	0.03%	28.79%	323	0.07%
800	97.37%	1254	0.26%	73.59%	765	0.16%	32.69%	1355	0.28%
900	96.05%	3619	0.74%	73.24%	2425	0.50%	36.80%	3864	0.79%
1000	95.53%	7781	1.60%	74.02%	5665	1.17%	39.32%	8289	1.70%
1100	94.87%	13893	2.86%	75.06%	11097	2.28%	42.10%	15285	3.14%

1200	93.79%	20985	4.32%	73.87%	18634	3.83%	43.24%	25247	5.19%
1300	92.67%	28576	5.88%	72.38%	28480	5.86%	43.50%	38310	7.88%
1400	91.34%	35990	7.40%	70.99%	39747	8.17%	43.14%	54534	11.22%
1500	89.92%	42948	8.83%	69.43%	51802	10.65%	42.34%	72350	14.88%
1600	88.41%	49014	10.08%	67.97%	63729	13.11%	41.42%	91126	18.74%
1700	87.00%	53945	11.09%	66.57%	74339	15.29%	40.45%	109133	22.44%
1800	85.82%	57767	11.88%	65.34%	83047	17.08%	39.66%	124386	25.58%
1900	84.88%	60307	12.40%	64.44%	89033	18.31%	38.98%	135447	27.86%
2000	84.29%	61763	12.70%	63.88%	92702	19.07%	38.54%	142728	29.35%

Table 4. Prediction accuracy with 1.0 Å criteria on three cluster groups under different distance threshold

	Excellent Group			Good Group			Fair Group		
Distance Threshold	Prediction accuracy (%)	#segment Predicted	Coverage(%)	Prediction accuracy (%)	#segment Predicted	Coverage(%)	Prediction accuracy (%)	#segment Predicted	Coverage(%)
600	95.00%	20	≈0%	42.86%	14	≈0%	3.23%	31	0.01%
700	96.46%	254	0.05%	45.81%	155	0.03%	6.50%	323	0.07%
800	93.14%	1254	0.26%	53.99%	765	0.16%	9.08%	1355	0.28%
900	90.80%	3619	0.74%	54.19%	2425	0.50%	10.61%	3864	0.79%
1000	89.37%	7781	1.60%	54.49%	5665	1.17%	11.01%	8289	1.70%
1100	88.36%	13893	2.86%	54.10%	11097	2.28%	12.16%	15285	3.14%
1200	86.71%	20985	4.32%	52.31%	18634	3.83%	12.65%	25247	5.19%
1300	85.08%	28576	5.88%	50.58%	28480	5.86%	12.81%	38310	7.88%
1400	83.48%	35990	7.40%	49.07%	39747	8.17%	12.70%	54534	11.22%
1500	81.86%	42948	8.83%	47.71%	51802	10.65%	12.36%	72350	14.88%
1600	80.22%	49014	10.08%	46.49%	63729	13.11%	11.96%	91126	18.74%
1700	78.69%	53945	11.09%	45.41%	74339	15.29%	11.48%	109133	22.44%
1800	77.46%	57767	11.88%	44.55%	83047	17.08%	11.13%	124386	25.58%
1900	76.44%	60307	12.40%	43.94%	89033	18.31%	10.89%	135447	27.86%
2000	75.85%	61763	12.70%	43.55%	92702	19.07%	10.68%	142728	29.35%

In this subsection, we indicate a successful prediction of local 3D structure if the average dmRMSD is less than 1.5 Å. A complete report on the prediction accuracy gen-erated from different cluster groups and the number of predicted segments is provided in Table3. The first column shows different distance thresholds (corresponding to step(4) in Figure 4) based on the distance calculation of "Distance measure together with Chou-Fasman parameter" described in 3.6. The second, fifth, and eighth column give the prediction accuracy based on given distance threshold under different cluster groups. The third, sixth, and ninth column illustrate the number of predicted sequence segments under different cluster groups. The prediction coverage (the fourth, seventh, and tenth column) is derived from the number of predicted segments divided by total number of testing sequence segments, which equals to 486,234.

