ISSN No. 0976-5697

Volume 2, No. 6, Nov-Dec 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Protein Local Structure Prediction through Improved Clustering Support Vector Machines (ICSVM)

Sasmita Rout*
Department of Computer Applications
ITER, SOA University
Bhubaneswar, India
rout_mca_sasmita@yahoo.co.in

Tripti Swarnkar
Department of Computer Applications
ITER, SOA University
Bhubaneswar, India
tswarnakar@iter.ac.in

Saswati Mahapatra
Department of Computer Applications
ITER, SOA University
Bhubaneswar, India
saswati@iter.ac.in

Debabrata Senapati
Department of Computer Applications
ITER, SOA University
Bhubaneswar, India
debabratasenapati@gmail.com

Abstract: Accurate protein secondary structure prediction plays an important role in direct tertiary structure modeling, and can also significantly enhance sequence analysis and sequence-structure threading for structure and function determination. Understanding the sequence-to-structure relationship is a central task in bioinformatics. Adequate knowledge about this can improve the accuracy for protein structure prediction. The conventional algorithm such as clustering and SVM can not reveal the complex nonlinear relationship adequately on a huge amount of data individually. So the model CSVMs (Clustering Support Vector Machines) was designed by merging the concept of both clustering and SVM. It has been seen that the generalization power for CSVMs is strong enough to recognize the complicated pattern of sequence-to-structure relationships. CSVMs can only predict the protein local structure with which it is being trained. But if a new type of protein segment comes then it may happen that the CSVMs will fail. That is, none of the cluster will treat it as a positive sample. So in this paper we introduced another robust method called Improved Support Vector Machines (ICSVM) which can handle the unseen example efficiently.

Keywords: Clustering algorithm, SVM, Protein structure prediction, CSVMs, ICSVM.

I. INTRODUCTION

Protein is a chain of simpler molecules called amino acid [1]. Accurate protein secondary structure prediction plays a vital role in direct tertiary structure modeling, and can also significantly enhance sequence analysis and sequence-structure threading for structure and function determination. No computational biology procedures have been able to accurately predict protein tertiary structure directly from the sequence. Secondary structure prediction remains an important step on the way to full tertiary structure prediction. The traditional clustering algorithms like the K-means and K-nearest neighbor algorithm can not reveal the sequence to structure relationship effectively as in these algorithms the distance function is not characterized properly, and also these algorithms cannot handle non linear data [7]. This problem can be solved using Support Vector Machine (SVM) [4] with any of the clustering techniques. Which popularly known as clustering support vector machine (CSVMs) [2]. In this model, one SVM is run for each cluster created by the clustering algorithm. CSVMs are modeled to learn the nonlinear relationship between protein sequences and their structures in each cluster.

However, CSVMs can be easily parallelized to speed up the modeling process. After gaining the knowledge about the sequence to structure relationship, CSVMs are used to predict distance matrices, torsion angles and secondary structures for backbone a-carbon atoms of protein sequence segments. Compared with the K-means clustering algorithm, CSVMs can estimate how close frequency profiles of protein sequences correspond with local 3D structures by using the nonlinear kernel [2]. CSVMs can easily predict the structure of a protein segment by assigning the segment to a cluster but if the segment is not assigned to any of the cluster then it simply cannot predict.

In this paper we have proposed a robust algorithm called Improved Clustering Support Vector Machines (ICSVM) which can handle the unseen protein segment efficiently.

II. PROTEIN STRUCTURE

Proteins are an important class of biological macromolecules present in all organisms. All proteins are polymers of amino acids. Classified by their physical size, proteins are nanoparticles. Each protein polymer – also known as a polypeptide – consists of a sequence of 20 different L-α-amino acids, also referred to as residues [3]. To be able to perform their biological function, proteins fold into one or more specific spatial conformations, driven by a number of non-covalent interactions such as hydrogen bonding, ionic interactions. To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure. A protein may undergo reversible structural changes in performing its biological function. The alternative structures of the same protein are referred to as different conformations, and transitions between them are called conformational changes. Protein amino acids are combined into a single polypeptide chain in a condensation reaction. This reaction is catalysed by the ribosome in a process known as translation [3].

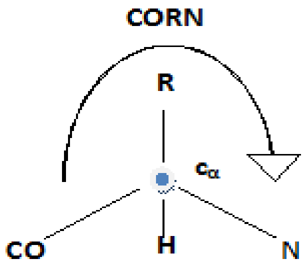

Figure 1.A α-amino acid. The $C_{\alpha}H$ atom is omitted in the diagram

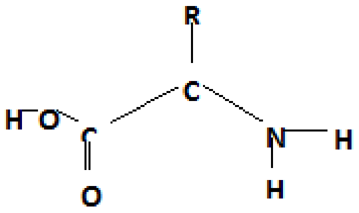

Figure 2. CO-R-N rule

The 20 naturally occurring amino acids have different physical and chemical properties, including their electrostatic charge, pKa, hydrophobicity, size and specific functional groups. These properties play a major role in molding protein structure.

