Further Improvements on SMART: Sequence Motif Analysis and Retrieval Tool

A. Ogiwara[1]
ogi@ims.u-tokyo.ac.jp

I. Uchiyama[2]
uchiyama@kuicr.kyoto-u.ac.jp

T. Takagi[1]
takagi@ims.u-tokyo.ac.jp

M. Kanehisa[2]
kanehisa@kuicr.kyoto-u.ac.jp

[1] Human Genome Center,
Institute of Medical Science, The University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

[2] Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611, Japan

Abstract

We present further improvements on the computer system SMART (sequence motif analysis and retrieval tool) that assists biological interpretation of sequence data by searching sequence motifs in a query sequence and annotating functional features associated with the motifs found. The new version of the system fully utilizes the network communication based on a client-server model so that users run only the client program on their workstations without any database resources locally. In the previous version, SMART could treat either PROSITE or MotifDic as a motif dictionary, but in the new release, a new motif dictionary characterizing structural groups derived from PDB is also available. SMART runs on Sun workstations using the XView graphical user interface.

Annotating biological functions of a newly determined sequence is an essential part for molecular biology of today and many computational tools and methods have been developed. Most of them utilize the existing biological knowledge or information stored in the databases. There are two major ways to perform annotation: (1) to find out similar examples in the database and (2) to consult rules or knowledge that are derived from the database. A typical one of the former is the homology search method using the FASTA[5] algorithm or the BLAST[1] algorithm. On the other hand, a motif search method presented here corresponds to the latter approach and PROSITE[2] and our MotifDic[3] are examples of sequence motif libraries.

Previously, we introduced the motif search tool SMART[4] to find out protein sequence motifs described in PROSITE or MotifDic. The program not only searches locations of motifs but also gives additional information related to the found motifs, which may help users to annotate the query sequence by presenting structural and functional examples of the motifs in the actual proteins. Thus, SMART can offer both the direct discovery of biological features and the indirect inference by examples.

We have refined the system to be more friendly for casual users. The SMART system now fully utilizes the network-based distributed processing mechanism for searching and retrieving the databases to allow the client program executable on a machine without any database resources. Currently, the

[1]荻原 淳，高木 利久：東京大学医科学研究所ヒトゲノム解析センター，〒108 東京都港区白金台 4-6-1
[2]内山 郁夫，金久 實：京都大学化学研究所，〒611 京都府宇治市五ヶ庄

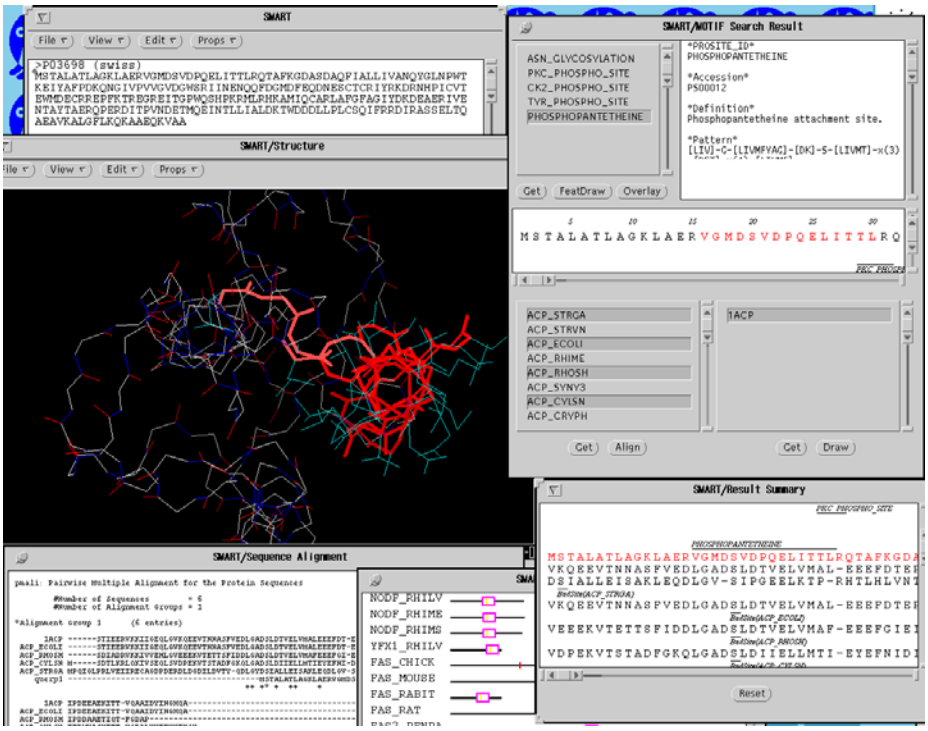

Figure 1: Various windows of the SMART system

growth of databases becomes too rapid to be manageable by individual researchers. Furthermore, the Internet becomes more and more popular. So, it is a good choice to adopt a network mechanism to obtain data from the central places such as the Human Genome Center.

The new version of SMART can also treat a new motif library that characterizes structurally related groups derived from PDB. Recently, structural database entries have also increased remarkably and it is of value to classify groups of structurally resolved proteins based on sequence information and to correlate with sequence databases.

The system runs on Sun workstations with the XView graphical user interface system. The client program may be obtained from the GenomeNet anonymous FTP site (ftp.genome.ad.jp).

This work was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Informatics', from the Ministry of Education, Science 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.*, vol. 215, pp. 403-410, 1990.

[2] Bairoch A. "PROSITE: a dictionary of sites and patterns in proteins" *Nucleic Acids Res.*, vol. 20, pp. 2013-2018, 1992.

[3] Ogiwara A., Uchiyama I., Seto Y. and Kanehisa M. "Construction of a dictionary of sequence motifs that characterize groups of related proteins" *Prot. Engng.*, vol. 5, pp. 479-488, 1992.

[4] Ogiwara A., Uchiyama I. and Kanehisa M. "Sequence Motif Analysis and Retrieval Tool" *Proc. Genome Informatics Workshop IV*, pp. 402-410, 1993.

[5] Pearson W.R. and Lipman D.J. "Improved tools for biological sequence comparison" *Proc. Natl. Acad. Sci. U.S.A.*, vol. 85, pp. 2444-2448, 1988.