A full comparison of the prediction accuracy between the Super GSVM model [29] and our newly proposed Super GSVM with Chou-Fasman parameter model is presented in Figure 8 and 9. Excellent (in Figure 8) and Good (in Figure 9) are the prediction results generated from four different groups in this paper. P-Excellent (in Figure 8) and P-Good (in Figure 9) are the prediction results re-ported in [29]. As we mentioned in section 2.3, different distance thresholds generate different prediction accuracy and prediction coverage. Since the distance is calculated differently in this research and in [29], it is not useful to directly compare the accuracy-vs.-distance threshold relationship.However, coverage is consistent between both experiments; as a result, we use coverage as X-axis in Figure 8 and 9 to show the direct comparison.

The new prediction results show a clear increase in accuracy while comparing with the previous work [29]. The prediction accuracy line of the excellent group stands alone at the top of the figure and always keeps above 84%. The best prediction accuracy result in [29] is 71.98% which covers a mere 0.14% of testing dataset. Comparing the above finding with this work, the prediction accuracy is approximately 97% at the same coverage. This is a 25% prediction accuracy improvement. Even the Good group in this experiment shows

better quality than the P-Excellent group. Since the fair group did not generate meaningful prediction results in both this research and [29], we just skip the comparison.

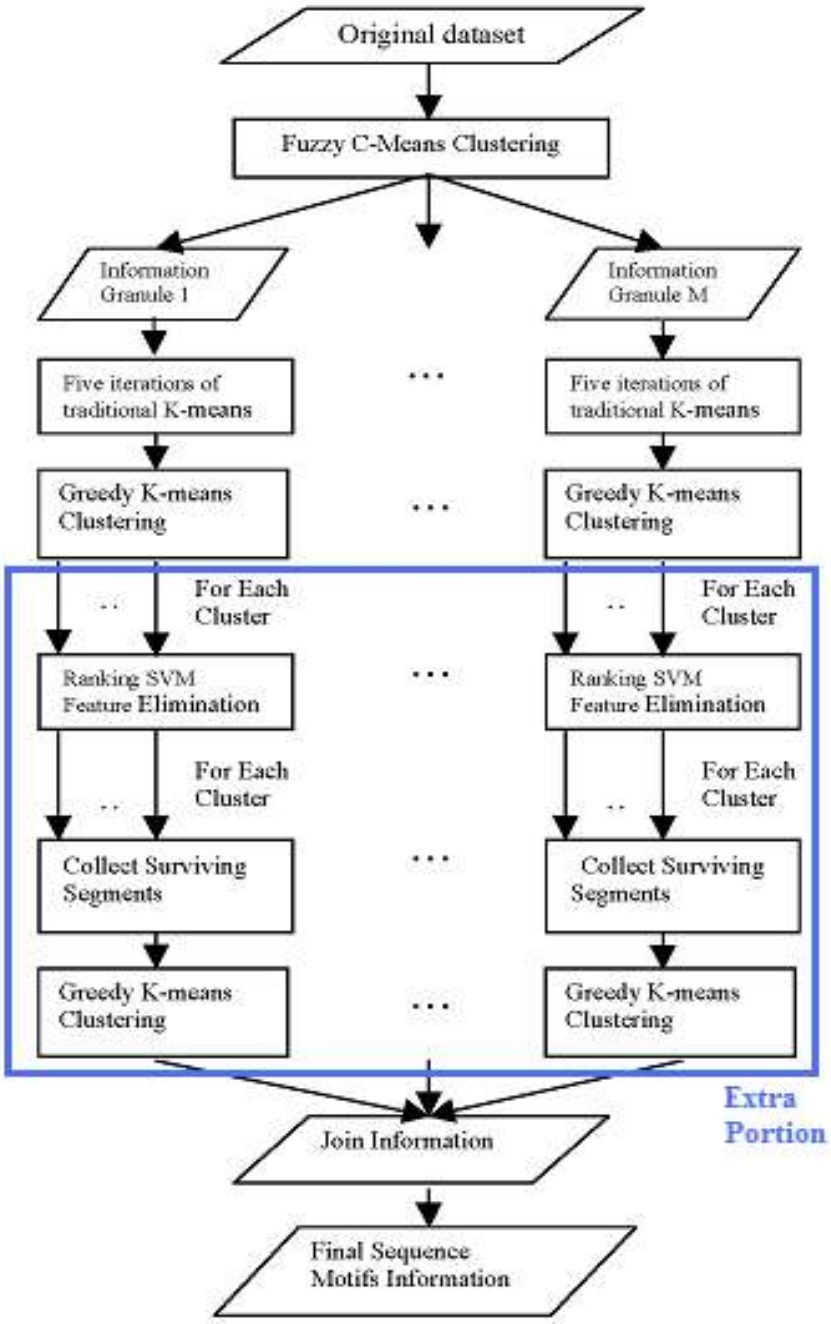

Fig. 2 The sketch of the Super GSVM-FE Model

4.3 A Stricter Prediction Criterion

In order to compare with the latest research result, we adopt the criterion that if the average dmRMSD is less than 1.5 Å, it indicates a successful prediction. However, we are also curious whether our research results are still outstanding under a stricter criterion? Therefore, we reevaluate our prediction accuracy by the standard that "if the average dmRMSD is less than 1.0 Å, it indicates a successful prediction." Table 4 (use the same format in Table 3) represents the prediction accuracy under different distance thresholds based on the new criterion. Figure 10 is derived from Table 4 for visually comparison.

