Article

PROSITE: recent developments

BAIROCH, Amos Marc, BUCHER, Philip

Abstract

PROSITE is a compilation of sites and patterns found in protein sequences; it can be used as a method of determining the function of uncharacterized proteins translated from genomic or cDNA sequences.

Reference

BAIROCH, Amos Marc, BUCHER, Philip. PROSITE: recent developments. *Nucleic acids research*, 1994, vol. 22, no. 17, p. 3583-9

PMID : 7937064

Nucleic Acids Research, 1994, Vol. 22, No. 17 3583–3589

PROSITE: recent developments

Amos Bairoch* and Philipp Bucher[1]
Department of Medical Biochemistry, University of Geneva, 1 rue Michel Servet, 1211 Geneva 4 and [1]Biocomputing Group, Swiss Institute for Experimental Cancer Research (ISREC), 1066 Epalinges s/Lausanne, Switzerland

ABSTRACT

PROSITE is a compilation of sites and patterns found in protein sequences; it can be used as a method of determining the function of uncharacterized proteins translated from genomic or cDNA sequences.

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 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 [1,2]. While there have been a number of reviews of published patterns [3,4,5], no attempt had been made until very recently [6,7] to systematically collect biologically significant patterns or to discover new ones. Based on these observations, we decided in 1988, to actively pursue the development of a database of patterns which would be used to search against sequences of unknown function. This database, called PROSITE, contains some patterns which have been published in the literature, but the majority have been developed in the last five years by the author.

LEADING CONCEPTS

The design of PROSITE follows four 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 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.

FORMAT

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 the SWISS-PROT protein sequence data bank [8]. 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 user manual (PROSUSER.TXT) is distributed with the database; it fully describes the format of both files. A sample textbook entry is shown (Figure 1a) with the corresponding data from the pattern file (Figure 1b).

EXTENSION OF PROSITE TO PROFILES

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

*To whom correspondence should be addressed

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.

Starting with the current release of PROSITE, we have introduced some profile entries. The profile structure used in PROSITE is similar to but slightly more general than the one introduced by Gribskov and co-workers [9]. Additional parameters allow representation of other motif descriptors, including the currently popular hidden Markov models. A technical description of the profile structure and of the corresponding motif search method is given in the file 'PROFILE.TXT' included in the PROSITE release.

Profiles can be constructed by a large variety of different techniques. The classical method developed by Gribskov and co-workers [10] requires a multiple sequence alignment as input and uses a symbol comparison table to convert residue frequency distributions into weights. The profiles included in the current PROSITE release were generated by this procedure applying recent modifications described by Luethy and co-workers [11]. In the future, we intend to apply additional profile construction tools including structure-based approaches and methods involving machine learning techniques. We also consider the possibility of distributing published profiles developed by others in PROSITE format along with locally produced documentation entries.

Unlike patterns, profiles 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. This can lead to specific problems not arising with PROSITE patterns. With a profile covering conserved as well as divergent sequence regions, there is a chance to obtain a significant similarity score even with a partially incorrect alignment. This possibility is taken into account by our quality evaluation procedures. In order to be acceptable, a profile must not only assign high similarity scores to true motif occurrences and low scores to false matches. In addition, it should correctly align those residues having analogous functions or structural properties according to experimental data.

An example of a PROSITE profile entry is shown in figure 2.

CONTENT OF THE CURRENT RELEASE

Release 12 of PROSITE (June 1994) contains 785 documentation entries describing 1029 different patterns, rules and profiles. The list of the entries which have been added since the publication of the previous article [12] describing PROSITE is provided in Appendix 1. The database requires about 4 Mb of disk storage space. The present distribution frequency is four releases per year. No restrictions are placed on use or redistribution of the data.

COMPUTER PROGRAMS THAT MAKE USE OF PROSITE

Many academic groups and commercial companies have developed computer programs that make use of the pattern entries in PROSITE. We list here some of these programs (a full descriptive list is included with the database and is stored in a file called 'PROSITE.PRG').

