Nucleic Acids Research, 2002, Vol. 30, No. 1 235–238

The PROSITE database, its status in 2002

Laurent Falquet, Marco Pagni, Philipp Bucher*, Nicolas Hulo[1], Christian J. A. Sigrist[1], Kay Hofmann[2] and Amos Bairoch[1]

Swiss Institute of Bioinformatics (SIB), Swiss Institute for Experimental Cancer Research (ISREC), CH-1066 Epalinges /Lausanne, Switzerland, [1]Swiss Institute of Bioinformatics (SIB), CMU, University of Geneva, 1 rue Michel Servet, CH-1211 Geneva 4, Switzerland and [2]MEMOREC, Stoffel GmbH, Stoeckheimer Weg 1, D-50829 Koeln, Germany

Received September 24, 2001; Accepted October 2, 2001

ABSTRACT

PROSITE [Bairoch and Bucher (1994) *Nucleic Acids Res.*, 22, 3583–3589; Hofmann *et al.* (1999) *Nucleic Acids Res.*, 27, 215–219] is a method of identifying the functions of uncharacterized proteins translated from genomic or cDNA sequences. The PROSITE database (http://www.expasy.org/prosite/) consists of biologically significant patterns and profiles designed in such a way that with appropriate computational tools it can rapidly and reliably help to determine to which known family of proteins (if any) a new sequence belongs, or which known domain(s) it contains.

BACKGROUND

In some cases the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by pairwise sequence alignment. However, relationships can be revealed by the occurrence in its sequence of a particular cluster of residue types, which is variously known as a pattern, motif, signature or fingerprint. These motifs arise because specific region(s) of a protein which may be important, for example, for their binding properties or for their enzymatic activity are conserved in both structure and sequence. Based on these observations, we decided in 1988, to actively pursue the development of a database of regular expression-like patterns, which would be used to search against sequences of unknown function.

But, while sequence patterns are very useful, there are a number of protein families as well as functional or structural domains that cannot be detected using patterns due to their extreme sequence divergence. Typical examples of important functional domains, which are weakly conserved, are the globins, and the SH2 and SH3 domains. In such domains there are only a few sequence positions which are well conserved.

The use of techniques based on profiles or weight matrices (the two terms are used synonymously here) allows the detection of such proteins or domains (1,2). A profile is a table of position-specific amino acid weights and gap costs. These numbers (also referred to as scores) are used to calculate a similarity score for any alignment between a profile and a sequence, or parts of a profile and a sequence. An alignment with a similarity score higher than or equal to a given cut-off value constitutes a motif occurrence. A distinguishing feature between a pattern and a profile is that the former is usually confined to a small region with high sequence similarity whereas the latter attempts to characterize a protein family or domain over its entire length.

We thus started in 1994 to complement the approach based on patterns by gradually adding profile entries to PROSITE (3,4). The profile structure (5,6) used in PROSITE is an extension of the profiles introduced by Gribskov *et al.* (7) and very similar to the HMM-profiles used in the Pfam database (8). Most profiles in PROSITE were generated from multiple sequence alignments using Gribskov's method (9) with modifications described previously (10,11).

LEADING CONCEPTS

The design of PROSITE follows five leading concepts:

1. Completeness. For such a compilation to be helpful in the determination of protein function, it is important that it contains as many biologically meaningful patterns and profiles as possible.
2. High specificity. In the majority of cases we have chosen patterns or profiles that are specific enough that they do not detect too many unrelated sequences, yet they will detect most, if not all, sequences that clearly belong to the set in consideration.
3. Documentation. Each of the entries in PROSITE is fully documented; the documentation includes a concise description of the protein family or domain that it is designed to detect, as well as a summary of the reasons leading to the development of the pattern or profile.
4. Periodic reviewing. It is important that each entry be periodically reviewed to insure that it is still valid.
5. A very close relationship with the SWISS-PROT protein sequence data bank (12). Updating of PROSITE and of the annotations of the relevant SWISS-PROT entries are very often carried out in parallel. Software tools based on PROSITE are used to automatically update the feature table lines of SWISS-PROT entries relevant to the presence and extent of specific domains.

*To whom correspondence should be addressed. Tel: +41 21 692 5892; Fax: +41 21 692 5945; Email: philipp.bucher@isrec.unil.ch

Table 1. Supplementary material on the web

Examples	
Pattern	http://www.expasy.org/cgi-bin/nicesite.pl?PS00041
Linear profile	http://www.expasy.org/cgi-bin/nicesite.pl?PS01124
Circular profile	http://www.expasy.org/cgi-bin/nicesite.pl?PS50297
Compositional profile (pre-release)	http://www.isrec.isb-sib.ch/cgi-bin/get_doc?format=text&db=prf&entry=PS50099
Documentation	http://www.expasy.org/cgi-bin/prosite-search-ac?PDOC00041
Other	
List of modified documentations	http://www.isrec.isb-sib.ch/prosite/list.html
Graphical representation of a match	http://hits.isb-sib.ch/doc/prf_graphics.shtml

FORMAT AND DOCUMENT FILES

The core of the PROSITE database is composed of two ASCII (text) files. The first file (PROSITE.DAT) is a computer-readable file that contains all the information necessary for programs that make use of PROSITE to scan sequence(s) for the occurrence of the patterns and/or profiles. This file also includes, for each of the entries described, statistics on the number of hits obtained while scanning for that pattern or profile in SWISS-PROT. Cross-references to the corresponding SWISS-PROT and PDB entries are also present in the file. The second file (PROSITE.DOC), which we call the textbook, contains textual information that documents each pattern.