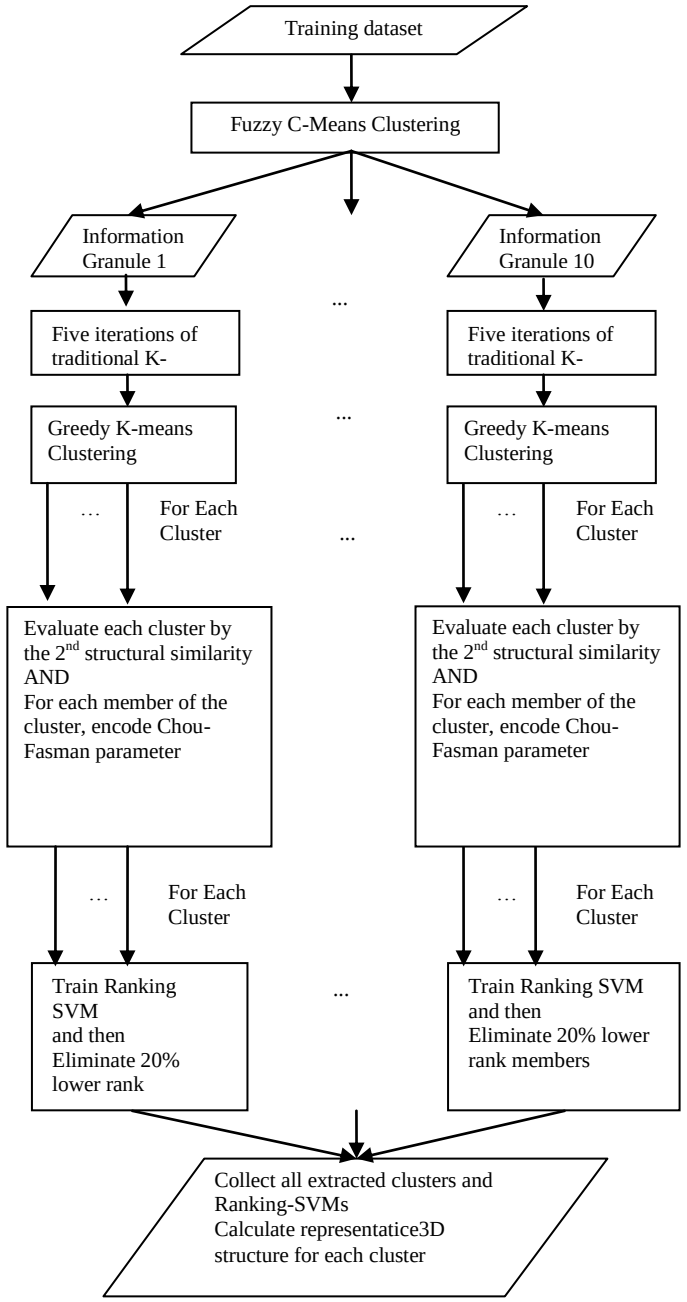

Fig. 3 The procedure of generating and extracting protein sequence motifs from primary sequence information with Chou-Fasman parameter

---

1) For each protein sequence segments with unknown 3D structure
2) Encode Chou-Fasman Parameter
3) Find the cluster (generated in Fig 3) with closest distance
4) **If** the distance is within the distance threshold:
5) Feed the unknown protein sequence segment into cluster's Ranking SVM
6) **If** Ranking SVM provides a good rank:
7) Predict protein sequence segment's 3D structure via cluster's representative 3D structure
8) **Else**
9) Find the next cluster with closest primary sequence
10) **Goto** step (3)
11) **Else**
12) Unable to predict the given protein sequence segment

Fig. 4 Pseudo code for the super Granule Support Vector Machine Model (Super GSVM)