Academic

Program	Operating system	Author
MacPattern	Apple MacIntosh	Rainer Fuchs [13]
prosite.c	IBM 3090–400E and Unix	Klaus Hartmuth
ProSearch	Unix and DOS (AWK)	Lee Kolakowski [14]
dbsite/mksite	Unix	J.-M. Claverie
PROINDEX	VAX VMS	Steve Clark
Quelsite	VAX VMS	Claude Valencien
Scrutineer	VAX VMS and Unix	Peter Sibbald [15]
PATTERN	Unix	Olivier Boulot
PIP and PIPL	VAX VMS and Unix	Rodger Staden [16]
PATMAT	OS and Unix	Steven Henikoff [17]
PROTOMAT	OS and Unix	Steven Henikoff [18]
PPS	DOS	Huiachun Wang [19]

Commercial

Program	Package	Supplier	Operating system
MOTIF	GCG	Genetics Comp. Group	Vax VMS and Unix
QUEST	IG-Suite	IntelliGenetics	Vax VMS and Unix
PROMOT	OML		Vax VMS and Unix
PROSITE	PC/Gene	IntelliGenetics	DOS
PROSITE	GeneWorks	IntelliGenetics	Apple MacIntosh
Protean	LaserGene	DNASTAR	Apple MacIntosh
PROTSITE		National Biosciences	DOS

As it is the first release of PROSITE to include profile entries, none of the above programs can currently make use of them. We are therefore distributing, with the PROSITE release, the source code (C language) of two programs that should help software developers to implement profile-specific routines in their application(s):

scan4prf Loads a sequence from a file and scans it with all (or one) of the PROSITE profiles.

srch4prf Loads a profile from a file and scans for that profile in a SWISS-PROT data base file.

EMAIL SERVERS

There is a wealth of email servers that are available to molecular biologists [20]. At least three of these servers can be used in conjunction with PROSITE:

Name: EMBL Mail-PROSITE Server
Organization: European Molecular Biology Laboratory (EMBL)/ Heidelberg/Germany
Description: Allows to rapidly compare a new protein sequence against all patterns stored in PROSITE.
Server email address: prosite@embl-heidelberg.de
Address to report problems: nethelp@embl-heidelberg.de
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 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 the MotifDic library [21].
Server email address:	motif@genome.ad.jp
Address to report problems:	motif-manager@genome.ad.jp

HOW TO OBTAIN PROSITE

PROSITE is distributed on magnetic tape and on CD-ROM by the EMBL Data Library. For all enquiries regarding the subscription and distribution of PROSITE one should contact:

EMBL Data Library
European Molecular Biology Laboratory
Postfach 10.22.09, Meyerhofstrasse 1
D-69012 Heidelberg, Germany
Telephone: (+49 6221) 387 258
Telefax : (+49 6221) 387 519 or 387 306
Electronic network address: datalib@EMBL-heidelberg.de

PROSITE can be obtained from the EMBL File Server [22]. 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@EMBL-heidelberg.de the following message:

HELP
HELP PROSITE

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:

EMBL anonymous FTP server

Internet address: ftp.EMBL-heidelberg.de (or 192.54.41.33)
NCBI Repository (National Library of Medicine, NIH, Washington D.C., U.S.A.)
Internet address: ncbi.nlm.nih.gov (130.14.20.1)
ExPASy (Expert Protein Analysis System) server, University of Geneva, Switzerland
Internet address: expasy.hcuge.ch (129.195.254.61)
National Institute of Genetics (Japan) FTP server
Internet address: ftp.nig.ac.jp (133.39.16.66)

INTERACTIVE ACCESS TO PROSITE

You can browse through PROSITE using various Internet Gopher servers that specialize in biosciences (biogophers) [23]. Gopher is a distributed document delivery service that allows a neophyte user to access various types of data residing on multiple hosts in a seamless fashion.

PROSITE is currently available on the ExPASy World-Wide Web (WWW) molecular biology server [24]. WWW, which originated at CERN in Geneva, 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 (including images, movies and sound) available on thousands of servers around the world, using network protocols such as FTP, WAIS, Gopher, X500, etc. as well as the WWW specific HyperText Transfer Protocol (HTTP). To access this server (or any other WWW server), one needs a WWW browser. Public domain browsers exist for a variety of computer systems, including Unix, MS-Windows and Macintoshes. One popular browser available for all three platforms is Mosaic, developed at the National Center for Supercomputing Applications (NCSA) of the University of Illinois at Champaign. It may be obtained by anonymous ftp from ftp.ncsa.uiuc.edu, in the directories /Mosaic, respectively /PC and /Mac. Using a WWW browser, one has access to all the hypertext documents stored on the ExPASy server (as well as other WWW servers).

