New and continuing developments at PROSITE

Christian J. A. Sigrist1,*, Edouard de Castro1, Lorenzo Cerutti1, Béatrice A. Cuche1, Nicolas Hulo1, Alan Bridge1, Lydie Bougueleret1 and Ioannis Xenarios1,2

1SIB Swiss Institute of Bioinformatics, Centre Médical Universitaire (CMU), 1 rue Michel Servet, 1211 Geneva 4 and 2SIB Swiss Institute of Bioinformatics, Vital-IT Group, Quartier Sorge, Bâtiment Génopode, 1015 Lausanne, Switzerland

Received September 20, 2012; Revised and Accepted October 11, 2012

ABSTRACT

PROSITE (http://prosite.expasy.org/) 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, which increases the discriminatory power of these profiles and patterns by providing additional information about functionally and/or structurally critical amino acids. PROSITE signatures, together with ProRule, are used for the annotation of domains and features of UniProtKB/Swiss-Prot entries. Here, we describe recent developments that allow users to perform whole-proteome annotation as well as a number of filtering options that can be combined to perform powerful targeted searches for biological discovery. The latest version of PROSITE (release 20.85, of 30 August 2012) contains 1308 patterns, 1039 profiles and 1041 ProRules.

INTRODUCTION

PROSITE is a resource for the identification and annotation of conserved regions in protein sequences. These regions are identified using two types of signatures: generalized profiles (weight matrices) that describe protein families and modular protein domains and patterns (regular expressions) that describe short sequence motifs often corresponding to functionally or structurally important residues (1). PROSITE signatures are linked to annotation rules, or ProRules, which define protein sequence annotations (such as active site and ligand-binding residues) and the conditions under which they apply (for example requiring specific amino acid residues) (2). ProRule is used for the annotation of protein families, domains and sequence features in UniProtKB/Swiss-Prot, the manually curated section of the UniProt KnowledgeBase (3), and currently provides annotation for >75% of the 1054 domains to be found there (release 2012_08, 5 September 2012). Part of the information stored in ProRule (e.g. active and binding sites, disulfide bonds) is also accessible to the ScanProsite user. PROSITE provides extensive documentation for each signature including information on nomenclature, function, sequence features, pointers to 3D structure(s), protein architectures in which the signature is found, its taxonomic distribution and important literature references (1). PROSITE signatures, ProRules and PROSITE documentation can be accessed from our website at http://prosite.expasy.org/ (4). PROSITE signatures are also made available through InterPro (http://www.ebi.ac.uk/interpro/index.html), an integrated database of protein signatures used for the classification and annotation of proteins and genomes (5). Through InterPro users can combine PROSITE classifications with those provided by other InterPro consortium members. Since our last report in the NAR database issue (6), PROSITE has increased the number of available signatures to 1308 patterns and 1039 profiles, which are associated with 1041 ProRules and 1650 documentation entries.

NEW DEVELOPMENTS: SCANPROSITE

The ScanProsite tool (http://prosite.expasy.org/scanprosite/) allows users to search protein sequences against all PROSITE signatures, and to search for matches to defined PROSITE signatures in the UniProtKB and PDB databases (4). PROSITE signatures are also made available through InterPro (http://www.ebi.ac.uk/interpro/index.html), an integrated database of protein signatures used for the classification and annotation of proteins and genomes (5). Through InterPro users can combine PROSITE classifications with those provided by other InterPro consortium members. Since our last report in the NAR database issue (6), PROSITE has increased the number of available signatures to 1308 patterns and 1039 profiles, which are associated with 1041 ProRules and 1650 documentation entries.

*To whom correspondence should be addressed. Tel: +41 22 379 58 68; Fax: +41 22 379 58 58; Email: prosite-group@isb-sib.ch

Present address:
Nicolas Hulo, Service for biomathematical and biostatistical analyses, Institute of Genetics and Genomics in Geneva, Centre Médical Universitaire (CMU), University of Geneva, 1 rue Michel Servet, 1211 Geneva 4, Switzerland.

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.
entire library of PROSITE signatures. We have since
relaxed this restriction and now offer users the possibility
to upload complete proteome sets in FASTA format to the
PROSITE server (subject to a size limitation of 16 Mb,
which is sufficient for the majority of proteomes).
A unique identifier is assigned to each uploaded set of
protein sequences and is returned to the user as a reference
for use in subsequent searches. The identifier remains valid
for 1 month, allowing users to perform multiple analyses
on the same set of sequences, if desired. These analyses are
performed on the high-performance cluster of the Vital-IT
facility (http://www.vital-it.ch/). It is possible to perform
combinatorial scans (see below), and users can perform searches against their own defined sequence patterns.

To demonstrate this application, we annotated the
complete proteome sequence of the fire ant Solenopsis
invicta at the ScanProsite server (7). The Official Gene
Set of S. invicta is predicted to encode 16,569 canonical
protein sequences. These were uploaded to the
ScanProsite server in FASTA format and run against all
PROSITE motifs, including both patterns and profiles.
The entire process took <30 min. A total of 14,562 hits
to 1248 distinct PROSITE signatures were found in 5496
protein sequences, giving total coverage at the protein
level of ~33% for this organism (Table 1). Users
wishing to obtain higher coverage may of course
combine the classification and annotation from
PROSITE with that provided by other annotation tools
and pipelines.

