Poster I-30

Automatically Clustering Protein Structure Using the Distribution of Atom Bonds in Backbone

Authors:
Sung Hee Park (*ETRI bioinformatics team*)
Chan Yong Park (*ETRI bioinformatics team*)
Dae Hee Kim (*ETRI bioinformatics team*)
Soo Jun Park (*ETRI bioinformatics team*)

Short Abstract: Recently, a protein structural similarity, 3D atom bond histogram, has proposed by the authors. the 3D atom bond histogram descriptor is enough efficient to retrieve neighbors within a second from huge protein structure database with more than 9700 proteins.In this paper, this descriptor is applied to automatically clustering proteins.

Long Abstract:

Automatically Clustering Protein Structure
Using the Distribution of Atom Bonds in Backbone

Sung Hee Park, Chan Yong Park, Dae Hee Kim, Soo Jun Park

Bioinformatics Team, Electronics and Telecommunications Research Institute, Daejeon, 305-700, Korea E mail: {sunghee, cypark, dhkim98, psj}@etri.re.kr

1 Introduction.

Automatically clustering protein structure is very important in drug design, molecular biology and structural genomics. During recent years, many efforts have been made to analyze the relation between structure and function. Most previous research work have focused on classifying protein families based on homology [1][2]. In this paper, we are involved in the problem of automatic clustering of protein structure.

By recent years, protein clustering has been carried out by protein primary sequence. To cluster proteins effectively, we must take into account structures of proteins. But due to the complex features of proteins, it is not easy to effectively and efficiently figure out their similarities in the aspects of structures and functions. To resolve these difficulties, most research on clustering focused on defining efficient similarity between proteins.

Recently, the authors of this poster have proposed an efficient protein structural similarity, the 3-dimensional (3D) atom bond histogram [3]. In previous research work, it was used to retrieve protein structure. The work showed 3D atom bond histogram descriptor is a very efficient descriptor that can retrieve neighbors within just one second from huge protein structure database with more than 9700 proteins

In this paper, we present a method to automatically cluster proteins using the 3D atom bond histogram.

2 Automatic Protein Clustering by Atom Bond Distribution.

Figure 1 shows our clustering algorithm using 3D atom bond histogram. Our system consists of following three modules: i) preprocessing module, ii) distance matrix computing module, and iii) clustering module.

Figure 1: Processes flow of automatic clustering system

3 Implementation and Result

Our clustering system was implemented in C++ programming language on Windows operating system. In this paper, we use the 3D atom bond histogram as similarity measure. Clustering algorithm used in this paper is K-Means clustering algorithm that has been applied to analyze expression profiles in several biomedical and systems biology studies. To cluster the proteins best we used a cluster validation measure, silhouette method that chooses the optimal number of clusters [4].

4 Discussion.

In this paper, we used the 3D atom bond histogram to represent protein structure. Since we just use a simple histogram to cluster structurally complex proteins, our system is very simple but accurate. We highly believe that if we use other information such as angles and/or types as well as sequence information of secondary structure element, the accuracy of our system would be better.

References

[1] Dorohonceanu, B. and Nevill-Manning, C.: Accelerating protein classification using suffix trees, Proc. of Intelligent Systems for Molecular Biology (2000)

[2] Bailey, T. and Grundy, W.: Classifying proteins by family using the product of correlated p-values, Proc. of ACM RECOMB (1999) 10-14

[3] S.H. Park, S.J. Park, S.H. Park: A Protein Structure Retrieval System Using 3D Edge Histogram, Key Engineering Materials, Vols. 277-279 (2005) 324-330.

[4] P.J. Rousseeuw: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comp App. Math, vol. 20. (1987) 53-65