The ExPASy WWW server may be accessed through its Uniform Resource Locator (URL - the addressing system defined in WWW), which is:

http://expasy.hcuge.ch/

REFERENCES

1. 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).
2. Lesk A.M. (In) Computational Molecular Biology, Lesk A.M., Ed., pp17–26, Oxford University Press, Oxford (1988).
3. Barker W.C., Hunt T.L., George D.G. Protein Seq. Data Anal. 1:363–373(1988).
4. Hodgman T.C. Comput. Appl. Biosci. 5:1–13(1989).
5. Taylor W.R., Jones D.T. Curr. Opin. Struct. Biol. 1:327–333(1991).
6. Bork P. FEBS Lett. 257:191–195(1989).
7. Smith H.O., Annau T.M., Chandrasegaran S. Proc. Natl. Acad. Sci. USA 87:826–830(1990).
8. Bairoch A., Boeckmann B. Nucleic Acids Res. 21:3093–3096(1993).
9. Gribskov M., McLachlan A.D., Eisenberg D. Proc. Natl. Acad. Sci. U.S.A. 84:4355–4358(1987).
10. Gribskov M., Luethy R., Eisenberg D. Meth. Enzymol. 183:146–159(1990).
11. Luethy R., Xenarios I., Bucher P. Protein Sci. 3:139–146(1994).
12. Bairoch A. Nucleic Acids Res. 21:3097–3103(1993).
13. Fuchs R. Comput. Appl. Biosci. 10:171–178(1994).
14. Kolakowski L.F. Jr., Leunissen J.A.M., Smith J.E. Biotechniques 13:919–921(1992).
15. Sibbald P.R., Sommerfeldt H., Argos P. Comput. Appl. Biosci. 7:535–536(1991).
16. Staden R. DNA Sequence 1:369–374(1991).
17. Wallace J.C., Henikoff S. Comput. Appl. Biosci. 8:249–254(1992).
18. Henikoff S., Henikoff J. Nucleic Acids Res. 19:6565–6572(1991).
19. Wang H., He Y. China High Tech. Lett. 4:30–33(1994).
20. Ogiwara A., Uchiyama I., Seto Y., Kanehisa M. Protein Eng. 5:479–488(1992).
21. Henikoff S. Trends Biochem. Sci. 18:267–268(1993).
22. Stoehr P.J., Omond R.A. Nucleic Acids Res. 17:6763–6764(1989).
23. Gilbert D. Trends Biochem. Sci. 18:107–108(1993).
24. Appel R.D., Bairoch A., Hochstrasser D.F. Trends Biochem. Sci. 19:258–260(1994).

a

{PDOC00107}
{PS00116; DNA_POLYMERASE_B}
{BEGIN}
****
* DNA polymerase family B signature *
****

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the accurate replication of DNA. They require either a small RNA molecule or a protein as a primer for the *de novo* synthesis of a DNA chain. On the basis of sequence similarities a number of DNA polymerases have been grouped together[1 to 7] under the designation of DNA polymerase family B. The polymerases that belong to this family are:

- Higher eukaryotes polymerases alpha.
- Higher eukaryotes polymerases delta.
- Yeast polymerase I/alpha (gene POL1), polymerase II/epsilon (gene POL2), polymerase III/delta (gene POL3), and polymerase REV3.
- Escherichia coli polymerase II (gene dinA or polB).
- Archaebacterial polymerases.
- Polymerases of viruses from the herpesviridae family.
- Polymerases from Adenoviruses.
- Polymerases from Baculoviruses.
- Polymerases from Chlorella viruses.
- Polymerases from Poxviruses.
- Bacteriophage T4 polymerase.
- Podoviridae bacteriophages Phi-29, M2, and PZA polymerase.
- Tectiviridae bacteriophage PRD1 polymerase.
- Polymerases encoded on mitochondrial linear DNA plasmids in various fungi and plants (Kluyveromyces lactis pGKL1 and pGKL2, Agaricus bitorquis pEM, Ascobolus immersus pAI2, Claviceps purpurea pCLK1, Neurospora Kalilo and Maranhar, maize S-1, etc).

