A System for Displaying Protein Structure Information Based on Sequence Fragment Similarity

Shouji Tatsumoto **Kenji Satou**
s-tatsu@jaist.ac.jp ken@jaist.ac.jp

School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

1 Introduction

Correlation analysis of protein sequence and structure is increasing its importance as the mainstream of genome analysis shifts from sequencing to functional analysis. In conjunction with such situation, algorithms for structure prediction are actively studied in these years as corroborated by the boom of a series of competitions called CASP. The accuracy of structure prediction is improving by the combination of promising approaches, for example, threading and comprehensive homology search using PSI-BLAST. However, it still remains unsatisfiable level against the huge amount of protein sequences which will be yielded by complete genome sequencing projects for model organisms. To devise an advanced and more accurate algorithm of structure prediction, it might be needed to scrutinize the sequence-structure correlation again from the viewpoint of short fragments. For this purpose, we are trying to develop a system for displaying and analyzing "how much sequence-structure correlation in the level of fragment are buried in the database of determined protein structures (PDB)".

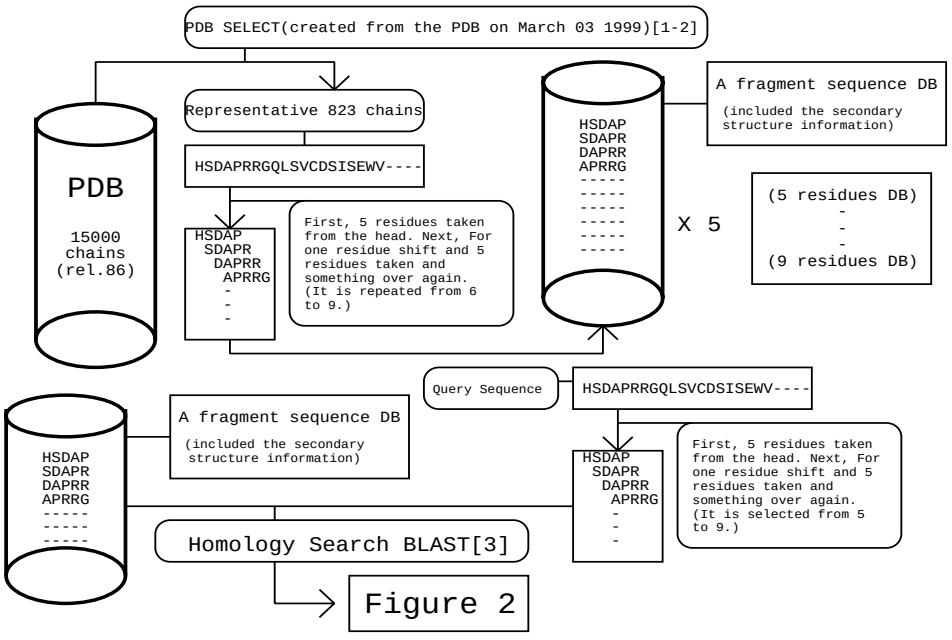

Figure 1: The design concept of system.

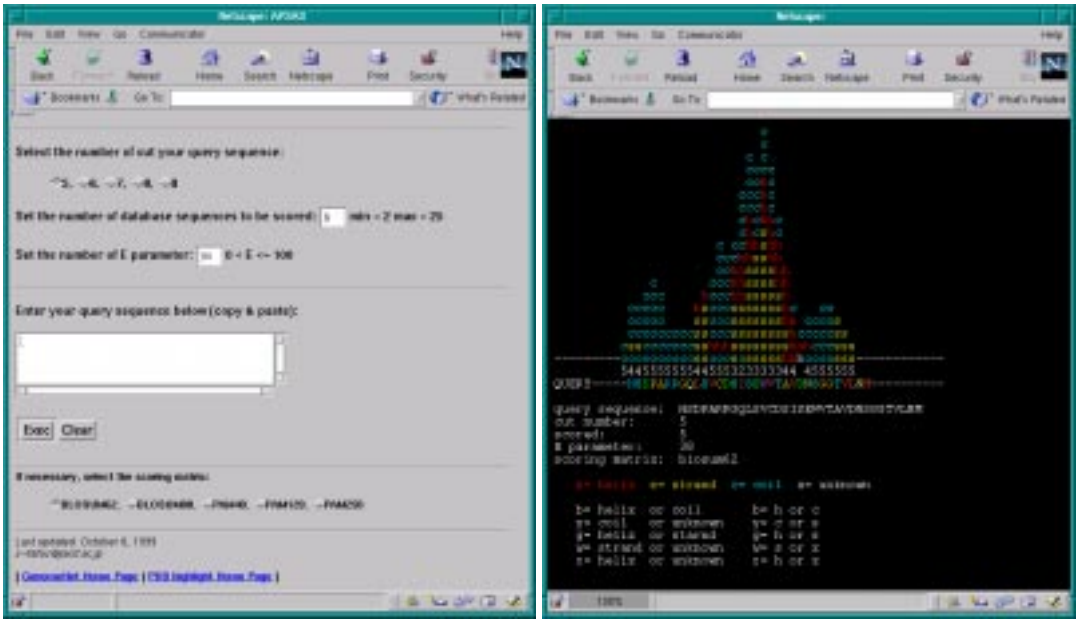

Figure 2: Screen shots of the web interface.

2 Summary and Future work

Fig. 1 shows the configuration of our prototype system. To normalize the biased distribution of PDB entries, PDB SELECT [2, 3] was used. Representative protein chains were decomposed into short fragments with secondary structure information, and stored in a fragment database. Using homology search [1], a query sequence given by a user is compared with the fragment database, and the results are visualized in a web browser (Fig. 2). It shows that secondary structure information drawable from normalized PDB entries is continuously fluctuating. In other words, a user can observe that there exists at least three types of regions, that is, 1) less informative, 2) much informative and uniform (e.g. most of the results suggest that a position forms random coil), 3) much informative but discordant. We believe that this information is useful to grasp the predictability of each positions in a query sequence. As the future work, now we are incorporating more information of protein structure, e.g. dihedral angle, surface or internal, and so on.

Acknowledgments

This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science", from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215:403–410, 1990.

[2] Hobohm, U. and Sander, C., Enlarged representative set of protein structures, *Protein Science*, 3:522-524, 1994.

[3] Hobohm, U., Scharf, M., Schneider, M., and Sander, C., Selection of a representative set of structures from the Brookhaven Protein Data Bank, *Protein Science*, 1:409–417, 1992.