Nucleic Acids Research, 1999, Vol. 27, No. 1 215–219

The PROSITE database, its status in 1999

Kay Hofmann, Philipp Bucher[1,*], Laurent Falquet[1] and Amos Bairoch[2]

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

Received October 16, 1998; Accepted October 21, 1998

ABSTRACT

The PROSITE database (http://www.expasy.ch/sprot/prosite.html) consists of biologically significant patterns and profiles formulated in such a way that with appropriate computational tools it can help to determine to which known family of protein (if any) a new sequence belongs, or which known domain(s) it contains.

BACKGROUND

PROSITE (1,2) is a method of identifying what is the function of uncharacterized proteins translated from genomic or cDNA sequences. It consists of a database of biologically significant patterns and profiles formulated in such a way that with appropriate computational tools it can rapidly and reliably determine to which known family of protein (if any) the new sequence belongs, or which known domain(s) it contains.

In some cases the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall 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. These structural requirements impose very tight constraints on the evolution of this small but important portion(s) of a protein sequence. The use of protein sequence patterns or profiles to determine the function of proteins is becoming very rapidly one of the essential tools of sequence analysis. Many authors (3,4) have recognized this reality. 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, the immunoglobulin, and the SH2 and SH3 domains. In such domains there are only a few sequence positions which are well conserved. Any attempt to build a consensus pattern for such regions will either fail to pick up a significant proportion of the protein sequences that contain such a region (false negatives) or will pick up too many proteins that do not contain the region (false positives).

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. 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. As with patterns, there may be several matches to a profile in one sequence, but multiple occurrences in the same sequences must be disjoint (non-overlapping) according to a specific definition included in the profile. Another feature that distinguishes patterns from profiles is that the latter are usually not confined to small regions with high sequence similarity. Rather they attempt to characterize a protein family or domain over its entire length.

We therefore started in 1994 to complement the approach based on patterns by gradually adding to PROSITE profile entries. The profile structure (5,6) used in PROSITE is similar to but slightly more general than the one introduced by Gribskov and co-workers (7); additional parameters allow representation of other motif descriptors, including the currently popular hidden Markov models (8). Profiles can be constructed by a large variety of different techniques. The classical method developed by Gribskov and co-workers (9) requires a multiple sequence alignment as input and uses a symbol comparison table to convert residue frequency distributions into weights. Most profiles included in PROSITE are generated by this procedure applying recently described modifications (10,11). In some cases we also applied alternative profile construction methods including structure-based approaches and methods involving hidden Markov modelling.

LEADING CONCEPTS

The design of PROSITE follows five leading concepts.

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.

*To whom correspondence should be addressed. Tel: +41 21 692 5892; Fax: +41 21 652 6933; Email: philipp.bucher@isrec.unil.ch

