*Published online 14 November 2007* *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D245–D249***
*doi:10.1093/nar/gkm977*

# The 20 years of PROSITE

**Nicolas Hulo[1,*], Amos Bairoch[1,2], Virginie Bulliard[1], Lorenzo Cerutti[3], Béatrice A. Cuche[1], Edouard de Castro[1], Corinne Lachaize[1], Petra S. Langendijk-Genevaux[1] and Christian J. A. Sigrist[1]**

[1]Swiss Institute of Bioinformatics (SIB), Centre Medical Universitaire, [2]Structural Biology and Bioinformatics Department, University of Geneva, 1 rue Michel Servet, CH-1211 Geneva 4 and [3]Swiss Institute of Bioinformatics (SIB), Génopode, UNIL-Sorge, CH-1015 Lausanne, Switzerland

Received September 14, 2007; Revised October 17, 2007; Accepted October 18, 2007



## ABSTRACT

**PROSITE consists of documentation entries describing protein domains, families and functional sites, as well as associated patterns and profiles to identify them. It is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids. In this article, we describe the implementation of a new method to assign a status to pattern matches, the new PROSITE web page and a new approach to improve the specificity and sensitivity of PROSITE methods. The latest version of PROSITE (release 20.19 of 11 September 2007) contains 1319 patterns, 745 profiles and 764 ProRules. Over the past 2 years, about 200 domains have been added, and now 53% of UniProtKB/Swiss-Prot entries (release 54.2 of 11 September 2007) have a PROSITE match. PROSITE is available on the web at: http://www.expasy.org/prosite/.**

## HISTORICAL BACKGROUND

PROSITE, initially a 'signature' or pattern database, was created in 1988 by Amos Bairoch. It was first distributed through PC/Gene, the sequence analysis software suite he was developing at the time. The first release of PROSITE was made available in PC/Gene in March 1988 and contained 58 patterns. Each pattern was accompanied by an abstract that described the corresponding protein family or domain (1). PROSITE was developed in parallel with Swiss-Prot and both databases benefited from each other. Many patterns were identified by annotating protein families in Swiss-Prot (often before any description in the literature). The patterns were then used to populate Swiss-Prot with new family members. PROSITE generated an immediate interest and it then grew regularly to reach 1000 entries 6 years later. The PROSITE pattern syntax is adapted for short well-conserved regions. Such regions are typically enzyme catalytic sites, prosthetic group-attachment sites (haem, pyridoxal phosphate, biotin, etc.), metal ion-binding amino acids, cysteines involved in disulfide bonds or regions involved in binding a molecule. But this syntax is very sensitive to any sequence 'exception', whether due to a *bona fide* divergence or to a sequencing error. Patterns are thus not adapted to identify less-conserved regions or whole domains.

In 1994, Philipp Bucher introduced in PROSITE 'generalized profiles' as new motif descriptors (2). All profile methods are more or less statistical descriptions of the consensus of a multiple sequence alignment. They use position-specific scores for amino acids and position-specific penalties for opening and extending an insertion or deletion. 'Generalized profiles', compared to previous profiles (3), use a more rigorous syntax for insertion, deletion and match states. Since the 'generalized profile' syntax is very similar to the HMM profile one, nearly all 'generalized profile' scores can be mapped to HMM parameters used by *HMMER* (4). It is thus possible to convert an HMM profile into a 'generalized profile' format and several PROSITE profiles are in fact HMM profiles that were converted with the *pftools* program *htop* (5). Currently, nearly all new PROSITE entries are profiles.

Since its creation, PROSITE has provided extensive documentation and detailed annotation of domains, families and functional sites. This information was mainly stored in free text and used by biologists who read the various documents and made their own decisions on the function of their protein according to the PROSITE matches. But with the rapid growth of sequence databases during the last 10 years, there was an increasing need for a reliable tool that could generate automatically precise and accurate functional annotation in standard format. In 2005, we decided to group some functional information stored in PROSITE in a database of rules

*To whom correspondence should be addressed: Tel: +4122 379 58 72; Fax: +4122 379 58 58; Email: Nicolas.Hulo@isb-sib.ch

that can easily be read by a program and applied on proteins that are recognized by PROSITE profiles (6). We named this complementary database ProRule, for PROSITE Rules. ProRule generates a variety of annotation in Swiss-Prot format. The main characteristic of ProRule is that it generates conditional annotation: the annotation is dependent on the presence of given amino acids at precise positions, on the occurrence of other domains or on taxonomic specificity. This information is only transferred if all the conditions are fulfilled. For example, an enzymatic active site is annotated only if the correct amino acid is found at the required position (for an example of ProRule, see Figure 1). As ProRule uses PROSITE profiles that are mainly directed against protein domains, it is well adapted to annotate modular proteins. The Swiss-Prot group has also developed a complementary database of rules (HAMAP), which uses the same format of rules but which is specific for well-conserved bacterial protein families (7).