A sample textbook entry is accessible at http://www.expasy.org/cgi-bin/nicesite.pl?PS00041; this particular entry is linked to two entries in the PROSITE.DAT file: a pattern, http://www.expasy.org/cgi-bin/prosite-search-ac?pdoc00040; and a profile, http://www.expasy.org/cgi-bin/nicesite.pl?PS01124.

Several document files are also distributed with the database: PROSUSER.TXT, the database user's manual; PROFILE.TXT, a detailed description of the syntax for the profiles; PROSITE.LIS, a list of PROSITE documentation entries; PROSITE.GET, a document on how to obtain a local copy of PROSITE; PROSITE.PRG, a description of programs and electronic mail servers that make use of PROSITE; PAUTINDX.TXT, an index of authors cited in the PROSITE.DOC file.

CONTENT OF THE CURRENT RELEASE

Release 16.46 of PROSITE (August 30, 2001) contains 1096 documentation entries describing 1483 different patterns, rules and profiles/matrices. In addition to these entries, a collection of 215 preliminary profiles is available in the pre-release distribution from the FTP server of the ISREC group (see below). The list of the documentation entries that have been added since the last release of PROSITE (15.0) is provided in Table 1; furthermore, many entries were updated.

SPECIAL PROFILES AND FUTURE DEVELOPMENT

Profiles can be used to represent a great variety of sequence features besides protein families and domains. We briefly mention a few special cases and how they can be recognized by software for different treatment.

CC /SKIP-FLAG=TRUE: as for patterns this means that the profile is a frequent hit producer. Such profiles may be skipped in order to keep the output short. Current profile belonging to this type: NLS_BP, bipartite nuclear localization signal.

MA TOPOLOGY=CIRCULAR and CC /MATRIX_TYPE=repeat_region: these two lines identify a circular profile which will produce one match for a repeat region consisting of several tandem repeated units. Current profiles belonging to this type: ANK_REP_REGION, ankyrin repeat region; COLLAGEN_REP, collagen repeat (G-x-x); TPR_REGION, TPR repeat region; WD40_REGION, WD40 repeat region; PUM_REPEATS, pumilio RNA-binding domain.

CC /MATRIX_TYPE=composition: this identifies a profile for compositionally biased regions; for example, PRO_RICH, proline-rich region. There is currently one such profile for each amino acid. We plan to add more such profiles in the future, especially profiles to identify subcellular localization signals and simple repeat domains.

OBTAINING A LOCAL COPY OF PROSITE

PROSITE is freely available to academic users. As of Release 16, the documentation entries are copyrighted. To obtain a license, commercial users should contact: The Swiss Institute of Bioinformatics by email: license@isb-sib.ch. Or its commercial representative: Geneva Bioinformatics (GeneBio) S.A., Case Postale 210, CH-1211 Geneva 12, Switzerland. Tel: +41 22 702 99 00; Fax: +41 22 702 99 99; Email: info@genebio.com.

PROSITE is distributed on CD-ROM and email server (4,13,14), or can be directly downloaded via anonymous FTP from the following FTP sites: ExPASy, ftp://ftp.expasy.org/databases/prosite/; EBI, ftp://ftp.ebi.ac.uk/pub/databases/prosite/.

A complete profile collection including pre-release profiles and corresponding preliminary documentation can be downloaded from: ISREC, ftp://ftp.isrec.isb-sib.ch/sib-isrec/profiles/. This site also offers so-called frame search versions of the profiles which can be used for searching DNA sequences for matches in open reading frames using an error-tolerant method (described in the manual pages of the *pftools* package discussed below).

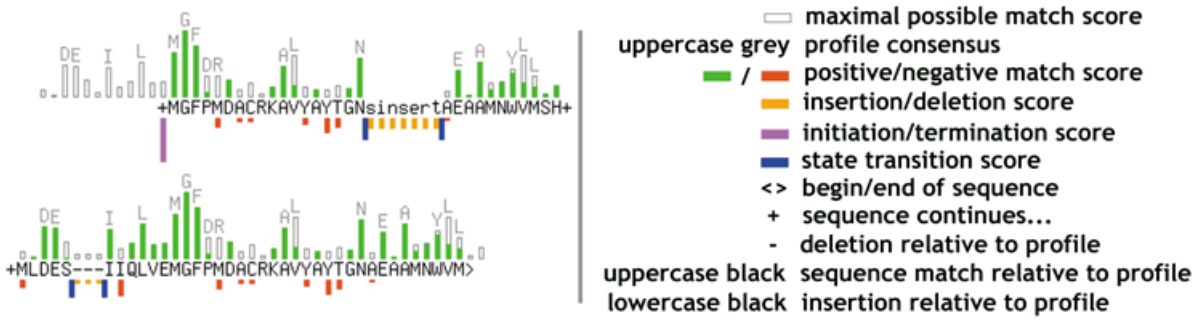

Figure 1. Graphical representation of a match.

