Sequence Motif Analysis and Retrieval Tool

† Atsushi Ogiwara, Ikuo Uchiyama, Minoru Kanehisa

ogi@kuicr.kyoto-u.ac.jp uchiyama@kuicr.kyoto-u.ac.jp

kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research
Kyoto University
Uji, Kyoto 611, Japan

Abstract

The Sequence Motif Analysis and Retrieval Tool (SMART) is a computer program package that searches sequence motifs in a query protein sequence and gives additional information of structural and functional features about the detected motifs. These features are represented in a graphical fashion, and for structural features, users can manipulate the three-dimensional structures containing the detected motifs using a graphical user interface. SMART aims at providing more useful information for interpreting biological functions of the query sequence than a simple database retrieval of homology searches. SMART works on Sun workstations with an X windows graphical user interface system.

---

† 荻原 淳, 内山 郁夫, 金久 實
京都大学化学研究所
〒611 京都府宇治市五ケ庄

Introduction

Computational methods are becoming indispensable tools for a researcher who has determined a new nucleotide or protein sequence to obtain whatever information to help classify, identify, and annotate the unknown sequence. Database searching is currently the most popular way to get such information. It is a standard procedure for biologists to search the databases for homologous sequences of their query sequences using the homology search programs like FASTA and BLAST[1,2]. In fact, there are a number of homology search mail servers in the Internet which accept query sequences by electric mails, run homology search programs and return the results by e-mails. However, there are inherent problems in this homology search method. First, what researchers want to know is not only whether someone else has already sequenced identical or similar sequences, but what kind of functions are expected to be possessed by their sequences. They are expecting surprising homologies. It is often the case that the search result contains a long list of sequences, and researchers must distinguish which are biologically meaningful according to their experiences. This is especially troublesome when there are many weak homologies in the so-called twilight zone. Another problem is that the homology search generally requires huge computation, especially when trying to find weak homologies. This is not really because of the efficiency of the algorithm, but because of the nature of the databases which contain a rapidly growing number of ill-organized data.

In our view, although the homology search will continue to be used for finding identical or very closely related sequences, it will be replaced by the motif search for the interpretation of sequence data. We regard sequence motifs as effective indices to classify sequences. Generally, sequence motifs are defined by the biological function such as in the zinc finger nucleotide binding motif. The best protein sequence motif library PROSITE[3] collects those sequence patterns with known biological functions. In such a case, motifs are good indices for biological functions. However, the collection and organization of published motifs require intensive human works. As the recent genome projects are producing huge amount of sequence data, it is urgently required to develop automatic methods to organize biological knowledge like motifs. Thus, we have been taking an alternative approach to automatically extract well-conserved and, hopefully unique, patterns in

groups of evolutionarily related sequences.[4] Because our motifs are computationally derived, they should be considered as candidates of functional sites, although we are trying to correlate with experimental information such as in PROSITE. In any case, our motifs are useful in classifying query sequences into known groups.

Motif Dictionary (MotifDic)

In the previous study, we developed a method to extract sequence motifs that characterize groups of protein sequences using the two criteria of conservation and uniqueness.[4] Namely, sequence patterns comprising motifs must be conserved within the group and they must not be observed outside the group. We then applied this method to the superfamily classification of the PIR1 database. While we are developing a new method to automatically extract sequence motifs from groups of homologous proteins, we also maintain the motif library called MotifDic according to the previous method. Based on the PIR release 36.0, MotifDic currently contains 165 motif entries corresponding to 165 superfamilies. This library is made available with a network mail server described in the next section.

Motif E-Mail Server (MotifFinder)

MotifFinder is a program which searches the motif library of either PROSITE or MotifDic for motif patterns that match the query protein sequence. It also returns additional information about functional and structural features of the detected motifs. Functional features are shown in a schematic diagram visualizing the size and location of motifs along the sequence in relation to biological functional sites annotated in the feature tables of the sequence databases: SWISS-PROT for motifs in PROSITE and PIR1 for motifs in MotifDic. Structural features are shown, if available, in a stereo diagram of three-dimensional protein structures having the motifs. Thus, MotifFinder provides the user with functional and structural meanings of the motif concerned in known sequences. It is up to the user to interpret the meaning of the same motif in the query sequence.