## ASSIGNING A STATUS TO PROSITE PATTERN MATCHES

A pattern or regular expression is a qualitative descriptor: it either matches or it does not. It does not produce a score that can help to estimate the significance of a match. There are currently various quantitative methods producing scores that are more efficient than regular expressions (8), but patterns are still very popular because of their intelligibility for users, and because, when used to scan protein databases, they are not CPU expensive compared to quantitative methods. We thus still maintain patterns, but to make them more accurate, we have developed a new method to estimate the significance of their matches. A profile has been constructed and associated to each pattern and used to assign a status to pattern matches. When a pattern matches a protein, the corresponding profile is run on this protein and, if it is also identified by the profile, the match is tagged as true positive, otherwise the status is unknown. The advantage of this approach is that it does not influence the rapidity of the search algorithm and it keeps the user-friendly format of patterns.

Based on the pattern match-list, 1309 profiles have been automatically generated. For each pattern, PROSITE maintains a manually curated match-list, in which a status is assigned to each pattern match on UniProtKB/Swiss-Prot entries. The status can be True Positive (TP), Unknown (?), False Positive (FP), False Negative (FN) or potential (P). For each pattern, we have extracted the sequence of each TP match. As the efficiency of a motif descriptor is partly dependent on the size of the sequences in the seed alignment (9), we have increased the length of each TP matched fragment. The average size of PROSITE patterns is 20 amino acids, and we noticed that it is difficult to construct good profiles smaller than 50 amino acids. We also did not want to increase too much the size of the profile as the scanning time of a profile is proportional to its size. Starting with multiple sequence alignments of an average size of 60 amino acids is thus a good compromise. We thus have increased each fragment by 20 amino acids on both sides. The sequences were then aligned with T-Coffee (10), the alignment was used to construct a profile for each pattern (hereafter miniprofile) and the miniprofile was then calibrated on a randomized protein database. If a PROSITE profile was already associated with a pattern and was of better quality than the automatically generated one, we used the PROSITE profile to assign the status.

Each automatically generated miniprofile was then tested on the corresponding pattern match-list and the cutoff calculated to recover all TP and no FP, and the maximum of FN. When a miniprofile did not recover all TP it was reconstructed with different profile construction parameters or different multiple sequence alignment programs (ClustalW and ProbCons) (11,12). When it was not possible to recover all TP, the cutoff was set just above the highest FP. The rule is that none of the miniprofile should recover UniProtKB/Swiss-Prot proteins tagged as FP with the corresponding pattern and indeed 95% of the profiles recover 100% of the TP and only 65 profiles do not recover all TP. Since a miniprofile is always run on a subset of the database (proteins matched by the pattern) the *e*-value is always very low ($<0.001$).

Even though miniprofiles were not designed to be scanned alone, and must be run on proteins matched by a pattern, 80% of the miniprofiles produce only matches with *e*-value bellow 0.01 (N_Score = 9) on UniProtKB. This subset of miniprofiles is thus safe enough to be run alone on a database of proteins.

In our search algorithm, miniprofiles are run only on proteins matched by patterns, and this procedure does not drastically impact the calculation time. It takes about 10 min on 1 CPU (Pentium 4, 3.4 GHz) to run all the PROSITE patterns on the whole *Escherichia coli* proteome (4339 protein sequences). The selection of the status option increases the calculation time by only 1 min. The status of pattern matches on UniProtKB/TrEMBL is accessible from the 'ScanProsite' web page. It is also possible to download the different tools to use it locally from the PROSITE ftp site (ftp://ftp.expasy.org/databases/prosite/tools/).