Six regions of similarity (numbered from I to VI) are found in all or a subset of the above polymerases. The most conserved region (I) includes a conserved tetrapeptide which contains two aspartate residues. The function of this conserved region is not yet known, however it has been suggested [3] that it may be involved in binding a magnesium ion. We selected this conserved region as a signature for this family of DNA polymerases.

- Consensus pattern: [YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-x- [LIVMSTAC]
- Sequences known to belong to this class detected by the pattern: ALL, except for yeast polymerase II/epsilon and Agaricus bitorquis pEM.
- Other sequence(s) detected in SWISS-PROT: 3 other proteins.
- Last update: June 1994 / Text revised.

[1] Jung G., Leavitt M.C., Hsieh J.-C., Ito J. Proc. Natl. Acad. Sci. U.S.A. 84:8287–8291(1987).
[2] Bernad A., Zaballos A., Salas M., Blanco L. EMBO J. 6:4219–4225(1987).
[3] Argos P. Nucleic Acids Res. 16:9909–9916(1988).
[4] Wang T.S.-F., Wong S.W., Korn D. FASEB J. 3:14–21(1989).
[5] Delarue M., Poch O., Todro N., Moras D., Argos P. Protein Eng. 3:461–467(1990).
[6] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045–4057(1991).
[7] Braithwaite D.K., Ito J. Nucleic Acids Res. 21:787–802(1993).
{END}

b

```
ID   DNA_POLYMERASE_B; PATTERN.
AC   PS00116;
DT   APR-1990 (CREATED); OCT-1993 (DATA UPDATE); JUN-1994 (INFO UPDATE).
DE   DNA polymerase family B signature.
PA   [YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-x-[LIVMSTAC].
NR   /RELEASE=29,38303;
NR   /TOTAL=56(56); /POSITIVE=53(53); /UNKNOWN=0(0); /FALSE_POS=3(3);
NR   /FALSE_NEG=3(3);
CC   /TAXO-RANGE=ABEPV; /MAX-REPEAT=1;
DR   P26019, DPOA_DROME, T; P09884, DPOA_HUMAN, T; P33609, DPOA_MOUSE, T;
DR   P28040, DPOA_SCHPO, T; P27727, DPOA_TRYBB, T; P13382, DPOA_YEAST, T;
DR   P28339, DPOD_BOVIN, T; P28340, DPOD_HUMAN, T; P30315, DPOD_PLAFK, T;
DR   P30316, DPOD_SCHPO, T; P15436, DPOD_YEAST, T; P14284, DPOX_YEAST, T;
DR   P21189, DPO2_ECOLI, T; P80061, DPOL_PYRFU, T; P26811, DPOL_SULSO, T;
DR   P03261, DPOL_ADE02, T; P04495, DPOL_ADE05, T; P05664, DPOL_ADE07, T;
DR   P06538, DPOL_ADE12, T; P03198, DPOL_EBV , T; P08546, DPOL_HCMVA, T;
DR   P27172, DPOL_MCMVS, T; P04293, DPOL_HSV11, T; P07917, DPOL_HSV1A, T;
DR   P04292, DPOL_HSV1K, T; P09854, DPOL_HSV1S, T; P07918, DPOL_HSV21, T;
DR   P28857, DPOL_HSV6U, T; P28858, DPOL_HSVEB, T; P28859, DPOL_HSVI1, T;
DR   P24907, DPOL_HSVSA, T; P09252, DPOL_VZVD , T; P30318, DPOL_NPVLD, T;
DR   P18131, DPOL_NPVAC, T; P20509, DPOL_VACCC, T; P06856, DPOL_VACCV, T;
DR   P33793, DPOL_VARV , T; P21402, DPOL_FOWPV, T; P30319, DPOL_CBEPV, T;
DR   P30320, DPOL_CHVN2, T; P30321, DPOL_CHVP1, T; P03680, DPOL_BPPH2, T;
DR   P06950, DPOL_BPPZA, T; P19894, DPOL_BPM2 , T; P10479, DPOL_BPPRD, T;
DR   P04415, DPOL_BPT4 , T; P09804, DPO1_KLULA, T; P05468, DPO2_KLULA, T;
DR   P22374, DPOM_ASCIM, T; P22373, DPOM_CLAPU, T; P10582, DPOM_MAIZE, T;
DR   P33537, DPOM_NEUCR, T; P33538, DPOM_NEUIN, T;
DR   P21951, DPOE_YEAST, N; P30317, DPOL_THELI, N; P30322, DPOM_AGABT, N;
DR   P17545, RPB1_TRYBB, F; P17546, RPB2_TRYBB, F; P09278, TEGU_VZVD , F;
DO   PDOC00107;
//
```