1a) A documentation (textbook) entry from the PROSITE.DOC file

```
{PDOC00040}
{PS00041; HTH_ARAC_FAMILY_1}
{PS01124; HTH_ARAC_FAMILY_2}
{BEGIN}
**********************************************************************
* Bacterial regulatory proteins, araC family signature and profile *
**********************************************************************

The many bacterial transcription regulation proteins which bind DNA  through a
'helix-turn-helix' motif can  be  classified into subfamilies  on the basis of
sequence similarities. One  of these subfamilies groups together the following
proteins [1,2]:

 - aarP,  a  transcriptional  activator  of the 2'-N-acetyltransferase gene in
   Providencia stuartii.
 - ada,  an  Escherichia  coli and Salmonella typhimurium bifunctional protein
   that repairs  alkylated  guanine  in DNA by transferring the alkyl group at
   the O(6)  position  to  a  cysteine  residue  in the enzyme. The methylated
   protein acts  a  positive  regulator  of its own synthesis and of the alkA,
   alkB and aidB genes.
 - adaA, a Bacillus subtilis  bifunctional  protein  that  acts  both  as  a
   transcriptional activator of the ada operon and as a methylphosphotriester-
   DNA alkyltransferase.
 - adiY, an Escherichia coli protein of unknown function.
 - aggR,  the  transcriptional  activator  of  aggregative adherence fimbria I
   expression in enteroaggregative Escherichia coli.
 - appY, a protein  which  acts  as  a  transcriptional  activator  of  acid
   phosphatase and other proteins during  the deceleration phase of growth and
   acts as a repressor for other proteins that  are synthesized in exponential
   growth or in the stationary phase.
 - araC,  the  arabinose operon  regulatory  protein,  which  activates  the
   transcription of the araBAD genes.
 - cafR, the Yersinia pestis F1 operon positive regulatory protein.
 - celD, the Escherichia coli cel operon repressor.
 - cfaD, a protein which is required for  the  expression of the CFA/I adhesin
   of enterotoxigenic Escherichia coli.
 - csvR,  a  transcriptional  activator  of  fimbrial genes in enterotoxigenic
   Escherichia coli.
 - envY, the porin thermoregulatory  protein, which is involved in the control
   of  the  temperature-dependent  expression  of  several  Escherichia  coli
   envelope proteins such as ompF, ompC, and lamB.
 - exsA, an activator of exoenzyme S synthesis in Pseudomonas aeruginosa.
 - fapR, the positive activator for the expression of  the  987P operon coding
   for the fimbrial protein in enterotoxigenic Escherichia coli.
 - hrpB, a  positive  regulator  of  pathogenicity  genes  in  Burkholderia
   solanacearum.
 - invF, the Salmonella typhimurium invasion operon regulator.
 - marA,  which  may  be  a transcriptional activator of genes involved in the
   multiple antibiotic resistance (mar) phenotype.
 - melR, the  melibiose  operon  regulatory  protein,  which  activates  the
   transcription of the melAB genes.
 - mixE, a Shigella flexneri protein necessary for secretion of ipa invasins.
 - mmsR, the transcriptional  activator  for  the mmsAB operon  in Pseudomonas
   aeruginosa.
 - msmR, the multiple  sugar  metabolism  operon  transcriptional activator in
   Streptococcus mutans.
 - pchR, a Pseudomonas aeruginosa  activator for  pyochelin and ferripyochelin
   receptor.
 - perA,  a  transcriptional  activator  of  the  eaeA  gene  for  intimin  in
   enteropathogenic Escherichia coli.
 - pocR, a  Salmonella  typhimurium  regulator  of the  cobalamin biosynthesis
   operon.
 - pqrA, from Proteus vulgaris.
 - rafR, the regulator of the raffinose operon in Pediococcus pentosaceus.
 - ramA, from Klebsiella pneumoniae.
 - rhaR,  the  Escherichia  coli and  Salmonella typhimurium L-rhamnose operon
   transcriptional activator.
 - rhaS, an Escherichia coli and Salmonella typhimurium positive activator of
   genes required for rhamnose utilization.
 - rns, a protein which is required for  the  expression  of  the  cs1 and cs2
   adhesins of enterotoxigenic Escherichia coli.
 - rob,  a protein which binds to the right arm of the replication origin oriC
   of the Escherichia coli chromosome.
 - soxS, a protein that, with the soxR protein, controls a superoxide response
   regulon in Escherichia coli.
 - tetD, a protein from transposon TN10.
 - tcpN or toxT, the Vibrio cholerae  transcriptional  activator  of  the  tcp
   operon involved in pilus biosynthesis and transport.
 - thcR,  a  probable  regulator  of the thc operon for the degradation of the
   thiocarbamate herbicide EPTC in Rhodococcus sp. strain NI86/21.
 - ureR, the transcriptional activator of the plasmid-encoded urease operon in
   Enterobacteriaceae.
 - virF and lcrF, the Yersinia virulence regulon transcriptional activator.
 - virF,  the  Shigella  transcriptional  factor  of  invasion related antigens
   ipaBCD.
 - xylR, the Escherichia coli xylose operon regulator.
 - xylS,  the  transcriptional activator of the Pseudomonas putida TOL plasmid
   (pWWO, pWW53 and pDK1) meta operon (xylDLEGF genes).
 - yfeG, an Escherichia coli hypothetical protein.
 - yhiW, an Escherichia coli hypothetical protein.
 - yhiX, an Escherichia coli hypothetical protein.
 - yidL, an Escherichia coli hypothetical protein.
 - yijO, an Escherichia coli hypothetical protein.
 - yuxC, a Bacillus subtilis hypothetical protein.
 - yzbC, a Bacillus subtilis hypothetical protein.

Except for celD, all of these  proteins seem  to  be  positive transcriptional
factors. Their size range from 107 (soxS) to 529 (yzbC) residues.

The helix-turn-helix  motif  is  located  in  the third quarter of most of the
sequences; the  N-terminal and  central regions of these proteins are presumed
to interact  with  effector molecules and may be involved in dimerization [3].
The minimal DNA binding domain, which spans roughly 100 residues and comprises
```

