Multi-level clustering support vector machine trees for improved protein local structure prediction

Wei Zhong*

Division of Math and Computer Science,
University of South Carolina Upstate,
Spartanburg, SC 29303, USA
E-mail: wzhong@uscupstate.edu
*Corresponding author

Jieyue He

School of Computer Science and Engineering,
Southeast University, Nanjing 210096, China
E-mail: jieyuehe@seu.edu.cn

Xiujuan Chen

CollegeNET, Inc,
805 SW Broadway, Suite 1600,
Portland, OR 97205, USA
E-mail: xchen8@gsu.edu

Yi Pan

Department of Computer Science,
Georgia State University,
34 Peachtree Street Room1417,
Atlanta, GA 30303, USA
E-mail: pan@cs.gsu.edu

Abstract: Local protein structure prediction is one of important tasks for bioinformatics research. In order to further enhance the performance of local protein structure prediction, we propose the Multi-level Clustering Support Vector Machine Trees (MLSVMTs). Building on the multi-cluster tree structure, the MLSVMTs model uses multiple SVMs, each of which is customized to learn the unique sequence-to-structure relationship for one cluster. Both the combined 5×2 CV F test and the independent test show that the local structure prediction accuracy of MLSVMTs is significantly better than that of one-level K-means clustering, Multi-level clustering and Clustering Support Vector Machines.

Keywords: clustering algorithm; support vector machine; protein local structure prediction; parallel algorithm; protein structure prediction; clustering support vector machines.

Reference to this paper should be made as follows: Zhong, W., He, J., Chen, X. and Pan, Y. (2014) 'Multi-level clustering support vector machine trees for improved protein local structure prediction', *Int. J. Data Mining and Bioinformatics*, Vol. 9, No. 2, pp.172–198.

Biographical notes: Wei Zhong is an Assistant Professor in the Division of Math and Computer Science at University of South Carolina Upstate. He is an elected fellow of International Society of Intelligent Biological Medicine. He received his PhD Degree in Computer Science from Georgia State University, USA, in 2006. His research interests include machine learning, data mining, bioinformatics and high performance computing.

Jieyue He received the BSc and MSc from Department of Computer Science and Technique at Nanjing University, China and received her PhD in Computer Science with specialisation in Bioinformatics from Southeast University, China. She is a Professor of the school of Computer Science and Engineering, Southeast University, China. Her current research interests include bioinformatics, data mining, machine learning and database system.

Xiujuan Chen is a Software Engineer at CollegeNET, Inc in Portland OR. She obtained her PhD Degree in Computer Science at Georgia State University in 2008. Her research interests include computational intelligence, machine learning, data mining and bioinformatics.

Yi Pan is the Chair and a Professor in the Department of Computer Science and a Professor in the Department of Computer Information Systems at Georgia State University. He received his PhD in Computer Science from the University of Pittsburgh, USA, in 1991. His research interests include parallel and distributed computing, optical networks, wireless networks and bioinformatics. He has published more than 100 journal papers with over 30 papers published in various IEEE journals. His recent research has been supported by NSF, NIH, NSFC, AFOSR, AFRL, JSPS, IISF and the states of Georgia and Ohio.

1 Introduction

Studying the protein sequence-to-structure relationship is one of the most active bioinformatics research areas. A better understanding about protein sequence-to-structure correspondence can improve effectiveness and efficiency of local protein structure prediction (Rahman and Zomaya, 2005). Many biochemical tests indicate that a sequence can determine its structure completely because all the information that is necessary to specify protein interactions with other molecules is embedded into its sequence (Karp, 2002). These studies provide the experimental support for exploring the protein sequence-to-structure relationship using the data mining techniques. In the previous work, the structure-cluster based approach and the sequence-cluster based approach are used to explore the sequence-to-structure relationship. Subsequently, knowledge generated from these approaches is utilised for local protein structure prediction.

For the structure-cluster based approach, protein structural segments are grouped into different structural clusters using multiple structural alignments (Yang and Wang, 2003) and

unsupervised clustering algorithms (De Brevern1 et al., 2004; Etchebest et al., 2005; Benros et al., 2006). Each cluster is associated with a representative local structural prototype. In these approaches, multiple structural alignments and unsupervised clustering algorithms alone may not be adequate to understand the complex nonlinear sequence-to-structure relationship since these approaches generally do not incorporate evolutionary information from homologous sequences during structural alignment and clustering process. For the sequence-cluster based approach, sequence segments are clustered into high quality sequence clusters with the one-level K-means clustering algorithm (Han and Baker, 1995, 1996; Zhong et al., 2005) and multiple sequence alignment (Hunter and Subramaniam, 2003).

The sequence-cluster based approach and the structure-cluster based approach described above utilise a set of structure-based sequence profiles generated from structure clusters and sequence clusters to predict the backbone torsion angles and protein secondary structure for local protein structure (Yang and Wang, 2003; Benros et al., 2006; Bystroff and Baker, 1998). In 2000, Hidden Markov Model (HMM) was set up based on high quality sequence clusters in order to predict the backbone torsion angles for local protein structure (Bystroff et al., 2000).

This study focuses on analysing the sequence-cluster based approach. Current sequence-cluster based approach (Han and Baker, 1995, 1996; Zhong et al., 2005; Bystroff and Baker, 1998) depends on the one-level K-means clustering algorithm. One-level clustering may not reflect optimal partitioning especially for very large and complex protein datasets. For example, the protein dataset in this study contains more than half million sequence segments. As a result, a number of clusters produced by the one-level clustering algorithm have poor structural similarity. A number of clusters with poor structural similarity can affect the performance of protein local structure prediction noticeably. Furthermore, the clustering algorithm is critical to explore how protein sequences correspond to local 3D protein structure in these approaches. The conventional clustering algorithm such as the K-means and K-nearest neighbor algorithm assumes that the distance between samples can be calculated with exact precision (Zhong et al., 2007). When the distance function for these clustering algorithms is not well defined, the clustering algorithm may not be effective to discover the complex sequence-to-structure relationship.

Support Vector Machine (SVMs) has shown superior classification performance in various bioinformatics applications due to strong generalisation capability (Wang and Wu, 2006; Xia et al., 2010; Vapnik, 1998). SVM can deal with the nonlinear relationship by implicitly mapping input samples from the input feature space into another high dimensional feature space with the nonlinear kernel function. Consequently, SVM is more favorable to explore the nonlinear protein sequence-to-structure relationship than the conventional clustering algorithm. Since its training time complexity is at least quadratic to the number of samples, SVM trainings for very large datasets are slow process (Vapnik, 1998). The task of learning the sequence-to-structure correlation using an SVM becomes more challenging when each subspace of the whole protein sample space corresponds to different local 3D structure (Zhong et al., 2005).

Many SVM training algorithms were proposed to enhance the efficiency of SVM trainings for large datasets while keeping reasonable performance. These algorithms can be divided into three major classes. The first class of algorithms is *decomposition algorithms* (Vapnik, 1998; Platt, 1999; Joachims, 1999). Although the decomposition algorithm can speed up the training process, they do not scale well with the size of datasets since the kernel matrix may grow beyond the available memory during the optimisation process. The second class of algorithms to handle large datasets is *selective sampling techniques*, which choose a small number of high quality training samples intelligently from the whole dataset in order to improve the learning capacity of SVM (Khan et al., 2007; Li et al., 2007, 2008). The

selective sampling techniques may reduce the classification performance of SVM when a single effective decision boundary is difficult to form for the protein datasets having multiple sequence-to-structure distribution patterns in different sample subspaces.

The third class of algorithms is the multiple SVM systems such as Bayesian Committee Machine (BCM) (Tresp, 2000), SVM ensembles (Valentini, 2005), Clustering Support Vector Machines (CSVMs) (Zhong et al., 2007) and Super Granule Support Vector Machines (GSVMs) (Chen and Johnson, 2009). The success of CSVMs and Super GSVMs depends on the greedy cluster assignment algorithm and one-level clustering. The greedy cluster assignment algorithm used by CSVMs and Super GSVMs assumes that the cluster distance function takes precedence over SVM's decision function (Zhong et al., 2007; Chen and Johnson, 2009). This assumption may not be correct. Furthermore, CSVMs and Super GSVMs are constructed based on the one-level clustering algorithm, which may not explore the sequence-to-structure relationship effectively.

To overcome weakness of one-level clustering and enhance SVM training for very large datasets, we proposed MLSVMTs for huge datasets. The construction of MLSVMTs is divided into four phases.

1 In the first phase, an improved K-means clustering algorithm is used to cluster the dataset into the one-level partition.

2 In the second phase, the hierarchical clustering algorithm is applied to each one-level cluster having low structural similarity in parallel. The root clusters of resulting cluster trees in the first step are merged until the structural similarity of the combined clusters falls below the given structural similarity threshold. At the end of the second phase, the cluster subtrees are generated.

3 In the third phase, a SVM is trained for each cluster in the cluster subtrees. Each SVM focuses on its cluster at a particular level of the tree so that a specific classifier is trained utilizing a particular high dimensional hyperspace.

4 In the final phase, the SVMs from different levels of the tree, operating in different hyperspaces, cooperatively decide the cluster assignment of a given sequence segment based on the combined distance score and the SVM decision score.