Figure 1. Sample data from PROSITE. (**a**) A documentation (textbook) entry from the PROSITE.DOC file. (**b**) The corresponding entry in the PROSITE.DAT file.

```
ID   HSP20; MATRIX.
AC   PS01031;
DT   JUN-1994 (CREATED); JUN-1994 (DATA UPDATE); JUN-1994 (INFO UPDATE).
DE   Heat shock hsp20 proteins family profile.
MA   /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY'; LENGTH=97;
MA   /DISJOINT: DEFINITION=PROTECT; N1=2; N2=96;
MA   /NORMALIZATION: MODE=1; FUNCTION=GLE_ZSCORE;
MA   R1=239.0; R2=-0.0036; R3=0.8341; R4=1.016; R5=0.169;
MA   /CUT_OFF: LEVEL=0; SCORE=400; N_SCORE=10.0; MODE=1;
MA   /DEFAULT: MI=-210; MD=-210; IM=0; DM=0; I=-20; D=-20;
MA   /M: SY='R'; M=-12,-44,-11,-13,-13,-22,-2,-7,18,-12,5,-3,-11,0,21,-6,-5,-11,-16,-34;
MA   /M: SY='D'; M=1,-41,17,16,-41,-3,3,-11,-1,-22,-12,8,-7,12,-7,0,-2,-19,-53,-36;
MA   /M: SY='D'; M=2,-37,15,13,-36,2,5,-15,-3,-26,-17,10,-6,7,-10,3,2,-17,-53,-28;
MA   /M: SY='P'; M=1,-41,6,8,-38,-4,2,-20,9,-30,-14,6,13,9,8,3,0,-22,-48,-45;
MA   /M: SY='D'; M=2,-43,23,20,-42,2,9,-18,2,-30,-18,14,-5,14,-6,2,0,-21,-57,-35;
MA   /M: SY='D'; M=4,-34,9,8,-34,6,0,-17,5,-29,-14,8,-1,5,1,5,2,-17,-47,-38;
MA   /M: SY='F'; M=-28,-32,-38,-38,50,-42,-1,2,-11,6,-6,-21,-35,-27,-27,-24,-23,-14,-3,47;
MA   /M: SY='Q'; M=0,-33,-2,-7,-26,-9,-4,1,1,-10,1,-1,-5,2,0,-2,1,0,-44,-37;
MA   /M: SY='L'; M=-13,-36,-34,-37,23,-31,-21,28,-15,29,24,-24,-25,-24,-27,-20,-10,22,-33,0;
MA   /M: SY='K'; M=-8,-32,-5,-5,-19,-16,3,-11,13,-19,-2,1,-9,2,12,-3,-3,-15,-32,-28;
MA   /M: SY='L'; M=-10,-39,-30,-32,15,-26,-20,20,-16,27,20,-21,-20,-21,-27,-17,-9,16,-32,-5;
MA   /M: SY='D'; M=3,-48,33,27,-51,4,6,-19,0,-35,-22,18,-10,13,-13,2,0,-16,-65,-41;
MA   /I: MI=-55; MD=-55; I=-5;
MA   /M: SY='V'; D=-5; M=-3,-33,-23,-32,-5,-19,-21,28,-16,26,30,-17,-14,-15,-19,-12,-1,30,-48,-28;
MA   /I: MI=-55; MD=-55; I=-5;
MA   /M: SY='P'; D=-5; M=1,-2,-1,0,-3,0,0,-1,-1,-2,-2,0,4,0,0,1,0,-1,-4,-4;
MA   /I: MI=-55; MD=-55; I=-5;
MA   /M: SY='P'; D=-5; M=5,-32,-2,-1,-33,-2,-2,-19,3,-27,-15,2,28,3,4,7,3,-15,-48,-44;
MA   /M: SY='G'; M=3,-35,6,3,-38,18,7,-20,-10,-29,-21,6,-4,4,-11,3,-4,-13,-56,-36;
..
...
....  Lot of lines omitted.
...
..
MA   /M: SY='T'; M=7,-20,4,2,-33,0,-8,-8,1,-24,-12,5,0,-2,-6,10,14,-10,-49,-30;
MA   /M: SY='V'; M=5,-20,-14,-24,-18,-3,-21,20,-21,4,6,-11,-8,-16,-23,-4,3,24,-58,-32;
MA   /M: SY='P'; M=9,-30,-6,-2,-45,-5,2,-20,-9,-25,-20,-2,50,5,-1,8,2,-14,-55,-46;
MA   /M: SY='K'; M=-11,-52,1,-1,-1,-17,2,-18,43,-28,3,9,-10,8,33,-2,-1,-23,-33,-43;
MA   /I: MI=*; MD=*; I=0;
NR   /RELEASE=29,38303;
NR   /TOTAL=117(116); /POSITIVE=117(116); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0(0);
CC   /TAXO-RANGE=??EP?; /MAX-REPEAT=2;
DR   P06904; CRAA_ALLMI, T; P02482; CRAA_ARTJA, T; P02474; CRAA_BALAC, T;
DR   P02470; CRAA_BOVIN, T; P02487; CRAA_BRAVA, T; P02472; CRAA_CAMDR, T;
DR   P02473; CRAA_CANFA, T; P02491; CRAA_CAVPO, T; P02479; CRAA_CERSI, T;
DR   P02504; CRAA_CHICK, T; P02486; CRAA_CHOHO, T; P02503; CRAA_DIDMA, T;
..
...
....  Lot of lines omitted.
...
..
DR   P19752; HS30_NEUCR, T; P29778; OV21_ONCVO, T; P29779; OV22_ONCVO, T;
DR   P29209; IBPA_ECOLI, T; P29210; IBPB_ECOLI, T; Q03928; HS18_CLOAB, T;
DR   Q06823; SP21_STIAU, T; P12812; P40_SCHMA , T;
DR   P30220; HS3E_XENLA, P;
DO   PDOC00791;
//
```