Figure 1. Sample data from PROSITE.

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.

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.

Periodic reviewing. It is important that each entry be periodically reviewed to ensure that it is still valid.

A very tight 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 done 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.

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 entry 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 entries are also present in the file. The second file (PROSITE.DOC), which we call the

```
the HTH motif contains another region with similarity to classical HTH domain.
However, it contains an insertion of one residue in the turn-region.

A signature pattern was derived from the region that follows the first HTH
domain and that includes the totality of the putative second HTH domain. A
more sensitive detection of members of the araC family is available through
the use of a profile which spans the minimal DNA-binding region of 100
residues.

-Consensus pattern: [KRQ]-[LIVMA]-x(2)-[GSTALIV]-{FYWPGDN}-x(2)-[LIVMSA]-
                    x(4,9)-[LIVMF]-x(2)-[LIVMSTA]-[GSTACIL]-x(3)-[GANQRF]-
                    [LIVMFY]-x(4,5)-[LFY]-x(3)-[FYIVA]-{FYWHCM}-x(3)-
                    [GSADENQKR]-x-[NSTAPKL]-[PARL]
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in SWISS-PROT: 34.

-Sequences known to belong to this class detected by the profile: ALL.
-Other sequence(s) detected in SWISS-PROT: NONE.

-Note: this documentation entry is linked to both a signature pattern and a
 profile. As the profile is much more sensitive than the pattern, you should
 use it if you have access to the necessary software tools to do so.

-Expert(s) to contact by email:
        Ramos J.L.: jlramos@samba.cnb.uam.es
        Gallegos M.-T.: mtrini@samba.cnb.uam.es

-Last update: November 1997 / Text revised.

[ 1] Gallegos M.-T., Michan C., Ramos J.L.
     Nucleic Acids Res. 21:807-810(1993).
[ 2] Henikoff S., Wallace J.C., Brown J.P.
     Meth. Enzymol. 183:111-132(1990).
[ 3] Bustos S.A., Schleif R.F.
     Proc. Natl. Acad. Sci. U.S.A. 90:5638-5642(1993).

+----------------------------------------------------------------------------+
| This PROSITE entry is copyright by the Swiss Institute of Bioinformatics |
| (SIB). There are no restrictions on its use by non-profit institutions as |
| long as its content is in no way modified and this statement is not |
| removed. Usage by and for commercial entities requires a license agreement |
| (See http://www.isb-sib.ch/announce/ or email to license@isb-sib.ch). |
+----------------------------------------------------------------------------+
{END}
```