The representative 3D local structure of the assigned cluster is given to the sequence segment. To the best of our knowledge, no researchers have predicted protein secondary structure, the backbone torsion angles and distance matrix for local protein structure simultaneously.

The multi-level clustering algorithm can explore subclusters of one-level clusters having poor structural similarity. This strategy can potentially discover some high quality sublcusters from these one-level clusters. Increasing number of clusters with high structural similarity can potentially improve the accuracy of local structure prediction. Building on the multi-level cluster trees, the SVM can filter out the noisy information from each cluster in the multi-level tree after learning the specific sequence-to-structure relationship for each cluster. Consequently, MLSVMTs can handle the complex protein sequence-to-structure relationship more effectively than the sequence-cluster based approach and the structural-cluster based approach described previously for local structure prediction. Unlike previous approaches for local protein structure prediction, MLSVMTs proposed in this work are utilized to predict protein secondary structure, the backbone torsion angles and distance matrix for local protein structure simultaneously, which is critical for deriving the global 3D protein structure.

In order to evaluate the effectiveness of MLSVMTs for local protein structure prediction, the performance of MLSVMTs is compared to three computational models:

- the improved K-meaning clustering algorithm
- the multi-level clustering algorithm
- Clustering Support Vector Machines (CSVMs).

The local structure prediction performance is measured by Accuracy One and Accuracy Two, which are defined in this work. Both the combined 5×2 Cross Validation (CV) F test (Alpaydin, 1999) and the independent test are conducted for rigorous performance evaluation.

Our paper is organised as follows. In the Section 2, four phases of MLSVMTs are discussed in details. In the Section 3, the training set, the testing set, accuracy definition and parallel algorithm are explained. In the Section 4, the experimental results and analysis are given. Finally, the conclusion and the future works are presented.

2 Multi-Level Clustering Support Vector Machine Trees (MLSVMTs)

Construction of the MLSVMTs model is divided into four phases. The detailed algorithm for constructing the MLSVMTs model is shown in Figure 1. The running example for MLSVMTs is shown in Figure 2.

2.1 Partitioning the whole dataset into multiple clusters using improved K-means clustering algorithm

Since the K-means clustering is computationally efficient for large data sets with both numeric and categorical attributes (Gupta et al., 1999), K-means clustering is selected to partition the whole dataset into multiple data subsets. Sequence segments of nine successive residues generated from protein sequences using the sliding window techniques are partitioned into different clusters with the improved K-means algorithm (Zhong et al., 2005). In order to compare performance of the improved K-means clustering reported by the previous work and multi-level clustering, the number of clusters, K, is selected as 800 in this work. During the clustering process, each sequence segment represented by the 20×9 HSSP frequency profiles matrix (Sander and Schneider, 1991) is assigned to the cluster with the lowest *distance score between the sequence segment and the cluster*. The distance score between a given sequence segment and a specified cluster is defined as (Zhong et al., 2005):

$$Dist(k, x) = \sum_{i=1}^{L} \sum_{j=1}^{N} \left| F_x(i, j) - F_k(i, j) \right| \tag{1}$$

where L is the window size and N is 20. $F_x(i, j)$ is the value of frequency profiles at row i and column j for the sequence segment x. $F_k(i, j)$ is the value of the matrix at row i and column j for the centroid of the cluster K. The centroid of the given cluster is the average of all HSSP frequency profiles of sequence segments belonging to this cluster (Zhong et al., 2005). The distance score between the sequence segment o and the sequence segment p is defined as:

$$Dist(o, p) = \sum_{i=1}^{L} \sum_{j=1}^{N} \left| F_o(i, j) - F_p(i, j) \right| \tag{2}$$

where F_o (i, j) is the value of frequency profiles at row i and column j for the sequence segment o and F_p (i, j) is the value of frequency profiles at row i and column j for the sequence segment p.

Figure 1 Four phases of MLSVMTs model

Multi-Level Clustering Support Vector Machine Trees

Phase 1. Partitioning the whole dataset into multiple clusters using the improved K-means clustering algorithm. The value of K is selected as 800 based on the previous study.

Phase 2. Generating multiple cluster subtrees

FOR each one-level cluster
{
Applying the agglomerative hierarchical clustering algorithm. Merging of two clusters stops when the structural similarity of the merged cluster falls below the given threshold. In the end, this step produces a cluster tree for one cluster.
}

Applying the agglomerative hierarchical clustering algorithm to the root clusters of each tree structure. Merging of two clusters stops when the structural similarity of the merged cluster falls below the given threshold.

Phase 3. Training SVM for each cluster in the cluster trees
Classifying clusters into different groups based on the structural similarity
FOR each cluster in the cluster subtree
{
Labelling each training sample as positive or negative respectively for different cluster groups
Training each SVM for each cluster by optimising RBF kernel parameters (j, γ, and C) with the grid search algorithm
}

Phase 4. Selecting the most suitable cluster from cluster trees and assigning the local 3D structure of this cluster to the sequence segment
FOR each subtree
{
Selecting the representative $cluster_k$ *s.t.* $decision\_value_{cluster\_k} = \max_j (decision\_value_{cluster\_j})$
where $decision\_value_{cluster\_j}$ *is the weighted_decision_value for cluster j in the subtree*
}
Selecting the representative $cluster_i$ *s.t* $decision\_value_{representative\_cluster\_i} = \max_j (decision\_value_{representative\_cluster\_j}$
where $decision\_value_{representative\_cluster\_j}$ *is the weighted decision value of the representative cluster from subtree* $_j$

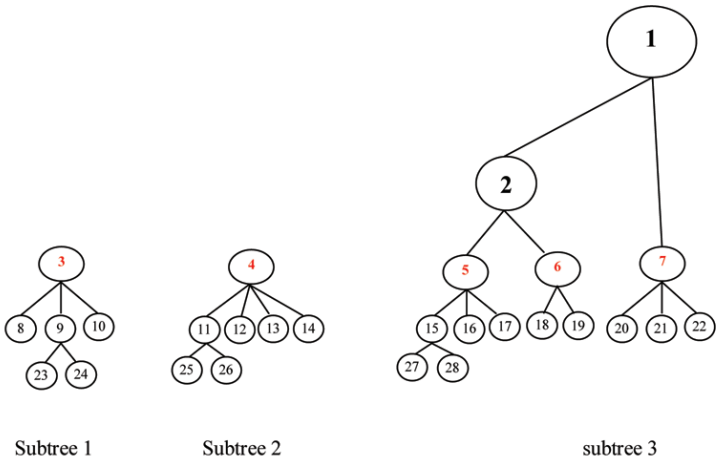
Assigning the 3D local structure of the representative $cluster_i$ *to the sequence segment*

2.2 Generation of multiple cluster subtrees

Since many one-level K-means clusters have low structural similarity, the multiple cluster subtree generation algorithm is proposed to explore high quality subclusters from these one-level clusters having low structural similarity.

In the first step of the multiple cluster subtree generation algorithm, the agglomerative hierarchical algorithm is applied to each of one-level clusters having low structural similarity using the distance score as defined in the equation (2). The cluster merging process repeats until the structural similarity of the merged cluster falls below the giving threshold. Essentially, a forest of cluster trees is generated after the first step. In the second step, the agglomerative hierarchical clustering algorithm is applied to the root clusters of each tree structure. The merging process repeats until the structural similarity of the

Figure 2 Running examples for the MLSVMTs model (see online versions for colours)



Running Example for MLSVMTs

Phase 1.
For simplicity, assume K be 5 for this running example. The whole dataset is divided into five subsets from cluster 3 to cluster 7 using the improved K-means clustering algorithm.

Phase 2.
Step1: The agglomerative hierarchical clustering algorithm is applied to each of one-level clusters from cluster 3 to cluster 7. As a result, five tree structures are produced with the root cluster from cluster 3 to cluster 7.
Step 2: The agglomerative hierarchical clustering algorithm is applied to the root clusters of these five tree structures. Consequently, three subtrees are generated.

Phase 3.
Step 1: 28 clusters are classified into different cluster groups based on structural similarity.
Step 2: Sequence segments for each cluster in three subtrees are labelled as positive or negative based on structural deviation from the representative structure of each cluster.
Step 3: One SVM is trained for one cluster in three subtrees. Totally 28 SVMs are constructed.

Phase 4.
Step 1: Cluster 9 with the 0.5 weighted decision value is selected as the representative cluster for subtree1.
Cluster 25 with the 0.3 weighted decision value is selected as the representative cluster for subtree2.
Cluster 6 with the 0.8 weighted decision value is selected as the representative cluster for subtree3.
Step 2: Cluster 6 is finally selected since it has the highest weighted decision value among three representative clusters
Step 3: The representative structure of cluster 6 is assigned to the sequence segment.

merged cluster falls below the given threshold. In the end, multiple cluster subtrees are generated.

If *secondary structural similarity* is below 60% or *Average_dmRMSD* is above 2.5 Å or *Average_taRMSD* is above 35 degrees for a given cluster, the structural similarity of this cluster falls below the threshold in the pseudo code. The threshold is based on our analysis of experimental results. Our analysis shows that the structures of sequence segments in one cluster are generally highly deviated from its representative structure if the structural similarity of the cluster falls below a given threshold. In other words, the structures of sequence segments belonging to one cluster are not compact. Clusters whose structural similarity is above this threshold belong to the average cluster group defined in this work.