The MotifFinder system may be accessed as an electric mail server. The user can search the motif library and get additional information about

motifs found by sending a query sequence in an e-mail. The address of this mail server is: motif@genome.ad.jp. The format of the query sequence should be in the FASTA format; that is, a query must start with a name/comment line of which the first is letter '>', and the following lines contain sequence data in a free format, but in one letter codes. Before the sequence data, the user may designate which motif library is to be used by the DATALIB statement. In the current version of the MotifFinder program, the default library is PROSITE.

The return mail needs some processing because it contains graphical images of functional and structural information, which are described in the PostScript graphic language. This language is encoded in readable plain texts, which is suitable for treating with any text processing computational tools including the e-mail system. In the current version of MotifFinder the return mail is a shell script which automatically splits itself into necessary PostScript files. They may then be displayed by a PostScript viewer such as PageView in Sun OpenWindows or printed by a PostScript printer. Further information about how to use this mail server can be obtained by sending an e-mail to the same address above with the one-word text of 'help' (but no quotes).

There are several programs which search motifs in PROSITE, but most of them simply search which motifs are found in the query sequence and return their names and locations[5-7]. In contrast, our MotifFinder system returns additional structural and functional information in known sequences containing the detected motifs. If PROSITE is considered a word dictionary, our system returns actual instances of the usage of the word, in addition to giving the nominal meaning of the word.

Sequence Motif Analysis and Retrieval Tool (SMART)

The MotifFinder mail server is a useful system to annotate a protein sequence, but frankly it is not so easy to access and to make use of its results. A mail server is basically a batch processing containing ordered steps of preparing and sending an e-mail and processing the reply. The last step is especially cumbersome in MotifFinder because the reply contains graphical results. Thus, we are in the process of developing another tool for the same purpose of motif search and analysis, but with much easier user interface and more powerful functions. This tool will be most effective in the server-

client mode, where the server in a centralized place maintains the motif library as well as the sequence and structural databases, and performs computationally intensive tasks.

The Sequence Motif Analysis and Retrieval Tool (SMART) consists of four parts: (1) motif search and retrieval, (2) multiple sequence alignment, (3) functional site analysis, and (4) three-dimensional structure analysis. (1), (3) and (4) are basically the same as MotifFinder but the capabilities have been widely expanded (Figure 1).

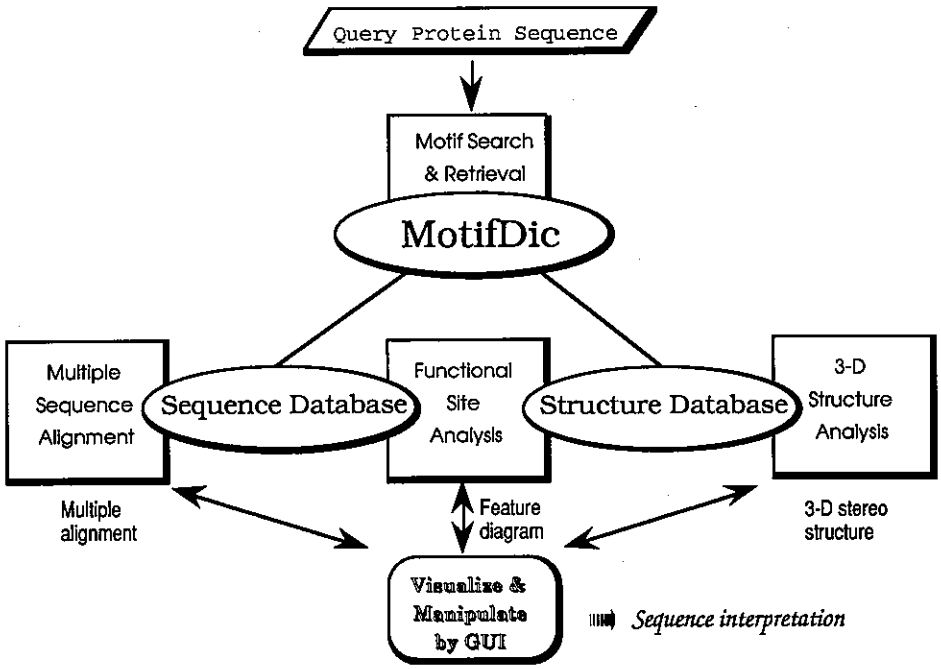


Fig. 1 The Sequence Motif Analysis and Retrieval Tool (SMART)

In the motif search and retrieval tool, a user can select either PROSITE or MotifDic as a sequence motif dictionary to be searched. These two dictionaries have different characteristics according to their construction method. PROSITE is a manual collection of biologically interesting sites and is useful to search known functional sites. MotifDic is automatically derived from the superfamily classification of proteins and is useful to determine the homology group of the query sequence.