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	0	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

Fig. 5 Part of 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	0	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

Fig. 6 An Example of the sliding window technique with a widow size of 9 applied on 1b25 HSSP file

Fig. 7 The visual description demonstrates the importance of protein local tertiary structure prediction in global structure prediction. Based on the colorful segments, which are generated by the local tertiary structure prediction model with high accuracy, we can simplify the complex exploration process from an astronomically large space by a reasonable extent.

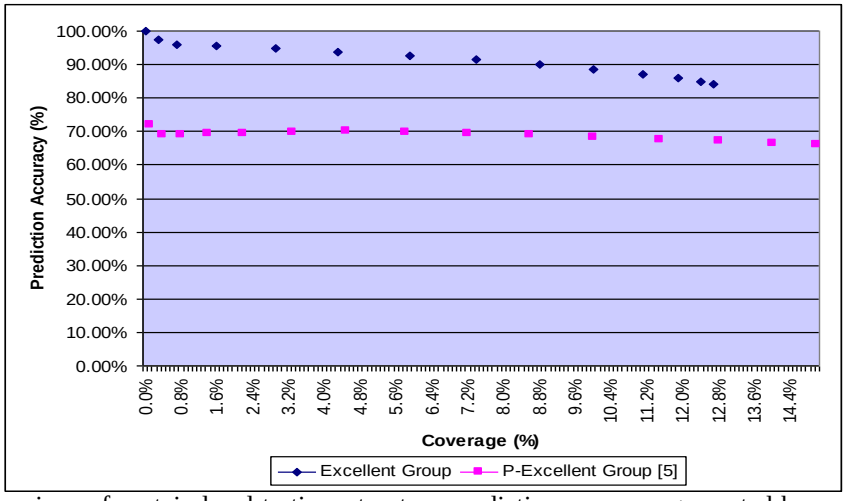

Fig. 8comparison of protein local tertiary structure prediction accuracy generated by protein sequence motifs with 2nd structure similarity > 80%