Clusters at different levels of the tree are capable of capturing local protein sequence-to-structure distribution at different levels. As demonstrated in the experimental results, the multi-level clustering algorithm is more capable of capturing the complex sequence-to-structure patterns in large protein datasets than the one-level clustering approach since the multi-level clustering algorithm may discover many high quality subclusters from one-level clusters having low structural similarity.

After the multi-level cluster trees are generated, the representative 3D structure of each cluster is calculated. The representative 3D structure for each cluster includes the average secondary structure, Average Distance Matrix (ADM) and representative torsion angles including φ and ψ defined in (Karp, 2002).

Average Distance Matrix (ADM) is defined as:

$$\alpha_{i\to j}^{ADM} = \sum_{K=1}^{N} \alpha_{i\to j}^{k} \Big/ N \tag{3}$$

where $\alpha_{i\to j}^{k}$ is the distance between α-carbon atom *i* and α-carbon atom *j* in the sequence segment *k* of the length *L* and *N* is the number of sequence segments of a given cluster. ADM basically calculates the average for the distance matrices of all the sequence segments in one cluster.

ϕ_i is the representative ϕ in the *ith* position of sequence segments for sequence clusters. All the ϕ values in the position *i* of sequence segments in a sequence cluster are put into a set. The representative ϕ_i is defined as the ϕ value that is taken from this set and has the minimum sum of modular distances to the other members of this set. In a sense, ϕ_i is the closest neighbor to the other members of this set. ψ_i is similarly defined.

2.3 Training svm for each cluster in the cluster subtrees

Multi-level partitioning can discover some subclusters with high structural similarity from one-level clusters. However, multi-level clustering can still introduce noisy and irrelevant information into each cluster, which may reduce the performance of local protein structure prediction. In order to identify noisy sequence-to-structure information, SVM is trained to evaluate the strength of the sequence-to-structure correspondence for each sequence segment belonging to the same cluster. After learning the relationship between the frequency profile distribution and local representative 3D structure for each cluster, SVM can filter out potentially unreliable structure prediction and select potentially reliable structure prediction for each cluster.

In each cluster, positive sequence segments are defined as those samples whose structure deviation from the representative structure is below a given threshold and negative sequence segments are defined as those samples whose structure deviation from the representative structure of this cluster is above a given threshold. Frequency profiles of positive sequence segments may be closely mapped to the representative 3D structure of the specified cluster. Labelling sequence segments for each cluster as positive or negative can provide important training patterns for SVM to learn the underlying sequence-to-structure relationship for each cluster. After SVM model construction, the SVM decision function can produce the decision score, which indicates the distance between the testing sequence segment and the optimal hyper-plane.

Since distribution patterns for frequency profiles in each cluster are quite different, SVM training is customised for clusters belonging to different cluster groups. The definition for different cluster groups is introduced in the datasets and experimental setup section. The SVMs trained for clusters belonging to the average cluster group are customised to recognise

sequence segments whose structure can be reliably predicted. The SVMs for clusters belonging to the good cluster group are trained to filter out sequence segments whose 3D structure cannot be reliably predicted.

The SVM decision function for the cluster K to classify the sequence segment x is formulated as:

$$f_{svm\_k}(x) = \left(\sum_{i=1}^{sv} \alpha_i y_i K_{svm\_k}(x, x_i) + b \right) \tag{4}$$

where sv is the number of support vectors and $K_{SVM\_k}(x,x_i)$ is the kernel function from svm_k trained for cluster K. In this work, the decision score reflects how closely the sequence segment corresponds to the representative 3D structure of this cluster.

2.4 Cluster assignment algorithm

After SVM for each cluster in the cluster subtrees is trained, the cluster assignment algorithm is used to select the most suitable cluster for local structure prediction.

First, the classification value, $f_{svm\_k}(x)$, of a SVM, svm_k, is normalised using the z-score for fair comparison of classification values from different SVMs. This normalisation step is necessary because decision boundaries of SVMs for different clusters in the tree structure are obtained in different high-dimensional sample spaces for tackling the classification problem in different sample subspaces. The *decision value* of svm_k for a sequence segment x is defined as the z-score of svm_k's classification value for a given sequence segment x:

$$Decision\_value_{svm\_k}(x) = \frac{(f_{svm\_k}(x) - mean_{svm\_k})}{\sigma_{svm\_k}} \tag{5}$$

where $mean_{svm\_k}$ is the mean classification values for svm_k in cluster k and σ_{SVM_k} is the standard deviation of classification values for svm_k in the cluster k. The higher the magnitude of the *decision value* of a svm_k, $|Decsion\_value_{svm\_k}(x)|$, the higher the SVM's *confidence level* for classifying a sequence segment x will be.

The confidence of the SVM decision value can be strongly affected by the distance between the sequence segment and the cluster associated with this SVM. Hence, the SVM decision value is weighted by *the distance score between the sequence segment and the given cluster*, as defined in equation (1). In the equation (6), the distance between the sequence segment x and cluster k is smoothed by the logistic function:

$$smooth\_dist(k,x) = \frac{1}{1 + e^{-dist(k,x)}} \tag{6}$$

where k is the cluster k and x is the given sequence segment. As a result, the *weighted decision value* for svm_k for a sequence segment x is defined as:

$$\psi(svm\_k, x) = Decision\_value_{svm\_k}(x) \times smooth\_dist(k,x) \tag{7}$$

The cluster assignment algorithm includes three steps. In the first step, the representative cluster having the highest weighted decision value is selected for each subtree. In the second step, the cluster with the highest weighted decision value among all subtree representative clusters is chosen. Finally, the 3D representative structure of the selected cluster is assigned to the sequence segment.

3 Datasets and experimental setup

In this section, datasets for the combined 5×2 CV F test and the independent test are described first. Then, details of cluster structural similarity, performance evaluation metrics and the parallel algorithm are explained.

3.1 Dataset for combined 5 × 2 Cross Validation F test

The dataset for the combined 5×2 CV F test has 2,952 protein sequences obtained from the Protein Sequence-Culling Server (PISCES) (Wang and Dunbrack, 2003). This protein dataset has 656,528 sequence segments. In this protein dataset, the percentage identity cutoff is 25%, the resolution cutoff is 1.8 and the R-factor cutoff is 0.25. No sequences of this dataset share more than 25% sequence identities. The structures of protein sequences in the training set and the testing set are available from Protein Data Bank (PDB) (Berman et al., 2000).

3.2 Dataset for independent test

To evaluate the performance of the new model more rigorously, the dataset for the combined 5×2 CV F test is used as the training set. 300 protein sequences from the recent release of PISCES are included into the independent test set. Any two sequences in the test set share less than 25% similarity.

3.3 Cluster structural similarity calculation

Secondary structure similarity, Average Distance Matrix Root Mean Square Deviation (average_dmRMSD) and Average Torsion angle RMSD (average_taRMSD) are three important metrics to evaluate structural similarity for each cluster.

3.3.1 Secondary structural similarity for a given cluster

Secondary Structural similarity for a given cluster is defined as (Zhong et al., 2005):

$$Seconary\_Structural\_Similarity = \frac{\sum_{i=1}^{ws} \max(P_{i,H}, P_{i,E}, P_{i,C})}{ws} \tag{8}$$

where ws is the window size. $P_{(i,H)}$ is the frequency of occurrence of helices among the sequence segments for the cluster in position i. $P_{(i,E)}$ and $P_{(i,C)}$ are similarly defined. The representative secondary structure in the given position is defined as the secondary structure having the maximum frequency. The results of the average maximum frequency from all positions indicate the secondary structural similarity for a given cluster.

3.3.2 Average Distance Matrix Root Mean Square Deviation for a given cluster

Distance Matrix Root Mean Square Deviation between a sequence segment *s1* and the representative structure of giving cluster is defined as:

$$dmRMSD(C, s1) = \sqrt{\frac{\sum_{i=1}^{L} \sum_{j=i+1}^{L} \left(\alpha_{i \to j}^{s1} - \alpha_{i \to j}^{ADM}\right)^2}{M}} \tag{9}$$

$$M = \frac{\left(L \times L - L\right)}{2} \tag{10}$$

where $\alpha_{i \to j}^{ADM}$ is the distance between α-carbon atom *i* and α-carbon atom *j* in the ADM for a cluster *C*. *M* is the number of distances in the distance matrix. Average Distance Matrix Root Mean Square Deviation (*average_dmRMSD*) is defined as:

$$average\_dmRMSD = \frac{\sum_{i=1}^{N} dmRMSD(C,i)}{N} \tag{11}$$

where *dmRMSD(C,i)* is Distance Matrix Root Mean Square Deviation between sequence segment *i* and the representative structure of a giving cluster *C*. *N* is the number of sequence segments in the given cluster.

