THE PROSITE DATABASE, ITS STATUS IN 1997

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

1 Department of Medical Biochemistry, University of Geneva, 1 rue Michel Servet, 1211 Geneva 4, Switzerland
2 Biocomputing Group, Swiss Institute for Experimental Cancer Research (ISREC), 1066 Epalinges s/Lausanne, Switzerland

Abstract

The PROSITE database 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 determining 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, but 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 these 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. This reality has been recognized by many authors [3,4]. 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, the SH2 and SH3 domain. 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 distinguish 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.

- 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 entry 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 insure 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 of the 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 textbook, contains textual information that documents each pattern.

A sample textbook entry is shown (Figure 1a); this particular entry is linked to two entries in the PROSITE.DAT file: a pattern and a profile (Figure 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 13.2 of PROSITE (October 1996) contains 936 documentation entries describing 1225 different patterns, rules and profiles. The list of the entries which have been added since the publication of the previous article [2] describing PROSITE is provided in table 1. The database requires about 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.

How to obtain a local copy of PROSITE

a) 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
United Kingdom

Telephone: (+44 1223) 494 400
Telefax  : (+44 1223) 494 468
Electronic network address: datalib@ebi.ac.uk
```

b) 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:

```
EBI anonymous FTP server
Internet address: ftp.ebi.ac.uk (or 192.54.41.33)

NCBI Repository (National Library of Medicine, NIH, Washington
D.C., U.S.A.)
Internet address: ncbi.nlm.nih.gov (or 130.14.20.1)

ExPASy (Expert Protein Analysis System) server, University of
```

```
Geneva, Switzerland
Internet address: expasy.hcuge.ch (or 129.195.254.61)

National Institute of Genetics (Japan) FTP server
Internet address: ftp2.ddbj.nig.ac.jp (or 133.39.3.6)
```

c) 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

a) 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.

To make use of profile entries, we are distributing, under the name 'pftools', the source code (in FORTRAN77) of two programs that should help software developers to implement profile-specific routines in their application(s):

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 data base file.

These tools are available by anonymous FTP from the server 'ulrec3.unil.ch' in the directory '/pub/pftools'..

b) Email servers

There are many email servers that are available to molecular biologists [14]. At least three of these servers can be used in conjunction with PROSITE:

Name:	EBI Mail-PROSITE Server
Organization:	European Bioinformatics Institute / Hinxton / UK
Description:	Allows to rapidly compare a new protein sequence against all patterns stored in PROSITE.
Server email address:	prosite@ebi.ac.uk
Address to report problems:	nethelp@ebi.ac.uk

Name:	BLOCKS e-mail searcher
Organization:	Fred Hutchinson Center / Seattle / USA
Description:	Compares a protein or DNA sequence to the database of protein

blocks.
Blocks are short multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The BLOCKS database [15] has been derived from PROSITE. This server can also be used to retrieve specifics blocks and PROSITE entries.

Server email address: blocks@howard.fhcrc.org
Address to report problems: henikoff@howard.fhcrc.org

Name: MOTIF E-Mail Server on GenomeNet
Organization: Supercomputer Laboratory / Kyoto Inst. 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 [16].
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 [17]. WWW is a global information retrieval system merging the power of world-wide networks, hypertext and multimedia. Through hypertext links, it gives access to documents and information available on thousands of servers around the world. To access a WWW server one needs a WWW browser (such as Mosaic(TM), Netscape Navigator(TM) or Microsoft Internet Explorer(TM)). 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 Uniform Resource Locator (URL - the addressing system defined in WWW), which is:

```
http://expasy.hcuge.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://expasy.hcuge.ch/sprot/prosite.html
```

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

- *ScanProsite*, which 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 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://expasy.hcuge.ch/sprot/scnpsite.html
  ```