Combinatorial search
In parallel to this work, we have developed and imple-
mented a number of search options that enhance the
power and flexibility of ScanProsite. The first of these
allows users to search for specific combinations of signa-
tures. This feature may be useful in fine-grained functional
inference, allowing users to search a given set of sequences
for instances of domains (profiles) that are associated with
particular functional residues (patterns) or to search for
specific combinations of domains that may confer particu-
lar functions (9,10). PROSITE descriptors are combined
using the logical operators ‘and’, ‘or’ and ‘not’, with
parentheses used to define the priority in which the oper-
ators are applied (Figure 1). Users can also define their
own sequence patterns and combine them with existing
PROSITE signatures. This may allow the further
discrimination of particular domain variants or
subfamilies that are not yet covered by existing
PROSITE signatures (2).

Targeted search with filters
The results of PROSITE searches on UniProtKB can be
further restricted using a variety of filtering options. Users
can limit the results to only those proteins that derived
from one or more taxa, according to the taxonomic clas-
sification of UniProtKB (http://www.uniprot.org/
taxonomy/), at any desired level in the taxonomy.
Taxonomic information is found in the ‘OC’ and ‘OS’
line(s) of the UniProtKB flat file. Users can also limit
their results to only those proteins having a particular
name (be it the recommended name or alternative
name), which can be a general class of protein such as
‘protease.’ Such nomenclature information is found in the ‘DE’ line(s) of the UniProtKB flat file. Users can also
limit their results to only those proteins that are
expressed in one of 56 adult tissues, using data from the
Bgee resource (http://bgee.unil.ch/bgee/bgee), a database
of gene expression and evolution (11). This particular filter
is applicable to proteins of Homo sapiens, Mus musculus,
Xenopus laevis and Danio rerio. Finally, users can also
limit their results to only those proteins having a certain
size or within a certain size range. Together, these filters
allow users to combine prior biological knowledge with
specific sequence features (or combinations of them) in
order to perform very powerful targeted searches.

We illustrate a typical application of these search
options using the alkylglycerol mono-oxygenase of
M. musculus as an example (12). Prior to the identifica-
tion of the sequence encoding this enzyme, a limited amount of
information was available regarding its biological and bio-
chemical characteristics. We used this information to
identify a number of possible candidate sequences for
experimental validation. It was known that this enzyme,
along with nitric oxide synthase and aromatic amino acid
hydroxylase, required tetrahydrobiopterin and iron to be
active. The enzyme was also known to have similar
iron-binding characteristics to aromatic amino acid
hydroxylase, suggesting a role for histidine residues in
this process (13). The protein was known to be present
in brain and liver, and its size was estimated at between
400 and 650 amino acids (14). We used this information to
perform a restricted search of the murine proteome using

Table 1. Results of the ScanProsite search of the 16,569 predicted Solenopsis invicta proteins against the complete set of PROSITE patterns and profiles

<table>
<thead>
<tr>
<th>Description</th>
<th>Patterns</th>
<th>Profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of PROSITE signature matches in all proteins</td>
<td>4903</td>
<td>9664</td>
</tr>
<tr>
<td>Number of distinct proteins matching PROSITE signatures</td>
<td>2696</td>
<td>4349</td>
</tr>
<tr>
<td>Number of distinct PROSITE signatures matched</td>
<td>626</td>
<td>622</td>
</tr>
<tr>
<td>Number of proteins annotated with one or more functional sites</td>
<td>520</td>
<td>1693</td>
</tr>
<tr>
<td>Total number of functional sites annotated</td>
<td>744</td>
<td>7022</td>
</tr>
<tr>
<td>Number of distinct PROSITE signatures providing annotation for functional sites</td>
<td>74</td>
<td>148</td>
</tr>
<tr>
<td>Total number of detected domains annotated with functional sites</td>
<td>606</td>
<td>3397</td>
</tr>
</tbody>
</table>

*Pattern hits are validated by automatically generated ‘miniprofiles’ that assign a status to pattern matches (8).
a degenerate pattern corresponding to the two iron-coordinating histidines of aromatic amino acid hydroxylase (H–X(3,5)–H). This reduced the list of UniProtKB protein entries matching this motif from over 1000 to only 31, corresponding to 22 genes. Following manual inspection of these sequences, we excluded a number of previously characterized proteins that were unlikely to be responsible for the specified activity, including transcription factors, transporters and enzymes. The remaining set of 16 proteins constituted a reasonable number of candidate sequences for experimental investigation. One of these was found to possess alkylglycerol mono-oxygenase activity, and this is described in UniProtKB/Swiss-Prot entry Q8BS35.

CONCLUSION

PROSITE provides a resource for the identification and annotation of conserved regions in protein sequences, covering protein families, domains and motifs. We will continue to develop new PROSITE profiles and ProRules as new proteins, domains and functions are characterized. We describe here improvements to ScanProsite that permit PROSITE to be applied by users for whole-proteome annotation, as well as a number of options that allow very fine-grained searches including prior biological knowledge. Our current software developments are addressed at further enhancing the speed of ScanProsite for improved proteome annotation. To achieve this, the original code of psfsearch is being rewritten and optimized to efficiently use modern multi-core processors and an heuristic implemented for further speed enhancements. This work will be described in a forthcoming publication (L. Cerutti and T. Schuepbach, personal communication).

ACKNOWLEDGEMENTS

The authors thank Frédéric Bastian for helpful discussions regarding gene expression library databases and for the selection of multi-species tissues that are used in ScanProsite.

FUNDING

An FNS project [315230-116864]. PROSITE activities are also supported by the Swiss Federal Government through the Federal Office of Education and Science. Funding for open access charge: Swiss Federal Office of Education and Science.

Conflict of interest statement. None declared.
REFERENCES


