

Mining Relationship between Structural Homology and Frequency Profile for Structure Clusters

Wei Zhong¹, Gulsah Altun¹, Robert Harrison^{1,2}, Phang C. Tai², Yi Pan¹

Keywords: protein structure cluster, K-means clustering algorithm, frequency profile

1 Introduction.

Understanding the relationship between sequence space information and structural space information is very important for research related to protein folding and protein structure prediction. Many researchers have made efforts to investigate the preference of amino acids for protein substructures [1][2]. However, previous research did not systematically analyze amino acid frequency profiles for structure clusters and did not study the underlying relationship between sequence space information and structural space information.

2. Method.

In our study, the sliding structural segments are classified into different structure clusters with the K-means clustering algorithm. Upon completion of the clustering process, the structural homology of these clusters is evaluated by a comprehensive method for the first time among similar research. The comprehensive measure includes coordinate root mean squared deviation for structure cluster (cRMSD_SC), distance matrix RMSD_SC, torsion angle RMSD_SC.

This is also the first systematic attempt to compare the frequency profiles between clusters with high structural homology and low structural homology. Comparison of frequency profiles establishes an important correlation between sequence space information and structural space information.

3. Result Analysis.

The frequency profiles are produced for recurrent structure clusters after the clustering process is complete. Special attention has been made to the number of prominent positions in the frequency profiles of these structure clusters. Five percent is the threshold frequency for an amino acid to be considered during the process of finding the prominent position. If the sum of the frequency of amino acids having their frequency above the threshold is greater than 60% and their average frequency is greater than 10% for the position in the frequency profile, this position is considered prominent. Prominent positions reveals that an average of five amino acids occupy 60% of the frequency space in that position of frequency profiles. Statistically, each of twenty amino acids may occur with the frequency of 5%. Therefore, five amino acids may occupy 25% of the frequency space. As a result, the distribution of amino acids is highly disproportionate in the prominent positions.

¹ Department of Computer Science, Georgia State University, Atlanta GA, 30303, USA
pan@cs.gsu.edu

² Department of Biology, Georgia State University, Atlanta GA, 30303, USA
biopct@langate.gsu.edu

Figure 2 shows the relationship between dmRMSD_SC and taRMSD_SC. In Figure 2, each point represents one recurrent cluster. Figure 2 is divided into the area A, B, C and D. Figure 1 displays the number of clusters with the specified number of prominent positions in the specified ranges of tertiary structural homology. In Figure 1, A represents the clusters in the area A, which is marked in the Figure 2. B, C and D are similarly defined. The clusters with the highest tertiary structural homology are located in the area A of the Figure 2. As the number of prominent position increases, the number of clusters in D and C shrinks rapidly. In contrast, the number of clusters in A is quite stable. None of clusters in D has four prominent positions. 31 clusters in A have four prominent positions.

Figure 1 reveals that the number of prominent positions greater than four in the frequency profile is a sufficient condition to predict the structural homology for a given cluster. In other words, the clusters with the number of prominent positions greater than 4 in its frequency profiles are likely to enjoy high structural homology. Analysis of frequency profiles of these clusters establishes an important rule to predict whether a given cluster has potential to be structurally homologous.

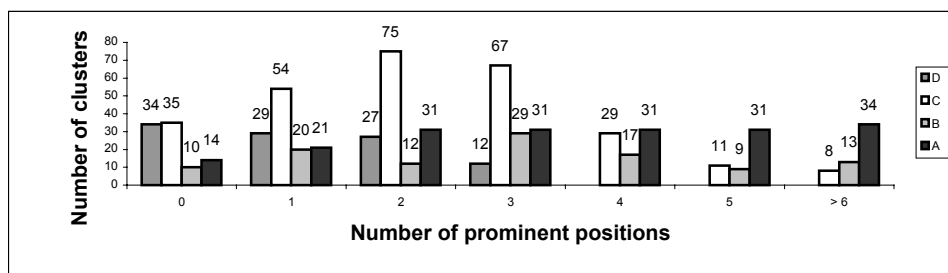


Figure 1: Relationship between the number of prominent positions and tertiary structural homology

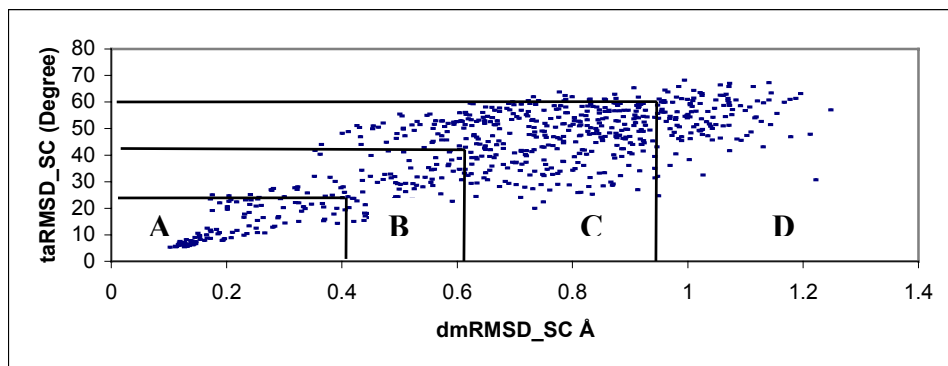


Figure 2: Relationship between taRMSD_SC and dmRMSD_SC

References

- [1] R. Aurora and G. D. Rose, 1998. Helix capping. In *Protein Sci.*, 7, 1:21-38.
- [2] A. J. Doig and R. L. Baldwin, 1995. N- and C-capping preferences for all 20 amino acids in alpha-helical peptides In *Protein Sci.*, 4:1325-1336.