## POST-PROCESSING PROSITE MATCHES

The sensitivity and specificity of descriptors can be enhanced by taking into account some contextual information, such as the co-occurrence of other domains, the position of a match in a protein, the taxonomic distribution, etc. (13–15). Such information can be used to promote some weak matches or to demote some irrelevant strong matches. PROSITE profiles normally use two cutoff levels, a reliable cutoff (LEVEL = 0) and a low confidence cutoff (LEVEL = −1). The low-level cutoff covers the twilight zone where few true positives that cannot be separated from false positives, might be present. By default, only matches higher than the reliable cutoff are shown. We have added a post-processing step in our scanning procedure, which allows the display of weak

**[?] General information about the entry**

| | |
|---|---|
| *Accession* | PRU00298 |
| *Dates* | 12-DEC-2003 (Created)<br>26-JUL-2006 (Last updated, Version 10) |
| *Data class* | Domain |
| *Predictors* | PROSITE; PS50292; PEROXIDASE_3 |
| *Names* | Animal heme peroxidase |
| *Function* | Removal of H(2)O(2) |

**[?] Comments**

- **[case <FTGroup:2>]**
  - ***CATALYTIC ACTIVITY***: Donor + H(2)O(2) = oxidized donor + 2 H(2)O.
  - ***COFACTOR***: Binds 1 calcium ion per heterodimer (By similarity).
  - ***COFACTOR***: Binds 1 heme B (iron-protoporphyrin IX) group covalently per heterodimer (By similarity).
  - ***SIMILARITY***: Belongs to the peroxidase family.

**[?] Cross-references**

PROSITE — PS00435; PEROXIDASE_1; 1 **[case <FTGroup:2>]**
PS00436; PEROXIDASE_2; 1 **[case <FTGroup:2>]**

**[?] Keywords and Gene Ontology**

- GO:0016491; Molecular function: oxidoreductase activity. **[case <FTGroup:2>]**
- GO:0004601; Molecular function: peroxidase activity. **[case <FTGroup:2>]**
- **Keyword: Oxidoreductase [case <FTGroup:2>]**
- **Keyword: Peroxidase [case <FTGroup:2>]**
- **Keyword: Iron [case <FTGroup:1>]**
- **Keyword: Heme [case <FTGroup:1>]**
- **Keyword: Calcium [case <FTGroup:3>]**
- **Keyword: Metal-binding [case <FTGroup:1> or <FTGroup:3>]**

**[?] Features**

| Key | From | To | Description | Condition [?] | Cases |
|---|---|---|---|---|---|
| **From: PS50292** | | | | | |
| CHAIN | 2 | 108 | light chain | | |
| CHAIN | 109 | Cter | heavy chain | | |
| BINDING | 95 | 95 | Heme (covalent) *(By similarity)* | D | FTGroup[1] |
| BINDING | 238 | 238 | Heme (covalent) *(By similarity)* | E | FTGroup[1] |
| ACT_SITE | 96 | 96 | Proton acceptor *(By similarity)* | H | FTGroup[2] |
| SITE | 235 | 235 | Transition state stabilizer *(By similarity)* | R | FTGroup[2] |
| METAL | 97 | 97 | Calcium *(By similarity)* | D | FTGroup[3] |
| METAL | 164 | 164 | Calcium *(By similarity)* | [ST] | FTGroup[3] |
| METAL | 166 | 166 | Calcium (via carbonyl oxygen) *(By similarity)* | [FY] | FTGroup[3] |
| METAL | 168 | 168 | Calcium *(By similarity)* | D | FTGroup[3] |
| METAL | 170 | 170 | Calcium *(By similarity)* | [ST] | FTGroup[3] |
| METAL | 332 | 332 | Iron (heme axial ligand) *(By similarity)* | H | |
| DISULFID | 4 | 17 | *By similarity* | C-x*-C | |
| DISULFID | 111 | 121 | *By similarity* | C-x*-C | |
| DISULFID | 115 | 139 | *By similarity* | C-x*-C | |
| DISULFID | 217 | 228 | *By similarity* | C-x*-C | |
| DISULFID | 435 | 491 | *By similarity* | C-x*-C | |
| DISULFID | 534 | 560 | *By similarity* | C-x*-C | |

**[?] Characteristics**

- Size range: 510-750 amino acids
- Related UniRules: None
- Repeats: 1-6
- Topology: Undefined
- Example: P11247

**Figure 1.** An example of a ProRule in 'niceview' format from the PROSITE web page showing the different types of annotation that can be generated by ProRule. The rule PRU00298 is used to annotate proteins matched by the animal peroxidase profile (PS50292) on the ScanProsite web page. It can annotate comment lines, KW, GO terms and various FT lines. In this rule, all types of annotation are conditionals. For example, the comment line 'catalytic activity' is generated only if the condition FTGroup(2) is fulfilled (a H in the sequence must align with position 96 of the profile and a R with position 235). The numbers in the 'From' and 'To' column in the features tables correspond to specific columns in the profile (for more details on the ProRule format see: ftp://ftp.expasy.org/databases/prosite/unirule.pdf).