MAKING USE OF PROSITE

Computer programs

Many academic groups and commercial companies have developed computer programs that make use of the pattern entries in PROSITE. The PROSITE.PRG file contains a full list of these programs, their operating system specificity, characteristics as well as information on how to obtain them.

Two software packages are distributed to make use of profile entries:

1. *pftools* (version 2.2 in FORTRAN77) written by Philipp Bucher. These tools are available by anonymous FTP from the server: ftp://ftp.isrec.isb-sib.ch/sib-isrec/pftools. Several versions are available, as well as executables compiled for many Unix platforms (including a recent MacOSX version) and for Windows 95, 98 or 2000.
2. *PrfLib* (version 1.0 in ANSI C) written by Nicolas Moeri. These tools are available from the server: ftp://ftp.isrec.isb-sib.ch/sib-isrec/PrfLib/.

A specialized hardware called GeneMatcher from Paracel Inc. allows very fast searches over entire databases using profiles or HMMs. The software distributed with this hardware supports PROSITE profiles although with some limitations: http://www.paracel.com.

One software package is distributed that makes use of pattern entries: *DeepView* [Swiss-PDB viewer (15)] allows you to search for and to highlight the different patterns in the structure http://www.expasy.org/spdbv/.

Interactive access to PROSITE using the World Wide Web

The most efficient and user-friendly way to browse interactively in PROSITE as well as to analyze a sequence for the occurrence of a pattern or a profile is to use the World Wide Web molecular biology server ExPASy (16).

You can directly access the 'top' page of the section of ExPASy that allows you to browse through the PROSITE documentation and data entries by going to http://www.expasy.org/prosite/.

To use the PROSITE patterns and profiles, you can make use of the following software tools:

ScanProsite. ScanProsite allows to either scan a protein sequence—from SWISS-PROT or provided by the user—for the occurrence of patterns stored in PROSITE or to scan the SWISS-PROT and/or TrEMBL database, including weekly releases, for the occurrence of a pattern that can originate from PROSITE or be provided by the user. The URL for *ScanProsite* is http://www.expasy.org/tools/scnpsite.html.

ProfileScan. ProfileScan allows to scan a protein sequence—from SWISS-PROT or provided by the user—for the occurrence of profiles stored in PROSITE and in the pre-release collection. Our new PFSCAN server page generates a graphical representation of a match (Fig. 1). We developed this tool to provide users with a much more informative display than a simple pairwise alignment between the query sequence and the consensus of the profile. It can be immediately visible that the scoring system is position-specific and that some regions of a sequence are scored differently and thus are responsible for a good or a bad overall score. The new URL for *ProfileScan* is http://hits.isb-sib.ch/cgi-bin/PFSCAN.

FrameProfileScan. FrameProfileScan allows one to scan a DNA sequence (translated on the fly into protein)—from EMBL or provided by the user—for the occurrence of profiles stored in PROSITE. The URL for *FrameProfileScan* is http://www.isrec.isb-sib.ch/software/PFRAMESCAN_form.html.

InterProScan. InterProScan allows one to scan a protein sequence provided by the user for the occurrence of profiles stored in InterPro resource of protein families, domains and sites (17). The URL for *InterProScan* is http://www.ebi.ac.uk/interpro/scan.html.

REFERENCES

1. Doolittle,R.F. (1986) *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, CA, pp. 3–36.
2. Lesk,A.M. (1988) PartII Sources of information. The NBRF protein sequence database. In Lesk,A.M. (ed.), *Computational Molecular Biology*, Oxford University Press, Oxford, UK, pp. 17–26.
3. Bairoch,A. and Bucher,P. (1994) PROSITE: recent developments. *Nucleic Acids Res.*, **22**, 3583–3589.
4. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
5. Bucher,P. and Bairoch,A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, pp. 53–61.
6. Bucher,P., Karplus,K., Moeri,N. and Hofmann,K. (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem.*, **20**, 3–23.

7. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
8. Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
9. Gribskov,M., Luethy,R. and Eisenberg,D. (1990) Profile analysis. *Methods Enzymol.*, **183**, 146–159.
10. Luethy,R., Xenarios,I. and Bucher,P. (1994) Improving the sensitivity of the sequence profile method. *Protein Sci.*, **3**, 139–146.
11. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.*, **10**, 19–29.
12. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
13. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Lombard,V., Lopez,R., Parkinson,H., Redaschi,N., Sterk,P., Stoehr,P. and Tuli,M.A. (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **29**, 17–21. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 21–26.
14. Henikoff,S. (1993) Sequence analysis by electronic mail server. *Trends Biochem. Sci.*, **18**, 267–268.
15. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
16. Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Biochem. Sci.*, **19**, 258–260.
17. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.