III. APPLICATIONS

The knowledge of the protein structure affords well-founded hypotheses of the function of the protein. If the structure of the relevant binding site of the protein is known, then it can help in disease detection and in designing drug for different disease. Protein structure prediction also helps in gene matching and evolutionary tree generation.

IV. ISSUES

No computational biology procedures have been able to accurately predict protein tertiary structure directly from the sequence. Secondary structure prediction remains an important step on the way to full tertiary structure prediction.

V. METHOD

In this section, the principle of granular computing, SVM, CSVM is introduced. Then how the ICSVM predicts the local structure of protein by the help of granular computing, SVM and CSVMs is explained.

A. *Granular Computing:*

Granular computing decomposes information into number of aggregate clusters and then solves the targeted problem in each cluster or granule. Granular construction and computing are two major part of granular computing; Granular computing visualizes the whole feature space at different granularities. i.e. granular computing uses divide and conquer strategy and breaks up the complex problem into smaller and computationally simpler problems and then it focuses on each small problem by avoiding unnecessary and irrelevant information. In this proposed work, the K-means clustering algorithm is used as the granulation method.

a. *K-Means Clustering:*

K-means clustering is efficient for large data sets with both numeric and categorical attributes. Appling K-means algorithm, similar types of data samples can be grouped together. As a result, the whole sample space is partitioned into subspaces intelligently and the complex data mining task is mapped into a series of computationally feasible simpler and smaller tasks [7].

a) ***Distance score and reliability score of a given sequence segment:*** The frequency profile for a given sequence segment can be compared with the centroid of each cluster in order to calculate the distance score of the given segment for each cluster. A smaller distance score shows that the frequency profile of the given sequence segment is closer to the centroid for the given cluster. The centroid of the related cluster is the average of all frequency profiles of sequence segments for this cluster. The following formula calculates the distance score of a given sequence segment for a specified cluster.

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

Where L is the window size.

F_k(i, j) is the value of frequency profile at row i and column j for the sequence segment k.

F_c(i, j) is the value of the matrix at row i and column j for the centroid of the cluster.

An average frequency profile summarizes how frequently amino acids occur in each position of a cluster. The following formula calculates the frequency of the amino acid of type R at the specified position of the average frequency profile for a sequence cluster:

$$f_R = \frac{\mathrm{Num_R}}{\mathrm{Total\ Number}}$$

Where $\mathrm{Num_R}$ is the number of amino acid of R in the specified position of the sequence cluster and ‘total number’ is the total number of amino acids in the specified position of the sequence cluster.

The reliability score of a given sequence segment for a cluster is determined by the sum of the frequency of the matched amino acid in the corresponding position of the average frequency profile of a cluster. Higher reliability scores indicate that prediction results by this cluster is more dependable since the amino acids of the given sequence segment match more frequently occurring amino acids in the corresponding position of a cluster for structure conservation.

$$Reliability\ Score = \sum_{i=1}^{L} \mathrm{F_c}\ \ (\mathrm{i,j})$$

Where F_c(i, j) is the value of the matrix at row i and column j for the average frequency profile of the cluster. The value of j is determined by the type of amino acid in the specified position of sequence segment.

b) ***Cluster membership assignment for each sequence segment:*** By sliding window method successive residues are generated from a given protein sequences. Then the segments are clustered into groups by K-means algorithm. The distance score of a given sequence segment for each cluster can be calculated in order to filter out some less significant clusters for cluster assignment. Some clusters with top smallest distance scores for the sequence segment can be selected. The rest can be ignored. The sequence segment is can be allocated to the cluster with highest reliability score among these clusters. If there is a tie, the cluster with the smaller distance score wins. The distance score efficiently narrows down the list of possible clusters based on similarity of the frequency profile between the given sequence and the centroid of each cluster. The reliability score assesses how well amino acids of a given sequence segment match frequently occurring amino acids in the important positions of the average frequency profile for each cluster in order to conserve a particular local structure. In CSVMs results, it has been shown that the combination of the distance score and the reliability score can improve efficiency of the clustering membership function noticeably since the distance score and the reliability score carry independent biological information.

B. SVM:

Since samples in the same subspace or in individual cluster are closely related, SVM can be modeled to capture natural data distribution for the samples efficiently.

C. CSVMs:

By using the advantage of granular computing and SVM, the method CSVMs [1] predicts the protein structure. In this method first it granulates the samples into clusters using K-means algorithm, then it runs SVM in each granule. As clustering by K-means algorithm may introduce noisy information into the cluster, so SVM is used to identify the strength of correspondence between frequency profiles and 3D local structure for each sequence segment belonging to the same cluster.