The multiple sequence alignment tool provides detailed comparison of the query sequence with related sequences sharing the same motifs. By this

tool, users can compare not only the sequence patterns of motifs but also whole sequences and examine more global similarity. We plan to implement various alignment algorithms including the most sensitive simulated annealing algorithm.[8]

The functional site analysis tool displays the size and location of the motif in relation to various biological functional sites, such as active sites and binding sites, in a similar way as MotifFinder. Users can also link this tool to the multiple alignment tool, which will make it easier to estimate the locations of functional sites on the query sequence.

The three-dimensional structure analysis tool allows visualization and manipulation of protein structures having specific motifs. Because only a limited number of protein structures are available, this tool is not always usable. But if structural data do exist for a given motif, the query protein sequence may be superimposed on the three-dimensional tracing of the known structure. Normally, the three-dimensional tool shows Cα backbone structure or whole backbone atoms using the wire model with motif regions marked. It is also possible to display side-chain atoms. The functional sites obtained in the functional site analysis tool can be visible on this three-dimensional tool as marked regions. Of course users can perform standard operations of rotation, translation, and resize. Printed hardcopy is available in two formats: the original PostScript code or the MOLSCRIPT[9] instructions. We also plan to implement the capability of searching structures locally similar to motif regions by simple partial structure matching.

Discussion

Though interpretation of biological functions is the most important part of sequence and structure analysis, it is not easy to computerize this process. This is because biological knowledge is still too fragmentary and dispersed in different, ill-organized databases. Motif libraries are relatively well-organized collection of biological knowledge, but there is not much that can be done when these libraries are utilized alone. The SMART system tries to integrate knowledge dispersed in various databases and obtained by various analysis tools. From the point of view of a user, SMART is a useful tool to get many information about the query sequence because it displays not only the description of found entries in the motif dictionary but also shows

related sequence, functional and structural information in various visualized fashions. A user can infer the biological meaning of the query sequence through these already annotated instances. From the viewpoint of a constructor of the motif dictionary, SMART is also a powerful workbench to refine the dictionary. The current version of our MotifDic is based on the classification information and is weak at biological semantics when compared with PROSITE. SMART will help to enrich the MotifDic with biological information.

Today there are so many redundant databases available that users are often perplexed to choose adequate ones. We feel it will not be possible, and in fact it is not necessary, to make, say, a single unified protein sequence database, for we think sequence databases will be utilized only for simple retrievals of specific entries or near identical sequences. We are assuming homology searches will no longer be used for sequence interpretation, but then we need to provide alternative methods. If we can incorporate into motif libraries all knowledge that may be obtained by most sensitive homology searches against all sequence databases available, then people will stop using homology searches for interpretation purposes. SMART is a start toward this goal.

SMART depends on the network access method of various databases, which guarantees most up-to-date data maintained at central sites, such as the Supercomputer Laboratory of Kyoto University and the Human Genome Center of the University of Tokyo. Under the Genome Informatics Research Project we maintain the GenomeNet, which in cooperation with other academic networks provides network access to a number of laboratories in Japan. Furthermore, it is difficult to maintain these databases in every researcher's laboratory because they require huge storage space and frequent time-consuming update procedures. CD-ROMs are indeed good media to distribute fixed release databases, but the access speed is limited and the data quickly become stale. Network data sharing is the solution to ever increasing and diverging biological data.

Acknowledgement

This work was supported by the grant-in-aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture, Japan. The computation time was provided by the

Supercomputer Laboratory, the Institute for Chemical Research, Kyoto University.

References

[1] Pearson W.R., Lipman D.J. *Proc. Natl. Acad. Sci. U.S.A.* **85**: 2444-2448 (1988).

[2] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. *J. Mol. Biol.* **215**: 403-410 (1990).

[3] Bairoch A. *Nucleic Acids Res.* **20**: 2013-2018 (1992).

[4] Ogiwara A., Uchiyama I., Seto Y., Kanehisa M. *Protein Eng.* **5**: 479-488 (1992).