1b) The corresponding pattern and profile entries in the PROSITE.DAT file

```
ID   HTH_ARAC_FAMILY_1; PATTERN.
AC   PS00041;
DT   APR-1990 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE   Bacterial regulatory proteins, araC family signature.
PA   [KRQ]-[LIVMA]-x(2)-[GSTALIV]-{FYWPGDN}-x(2)-[LIVMSA]-x(4,9)-[LIVMF]-
PA   x(2)-[LIVMSTA]-[GSTACIL]-x(3)-[GANQRF]-[LIVMFY]-x(4,5)-[LFY]-x(3)-
PA   [FYIVA]-{FYWHCM}-x(3)-[GSADENQKR]-x-[NSTAPKL]-[PARL].
NR   /RELEASE=36,74019;
NR   /TOTAL=115(107); /POSITIVE=78(73); /UNKNOWN=0(0); /FALSE_POS=37(34);
NR   /FALSE_NEG=8; /PARTIAL=0;
CC   /TAXO-RANGE=???P?; /MAX-REPEAT=1;
DR   P43463, AARP_PROST, T; P19219, ADAA_BACSU, T; Q10630, ADA_MYCTU , T;
DR   P33234, ADIY_ECOLI, T; P43464, AGGR_ECOLI, T; P05052, APPY_ECOLI, T;
DR   P11765, ARAC_CITFR, T; P03021, ARAC_ECOLI, T; P07642, ARAC_ERWCH, T;
DR   P03022, ARAC_SALTY, T; Q03320, ARAL_STRAT, T; P35319, ARAL_STRLI, T;
DR   P17410, CELD_ECOLI, T; P25393, CFAD_ECOLI, T; P43460, CSVR_ECOLI, T;
DR   P10805, ENVY_ECOLI, T; P26993, EXSA_PSEAE, T; P23774, FAPR_ECOLI, T;
DR   P31778, HRPB_BURSO, T; P39437, INVF_SALTY, T; P28808, LCRF_YERPE, T;
DR   Q51872, LUMQ_PHOLE, T; P27246, MARA_ECOLI, T; Q56070, MARA_SALTY, T;
DR   P10411, MELR_ECOLI, T; P28809, MMSR_PSEAE, T; Q00753, MSMR_STRMU, T;
DR   Q04642, MXIE_SHIFL, T; Q55292, MXIE_SHISO, T; P40883, PCHR_PSEAE, T;
DR   P43459, PERA_ECOLI, T; Q05587, POCR_SALTY, T; Q52620, PQRA_PROVU, T;
DR   P43465, RAFR_PEDPE, T; Q48413, RAMA_KLEPN, T; P09378, RHAR_ECOLI, T;
DR   P40865, RHAR_SALTY, T; P09377, RHAS_ECOLI, T; P27029, RHAS_SALTY, T;
DR   P16114, RNS_ECOLI , T; P22539, SOXS_ECOLI, T; Q56143, SOXS_SALTY, T;
DR   P29492, TCPN_VIBCH, T; P43462, THCR_RHOSO, T; P32326, URER_ECOLI, T;
