Automatic Classification of Protein Structures with Deductive Database System

Yoshihisa Yamamura [1]
yamamura@grt.kyushu-u.ac.jp
Yukiko Tsukamoto [2]
yukiko@mlg.co.jo
Kenji Satou [3]
ken@ims.u-tokyo.ac.jp
Emiko Furuichi [4]
emiko@grt.kyushu-u.ac.jp
Toshihisa Takagi [3]
takagi@ims.u-tokyo.ac.jp
Satoru Kuhara [1]
kuhara@grt.kyushu-u.ac.jp

[1] Graduate School of Genetic Resources Technology, Kyushu University Fukuoka 812, Japan

[2] Life Science Section, Science System Dept., Teijin System Technology Yokohama 231, Japan

[3] Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo 108, Japan

[4] Fukuoka Women's Junior College, Fukuoka 818-01, Japan

1 Introduction

Protein structure classification plays an important role in understanding the relationship between structure and function[1, 2, 3]. With the recent growth in the Protein Data Bank(PDB) there are sufficient examples of many families to allow detail analysis. The resulting data provide important information for protein engineering by revealing tolerances to mutations with a sequence. We have developed PACADE(Protein Atomic Coordinate Analyzer with Deductive Engine)[4, 5] for clustering the structure databank automatically into structural family. With comparison of super-secondary structures described in logical and declarative rules, homologous super-secondary structures could be searched for among protein structures in deductive database system.

2 Method and System

We used the dataset that consisted 186 proteins. To avoid redundant data in PDB, we limited the kinds of protein to 186. This limitation was made based on the following criterion: 1) Eliminate data of non-polypeptide and select data of non-model protein. 2) Choose data with the best resolution in case that there are data on identical proteins, in different resolution. 3) Eliminate incomplete data. 4) Choose data for which the sequence has less than 80% identity to any other sequence. A modified n-stranded meander rule was used for describing the all super-secondary structure topologies. Structural comparisons were performed as a round robin. The ratio of the number of similar super-secondary structures to the number of the all super-secondary structures was used as a degree of structure relatedness. Average linkage clustering method is applied to the structure relatedness matrix for making clusters of similar super-secondary structures.

3 Conclusions

A hundred and eighty-six proteins are clustered mainly in alpha, beta, alpha/beta and alpha+beta structure. Most of the proteins in some cluster have the same or related functions. For example, calcium-binding proteins, sugar-binding proteins, serene proteinases, metallo proteases, immunoglobulin family, aspartic proteinases are grouped in respective clusters.

Acknowledgement

This work was supported by the Grant-in-Aid for scientific on the priority area 'Genome Science for the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] C.A.Orengo, D.T. Jones and J.M. Thornton, "Protein superfamily and domain superfolds," *Nature*, Vol. 375, pp. 631-634, 1994.

[2] L. Holm and C. Sander, "Proetin structure comparison by alignment of distance matrix," *J. Mol. Biol.*, Vol. 233, pp. 123-138, 1993

[3] A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia, "scop: a structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, Vol. 247, pp. 536-540, 1995

[4] K. Satou, E. Furuichi, K. Takiguchi, T.Takagi and S. Kuhara, "A deductive database system PACADE for analyzing 3-D and secondary strucutres of protein," *CABIOS*, Vol. 9, pp. 259-265, 1993

[5] K. Satou, E. Furuichi, S. Hashimoto, Y. Tsukamoto, S. Kuahra, T. Takagi and K. Ushijima, "Development of a deductive database system for computing closures of similarity relationships among protein structures," *Jinkouchinou Gakkaishi*, Vol. 11, pp. 440-450, 1996