- *ProfileScan*, which allows 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://ulrec3.unil.ch/software/profilescan.html
  ```

References

1 Bairoch A., Bucher P.
Nucleic Acids Res. 22:3583-3589(1994).

2 Bairoch A., Bucher P., Hofmann K.
Nucleic Acids Res. 24:189-196(1996).

3 Doolittle R.F.
(In) Of URFs and ORFs: a primer on how to analyze derived amino acid sequences., University Science Books, Mill Valley, California, (1986).

4 Lesk A.M.
(In) Computational Molecular Biology, Lesk A.M., Ed., pp17-26, Oxford University Press, Oxford (1988).

5 Bucher P., Bairoch A.
ISMB-94; Proceedings Second International Conference on Intelligent Systems for Molecular Biology, Altman R., Brutlag D., Karp P., Lathrop R., Searls D., Eds., pp. 53-61, AAAI Press, Menlo Park, (1994).

6 Bucher P., Karplus K., Moeri N., Hofmann K.
Comput. Chem. 20:3-23(1996).

7 Gribskov M., McLachlan A.D., Eisenberg D.
Proc. Natl. Acad. Sci. U.S.A. 84:4355-4358(1987).

8 Eddy S.R.
Curr. Opinion Struct. Biol. 6:361-365(1996).

9 Gribskov M., Luethy R., Eisenberg D.
Meth. Enzymol. 183:146-159(1990).

10 Luethy R., Xenarios I., Bucher P.
Protein Sci. 3:139-146(1994).

11 Thompson J.D., Higgins D.G., Gibson T.J.
Comput. Appl. Biosci. 10:19-29(1994).

12 Bairoch A., Apweiler R.
Nucleic Acids Res. 24:21-25(1996).

13 Rodriguez-Tome P., Stoehr P.J., Cameron G.N., Flores T.P.
Nucleic Acids Res. 24:6-12(1996).

14 Henikoff S.
Trends Biochem. Sci. 18:267-268(1993).

15 Henikoff S., Henikoff J.G.
Nucleic Acids Res. 19:6565-6572(1991).

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

17 Appel R.D., Bairoch A., Hochstrasser D.F.
Trends Biochem. Sci. 19:258-260(1994).

Figure 1. Sample data from PROSITE

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.
- rafR, the regulator of the raffinose operon in Pediococcus pentosaceus.
- rhaR, the Escherichia coli and Salmonella thypimurium 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 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: 13.

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

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

-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.

-Last update: September 1995 / Pattern and text revised; profile added.

[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. USA 90:5638-5642(1993).
{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); SEP-1995 (DATA UPDATE); SEP-1995 (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=29,38303;
NR   /TOTAL=76(76); /POSITIVE=63(63); /UNKNOWN=0(0); /FALSE_POS=13(13);
NR   /FALSE_NEG=0(0);
CC   /TAXO-RANGE=???P?; /MAX-REPEAT=1;
DR   P43463, AARP_PROST, T; P19219, ADAA_BACSU, T; P06134, ADA_ECOLI , T;
DR   P26189, ADA_SALTY , T; P33234, ADIY_ECOLI, T; P43464, AGGR_ECOLI, T;
DR   P05052, APPY_ECOLI, T; P11765, ARAC_CITFR, T; P03021, ARAC_ECOLI, T;
DR   P07642, ARAC_ERWCH, T; P03022, ARAC_SALTY, T; Q03320, ARAL_STRAT, T;
DR   P35319, ARAL_STRLI, T; P26950, CAFR_YERPE, T; P17410, CELD_ECOLI, T;
DR   P43460, CSVR_ECOLI, T; P25393, CFAD_ECOLI, T; P10805, ENVY_ECOLI, T;
DR   P26993, EXSA_PSEAE, T; P23774, FAPR_ECOLI, T; P31778, HRPB_BURSO, T;
DR   P39437, INVF_SALTY, T; P28808, LCRF_YERPE, T; P27246, MARA_ECOLI, T;
DR   P10411, MELR_ECOLI, T; P28809, MMSR_PSEAE, T; Q00753, MSMR_STRMU, T;
DR   Q04642, MXIE_SHIFL, T; P40883, PCHR_PSEAE, T; P43459, PERA_ECOLI, T;
DR   Q05587, POCR_SALTY, T; P43465, RAFR_PEDPE, T; P09378, RHAR_ECOLI, T;
DR   P40865, RHAR_SALTY, T; P09377, RHAS_ECOLI, T; P27029, RHAS_SALTY, T;
DR   P16114, RNS_ECOLI , T; P27292, ROB_ECOLI , T; P22539, SOXS_ECOLI, T;
DR   P29492, TCPN_VIBCH, T; P28816, TETD_ECOLI, T; P43462, THCR_RHOSO, T;
DR   P32326, URER_ECOLI, T; Q02458, URER_PROMI, T; Q04248, VIRF_SHIDY, T;
DR   P13225, VIRF_YEREN, T; P37390, XYLR_ECOLI, T; P45043, XYLR_HAEIN, T;
DR   P07859, XYLS_PSEPU, T; Q04710, XYS1_PSEPU, T; Q05092, XYS2_PSEPU, T;
DR   Q05335, XYS3_PSEPU, T; Q04713, XYS4_PSEPU, T; P36547, YFEG_ECOLI, T;
DR   P37638, YHIW_ECOLI, T; P37639, YHIX_ECOLI, T; P31449, YIDL_ECOLI, T;
DR   P32677, YIJO_ECOLI, T; P40331, YUXC_BACSU, T; P40408, YZBC_BACSU, T;
DR   P45008, YA52_HAEIN, T; P43461, YCGK_ALTCA, T; P43458, YMCR_STRLA, T;
DR   P28647, AA3R_RAT  , F; P23577, CYF_CHLRE , F; P23969, MEND_BACSU, F;
DR   P35349, MGR6_RAT  , F; P40931, MPL_MPLV  , F; P29801, NU2C_SYNP7, F;
DR   P40238, TPOR_HUMAN, F; Q08351, TPOR_MOUSE, F; P28531, RL5_CHLTR , F;
DR   P33983, RP54_ACICA, F; P23626, V3A_TAV   , F; P15911, VFP3_FOWPV, F;
DR   P29940, YCB7_PSEDE, F;
DO   PDOC00040;
//

ID   HTH_ARAC_FAMILY_2; MATRIX.
AC   PS01124;
DT   SEP-1995 (CREATED); SEP-1995 (DATA UPDATE); SEP-1995 (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=-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=32,?;
NR    /TOTAL=53(53); /POSITIVE=53(53); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR    /FALSE_NEG=0(0);
CC    /TAXO_RANGE=???P?; /MAX-REPEAT=1;
DR    P43463, AARP_PROST, T; P19219, ADAA_BACSU, T; P06134, ADA_ECOLI , T;
DR    P26189, ADA_SALTY , T; P33234, ADIY_ECOLI, T; P43464, AGGR_ECOLI, T;
DR    P05052, APPY_ECOLI, T; P11765, ARAC_CITFR, T; P03021, ARAC_ECOLI, T;
..
...
.... Lot of lines omitted.
...
..
DR    P37638, YHIW_ECOLI, T; P37639, YHIX_ECOLI, T; P31449, YIDL_ECOLI, T;
DR    P32677, YIJO_ECOLI, T; P40331, YUXC_BACSU, T; P40408, YZBC_BACSU, T;
DR    P45008, YA52_HAEIN, T; P43461, YCGK_ALTCA, T; P43458, YMCR_STRLA, T;
DO    PDOC00040;
//
```