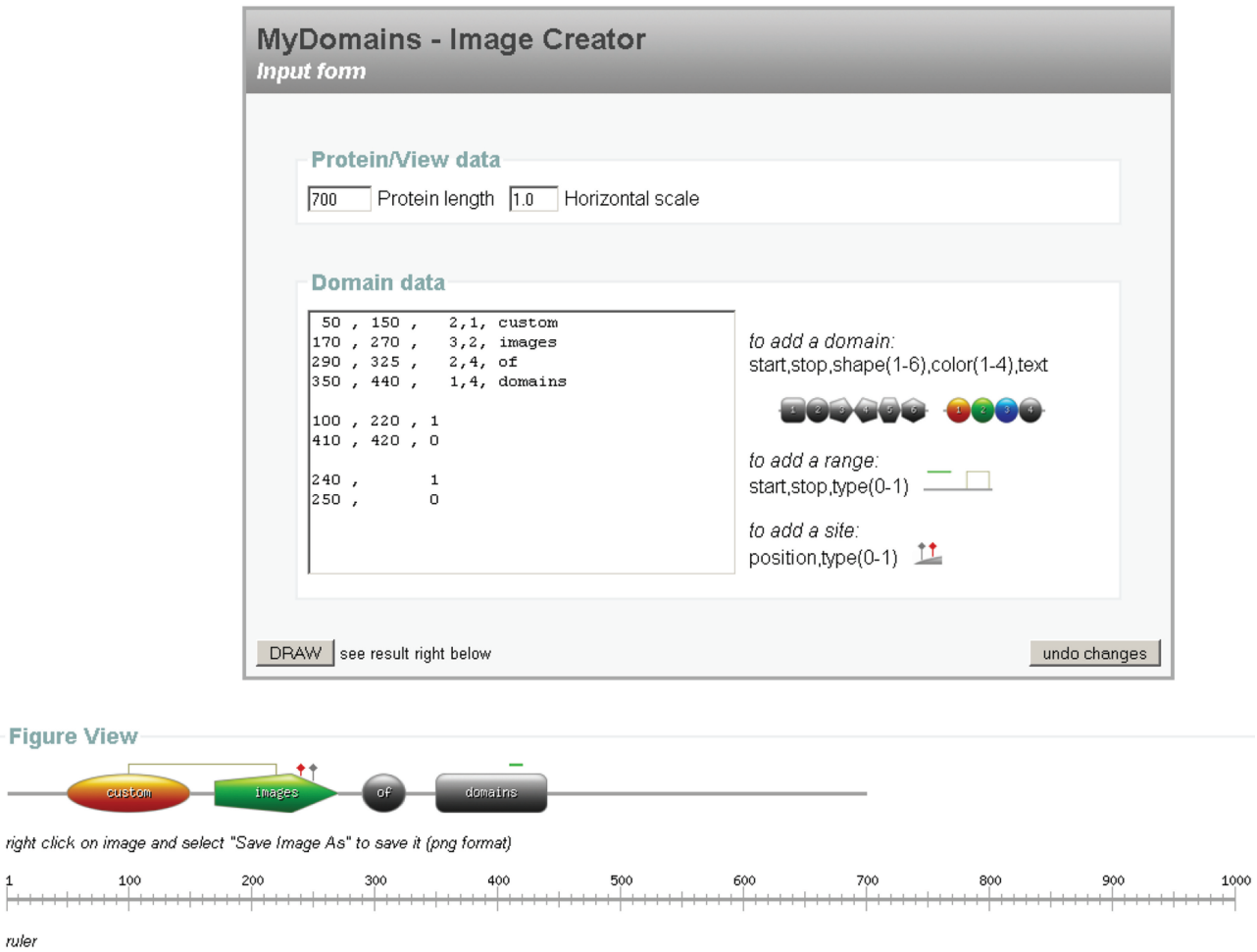


**Figure 2.** The web form and the output of the PROSITE 'MyDomains' image creator tool. A very simple syntax allows the user to define the shape, colour, size and name of one or several domains. Specific residues and ranges can also be marked.

matches or the masking of strong ones according to the occurrence of other specific features in the protein. The required features for the post-processing step are stored in a new line type in the profile (PP). We have defined three types of post-processing:

(i) When at least two matches overlap, only the one that has the highest score is reported. This is mainly used when two or more families are very closely related, like the different types of HTH DNA-binding domains, or when it makes sense to define subfamilies of a larger protein family; for example the ABC transporter family was subdivided into subfamilies to predict the type of transported substrate. A PROSITE profile can compete with one or several other profiles. The format is:
`PP /COMPETES_HIT_WITH:` *PS-accession*`;`