3.3.3 Average torsion angle Root Mean Square Deviation for a given cluster

Torsion angle RMSD between a sequence segment *s1* and the representative structure of a giving cluster *C* is defined as:

$$taRMSD(C,s1) = \sqrt{\frac{\sum_{k=1}^{L}\left\{\left(\varphi_{ki} - \varphi_{kj}\right)^2 + \left(\psi_{ki} - \psi_{kj}\right)^2\right\}}{2L}} \tag{12}$$

where ϕ_{kj} is ϕ in the position *k* of the representative angle for a cluster *C* and ψ_{kj} is ψ in the position *k* of the representative angle for a cluster *C*. ϕ and ψ are defined in (Karp, 2002). Average Torsion Angle Root Mean Square Deviation (*average_taRMSD*) is defined as:

$$average\_taRMSD = \frac{\sum_{i=1}^{N} taRMSD(C,i)}{N} \tag{13}$$

where *taRMSD*(*C,i*) is torsion angle Root Mean Square Deviation between sequence segment *i* and the representative structure of a giving cluster *C*. *N* is the number of sequence segments in the given cluster.

3.3.4 Classification of clusters into different groups based on structural similarity

Table 1 shows the criteria to classify clusters into different groups based on structural similarity in the training set. The clusters produced by the clustering algorithm are divided into three groups based on structural similarity in the training set. The excellent cluster group includes all clusters having *secondary structure similarity* greater than 80%, *average_dmRMSD* less than 1 Å and *average_taRMSD* less than 25 degrees. The average cluster group and the good cluster group are similarly defined. As a result, all the clusters in the good cluster group have high structural similarity. All the clusters in the average cluster group have average structural similarity.

Table 1 Standard to classify clusters into different groups

	Secondary structure similarity	*Average_dmrmsd*	*Average_tarmsd*
Average Cluster Group	between 60% and 70%	between 1.6 Å and 2.5 Å	Between 31 degree and 35 degrees
Good Cluster Group	between 70% and 80%	between 1 Å and 1.5 Å	between 25 and 30 degrees
Excellent Cluster Group	greater than 80%	less than 1 Å	less than 25 degrees

3.4 Performance evaluation metrics for local 3D structural prediction

Secondary structure accuracy called *Q3*, Distance Matrix Root Mean Square Deviation (dmRMSD) and Torsion angle RMSD (taRMSD) are three important metrics to evaluate accuracy for protein structure prediction.

Q3 representing the three-state overall percentage of correctly predicted residues is one of the popular performance evaluation measures in protein secondary structure prediction. *Q3* is defined as (Hu et al., 2004):

$$Q3 = \frac{\sum_{i\in\{H,E,C\}} \#\ of\ residues\ correctly\ predicted_i}{\sum_{i\in\{H,E,C\}} \#\ of\ residues\ in\ class\ i} \tag{14}$$

dmRMSD is defined as (Kolodny and Linial, 2004; Zagrovic and Pande, 2004):

$$dmRMSD = \sqrt{\frac{\sum_{i=1}^{L}\sum_{j=i+1}^{L}\left(\alpha_{i\to j}^{predicted} - \alpha_{i\to j}^{target}\right)^2}{M}} \tag{15}$$

$$M = \frac{\left(L\times L - L\right)}{2} \tag{16}$$

where $\alpha_{i\to j}^{target}$ is the distance between α-carbon atom *i* and α-carbon atom *j* in the target distance matrix of a sequence segment. $\alpha_{i\to j}^{predicted}$ is the distance between α-carbon atom *i* and α-carbon atom *j* in the predicted distance matrix of a sequence segment. *M* is the number of distances in the distance matrix. *taRMSD* is defined as:

$$taRMSD = \sqrt{\frac{\sum_{k=1}^{L}\left\{\left(\phi_{ki} - \phi_{kj}\right)^2 + \left(\psi_{ki} - \psi_{kj}\right)^2\right\}}{2L}} \tag{16}$$

where ϕ_{kj} is ϕ in the position *k* of the target angle for a cluster and ψ_{kj} is ψ in the position *k* of the target angle for a cluster. ϕ and ψ are defined in (Karp, 2002).

Only combined information of secondary structure, torsion angle and distance matrix can represent 3D protein structure precisely. In order to compare the 3D local structure

prediction performance of several computational models rigorously, two sets of accuracy criteria including Accuracy One and Accuracy Two are defined in this work in order to evaluate secondary structure accuracy, dmRMSD and taRMSD simultaneously. Table 2 provides the threshold for evaluating Accuracy One and Accuracy Two for local structure prediction. Accuracy Two is the percentage of sequence segments with secondary structure accuracy greater than 70%, dmRMSD less than 1.5 Å and taRMSD less than 30 degree in the test set for a given cluster. Accuracy Two indicates the percentage of sequence segments whose 3D structure can be predicted reliably. Accuracy One is similarly defined. Accuracy One indicates the percentage of sequence segments with acceptable 3D structure prediction.

Table 2 Threshold for evaluating 3d local structure prediction accuracy one and accuracy two

	Secondary structure accuracy	*dmRMSD*	*taRMSD*
Accuracy One	> 60%	< 2.5 Å	< 35 degrees
Accuracy Two	> 70%	< 1.5 Å	< 30 degrees

3.5 Parallel algorithm for multi-level clustering, CSVMs and MLSVMTs

Model construction is time consuming especially for a very large protein dataset containing 656,528 sequence segments. However, the multi-level clustering algorithm is inherently parallelisable since the agglomerative hierarchical clustering applied to each cluster can be performed in parallel. SVMs modeled for each cluster in the multi-level cluster tree can be constructed in parallel as well. Consequently, the parallel algorithm is applied to multi-level clustering, CSVMs and MLSVMTs. In this work, sixty four desktop computers using a Core 2 Duo 2.4 GHz Processor are used for the parallel experiment.

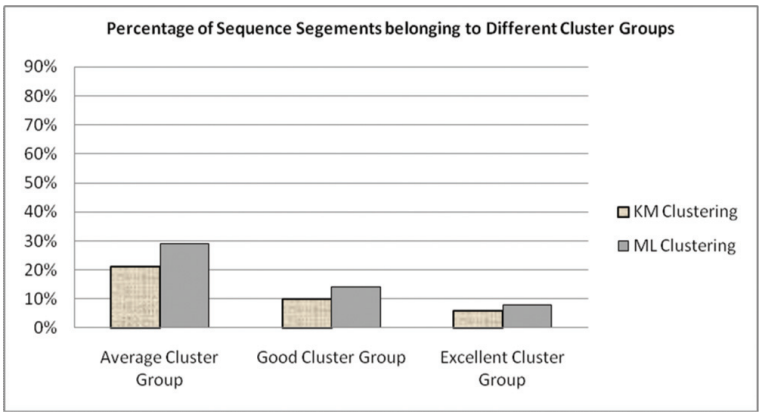
4 Experimental results and analysis

In this section, the percentage of sequence segments belonging to high quality clusters for the one-level clustering algorithm and the multi-level clustering algorithm is compared. Experimental results for the combined 5×2 CV F test and the independent test are used to compare the performance of four computational models. The running time for four computational models is also reported. Finally, sample cluster subtrees and their biological significance are discussed.

4.1 Structural similarity comparison between multi-level clustering algorithm and one-level clustering algorithm

Figure 3 compares the average percentage of sequence segments belonging to different cluster groups between the one-level improved K-means algorithm and the multi-level clustering algorithm. The label ‘KM clustering’ denotes the one-level improved K-means clustering algorithm. The label ‘ML Clustering’ denotes the multi-level clustering algorithm proposed in this work.

Figure 3 Comparison of the percentage of sequence segments belonging to clusters with high structural similarity (see online version for colours)



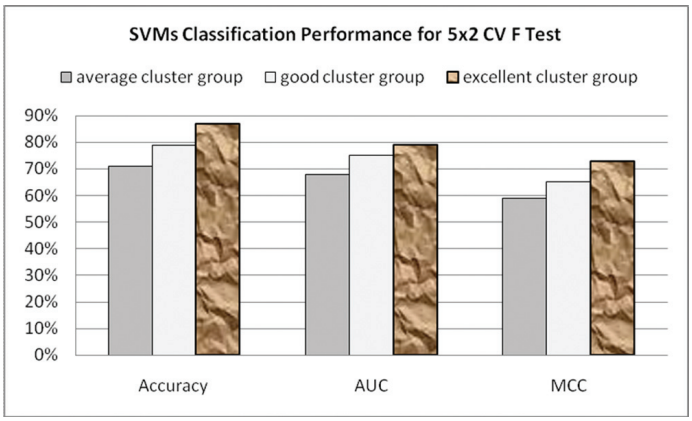
The multi-level clustering algorithm increases the average percentage of sequence segments belonging to the average cluster group by almost 8% and improves the average percentage of sequence segments belonging to the good cluster group by 4%. The increased average percentage of sequence segments belonging to high quality clusters suggest that the multi-level clustering algorithm can find some higher quality subclusters from clusters generated from the one-level improved K-means clustering algorithm. The solid results from the multi-level clustering algorithm make strong foundation for better local protein structure prediction.

