Mining Protein Sequence Motifs Representing Common 3D Structures

Wei Zhong, Gulsah Altum, Robert Harrison, Phang C. Tai, and Yi Pan*
Department of Computer Science and Department of Biology,
Georgia State University, Atlanta GA, 30303, USA
E-mail: pan@cs.gsu.edu

Abstract

Understanding the relationship between protein structure and its sequence is one of the most important tasks of current bioinformatics research. In this work, recurring protein sequence motifs are explored with a K-means clustering algorithm. No structural information is used during the clustering process so that the relationship between sequence similarity and structural similarity for sequence-based clusters can be studied. This work focuses on characterizing structural similarity so that the quality of sequence clusters can be assessed accurately. Analysis of results reveals that the combined metric of distance matrix root mean squared deviation for sequence cluster (dmRMSD_SC) and torsion angle RMSD_SC (taRMSD_SC) can provide the reliable indication of structural similarity for sequence clusters. Based on our combined metric, the recurrent sequence clusters with high structural similarity are used to generate sequence motifs. The common 3D structure of a sequence motif is represented by both representative backbone torsion angles and average distance matrices of the sequence cluster used to produce this motif. These motifs provide the foundation to develop a protein vocabulary reflecting sequence-structure correspondence.

1. Introduction

Bystroff and Baker have used the K-means clustering algorithm to find local sequence motifs for proteins and to assess structural similarity for these motifs[1].

Unlike this previous work, during the process of generating sequence clusters, no structural information is used so that dependence of structural similarity on sequence similarity can be evaluated accurately. This work focus on characterizing the structural similarity in the sequence clusters so that the significance of sequence clustering can be assessed. The metrics to evaluate structural similarity of sequence clusters are studied rigorously and systematically. Average distance matrices and representative torsion angles are a novel representation for representative structure of sequence clusters.

2. Data Set

The dataset used in this work includes 2290 protein sequences obtained from the protein sequence-culling server (PISCES)[5]. No sequences of this database share more than 25% identity. The structures of these protein sequences are available from Protein Data Bank[2].

3. Clustering of Sequence Segments in the Sequence Space

The sliding windows with nine successive residues are generated from protein sequences. Each window represents one sequence segment. The sequence segments with the length of nine are long enough to have some structural features and are short enough to have a statistically significant number of samples. These sequence segments are classified into different clusters with the K-means clustering algorithm. The frequency profiles defined in the similarity-derived secondary structure of proteins (HSSP) [4] are chosen as the numerical representations for sequence segments.

3. Four Metrics to Evaluate Structural Similarity for Sequence Clusters after the Clustering Process

After the K-means clustering algorithm is complete, the following four metrics are used to evaluate the structural similarity of sequence clusters. These four metrics includes secondary structure similarity, dmRMSD_SC, taRMSD_SC, and cRMSD_SC. The

combined metric of dmRMSD_SC and taRMSD_SC provides an important safeguard against the disadvantages of four individual metrics and gives accurate indication of structural similarity from four different perceptive.

4. Results

After the reliable combined metric to evaluate structural similarity is determined, more than 130 local sequence motifs are discovered in this study. Analysis of related biochemical studies published in the literature indicates that the patterns obtained in this work may play vital roles in intramolecular interactions, which decide the structure and function of proteins. Motif 1 indicates the sheet-coil with the clear hydrophobicity transition. Transitional patterns for hydrophobicity found in our sequence motifs are reasonable because hydrophobic amino acids are preferred for sheets and hydrophilic amino acids frequently occur in coils[3].

5. Future Work

In this work, the sliding window size is nine. For our protein vocabulary, the sequence profile of sequence clusters including sequence segments with the length of nine is considered to be the word with the length of nine. In the next step, the size of sliding windows can be changed in order to produce sequence segments with lengths ranging from 5 to 15 residues in order to make further analysis of similarity metrics for sequence clusters.

4. Reference

[1] C. Bystroff, and D. Baker, "Prediction of local structure in proteins using a library of sequence-structure motifs," *J. Mol. Biol*., vol 281. pp. 565-577, 1998.

[2] H. M. Berman, J. Westbrook … and P.E Bourne, "The protein data bank," *Nucleic Acids Res*, vol 28, pp. 235-242, 2000.

[3] E. G. Hutchinson and J. M. Thornton, "A revised set of potentials for ß-turn formation in proteins," *Protein Sci.,* vol. 3, no. 12, pp. 2207-2216, 1994.

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

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

MOTIF 1

Sheet-coil with clear hydrophobicity transition

Frequency Profile

Number of segments : 640				
Structural homology : 73.3%				
cRMSD_SC : 2.48				
dmRMSD_SC : 1.48				
taRMSD_SC : 36.01				
P	Patterns	V	H	S
1	agk	10	0.28	C
2	dGk	9	0.24	C
3	agKP	9	0.27	C
4	RK	8	0.18	C
5	ILV	5	0.74	E
6	AILV	6	0.62	E
7	ILV	4	0.82	E
8	ILV	3	0.86	E
9	AT	5	0.45	E