D. The Proposed method ICSVM:

In this section, the principle of granular computing, SVM, CSVM is introduced. Then how the ICSVM predicts the local structure of protein by the help of granular computing, SVM and CSVM is explained. As we have seen that CSVM can work well in predicting the local protein structure but it does not give any idea regarding an unseen protein structure, that means the SVM in each cluster checks for the unseen structure and predicts it as a negative sample. Whereas the advantage of ICSVM over CSVM is that if the unseen sample is not classified as a positive sample by any of the cluster, then ICSVM creates a new cluster having the unseen sample itself. And next time onwards if the same type of protein structure comes then the CSVM can classify it as a positive sample. That means the ICSVM can predict the protein structure as that of CSVMs and also it can train the model each time a new segment comes for classification.

a) Local protein segment prediction by ICSVM: Local protein structure prediction by ICSVM is based on the prediction method from the clustering algorithm. At first, the protein sequence segments whose structures to be predicted are assigned to a specific cluster in the cluster group by the clustering algorithm. Then ICSVM modeled for this specific cluster is used to identify how close the frequency profile of this sequence segment is nonlinearly correlated to the representative structure of this cluster. If the sequence segment is predicted as the positive sample by ICSVM, then the frequency profile of this segment has the potential to be closely mapped to representative structure for this cluster. Consequently, the 3D local structure of this cluster can be safely assigned to this sequence segment. The method to decide the representative structure of each cluster can be found in [5]. If the sequence segment is predicted as the negative sample by ICSVM, the frequency profile of this segment does not closely corresponds to this cluster. The structure cannot be predicted by this cluster. This cluster is removed from the cluster group. The cluster membership function calculating distance scores and reliability scores is used to select the next cluster from the remaining clusters of the cluster group. The aforesaid procedure will be repeated until one SVM modeled for the selected cluster predict the given segment as positive sample. ICSVM can be used to reclassify sequence segments which are misclassified by the clustering algorithms. If any unseen segment comes for classification then it will be predicted as negative sample by all the clusters and then it trains the ICSVM with the unseen sample, so that it can be classified as a positive sample henceforth.

Algorithm for ICSVM

```
Step 1:  Apply K-means algorithm on the sequence feature space
WHILE (the training error > threshold)
{
    Convert sequences into segments (applying sliding window method)
    Find the membership function of each segment.
    Assign the segment to the cluster based on membership function.
    Accordingly update the centroid and the frequency profile for each cluster
}
Step 2:  Training ICSVM for each granule
For each cluster
{
    Label each training sample as positive or negative respectively for different cluster groups
    Modeling each ICSVM for each cluster by optimizing RBF kernel parameters (j, $\gamma$, and c) with the grid search algorithm
}
Step 3: Predicting protein structure by the ICSVM algorithm
While (there are clusters in the cluster group)
{
```

```
Allocate a given sequence segment to a cluster in the cluster group by membership functions
Predicting the property of the given sequence segment by ICSVM modeled for the selected cluster.
If (the given sequence segment is predicted as positive)
{
Assigning the representative structure of the selected cluster to this sequence segment
leave the loop
}
remove the selected cluster from the cluster group
}
Assigning the representative structure of the selected cluster in the first iteration to the sequence segment
Step 4: If the given sequence segment is not assigned to any cluster then train (create a new cluster and make the unseen segment as the centroid) the model with the new segment.
```

VI. SUMMURY

The conventional clustering algorithm is used to divulge the sequence-to-structure relationship. The clustering membership functions may not able to find the nonlinear relationship efficiently. This problem can be nicely handled by the clustering support vector machines but again it cannot handle the unseen segment, the segment will be simply discarded by each SVM in CSVMs. But the ICSVM will create a new cluster with a single element and assign the membership function of the segment as the value of the cluster. In future if a new segment comes with membership value nearer to this then the ICSVM can predict it easily.

VII. REFERENCES

[1]. Karp, G. “Cell and molecular biology (concepts and experiments), 3rd ed., John Wiley & Sons Inc., 2002, pp. 52–65.

[2]. W. Zhong, J. He, R. Harrison, Phang C. Tai, Y. Pan. “Clustering support vector machines for protein local structure prediction,” Expert Systems with Applications, Vol. 32, 2007, pp. 518–526.

[3]. D. Krane, M. l. Raymer. Fundamental Concepts of Bioinformatics, Pearson Education, 2008.

[4]. C. W. Hsu, C. C. Chang, C. J. Lin. “A practical guide to support vector classification”, 2005. Available from http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf.

[5]. W. Zhong, G. Altun, R. Harrison, P. C. Tai, Y. Pan. “Mining relationship between structural homology and frequency profile for structure clusters”. The Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB2005), Boston, (Poster paper).

[6]. Y. Y. Yao, “Granular computing”. In J. Suan, & J. K. Xue (Eds.), Computer science, Proceedings of the 4th Chinese national conference on rough sets and soft computing, Vol. 31, 2004, pp. 1–5.

[7]. Jiawei Han and Micheline Kamber “Data Mining: Concepts and Techniques”, 2nd ed. Morgan Kaufmann Publishers, March 2006.ISBN 1-55860-901-6.