4.2 Experimental results for the combined 5 × 2 Cross-Validation F test

4.2.1 SVM classification performance for different cluster groups

Figure 4 shows average accuracy, the Area Under the Receiver Operating Characteristic Curve (AUC) (Baldi et al., 2000) and Matthews Correlation Coefficient (MCC) (Baldi et al., 2000) of SVMs for different clustering groups in the 5×2 CV F test. Besides accuracy, AUC and MCC is also the important indicator for the generalisation power of SVMs especially for the imbalanced dataset. Figure 4 indicates that SVMs modeled for different cluster groups display strong capability to discriminate between positive samples and negative samples. Satisfactory performance of SVMs for the average cluster group reveals that SVMs for the average cluster group are able to select frequency profiles of sequence segments whose structure can be reliably predicted. Strong performance of SVMs for the good cluster group demonstrates that these SVMs obtain the capability to filter out frequency profiles of sequence segments whose structure cannot be reliably predicted. Experimental analysis indicates that distribution patterns of frequency profiles for the average cluster group are more diverse while distribution patterns of frequency profiles for the good cluster group tend to be more compact. As a result, learning tasks of SVMs in different cluster groups are different and the customised SVMs can learn the unique sequence-to-structure relationships for different cluster groups more specifically.

Figure 4 SVMs classification performance for different cluster groups in the 5 × 2 CV F test (see online version for colours)



4.2.2 3D local structure prediction results for sequence segments in the 5 × 2 CV F test

At first, experimental results for comparing 3D local structure prediction accuracy of the four models are discussed. In Figure 5, average 3D local structure prediction Accuracy One for four models using the 5 × 2 CV F test is compared. The label 'KM clustering' denotes the improved K-means clustering algorithm. The label 'ML Clustering' denotes the multi-level clustering algorithm proposed in this work. The label 'CSVMs' denotes the Clustering Support Vector Machines. The label 'MLSVMTs' denotes Multi-level Clustering Support Vector Machines trees. Compared with the improved K-means clustering algorithm, the multi-level clustering algorithm improves Accuracy One by 3, 4 and 4 percentage points for the average cluster group, the good cluster group and the excellent cluster group respectively. This demonstrates that subclusters with high structural similarity play an important role in improving the local structural accuracy of the one-level improved K-means clustering algorithm. Compared with the multi-level clustering algorithm, the MLSVMTs improve Accuracy One by 6, 3 and 4 percentage points for the average cluster group, the good cluster group and the excellent cluster group respectively.

In Figure 6, average 3D local structure prediction Accuracy Two for four models using the 5 × 2 CV F test is compared. Compared with the improved K-means clustering algorithm, the multi-level clustering algorithm improves Accuracy Two by 2, 5 and 3 percentage points for the average cluster group, the good cluster group and the excellent cluster group respectively. Compared with the multi-level clustering algorithm, MLSVMTs improve Accuracy Two by 6, 2 and 3 percentage points for the average cluster group, the good cluster group and the excellent cluster group respectively.

The combined 5 × 2 CV F test is conducted to verify that the 3D local structure prediction performance improvement of MLSVMTs over other three computational models is statistically significant. The p-value produced by the combined 5 × 2 CV F test indicates the significant level at which the null hypothesis that algorithms have the same error rate can be rejected. A lower p-value implies that a more statistically significant improvement of MLSVMTs over the other three computational models. In this work, the significant level for p-value is set to 1%, which is more rigorous than the 5% commonly chosen by statistician.

Figure 5 Average 3D local structure prediction Accuracy One of four computational models for Combined 5×2 CV F test (see online version for colours)

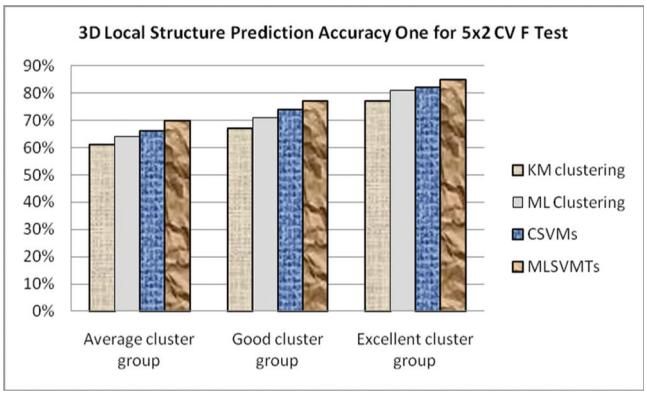
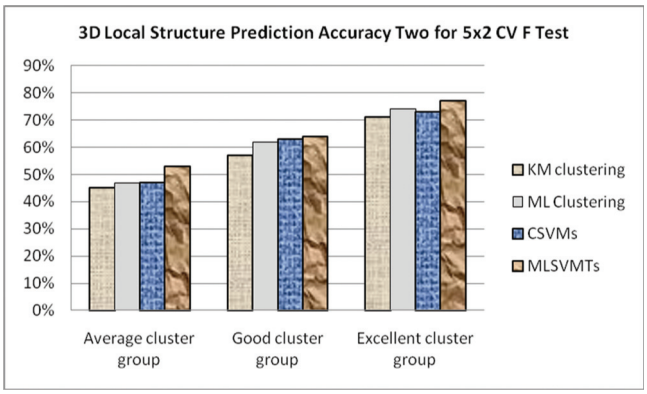


Figure 6 Average 3D local structure prediction Accuracy Two of four computational models for Combined 5×2 CV F test (see online version for colours)



The Table 3 shows 'p value by F test' when the four computational models are compared in terms of 3D local structure prediction Accuracy One during the combined 5×2 CV F test. The Table 4 shows 'p value by F test' when the four computational models are compared in terms of 3D local structure prediction Accuracy Two during the combined 5×2 CV F test. Experimental results from Table 3 and Table 4 show that both 3D local structure prediction Accuracy One and Accuracy Two improvement of MLSVMTs over the other three computational models are statistically significant.

Table 3 'P value by F test' in terms of 3D local structure prediction Accuracy One of four models

Model	*Average cluster group*	*Good cluster group*	*Excellent cluster group*
KM Clustering	<0.1%	<0.1%	<0.1%
ML Clustering	<0.1%	<0.1%	0.5%
$CSVM_S$	<0.1%	0.7%	0.9%

Table 4 'P value by F test' in terms of 3D local structure prediction accuracy two of four models

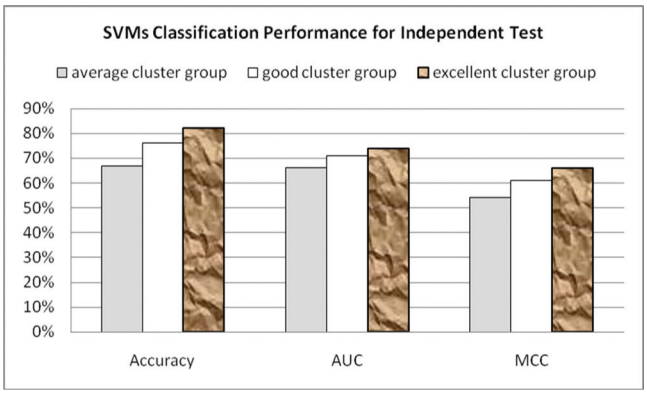
Model	*Average cluster group*	*Good cluster group*	*Excellent cluster group*
KM clustering	<0.1%	<0.1%	<0.1%
ML clustering	<0.1%	0.9%	0.7%
CSVMs	0.2%	0.9%	0.8%

4.3 Experimental results for independent test

4.3.1 SVM classification performance of different cluster groups

Figure 7 shows average accuracy, AUC and MCC of SVMs for different cluster groups in the independent test. Figure 7 demonstrates that SVMs trained for different cluster groups have achieved strong classification performance to recognise the positive sequence segments and negative sequence segments.

Figure 7 SVMs classification performance for different cluster groups in the independent test (see online version for colours)



4.3.2 3D local structure prediction results for independent test

To evaluate the effectiveness of the new computational model rigorously, the dataset for 5×2 CV F test is used as the training set and 300 newly released protein sequences are used as the independent testing set. Figure 8 compares 3D local structure predictions Accuracy One of four models on the independent testing set. Compared with the improved K-means clustering algorithm, the multi-level clustering algorithm improves the Accuracy One by 3, 3 and 4 percentage points for the average cluster group, the good cluster group and the excellent cluster group respectively. Compared with the multi-level clustering algorithm, MLSVMTs improve Accuracy One by 7, 4 and 3 percentage points for the average cluster group, the good cluster group and the excellent cluster group respectively. Figure 9 compares 3D local structure prediction Accuracy Two of four models on the independent testing set. Similar performance improvement for MLSVMTs is observed compared with other models.