(ii) A weak match is promoted by the presence of another PROSITE profile in the protein. This may happen, for example, when two domains are known to be frequently associated, like for example the ATP-binding helicase domain (PS51192) and the C-terminal helicase domain (PS51194). A PROSITE profile can be promoted by one or several other profiles. The format is:
`PP /PROMOTED_BY:` *PS-accession*`;`

(iii) A strong match is demoted by the presence of another domain in the protein. A PROSITE profile can be demoted by one or several other profiles. The format is:
`PP /DEMOTED_BY:` *PS-accession*`;`

In a given entry, different PP line types can be combined.

## WEB PAGE DEVELOPMENT

The PROSITE web page has been redesigned and new functionalities have been implemented. PROSITE can now be browsed by taxonomic scope, by ProRule description, by the number of positive hits or by matched proteins. We have reorganized the data presentation, which are now grouped into five different sections ('ScanProsite', ProRule, Documents, Downloads and Links) besides the home page. A new ProRule section has been created, which allows the visualization of the different rules that are used to generate annotation on the 'ScanProsite' web page (Figure 1).

The ftp site has also been reorganized, and new files can now be downloaded. As we have introduced PROSITE version numbers for each pattern/profile entries, we now distribute the old versions of PROSITE to allow the recovery of previous versions of entries. We also distribute a file that contains all the multiple sequence alignments of matched regions by patterns and profiles on UniProtKB/Swiss-Prot database.

We have made available to PROSITE users our tool that generates domain images to represent protein architectures. For a given protein, the user can enter in a web form the size of the protein, the positions of the domains, their names and, for each domain, the colour and the shape of the wanted image. The web form returns an image in Portable Network Graphics (png) format that can be integrated into any publication (Figure 2). The tool is accessible at the following address: http://www.expasy.org/tools/mydomains/.

## PROGRAMMATIC ACCESS TO 'SCANPROSITE'

The ScanProsite tool can be accessed programmatically through a simple web HTTP service where the naked data (without any message exchange envelope) is retrieved directly as the content of an HTTP query response ('low' REST service).

When a client sends an HTTP GET or POST query to the service; the response content will contain the results in XML or in the lightweight data-interchange format json (JavaScript Object Notation).

For details see http://www.expasy.org/tools/scanprosite/ScanPrositeREST.html.

Note: to avoid timeout problems with long jobs, we will soon introduce a queuing system.

It is also possible to access PROSITE entries in raw text/plain format using the following url: http://www.expasy.org/cgi-bin/get-prosite-raw.pl?PDOC-accession or PROSITE alignments: http://www.expasy.org/cgi-bin/aligner?psa = PS-accession.

## ACKNOWLEDGEMENTS

We wish to thank Tania Lima for critical reading of the manuscript and Laurent Falquet for the design of the ProRule logo. This work was funded by an FNS project grant (315200-116864). PROSITE activities are also supported by the Swiss Federal Government through the Federal Office of Education and Science. Funding to pay the Open Access publication charges for this article was provided by the FNS.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Bairoch,A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.*, **19**, 2241–2245.
2. Bairoch,A. and Bucher,P. (1994) PROSITE: recent developments. *Nucleic Acids Res.*, **22**, 3583–3589.
3. Gribskov,M., Luthy, and R. Eisenberg,D. (1990) Profile analysis. *Meth. Enzymol.*, **183**, 146–159.
4. Bucher,P., Karplus,K., Moeri,N. and Hofmann,K. (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem.*, **20**, 3–23.
5. http://www.isrec.isb-sib.ch/ftp-server/pftools/pft2.3/
6. Sigrist,C.J.A., De Castro,E., Langendijk-Genevaux,P.S., Le Saux,V., Bairoch,A. and Hulo,N. (2005) ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics*, **21**, 4060–4066.
7. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J.A. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
8. Hofmann,K. (2000) Sensitive protein comparisons with profiles and hidden Markov models. *Brief. Bioinform.*, **2**, 167–178.
9. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
10. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
11. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
12. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
13. Coin,L., Bateman,A. and Durbin,R. (2003) Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl Acad. Sci. USA.*, **100**, 4516–4520.
14. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
15. Coin,L., Bateman,A. and Durbin,R. (2004) Enhanced protein domain discovery using taxonomy. *BMC Bioinformatics*, **5**, 56–66.