DR   Q02458, URER_PROMI, T; Q04248, VIRF_SHIDY, T; P13225, VIRF_YEREN, T;
DR   P37390, XYLR_ECOLI, T; P45043, XYLR_HAEIN, T; P07859, XYLS_PSEPU, T;
DR   Q04710, XYS1_PSEPU, T; Q05092, XYS2_PSEPU, T; Q05335, XYS3_PSEPU, T;
DR   Q04713, XYS4_PSEPU, T; P55449, Y4FK_RHISN, T; P45008, YA52_HAEIN, T;
DR   P40408, YBBB_BACSU, T; P43461, YCGK_ALTCA, T; P76241, YEAM_ECOLI, T;
DR   P36547, YFEG_ECOLI, T; P54722, YFIF_BACSU, T; P37638, YHIW_ECOLI, T;
DR   P37639, YHIX_ECOLI, T; P31449, YIDL_ECOLI, T; P32677, YIJO_ECOLI, T;
DR   P40331, YISR_BACSU, T; P43458, YMCR_STRLA, T; P06134, ADA_ECOLI , T;
DR   P26189, ADA_SALTY , T; P26950, CAFR_YERPE, T; P27292, ROB_ECOLI , T;
DR   P28816, TETD_ECOLI, T;
DR   Q47129, FEAR_ECOLI, N; P55922, RAMA_ENTCL, N; Q06861, V38K_MYCTU, N;
DR   P77634, YBCM_ECOLI, N; P77601, YKGA_ECOLI, N; P77379, YKGD_ECOLI, N;
DR   Q46855, YQHC_ECOLI, N; P71663, YR12_MYCTU, N;
DR   P28647, AA3R_RAT , F; P74985, ARSB_YEREN, F; Q54468, CHB_SERMA , F;
DR   P23577, CYF_CHLRE , F; P43712, FABD_HAEIN, F; Q44763, FLIP_BORBU, F;
DR   P74930, FLIP_TREPA, F; O42101, FTF_CHICK , F; Q04952, GLS3_YEAST, F;
DR   P23970, MEND_BACSU, F; O15303, MGR6_HUMAN, F; P35349, MGR6_RAT , F;
DR   P40931, MPL_MPLV , F; P55015, NKC2_RABIT, F; P55016, NKC2_RAT , F;
DR   P29801, NU2C_SYNP7, F; P72714, NU2C_SYNY3, F; P28531, RL5_CHLTR , F;
DR   P33983, RP54_ACICA, F; P54991, SNAA_STRPR, F; P40238, TPOR_HUMAN, F;
DR   Q08351, TPOR_MOUSE, F; P23626, V3A_TAV , F; P15911, VFP3_FOWPV, F;
DR   P47571, Y329_MYCGE, F; P77400, YBAT_ECOLI, F; P75826, YBJE_ECOLI, F;
DR   P77744, YDAK_ECOLI, F; P87293, YDN1_SCHPO, F; O22288, YI79_ARATH, F;
DR   Q60306, YY07_METJA, F; Q45980, FLIP_CAUCR, F; P55719, Y4YK_RHISN, F;
DR   P29940, YCB7_PSEDE, F;
3D   2AAC; 2ARA; 2ARC; 1ADN; 1SFE;
DO   PDOC00040;
//
```