Figure 8 3D local structure prediction Acuracy One of four models for independent test (see online version for colours)

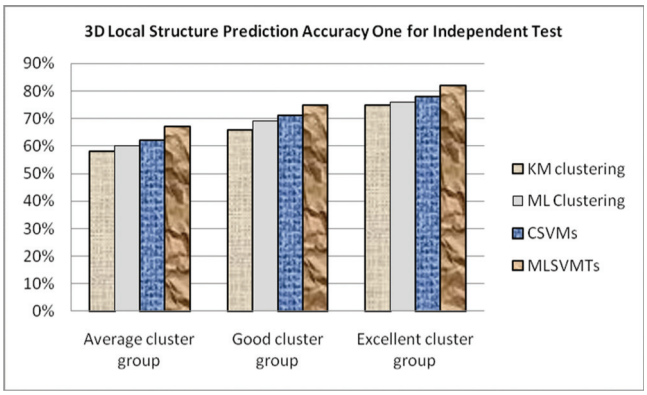
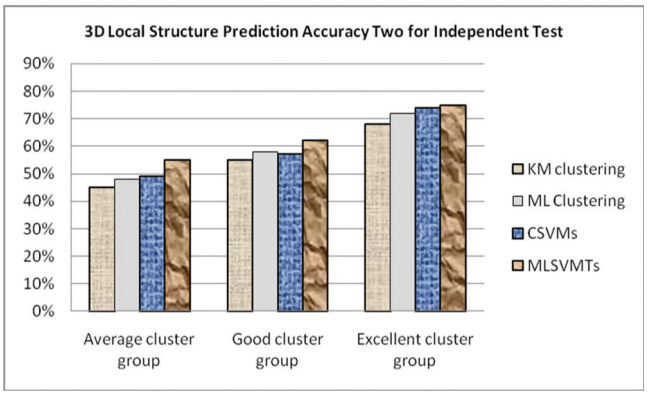


Figure 9 3D local structure prediction Acuracy Two of four models for independent test (see online version for colours)



4.4 Comparing running time for four computational models

The improved K-means clustering algorithm is the most efficient in terms of the running time compared with other three computational models used in this work. Since SVM training and multi-level clustering are the slow computational process for very large datasets, multi-level clustering, CSVMs and MLSVMTs are parallelised to speed up the training process. Figure 10 indicates the average program execution time (in hours) when different numbers of threads are used for 5×2 CV F Test. The experimental results demonstrate that the running time for the multi-level clustering is 60 hours and the running time for MLSVMTs is 79 hours when the 64 threads are used. The experimental results also show that the running time for the multi-level clustering algorithm, CSVMs and MLSVMTs has been reduced substantially when multiple threads are used. Figure 11 compares the average running time of four computational models for 5×2 CV F Test. The running time for multi-level clustering, CSVMs and MLSVMTs is based on results obtained from 64 threads computation. The experimental results show that the running time of MLSVMTs only doubles that of the improved K-means clustering while the significant performance gains have been achieved.

Figure 10 Average running time of four computational models for 5 × 2 CV F Test (see online version for colours)

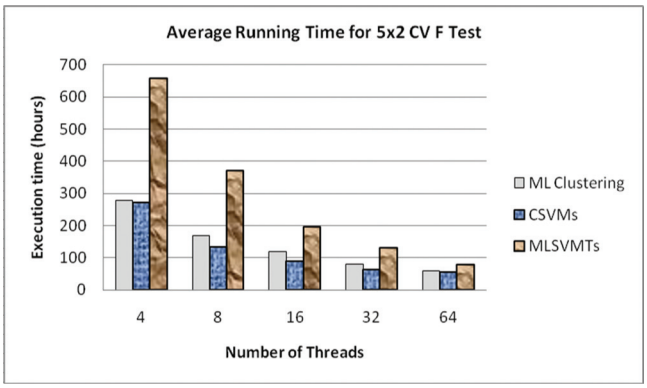
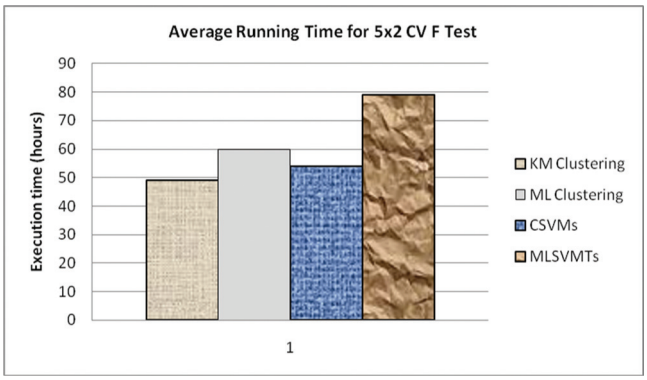


Figure 11 Average running time of four computational models for 5 × 2 CV F Test (see online version for colours)



4.5 Sample cluster subtrees and their biological significance

In this section, three sample cluster subtrees are shown to illustrate the advantage of the multi-level clustering algorithm compared to the one-level clustering algorithm. For comparison purpose, the root clusters of three subtrees are generated using the one-level improved K-means clustering algorithm. The subclusters are produced by the multi-level clustering algorithm.

Figure 12 indicates the coil substree. In Figure 12, the cluster 1 is labeled as C1. In the coil subtree, the structural similarity of subcluster 2 and subcluster 3 generated by the multi-level clustering has improved significantly as compared to the root cluster 1. Figure 13 shows the coil-sheet subtree and Figure 14 shows the helices subtree. Results show that multi-level clustering can discover some subclusters having much higher structural similarity compared to the root cluster.

Results obtained from related biochemical studies show that clusters discovered by the multi-level clustering algorithm may be involved in critical intramolecular and intermolecular interactions, which determine the structure and activities of protein. Furthermore, analysing

sequence profiles of these clusters can provide important insights into structural conservative substitutions of 20 amino acids during the evolutionary process. In order to analyse the sequence profiles and biological properties of these clusters systematically, the following format is used to represent biological and structural characteristic of each cluster.

Figure 12 Subtree 1

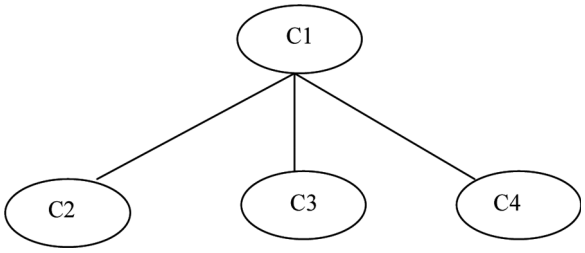


Figure 13 Subtree 2

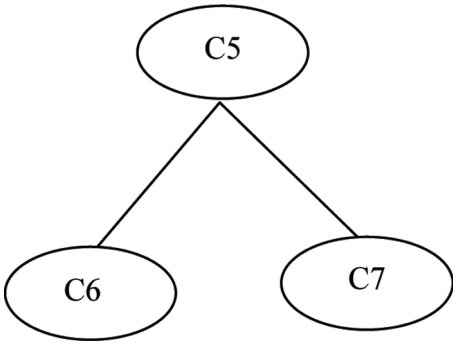
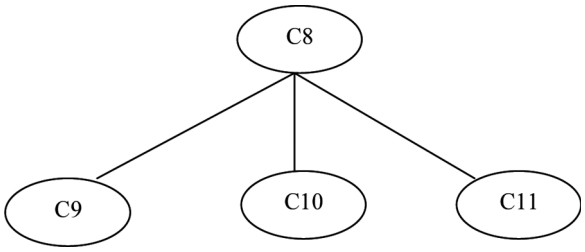


Figure 14 Subtree 3



The number of sequence segments, the *secondary structural similarity*, *average_dmRMSD* and *average_taRMSD* for a given cluster are indicated above the columns of the frequency profile.

- The first column of each frequency profile shows the position of amino acid profiles in each cluster with nine consecutive positions.

- The second column of each frequency profile shows the types of amino acids in the given position. The amino acid appearing with the frequency greater than 0.1 are indicated by the upper case. The amino acid with the upper case emphasises its high occurrence rate in that position. The amino acids appearing with the frequency between 0.08 and 0.1 are indicated by the lower case.
- The third column shows the variability. Variability indicates the number of amino acids occurring with the frequency greater than 0.05.
- The fourth column indicates the hydrophobicity index. The hydrophobicity index is the sum of the frequencies of occurrence of alanine, valine, isoleucine, leucine, methionine, proline, phenylalanine and tryptophan.
- The fifth column indicates the representative secondary structure in that position. In this work, H represents helices; E represents sheets and C represents coils.

AveΦ and Aveφ *are used to represent* torsion angles of each cluster. *The ADM as defined in the equation (3) is used to represent the distance matrix of each cluster. 3D visualisation of several representative clusters is also shown. average_dmRMSD* and *average_taRMSD* can indicate the reliability of these representations. Smaller *average_dmRMSD* and *average_taRMSD* indicate that average torsion angles and the distance matrices are closer to real structures of sequence segments for a sequence cluster.

In this paper, three sample groups of clusters are discussed. The first group of clusters in the subtree 1 is associated with the coil. The second group of clusters in the subtree 2 is associated with the coil-sheet. The third group of clusters in the subtree 3 is associated with helices. Since the pattern of hydrophobicity in the sequence profile of the clusters plays important roles in influencing the structure and activities of proteins, we make extra efforts to analyse the hydrophobic and hydrophilic patterns from these clusters.

Cluster 1 shows the coil with conserved Serine and Threonine and cluster 2 shows the coil with conserved Glycine and Serine. Clusters associated with coils display low hydrophobicity. Coils positioned on the surfaces of proteins often take part in chemical interactions between proteins and other molecules (Berg et al., 2002; Hutchinson and Thornton, 1994; Byrd et al., 1994). Consequently, the hyrdophobicity of coils exposing to the surfaces of proteins are low.