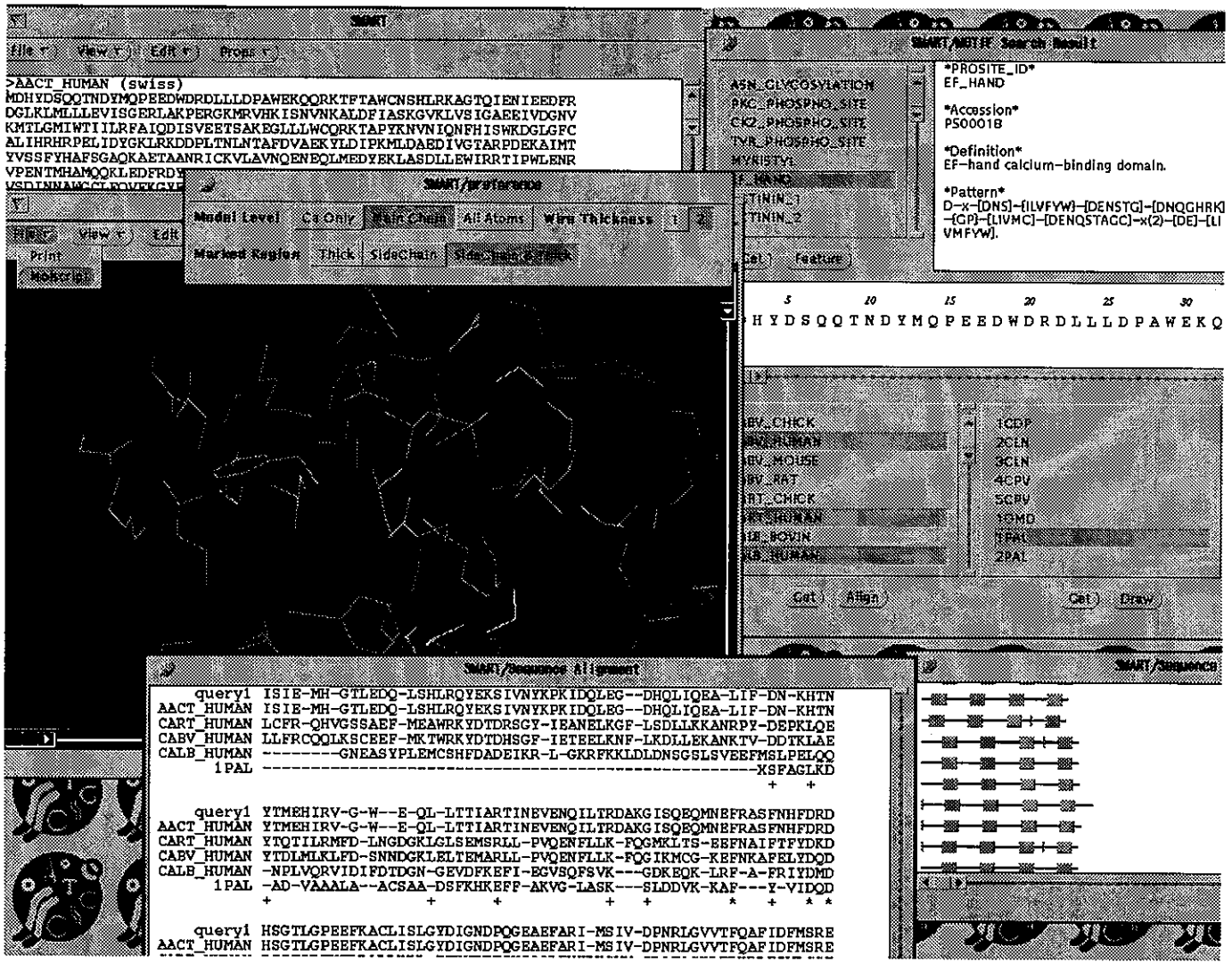
[5] Fuchs R. *Comput. Appl. Biosci.* **7**: 105-106 (1991).

[6] Sternberg M.J. *Comput. Appl. Biosci.* **7**: 257-260 (1991).

[7] Sibbald P.R., Sommerfeldt H., Argos P. *Comput. Appl. Biosci.* **7**: 535-536 (1991).