Table 1. List of patterns documentation entries which have been added to PROSITE since the last publication of the NAR database issue

Anaphylatoxin domain signature and profile
C-terminal cystine knot signature and profile
CUB domain profile
Calcium-binding EGF-like domain signature
LDL-receptor class A (LDLRA) domain signature
Phosphotyrosine interaction domain (PID) profile
VWFC domain signature
NF-kappa-B/Rel/dorsal family signature
Ribosomal protein L1 signature
Ribosomal protein L17 signature
Ribosomal protein L21 signature
Ribosomal protein L6e signature
Ribosomal protein L15e signature
Ribosomal protein L21e signature
Ribosomal protein L36e signature
Ribosomal protein L44e signature
Ribosomal protein S21 signature
Ribosomal protein S3Ae signature
Ribosomal protein S8e signature
Ribosomal protein S12e signature
Ribosomal protein S27e signature
Short-chain dehydrogenases/reductases family signature
N-acetyl-gamma-glutamyl-phosphate reductase active site
Gamma-glutamyl phosphate reductase signature
Copper amine oxidase signatures
RNA polymerases beta chain signature
Lipolytic enzymes "G-D-X-G" family, putative active sites signatures
Peptidyl-tRNA hydrolase signatures
Ribonuclease II family signature
Glycosyl hydrolases family 35 putative active site
Iodothyronine deiodinases active site
Protozoan/cyanobacterial globins signature
Phosphatidylethanolamine-binding protein family signature
Amiloride-sensitive sodium channels signature
Ammonium transporters signature
GNS1/SUR4 family signature
Caveolins signature
ATP P2X receptors signature
Initiation factor 2 signature
PMP-22 / EMP / MP20 family signatures
Glypicans signature
Tub family signatures
Mrp family signature
Hypothetical YBL036c/yggS family signature
Hypothetical YBR187w/SLL0615 family signature
Hypothetical YML110c/yigO family signatures
Hypothetical yigU/ycbT/ycf43 family signature