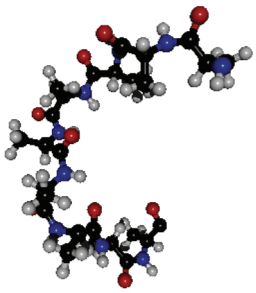
Cluster 5 and cluster 6 show the coil-sheet with clear hydrophobicity transition. Many clusters related to sheets display high levels of hydrophobicity since hydrophobic amino acids are statistically favored for the sheet structure (Hutchinson and Thornton, 1994; Lifson and Sander, 1979). Experimental results show that a simple hydrophobic hexapeptide is used to understand the principles of sheet formation in membranes (Wimley et al., 1998). Other research found out that many amino acids of the adjacent side chains on one side of the sheet are hydrophobic and many amino acids on the alternate side of the sheet are hydrophilic. These arrangements are useful if the sheet is to form the boundary between watery and greasy environments (Zhang et al., 1993).

Cluster 8 and cluster 10 show pronounced amphipathicity since amphipathic helices are one of common structural features found in many proteins and biologically active peptide (Segrest et al., 1990; Pérez-Payá et al., 1995). Amphipathic helices have been found to play important roles in protein folding, protein-membrane interaction and other important protein and peptide biological activity (Pérez-Payá et al., 1995). Studying biological activities of amphipathic helices discovered in this study can help understand the folding, self-association and stability of protein.

Cluster 1
Coil with conserved S and T
Frequency Profile

Number of segments	: 631
Structural Similarity	: 62%
average_dmRMSD	: 2.3
average_taRMSD	: 36.7

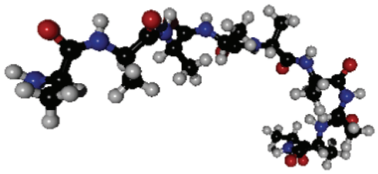
P	Patterns	V	H	S
1	nGSt	7	0.17	C
2	AGST	8	0.21	C
3	anST	9	0.24	C
4	aGST	6	0.22	C
5	FY	4	0.36	C
6	ST	9	0.17	C
7	NDGS	7	0.21	C
8	aGt	11	0.25	C
9	angS	8	0.26	C



Cluster 5
Coil-Sheet
Frequency Profile

Number of segments	: 932
Structural Similarity	: 63%
average_dmRMSD	: 1.8
average_taRMSD	: 49.2

P	Patterns	V	H	S
1	G	1	0.05	C
2	adEk	9	0.27	C
3	ektv	11	0.30	E
4	IlV	3	0.84	E
5	reTv	9	0.44	E
6	ILV	5	0.75	E
7	eSt	8	0.27	E
8	asv	11	0.38	C
9		11	0.33	C



Cluster 8
Amphipathic helices
Frequency Profile

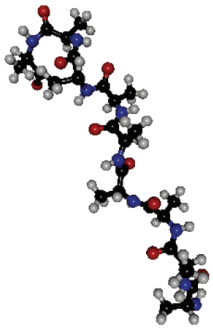
Number of segments	:1413
Structural Similarity	: 66%
average_dmRMSD	: 1.7
average_taRMSD	: 33.2

P	Patterns	V	H	S
1	ak	10	0.31	H
2	dEk	8	0.25	H
3	A	11	0.37	H
4	IlV	3	0.88	H
5	rEK	8	0.28	H
6	adEk	9	0.29	H
7	L	1	0.91	H
8	kv	10	0.33	H
9	adEK	7	0.24	H

Cluster 2
Coil with conserved G and S
Frequency Profile

Number of segments	:108
Structural Similarity	: 72%
average_dmRMSD	: 1.4
average_taRMSD	: 30.6

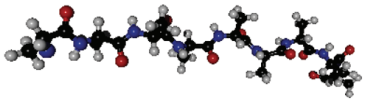
P	Patterns	V	H	S
1	aGS	8	0.30	C
2	AGs	5	0.27	C
3	aGsT	9	0.21	C
4	AqGt	7	0.24	C
5	G	6	0.27	C
6	GsT	9	0.31	C
7	aGs	6	0.23	C
8	anQG	6	0.26	C
9	dqGps	8	0.22	C



Cluster 6
Coil-sheet with clear transition
Frequency Profile

Number of segments	:329
Structural Similarity	: 71%
average_dmRMSD	: 1.5
average_taRMSD	: 31.8

P	Patterns	V	H	S
1	ILkV	6	0.80	E
2	ILV	4	0.85	E
3	EsT	10	0.25	E
4	ILFV	4	0.87	E
5	EsTV	9	0.48	E
6	ILV	6	0.75	E
7	dEs	9	0.22	E
8	nDGs	9	0.23	C
9	DEGs	8	0.20	C



Cluster 3
Coil-sheet with clear transition
Frequency Profile

Number of segments	:313
Structural Similarity	: 68%
average_dmRMSD	: 1.7
average_taRMSD	: 32.1

P	Patterns	V	H	S
1	aE	10	0.24	C
2	egst	12	0.25	C
3	anDEGs	9	0.20	C
4	anDEG	9	0.19	C
5	RgY	10	0.18	C
6	LV	12	0.67	E
7	ILV	5	0.82	E
8	rtv	11	0.41	E
9	ILV	4	0.83	E

Cluster 4
Coil with Low Structural Similarity
Frequency Profile

Number of segments	: 210
Structural Similarity	: 49%
average_dmRMSD	: 3.2
average_taRMSD	: 75.9

P	Patterns	V	H	S
1	gS	3	0.26	C
2	KSt	10	0.27	C
3	ADEpS	8	0.22	C
4	DE	9	0.24	C
5	Aest	9	0.30	C
6	Dps	9	0.22	C
7	L	4	0.89	C
8	AnDek	9	0.25	C
9	adek	10	0.28	H

Cluster 7
Coil-Sheet with low structure similarity

Frequency Profile

Number of segments :603
Structural Similarity : 52%
average_dmRMSD : 2.4
average_taRMSD : 70.1

P	Patterns	V	H	S
1	dGKs	9	0.21	C
2	aDgS	7	0.20	C
3	ILV	5	0.67	E
4	eKT	8	0.23	E
5	ILV	4	0.86	E
6	aEKSt	7	0.25	E
7	NDe	8	0.17	E
8	ILV	3	0.86	E
9	ekST	10	0.25	E

Cluster 9
Amphipathic helices with conserved E and K

Frequency Profile

Number of segments : 755
Structural Similarity : 85%
average_dmRMSD : 0.8
average_taRMSD : 25.2

P	Patterns	V	H	S
1	ILV	6	0.79	H
2	REK	9	0.24	H
3	AdEk	8	0.26	H
4	ILV	6	0.74	H
5	ILV	5	0.78	H
6	arDEK	8	0.19	H
7	ArEK	8	0.29	H
8	ILV	6	0.79	H
9	arLk	9	0.41	H

Cluster 10
Amphipathic helices with conserved E and K

Frequency Profile

Number of segments :423
Structural Similarity :77%
average_dmRMSD_ :1.1
average_taRMSD :28.2

P	Patterns	V	H	S
1	Al	10	0.41	H
2	L	2	0.89	H
3	ArdEK	7	0.25	H
4	AdqEk	8	0.26	H
5	aILV	5	0.70	H
6	arlk	9	0.42	H
7	ADEk	8	0.22	H
8	aElK	9	0.34	H
9	ILV	4	0.79	H

Cluster 11
Helices with low structural similairty

Frequency Profile

Number of segments : 235
Structural Similarity : 58%
average_dmRMSD : 2.8
average_taRMSD : 54.8

P	Patterns	V	H	S
1	ak	10	0,34	H
2	l	12	0.38	H
3	Lv	10	0.41	H
4	el	10	0.38	H
5	ael	8	0.42	H
6	L	1	0.91	H
7	aEl	10	0.40	H
8	E	1	0.08	H
9	Alk	10	0.40	H

4 Conclusion

In this work, MLSVMTs are proposed to predict the secondary structure, backbone torsion angle and backbone distance matrix for local protein structure at the same time. Our local protein structure prediction results can potentially provide more valuable information to derive the global 3D protein structure. Since MLSVMTs take advantages of multiple SVMs in different

levels, MLSVMTs are more effective in capturing complex sequence-to-structure distribution patterns for large protein datasets than the conventional clustering algorithm. Experimental results demonstrate that protein structure prediction performance of MLSVMTs is much superior to that of the one-level K-means clustering algorithm, the multi-level clustering algorithm and CSVMs. Furthermore, MLSVMTs are parallelised to speed up the training process. The experimental analysis indicates that the running time of MLSVMTs is reduced substantially when the parallel algorithm is applied. In this work, the multi-level clustering algorithm reveals larger number of hydrophobicity patterns for helices, sheets and coils than the previous studies. These detailed hydrophobicity patterns are supported by related biochemical studies in the literature. The sequence clusters discovered in this work may provide some additional important information about structurally conservative substitutions during the evolutionary process. For the future work, the advanced algorithm need be developed to derive the complete global 3D structure based on local protein structure prediction results obtained from this work.

Acknowledgements

This research was supported in part by Magellan Scholar Program from University of South Carolina and Science Foundation of Jiangsu Province of China (BK2007105). This research was also supported in part by Healthy Living Initiative Faculty Research Grant from the ReGenesis Community Health Center (RCHC).