[8] Ishikawa, M., Toya, T., Hoshida, M., Nitta, K., Ogiwara, A., Kanehisa, M. *Comput. Appl. Biosci.* **9**: 267-273 (1993).

[9] Kraulis P.J. *J. Appl. Cryst.* **24**: 946-950 (1991).


SMART
file
View
Edit
Print
Postscript
>AACT_HUMAN (swiss)
MDHYDSQQTNDYMQPEEDWDRDLLLDPAWEKQQRKTFTAWCNSHLRKAGTQIENIEEDFR
DGLKLMLLLEVISGERLAKPERGKMRVHKISNVNKALDFIASKGVKLVSIGAEEIVDGNV
KMTLGMIWTIILRFAIQDISVEETSAKEGLLLWCQRKTAPYKNVNIQNFHISWKDGLGFC
ALIHRHRPELIDYGKLRKDDPLTNLNTAFDVAEKYLDIPKMLDAEDIVGTARPDEKAIMT
YVSSFYHAFSGAQKAETAANRICKVLAVNQENEQLMEDYEKLASDLLEWIRRTIPWLENR
VPENTMHAMQQKLEDFRDY
SMART/preference
Model Level
Cα Only
Main Chain
All Atoms
Wire Thickness
Marked Region
Thick
SideChain
Print
Redraw
SMART/METIF Search Result
ASN_GLYCOSYLATION
PKC_PHOSPHO_SITE
CK2_PHOSPHO_SITE
TYR_PHOSPHO_SITE
MYRISTYL
EF_HAND
TININ_1
TININ_2
Cat
Feature
PROSITE_ID
EF_HAND
Accession
PS00018
Definition
EF-hand calcium-binding domain.
Pattern
D-x-[DNS]-[ILVFYW]-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW].
5 10 15 20 25 30
HYDSQQTNDYMQPEEDWDRDLLLDPAWEKQ
BV_CHICK
BV_HUMAN
BV_MOUSE
BV_RAT
BAT_CHICK
PVALB_HUMAN
LE_BOVIN
LE_HUMAN
1CDP
2CLN
3CLN
4CPV
5CPV
1PMD
1PAL
2PAL
Cat
Align
Draw
SMART/Sequence Alignment
query1 ISIE-MH-GTLEDQ-LSHLRQYEKSIVNYKPKIDQLEG--DHQLIQEA-LIF-DN-KHTN
AACT_HUMAN ISIE-MH-GTLEDQ-LSHLRQYEKSIVNYKPKIDQLEG--DHQLIQEA-LIF-DN-KHTN
CART_HUMAN LCFR-QHVGSSAEF-MEAWRKYDTDRSGY-IEANELKGF-LSDLLKKANRPY-DEPKLQE
CABV_HUMAN LLFRCQQLKSCEEF-MKTWRKYDTDHSGF-IETEELKNF-LKDLLEKANKTV-DDTKLAE
CALB_HUMAN --------GNEASYPLEMCSHFDADEIKR-L-GKRFKKLDLDNSGSLSVEEFMSLPELQQ
1PAL ----------XSFAGLKD
query1 YTMEHIRV-G-W--E-QL-LTTIARTINEVENQILTRDAKGISQEQMNEFRASFNHFDRD
AACT_HUMAN YTMEHIRV-G-W--E-QL-LTTIARTINEVENQILTRDAKGISQEQMNEFRASFNHFDRD
CART_HUMAN YTQTILRMFD-LNGDGKLGLSEMSRLL-PVQENFLLK-FQGMKLTS-EEFNAIFTFYDKD
CABV_HUMAN YTDLMLKLFD-SNNDGKLELTEMARLL-PVQENFLLK-FQGIKMCG-KEFNKAFELYDQD
CALB_HUMAN -NPLVQRVIDIFDTDGN-GEVDFKEFI-EGVSQFSVK---GDKEQK-LRF-A-FRIYDMD
1PAL -AD-VAAALA--ACSAA-DSFKHKEFF-AKVG-LASK---SLDDVK-KAF---Y-VIDQD
query1 HSGTLGPEEFKACLISLGYDIGNDPQGEAEFARI-MSIV-DPNRLGVVTFQAFIDFMSRE
AACT_HUMAN HSGTLGPEEFKACLISLGYDIGNDPQGEAEFARI-MSIV-DPNRLGVVTFQAFIDFMSRE
SMART/Sequence