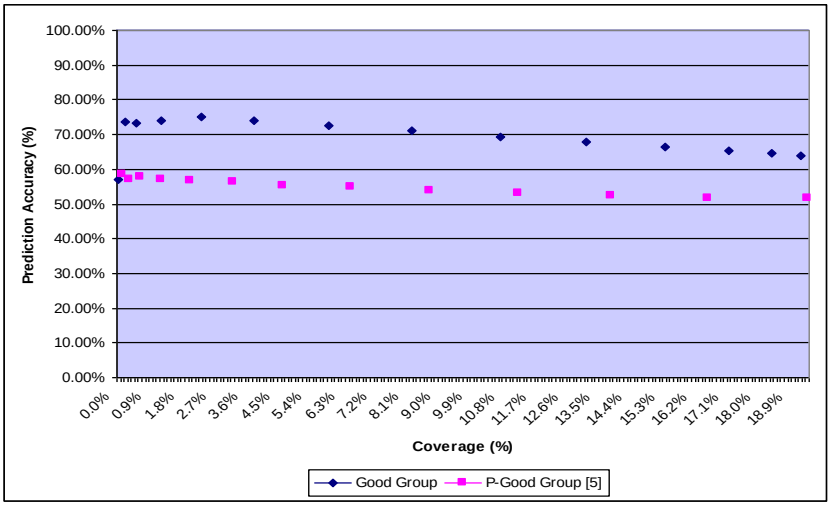

Fig. 9 Comparison of protein local tertiary structure prediction accuracy generated by protein sequence motifs with 2nd structure similarity between 70%~80%

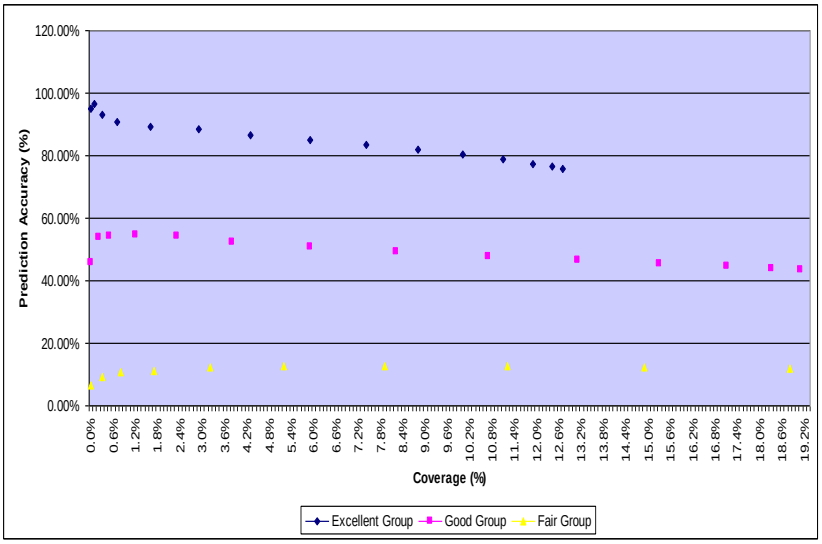

Fig. 10 Comparison of protein local tertiary structure prediction accuracy generated by different group under 1.0Å criterion

Under a much more strenuous criterion, the excellent group still able to generate prediction accuracy greater than 90% with distance thresholds equals to 900 (0.74% coverage) and generate prediction accuracy greater than 80% with distance thresholds equals to 1600 (10% coverage). At the 1.0Å criterion, all groups except the excellent group seem to fall approximately 20~30% in their prediction accuracies. The Excellent group, however, experiences a fall of < 10%. This result supports the idea that the sequence patterns with the highest secondary structural homology result in the highest prediction accuracies. This observation can also be supported by that when we change the criterion from 1.5Å to 1.0Å, the effects on the prediction accuracy is minimum.

5. FUTURE WORKS

Without any parallelization, it took our team 18 days to generate sequence clusters from our 500MB training dataset and another three months to train Ranking-SVM on all 799 clusters. Currently, we are adapting our model to support high performance computing so that we can feasibly try many different parameters and adopt the latest data.

Multiple experiments naturally follow from this study. Firstly, we can compare the newly generated clusters with the clusters from the previous study. This could reveal a new metric for cluster quality as well as increase our understanding of the impact that slight structural modifications at the primary structural level have on the overall tertiary structure. Secondly, we might discover the best weight (in equation (2)) between the protein sequence and Chou-Fasman parameter to calculate the optimal distance between two sequence segments. Last but not least, an intelligent voting mechanism can be included for better prediction accuracy generation.