References

Alpaydin, E. (1999) 'Combined 5×2 cv F test for comparing supervised classification learning algorithms', *Neural Computation*, Vol. 11, No. 8, pp.1885–1892.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. and Nielsen, H. (2000) 'Assessing the accuracy of prediction algorithms for classification: an overview', *Bioinformatics*, Vol. 16, No. 5, pp.412–424.

Benros, C., de Brevern, A.G., Etchebest, C. and Hazout, S. (2006) 'Assessing a novel approach for predicting local 3D protein structures from sequence', *PROTEINS: Structure, Function and Bioinformatics*, Vol. 62, pp.865–880.

Berg, J.M., Tymoczko, J.L. and Stryer, L. (2002) *Biochemistry*, 5th ed., Freeman, W.H., New York, USA, pp.53–70.

Berman, H.M., Westbrook, J. and Bourne, P.E. (2000) 'The protein data bank', *Nucleic Acids Research,* Vol. 28, pp.235–242.

Byrd, D.A., Sweet, D.J., Panté, N., Konstantinov, K.N., Guan, T., Saphire, A.C., Mitchell, P.J., Cooper, C. S., Aebi, U. and Gerace, L. (1994) 'Tpr, a large coiled coil protein whose amino terminus is involved in activation of oncogenic kinases, is localized to the cytoplasmic surface of the nuclear pore complex', *Journal of Cell Biology*, Vol. 127, pp.1515–26.

Bystroff, C. and Baker, D. (1998) 'Prediction of local structure in proteins using a library of sequence-structure motifs', *J. Mol. Biol*, Vol. 281, pp.565–577.

Bystroff, C., Thorsson, V. and Baker, D. (2000) 'HMMSTR: a hidden markov model for local sequence-structure correlations in proteins', *J. Mol. Biol*, Vol. 301, pp.173–190.

Chen, B. and Johnson, M. (2009) 'Protein local 3D structure prediction by super granule support vector machines (Super GSVM)', *BMC Bioinformatics*, Vol. 10, (Suppl 11), S15.

De Brevern1, A.G., Benros1, C., Gautier, R., Valadié, H., Hazout1, S. and Etchebest, C. (2004) 'Local backbone structure prediction of proteins', *Silico Biol*, Vol. 4, No. 3, pp.381–386.

Etchebest, C., Benros, C., Hazout, S. and de Brevern, A.G. (2005) 'A structural alphabet for local protein structures: improved prediction methods', *PROTEINS: Structure, Function and Bioinformatics*, Vol. 59, pp.810–827.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Publishing Company, Inc., pp.10–92.

Gupta, S.K., Rao, K.S. and Bhatnagar, V. (1999) 'K-means clustering algorithm for categorical attributes', *Data Warehousing and Knowledge Discovery DaWaK-99*, Florence, Italy, pp.203–208.

Han, J.W. and Kamber, M. (2006) *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufmann, San Fransisco, CA 94104, USA.

Han, K.F. and Baker, D. (1995) 'Recurring local sequence motifs in proteins', *J. Mol. Biol.*, Vol. 251, No. 1, pp.176–187.

Han, K.F. and Baker, D. (1996) 'Global properties of the mapping between local amino acid sequence and local structure in proteins', *Proc. Natl. Acad. Sci. USA*, Vol. 93, No. 12, pp.5814–5818.

Hu, H., Pan, Y., Harrsion, R. and Tai, P.C. (2004) 'Improved protein secondary structure prediction using support vector machine with a new encoding scheme and advanced tertiary classifier', *IEEE Transactions on NanoBioscience*, Vol. 2, No. 4, pp.265–271.

Hunter, C.G. and Subramaniam, S. (2003) 'Protein local structure prediction from sequence', *PROTEINS: Structure, Function and Genetics*, Vol. 50, pp.572–579.

Hutchinson, E.G. and Thornton, J.M. (1994) 'A revised set of potentials for ß-turn formation in proteins', *Protein Sci.*, Vol. 3, No. 12, pp.2207–2216.

Joachims, T. (1999) 'Making large-scale support vector machine learning practical', *Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, MA, USA, pp.169–184.

Karp, G. (2002) *Cell and Molecular Biology* (*Concepts and Experiments*), 3rd ed., John Wiley & Sons Inc., New York, USA, pp.52–65.

Khan, L., Awad, M. and Thuraisingham, B. (2007) 'A new intrusion detection system using support vector machines and hierarchical clustering', *The International Journal on Very Large Data Bases*, Vol. 16 , No. 4, pp.507–521.

Kolodny, R. and Linial, N. (2004) 'Approximate protein structural alignment in polynomial time' *Proc Natl. Acad. Sci*., USA, Vol. 101, pp.12201–12206.

Li, X.O., Cervante, J. and Yu, W. (2008) 'Support vector classification for large data sets by reducing training data with change of classes', *Proc. of 2008 IEEE International Conference on Systems, Man and Cybernetics*, Singapore, pp.2609–2614.

Li, X.O., Cervante, J. and Yu,W. (2007) 'Two-stage SVM classification for large data sets via randomly reducing and recovering training data', *Proc. of 2007 IEEE International Conference on Systems, Man and Cybernetics*, Quebec, Canada, pp.3633–3638.

Lifson, S. and Sander, C. (1979) 'Antiparallel and parallel beta-strands differ in amino acid residue preferences', *Nature*, Vol. 282, No. 5734, pp.109–111.

Magdalena, L., Cordon, O., Gomide, F., Herrera, F. and Hoffmann, F. (2004) 'Ten years of genetic fuzzy systems: current framework and new trends', *Fuzzy Sets & Systems*, Vol. 141, No. 1, pp.5–31.

Pérez-Payá, E., Houghten, R.A. and Blondelle, S.E. (1995) 'The role of amphipathicity in the folding, self-association and biological activity of multiple subunit small proteins', *Journal of Biological Chemistry*, Vol. 270, No. 3, pp.1048–56.

Platt, J. (1999) 'Fast training of support vector machines using sequential minimal optimization', *Advances in Kernel Methods: Support Vector Learning*, pp.185–208.

Rahman, A. and Zomaya, A.Y. (2005) 'An overview of protein-folding techniques: issues and perspectives', *International Journal of Bioinformatics Research and Applications*, Vol. 1, pp.121–143.

Sander, C. and Schneider, R. (1991) 'Database of homology-derived protein structures and the structural meaning of sequence alignment', *Proteins: Struct. Funct. Genet.,* Vol. 9, No. 1, pp.56–68.

Segrest, J.P., De Loof, H., Dohlman, J.G., Brouilette, C.G. and Anantharamaiah, G.M. (1990) 'Amphipathic helix motif: classes and properties', *Proteins: Struct. Funct. Genet*., Vol. 8, No. 2, pp.103–117.

Tresp, V. (2000) 'A bayesian committee machine', *Neural Computation*, Vol. 12, No. 11, pp.2719–2741.

Valentini, G. (2005) 'An experimental bias-variance analysis of SVM ensembles based on resampling techniques', *IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics*, Vol. 35, No. 6, pp.1252–1271.

Vapnik, V. (1998) *Statistical Learning Theory*, John Wiley & Sons, Inc., New York.

Wang, G. and Dunbrack, Jr., R. L. (2003) 'PISCES: a protein sequence-culling server', *Bioinformatics*, Vol. 19, No. 12, pp.1589–1591.

Wang, J.T.L. and Wu, X.M. (2006) 'Kernel design for RNA classification using Support Vector Machines', *International Journal of Data Mining and Bioinformatics*, Vol. 1, No. 1, pp.57–76.

Wimley, W.C., Hristova, K., Ladokhin, A. S., Silvestro, L., Axelsen, P.H. and White, S.H. (1998) 'Folding of beta-sheet membrane proteins: a hydrophobic hexapeptide model', *Journal of Cell Biology*, Vol. 277, No. 5, pp.1091–110.

Xia, J., Caragea, D. and Brown, S.J. (2010) 'Prediction of alternatively spliced exons using Support Vector Machines', *International Journal of Data Mining and Bioinformatics*, Vol. 4, No. 4, pp.411–430.

Yang, A.S. and Wang, L.Y. (2003) 'Local structure prediction with local structure-based sequence profiles', *Bioinformatics*, Vol. 19. No. 10, pp.1267–1274.

Zagrovic, B. and Pande, V.S. (2004) 'How does averaging affect protein structure comparison on the ensemble level?', *Biophysical Journal*, Vol. 87, pp.2240–2246.

Zhang, S.G., Holmes, T., Lockshin, C. and Rich, A. (1993) 'Spontaneous assembly of a self-complementary oligopeptide to form a stable macroscopic membrane', *Proc. Natl. Acad. Sci*., USA, Vol. 90, pp.3334–3338.

Zhong, W., Altun, G., Harrison, R., Tai, P.C. and Pan, Y. (2005) 'Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property', *IEEE Transactions on Nanobioscience*, Vol. 4, No. 3, pp.255–265.

Zhong, W., He, J., Harrison, R., Tai, P.C. and Pan, Y. (2007) 'Clustering support vector machines for protein local structure prediction', *Expert Systems With Applications*, Vol. 32, No. 2, pp.518–526.