

Automatic Protein Structure Clustering Using Secondary Structure Elements

Sung Hee Park¹, Chan Yong Park¹, Dae Hee Kim¹, Seon Hee Park¹,
Jeong Seop Sim²

Keywords: automatic clustering, sequence of secondary structure elements, K-means clustering

1 Introduction.

During recent years, many efforts have been made to analyze the relation between structure and function. Most previous research work focused on classifying protein families based on homology [1][2][3]. A major assumption of previous works is that the protein families or functional categories are known in advance and the protein features like sequence or structural features used to make the classification model are labeled with the corresponding families or categories. As a well known technique in statistics and computer science, clustering has been proven very useful in detecting unknown object categories and revealing hidden correlations and pattern among objects. In this paper, we are involved in the problem of automatic clustering of protein structure.

Protein clustering is very important and has applications in such diverse fields as drug design, molecular biology, and environmental industry. By recent years, protein clustering has been carried out by protein primary sequence [4][5][6]. 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 [7][8][9][10][11]. Holm and Sander tried to calculate similarity by alignment of residue-residue ($C\alpha$ - $C\alpha$) distance matrices in [7]. But this approach is computationally very complex and sensitive to errors. Recently, some efficient similarities were proposed. Schwarzer and Lotan proposed a fast similarity that is calculated using segments-segment distance matrices in [8] instead of residue-residue distance matrices in [7], where segment consists of several residues. Another approach, by Singh and Brutlag, was proposed that calculate similarities with the vector representations represent vectors of secondary structures [9]. We found a trend of abstraction of protein structure from previous work.

In this paper, we present a method to automatically cluster proteins using a well known abstract descriptor.

2 Method.

In this paper, we use the sequence of secondary structure elements(for short, SSES) and with generic sequence comparing algorithms 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[12]. And, automatic K-means clustering should provide best cluster partition. To cluster the proteins automatically and best, we optimize the number of clusters using cluster validation measure, silhouette method[13].

3 Clustering Algorithm Using SSES.

¹ Bioinformatics Team, Electronics and Telecommunication Research Institute, Daejeon, Korea. E-mail: {sunghee, cypark, dhkim98, shp}@etri.re.kr

² School of Computer Science and Engineering, Inha University, Seoul, Korea. E-mail: jssim@inha.ac.kr

Now we explain our clustering algorithm using SSES. Our system consists of following three modules: i) preprocessing module, ii) distance matrix computing module, and iii) clustering module. Details are shown below. See Figure 1.

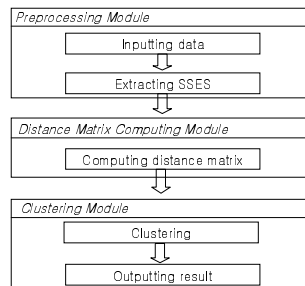


Figure 1: Processes flow of automatic clustering system

4 Discussion.

In this paper, we used SSES to represent protein structure. Since we just use the sequence data (in the aspect of data type, actually, including structural information) to cluster proteins with complex structure, 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 SSES information, the accuracy of our system would be better.

References

- [1] Apostolico, A. and Bejerano, G.: Optimal amnesic probabilistic automata or how learn and classify proteins in linear time and space.: Proc. of ACM RECOMB (2000) 25-32
- [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] Dorohonceanu, B. and Nevill-Manning, C.: Accelerating protein classification using suffix trees. Proc. of Intelligent Systems for Molecular Biology (2000)
- [4] N. Bolshakova, F. Azuaje: Improving expression data mining through cluster validation. Fourth Annual IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine (2003)
- [5] Jiong Yang, Wei Wang: Towards Automatic Clustering of Protein Sequences. CSB 2002 (2002) 175-186
- [6] Dubey, A., S. Hwang, C. Rangel, C. E. Rasmussen, Z. Ghahramani and D. L. Wild: Clustering Protein Sequence and Structure Space with Infinite Gaussian Mixture Models. Pacific Symposium on Biocomputing 2004 (2003)
- [7] L. Holm and C.Sander: Protein Structure Comparison by alignment of distance matrices. Journal of Molecular Biology, Vol. 233 (1993) 123-138
- [8] Rabian Schwarzer and Itay Lotan: Approximation of Protein Structure for Fast Similarity Measures. Proc. 7th Annual International Conference on Research in Computational Molecular Biology(RECOMB) (2003) 267-276
- [9] Amit P. Singh and Douglas L. Brutlag: Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representation. Proc. Intelligent Systems for Molecular Biology (1993)
- [10] 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.
- [11] T. Ohkawa, S. Hirayama, and H. Nakamura: A method of comparing protein structures based on matrix representation of secondary structure pairwise topology. In 4th IEEE Symposium on Intelligence in Neural and Biological Systems (2001) 10-15
- [12] Quackenbush: Computational analysis of microarray data. Nature Reviews Genetics. Vol. 2. (2001) 418-427
- [13] P.J. Rousseeuw: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comp App. Math, vol. 20. (1987) 53-65