Figure 1. *continued*

textbook, contains textual information that documents each pattern.

A sample textbook entry is shown (Fig. 1a); this particular entry is linked to two entries in the PROSITE.DAT file: a pattern and a profile (Fig. 1b).

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 15.0 of PROSITE (July 1998) contains 1014 documentation entries describing 1352 different patterns, rules and profiles/matrices. In addition to these entries, a collection of 241 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 (14.0) is provided in Table 1, furthermore, many entries were updated. The database requires ~5 Mb of disk storage space. The present distribution frequency is two releases per year. No restrictions are placed on use or redistribution of the data. Future releases of PROSITE will be copyright (releases up to number 15.0 are not).

HOW TO OBTAIN A LOCAL COPY OF PROSITE

By CD-ROM

PROSITE is distributed on CD-ROM by the EMBL Outstation—the European Bioinformatics Institute (EBI) (13). For all enquiries regarding the subscription and distribution of PROSITE one should contact: The EMBL Outstation—The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Tel: +44 1223 494 444; Fax: +44 1223 494 468; Email: datalib@ebi.ac.uk

```
ID   HTH_ARAC_FAMILY_2; MATRIX.
AC   PS01124;
DT   NOV-1995 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE   Bacterial regulatory proteins, araC family DNA-binding domain profile.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=99;
MA   /DISJOINT: DEFINITION=PROTECT; N1=6; N2=94;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=1.5162; R2=0.0218; TEXT='OrigScore';
MA   /CUT_OFF: LEVEL=0; SCORE=320; N_SCORE=8.5; MODE=1;
MA   /DEFAULT: D=-20; I=-20; B1=-70; E1=-70; MI=-105; MD=-105; IM=-105; DM=-105;
MA   /I: B1=0; BI=-105; BD=-105;
MA   /M: SY='D'; M=-10,11,-25,14,13,-25,-12,2,-25,4,-22,-15,7,-13,8,4,0,-7,-23,-25,-10,10;
MA   /M: SY='R'; M=-7,-1,-26,-1,5,-24,-15,-3,-24,15,-19,-11,0,-11,8,20,-3,-6,-18,-22,-11,5;
MA   /M: SY='V'; M=8,-24,-17,-29,-22,-2,-24,-25,21,-21,15,9,-22,-22,-20,-22,-11,-3,22,-23,-
8,-22;
MA   /M: SY='V'; M=-7,-18,-10,-21,-13,-7,-25,-14,4,-11,5,4,-15,-23,-10,-5,-11,-2,6,-23,-5,-
13;
MA   /M: SY='Q'; M=-2,-1,-20,-3,2,-23,-10,1,-19,0,-14,-7,0,-15,7,1,-1,-4,-16,-26,-12,3;
..
...
.... Lot of lines omitted.
...
..
MA   /M: SY='R'; M=-13,-14,-25,-15,-6,-9,-21,-7,-13,8,-3,-2,-7,-21,0,32,-11,-8,-9,-20,-6,-6;
MA   /M: SY='R'; M=-4,-4,-24,-6,2,-20,-16,-1,-17,9,-14,-6,-1,-15,7,11,-3,-4,-14,-22,-9,3;
MA   /M: SY='R'; M=-9,-7,-26,-8,-3,-17,-10,-2,-13,0,-12,-4,-2,-17,3,5,-4,-5,-12,-21,-6,-2;
MA   /I: E1=0; IE=-105; DE=-105;
NR   /RELEASE=36,74019;
NR   /TOTAL=81(81); /POSITIVE=81(81); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0; /PARTIAL=0;
CC   /TAXO-RANGE=???P?; /MAX-REPEAT=1;
DR   P43463, AARP_PROST, T; P19219, ADAA_BACSU, T; P06134, ADA_ECOLI , T;
DR   Q10630, ADA_MYCTU , T; P26189, ADA_SALTY , T; P33234, ADIY_ECOLI, T;
DR   P43464, AGGR_ECOLI, T; P05052, APPY_ECOLI, T; P11765, ARAC_CITFR, T;
..
...
.... Lot of lines omitted.
...
..
DR   P37639, YHIX_ECOLI, T; P31449, YIDL_ECOLI, T; P32677, YIJO_ECOLI, T;
DR   P40331, YISR_BACSU, T; P77601, YKGA_ECOLI, T; P77379, YKGD_ECOLI, T;
DR   P43458, YMCR_STRLA, T; Q46855, YQHC_ECOLI, T; P71663, YR12_MYCTU, T;
3D   1ADN; 1SFE; 2AAC; 2ARA; 2ARC;
DO   PDOC00040;
//
```

Figure 1. *continued*

Table 1. List of patterns documentation entries that have been added since the last release of PROSITE (14.0)

DNA repair protein radC family signature
recR protein signature
ubiH/COQ6 monooxygenase family signature
ATP phosphoribosyltransferase signature
Prolipoprotein diacylglyceryl transferase signature
Phosphatidate cytidylyltransferase signature
Lipoate-protein ligase B signature
moaA / nifB / pqqE family signature
BCCT family of transporters signature
Flagellar motor protein motA family signature
Protein secA signatures
ATP1G1 / PLM / MAT8 family signature
Protein smpB signature
Uncharacterized protein family UPF0044 signature
Uncharacterized protein family UPF0047 signature
Uncharacterized protein family UPF0054 signature
Uncharacterized protein family UPF0057 signature

By anonymous FTP

If you have access to a computer system linked to the Internet you can obtain PROSITE using FTP (File Transfer Protocol), from the following file servers:

ExPASy (Expert Protein Analysis System) server, Swiss Institute of Bioinformatics (SIB); Internet address: ftp://www.expasy.ch/databases/prosite/

ISREC (Swiss Institute for Experimental Cancer Research) anonymous FTP server, Swiss Institute of Bioinformatics (SIB); Internet address: ftp://ftp.isrec.isb-sib.ch/sib-isrec/profiles/

EBI (European Bioinformatics Institute) anonymous FTP server; Internet address: ftp://ftp.ebi.ac.uk/pub/databases/prosite/

The pre-release collection of profiles is only available from the ISREC FTP server.