Figure 2. Example of a profile entry

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

2'-5'-oligoadenylate synthetases signatures
3-hydroxyisobutyrate dehydrogenase signature
6-hydroxy-D-nicotine oxidase and reticuline oxidase FAD-binding
6-pyruvoyl tetrahydropterin synthase signatures
ABC-2 type transport system integral membrane proteins signature
Acyl-CoA-binding protein signature
ADP-glucose pyrophosphorylase signatures
ADP-ribosylation factors family signature
Adrenodoxin family, iron-sulfur binding region signature
Alanine dehydrogenase and pyridine nucleotide transhydrogenase

Alpha-isopropylmalate and homocitrate synthases signatures
Antenna complexes alpha and beta subunits signatures
Aspartate and glutamate racemases signatures
Bacterial export FHIPEP family signature
Bacterial formate and nitrite transporters signatures
Bacterial regulatory proteins, arsR family signature
Bacterial regulatory proteins, deoR family signature
Bacterial type II secretion system protein D signature
Bacterial type II secretion system protein F signature
Bacterial-type phytoene dehydrogenase signature
Beta-eliminating lyases pyridoxal-phosphate attachment site
BTG1 family signature
Calsequestrin signatures
CAP-Gly domain signature
Carbamoyl-phosphate synthase subdomain signatures
Chaperonins clpA/B signatures
Clathrin adaptor complexes medium chain signatures
Clathrin adaptor complexes small chain signature
Coproporphyrinogen III oxidase signature
Cyclin-dependent kinases regulatory subunits signatures
Cys/Met metabolism enzymes pyridoxal-phosphate attachment site
Cysteine synthase/cystathionine beta-synthase P-phosphate
Cytidine & deoxycytidylate deaminases zinc-binding region signature
Cytochrome c and c1 heme lyases signatures
Cytochrome c oxidase assembly factor COX10/ctaB/cyoE signature
Cytochrome c oxidase subunit VB, zinc binding region signature
D-alanine--D-alanine ligase signatures
Dehydroquinase class I active site
Dehydroquinase class II signature
Deoxyribonuclease I signatures
Dihydroorotate dehydrogenase signatures
Dihydroxy-acid and 6-phosphogluconate dehydratases signatures
DnaA protein signature
Dps protein family signatures
Elongation factor 1 beta/beta'/delta chain signatures
Ependymins signatures
Epimorphin family signature
ER lumen protein retaining receptor signatures
Ergosterol biosynthesis ERG4/ERG24 family signatures
Erythropoietin signature
Eukaryotic initiation factor 4E signature
Eukaryotic RNA polymerases 15 subunits signature
Extracellular proteins SCP/Tpx-1/Ag5/PR-1/Sc7 signatures
FAD-dependent glycerol-3-phosphate dehydrogenase signatures
Folylpolyglutamate synthase signatures
Fungal hydrophobins signature
G10 protein signatures
Galanin signature
Gamma-thionins family signature
Globins profile
glpT family of transporters signature
Glucoamylase active site region signature
Glutamate 5-kinase signature
Glycine radical signature
Glycoprotease family signature
Glycosyl hydrolases family 25 active sites signature
Glycosyl hydrolases family 39 putative active site