6. CONCLUSION

In conclusion, it appears that the inclusion of the Chou-Fasman parameter in the training set presented to the Super GSVM significantly increases prediction accuracies. The increase is experienced without a significant rising in the quality of the clusters as measured by secondary structure homology. This suggests that the Chou-Fasman parameter (used in the prediction of secondary structure) may hold some value in the prediction of tertiary structure that is outside of that held by secondary homology. To the best of our knowledge, it is the first time that Chou-Fasman parameter is adopted into the mechanism of protein local tertiary structural prediction. Above 90% of local tertiary structure prediction is achieved by our excellent protein sequence pattern group. The high prediction accuracy implies that it is feasible to predict local tertiary structure information based on purely sequence information.

REFERENCES

[1] Bork P, Gibson TJ: Applying motif andprofile searches. Methods Enzymo/1996, 266:162-184.
[2] Bairoch A, 8ucher P, Hofmann K: PROSITE: new developments.Nucleic Acids Res 1996, 24:189-196.
[3] Pietrokovski S, Henikoff JG, Henikoff S: The BLOCKS database - a system for protein classification. Nucleic Acids Res 1996,24:19?-200.
[4] Attwood T. K., Beck M. E., Bleasby A. J., Degtyarenko K, Smityh D. J. P.: Progress with the PRINTS protein fingerprint database. Nucleic Acids Res 1996, 24:182-183.
[5] Murval J, Gabrielian A, Fabian P, Hatsagi Z, Degtyarenko K, Hegyi H, Pongor S: The SBASE protein domain library, release 4.0: a collection of annotated protein sequence sgments. Nucleic Acids Res 1996, 24:210-213.
[6] Bateman, A. et al. The Pfam protein families database. Nucleic Acids Res. 28, 263±266 (2000).
[7] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW and Noble WS: MEME SUITE: Tools for motif discovery and searching. Nucleic Acids Research 2009.
[8] Bhattacharya, S. (2009). Gibbs Sampling Based Bayesian Analysis of Mixtures with Unknown Number of Components. Sankhya.Series B.To appear.
[9]Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science, 62, 208-214.
[10]Eskin E, Pevzner P. Finding composite regulatory patterns in DNA sequences. Bioinformatics (2002) 18:S354–S363.
[11]Price, A., Ramabhadran, S. and Pevzner, P. A. (2003), 'Finding subtle motifs by branching from sample strings', Bioinformatics, Vol. 19, Suppl. 2, pp. II149–II155.
[12]PENSA, R.G., ROBARDET, C., AND BOULICAUT, J.F. 2005.A bi-clustering framework for categorical data.In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD) (Porto, Portugal).643–650.
[13]K. F. Han and D. Baker, "Recurring local sequence motifs in proteins," J. Mol. Biol., vol. 251, no. 1, pp. 176–187, 1995.
[14]Chen, B., Tai, P.C., Harrison, R. and Pan, Y., "FIK model: A Novel Efficient Granular Computing Model for Protein Sequence Motifs and Structure Information Discovery". IEEE BIBE 2006 proceeding, 2006: p. 20-26
[15]Sander C. and Schneider R., "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.