By Email through the EBI network fileserver

PROSITE can be obtained from the EBI network fileserver. Detailed instructions on how to make the best use of this service, and in particular on how to obtain PROSITE, can be obtained by sending to the network address netserv@ebi.ac.uk the following message:

HELP
HELP PROSITE

HOW TO MAKE 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:

(i) *pftools* (version 2.1 in FORTRAN77) written by Philipp Bucher. pfscan loads a sequence from a file and scans it with all (or one) of PROSITE profiles; pfsearch loads a profile from a file and scans for it in a SWISS-PROT database file. 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 and for Windows 95/98.

(ii) *PrfLib* (version 1.0 in ANSI C) written by Nicolas Moeri. scan4prf loads a sequence from a file and scans it with all (or one) of PROSITE profiles; srch4prf loads a profile from a file and scans for it in a SWISS-PROT database file. These tools are available from the server: http://mamac29.epfl.ch/

Email servers

There are many Email servers that are available to molecular biologists (14). This an example of a server taking advantage of the PROSITE database:

Name:	MOTIF E-Mail Server on GenomeNet
Organization:	Supercomputer Laboratory, Kyoto Institute for Chemical Research, Japan
Description:	Allows to rapidly compare a new protein sequence against all patterns stored in PROSITE as well as in the MotifDic library (15).
Server email address:	motif@genome.ad.jp
Address to report problems:	motif-manager@genome.ad.jp

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 (WWW) molecular biology server ExPASy (16). Using a WWW browser, one has access to all the hypertext documents stored on the ExPASy server (as well as many other WWW servers) and also can make use of many sequence analysis software tools.

The ExPASy server may be accessed through its URL which is: http://www.expasy.ch/ . You can directly access to the 'top' page of the section of ExPASy that allows you to browse through the PROSITE documentation and data entries by opening the URL: http://www.expasy.ch/sprot/prosite.html

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

ScanProsite. Allows the user 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.ch/sprot/scnpsite.html

ProfileScan. Allows the user to scan a protein sequence—from SWISS-PROT or provided by the user—for the occurrence of profiles stored in PROSITE. The URL for ProfileScan is: http://www.isrec.isb-sib.ch/software/PFSCAN_form.html

FrameProfileScan. Allows the user 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

REFERENCES

1 Bairoch,A. and Bucher,P. (1994) *Nucleic Acids Res.*, **22**, 3583–3589.
2 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
3 Doolittle,R.F. (1986) *Of URFs and ORFs: A Primer On How To Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, California.
4 Lesk,A.M. (1988) In Lesk,A.M. (ed.), *Computational Molecular Biology*. Oxford University Press, Oxford, pp. 17–26.
5 Bucher,P. and Bairoch,A. (1994) In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *ISMB-94; Proceedings 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) *Comput. Chem.*, **20**, 3–23.
7 Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
8 Eddy,S.R. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361–365.
9 Gribskov,M., Luethy,R. and Eisenberg,D. (1990) *Methods Enzymol.*, **183**, 146–159.
10 Luethy,R., Xenarios,I. and Bucher,P. (1994) *Protein Sci.*, **3**, 139–146.
11 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Comput. Applic. Biosci.*, **10**, 19–29.
12 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
13 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res.*, **26**, 8–15.
14 Henikoff,S. (1993) *Trends Biochem. Sci.*, **18**, 267–268.
15 Ogiwara,A., Uchiyama,I., Seto,Y. and Kanehisa,M. (1992) *Protein Engng*, **5**, 479–488.
16 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci.*, **19**, 258–260.