Glycosyl hydrolases family 8 signature
Glyoxalase I signatures
GTP cyclohydrolase I signatures
GTP1/OBG family signature
Guanylate kinase signature
Heat shock hsp20 proteins family profile
HIT family signature
Hypothetical YCR59c/yigZ family signature
Imidazoleglycerol-phosphate dehydratase signatures
Indoleamine 2,3-dioxygenase signatures
Initiation factor 3 signature
Interleukins -4 and -13 signature
LacY family proton/sugar symporters signatures
Ly-6 / u-PAR domain signature
Lysyl oxidase putative copper-binding region signature
MAM domain signature
Mandelate racemase / muconate lactonizing enzyme family signatures

Mannitol dehydrogenases signature
MARCKS family signatures
MCM2/3/5 family signature
mutT domain signature
Myristoyl-CoA:protein N-myristoyltransferase signatures
NAD-dependent glycerol-3-phosphate dehydrogenase signature
Neuromedin U signature
Neutrophil bactenecins signatures
Nickel-dependent hydrogenases b-type cytochrome subunit signatures
Nitrilases / cyanide hydratase signatures
Nuclear transition protein 2 signatures
OHHL biosynthesis luxI family signature
Oleosins signature
Orn/DAP/Arg decarboxylases family 2 signatures
Osteopontin signature
Oxysterol-binding protein family signature
Peripherin / rom-1 signature
Phosphatidylinositol 3-kinase signatures
Phosphoglucomutase & phosphomannomutase phosphoserine signature
Phosphomannose isomerase type I signatures
Photosystem I psaG and psaK proteins signature
Pollen proteins Ole e I family signature
Prephenate dehydratase signatures
Prokaryotic transcription elongation factors signatures
Prokaryotic transglycosylases signature
Proteasome B-type subunits signature
Protein phosphatase 2A regulatory subunit PR55 signatures
Protein phosphatase 2C signature
Protein prenyltransferases alpha subunit repeat signature
Protein splicing signature
PTR2 family proton/oligopeptide symporters signatures
Renal dipeptidase active site
Ribosomal protein L1e signature
Ribosomal protein L20 signature
Ribosomal protein L27 signature
Ribosomal protein L35 signature
Ribosomal protein L36 signature
Ribosomal protein S2 signatures
Ribosomal protein S21e signature
Ribosomal protein S28e signature
Ribosomal protein S7e signature
SAR1 family signature
Serine proteases, omptin family signatures
Serum amyloid A proteins signature
Signal peptidases II signature
Sodium:alanine symporter family signature
Sodium:galactoside symporter family signature
Spermadhesins family signatures
Streptomyces subtilisin-type inhibitors signature
Succinate dehydrogenase cytochrome b subunit signatures
Syndecans signature
Transaldolase active site
Transcription termination factor nusG signature
Translationally controlled tumor protein signatures
Transposases, Mutator family, signature
Trehalase signatures
Ubiquitin carboxyl-terminal hydrolases family 2 signatures
Uroporphyrin-III C-methyltransferase signatures
Uroporphyrinogen decarboxylase signatures
Urotensin II signature
XPGC protein signatures
Yeast PIR proteins repeats signature