[16]Chen, B., Tai, P.C., Harrison, R. and Pan, Y., "FGK model: A Efficient Granular Computing Model for Protein Sequence Motifs Information Discovery", IASTED CASB 2006, Dallas, proceeding pp56-61.
[17]Lin, T.Y. 'Data mining and machine oriented modeling: a granular computing approach',Journal of Applied Intelligence, Kluwer, Vol. 13, No. 2, pp.113–124, 2002.
[18]Yao, Y.Y. 'On modeling data mining with granular computing', Proceedings of COMPSAC2001, pp.638–643, 2001.
[19]Wang, G. & Dunbrack, R. L. (2003) PISCES: a protein sequence culling server in Bioinformatics pp. 1589-1591, Oxford Univ Press.
[20]Zhong, W., Altun, G., Harrison, R., Tai, P. C. & Pan, Y. (2005) Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property, NanoBioscience, IEEE Transactions on. 4, 255-265.
[21]Han KF and Baker D: Global properties of the mapping between local amino acid sequence and local structure in proteins. Proceedings of the National Academy of Sciences of the United States of America 1996, 93(12):5814–5818.
[22]Bernard Chen, Stephen Pellicer, Phang C. Tai, Robert Harrison and Yi Pan, "Novel efficient granular models for protein sequence motifs and structure discovery", International Journal of Computational Biology and Drug Design, Volume 2 - Issue 2 - 2009, pp. 168-186
[23]Bernard Chen, Stephen Pellicer, Phang C. Tai, Robert Harrison and Yi Pan, "Efficient Super Granular SVM Feature Elimination (Super GSVM-FE) Model for Protein Sequence Motif Information Extraction", International Journal of Functional Informatics and Personalised Medicine, 2008 Vol. 1. No. 1, pp. 8-25.
[24]Bernard Chen, and Sinan Kockara, "Mining Positional Association Super-Rules on Fixed-Size Protein Sequence motifs", IEEE BIBE 2009, Taichung, Taiwan, proceeding pp. 1-8.
[25]Bernard Chen, Jieyue He, Stephen Pellicer, and Yi Pan, "Protein Sequence Motif Super-Rule-Tree (SRT) Structure Constructed by Hybrid Hierarchical K-means Clustering Algorithm", IEEE BIBM 2008, Philadelphia, proceeding pp. 98-103
[26]Bernard Chen, Stephen Pellicer, Phang C. Tai, Robert Harrison and Yi Pan, "Super Granular Shrink-SVM Feature Elimination (Super GS-SVM-FE) Model for Protein Sequence Motif Information Extraction", IEEE BIBE 2007,Boston, proceeding pp. 379-386
[27]Bernard Chen, Stephen Pellicer, Phang C. Tai, Robert Harrison and Yi Pan, "Super Granular SVM Feature Elimination (Super GSVM-FE) Model for Protein Sequence Motif Information Extraction", IEEE CIBCB 2007, Hawaii, proceeding pp.317-323
[28]Bystroff C, Thorsson V and Baker D: HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. Journal of Molecular Biology 2000, 301:173–190.
[29]Bernard Chen and Matthew Johnson, "Protein Local 3D Structure Prediction by Super Granule Support Vector Machines (Super GSVM)", BMC Bioinformatics 2009, 10(Suppl 11):S15
[30]ZhongW, He J, Harrison R, Tai PC and Pan Y: Clustering SupportVector Machines for Protein Local Structure Prediction. Expert Systems with Applications 2007, 32(2):518–526.
[31]Cortes C and Vapnik V: Support-Vector Networks. Machine Learning 1995, 20(3):273–297.
[32]P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," Biochemistry, vol. 13, no. 2, pp. 222–245, 1974.
[33]P. Y. Chou and G. D. Fasman, "Prediction of the secondary structure of proteins from their aminoacid sequence," Adv Enzyol Relat Areas Mol. Biol., vol. 47, pp. 45–148, 1978.
[34]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.
[35] R. Schneider, A. Daruvar, and C. Sander, "The HSSP database of protein structure-sequence alignments," Nucleic Acids Research, Vol 25, No. 1, pp. 226-230, 1997.
[36] R. Kolodny, and N. Linial, "Approximate protein structural alignment in polynomial time," Proceedings of the National Academy of Science of the United States of America, 101, 12201-12206, 2004
[37] B. Zagrovis and V. S. Pande, "How does a veraging affect protein structure comparison on the ensemble level" Biophysical Journal, 87, 2240-2246, 2004.