APPLICATION NOTE

ScanProsite: a reference implementation of a PROSITE scanning tool

Alexandre Gattiker, Elisabeth Gasteiger, Amos Bairoch

Swiss Institute of Bioinformatics, Geneva, Switzerland

Abstract: Many different software tools are available publicly to scan the PROSITE database of protein families. However, none of them, to our knowledge, wholly implements the PROSITE syntax, or satisfies all the rules for scanning a pattern against a sequence. We hereby propose a strict definition of how a PROSITE pattern is to be scanned against a sequence, and provide a reference implementation of a tool to scan PROSITE patterns, rules and profiles against protein sequences.

Keywords: sequence analysis, pattern, regular expression, profile

AUTHOR website version

Availability: Licensed under the GNU Public License, at ftp://ftp.expasy.org/databases/prosite/tools/ps_scan/. Online version at http://www.expasy.org/tools/scanprosite
Contact: prosite@isb-sib.ch

The PROSITE database (http://www.expasy.org/prosite) of protein families and domains consists of motif descriptors for biologically significant sites and regions (Sigrist et al 2002). The database contains three types of motifs: patterns (regular expressions), rules and profiles. Rules are logical assertions written in English that complement or replace a pattern to identify sites that cannot be described by the simple regular expression syntax of pattern. Profiles are an extension of weight matrices. An implementation of profile scanning tools written in Fortran is provided in the PFTOOLS package (Bucher et al 1996), which is the reference implementation for profiles. On the other hand, strict rules on how patterns and rules are to be applied to sequences have never been defined in the PROSITE user manual and documents. Therefore, implementations have diverged on the extent of their coverage of the PROSITE syntax, and on several critical aspects of the implementation of the pattern scanning engine.

PROSITE patterns describe sites on protein sequences using a simple regular expression syntax. Tokens can be fixed amino acids, ambiguous amino acids ([ST] for Ser or Thr, {AM} for any amino acid except Ala or Met, x for any amino acid), or anchors at the start (<) or end (>) of the sequence; and tokens for amino acids may be postfixed by a fixed- or variable-length range. For example, the pattern '<A-x-[ST](2)-x(0,1)-V' matches the subsequence Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val at the N-terminus of a sequence. Although these specifications appear quite simple to follow, they are not fully implemented by all packages.

PROSITE contains four entries of the type 'RULE' describing three post-translational modifications and a bipartite nuclear targeting sequence. Since the text of the rules is written in free-text English, the only way to scan these rules against a sequence is to implement the rules directly in the program code.

An issue encountered with patterns is that the PROSITE documentation does not specify how variable-length residue stretches are to be matched when there are several possible alignments of the sequence and the pattern ('greediness'), nor whether overlapping pattern matches should count as a single or as multiple hits. The setting of these parameters can alter the number and position of reported matches of a pattern on a sequence. In our program, this setting can be controlled by the user, and the default is to be greedy and allow overlaps which are not completely included in another match. Another issue lies in the treatment of sequence ambiguities: in some sequences in SWISS-PROT and other databases, the character X is used in place of an unknown residue, and the characters B and Z are used when a residue may be either Asp or Asn, or Glu or Gln, respectively. Our implementation handles the B and Z characters, and allows at most one X in the sequence to match one conserved position in a pattern, in order to identify more matches.

The ps_scan program and the ScanProsite online application provide a reference implementation of a PROSITE scanning tool. Together with the PROSITE user

Correspondence: Elisabeth Gasteiger, Swiss Institute of Bioinformatics, 1 rue Michel-Servet, CH-1211 Geneva 4, Switzerland; tel +41 22 702 5875; fax +41 22 702 5858; email Elisabeth.Gasteiger@isb-sib.ch

© 2002 Open Mind Journals Limited. All rights reserved.

Table 1 Comparison of the output of several PROSITE scanning tools[a]

PROSITE entry	*PS01253 (Fibronectin pattern)*	*PS00539 (Pyrokinin pattern)*	*PS00262 (Insulin pattern)*	*PS00003 (Tyr-sulfation rule)*	*PS00013 (Lipoprotein rule)*
SWISS-PROT sequence	*FINC_MOUSE (P11276)*	*PHPT_PSESE (P25271)*	*INS_KATPE (P01340)*	*CAE5_XENLA (P05225)*	*NLPB_ECOLI (P21167)*
InterProScan	3+1 wrong	1	none	None	None
Patsearch	3	None	none	None	None
PROSCAN	3	None	none	None	None
PROSCAN/1	23	None	1	None	None
Old ScanProsite	3	1	none	None	1
ScanProsite	4	1	1	4	1

[a] Number of reported matches between the specified PROSITE entry and SWISS-PROT sequence.
InterProScan = http://www.ebi.ac.uk/interpro/scan.html
Patsearch = HITS server, http://hits.isb-sib.ch/cgi-bin/hits_protein_query
Old ScanProsite = http://www.expasy.org/tools/scanprosite.old/scnpsit1.html
PROSCAN = PBIL Prosite scanner, http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_prosite.html
PROSCAN/1 = PROSCAN allowing 1 mismatch
ScanProsite = http://www.expasy.org/tools/scanprosite.html
All websites accessed 25 June 2002.

manual (http://www.expasy.org/prosite/prosuser.html) it aims to clarify ambiguities regarding the implementation of a PROSITE regular expression matching engine, implement the rules described in PROSITE, handle ambiguous sequence characters in protein databases and provide a single user interface to scan PROSITE patterns, rules and profiles as well as user-defined patterns against one or more sequences. It also provides sequence-to-sequence alignments guided by patterns and profiles (where all match positions in all sequences are aligned together like in a multiple sequence alignment, by extending all inserts with dots up to the maximal insert length), in addition to sequence-to-pattern alignments. The alignments are given in an attractive format, with residues corresponding to unspecified regions such as 'X(2,4)' in lowercase, and match positions corresponding to a deleted protein residue indicated with dashes. The program is implemented in Perl and provided in two separate files: a module Prosite.pm containing functions which can be reused by any Perl program, and an application called ps_scan.pl which can be called from a DOS or UNIX command line to scan one or more sequences against one or more patterns, rules and/or profiles and report the hits in a range of formats.

ScanProsite in its new implementation considerably differs from the older version previously available, as it has a completely remodelled interface and code, and many more options and output formats. A comparison of the results of ScanProsite and 5 other online PROSITE scanning tools in selected examples is shown in Table 1. No other implementation supports rules. Most of the other tools do not support the pyrokinin signature 'F-[GSTV]-P-R-L-[G>]', and none detected the ambiguous amino acid B in the insulin signature of INS_KATPE or identified the fourth apparent fibronectin type I domains in FINC_MOUSE where one conserved cysteine is occupied by the residue X. It was possible in those two latter cases to retrieve the matches (along with a huge number of false positive matches to other PROSITE entries) by running the PROSCAN tool and allowing 1 mismatch. However, in the case of fibronectin the tool reported many overlapping matches.

References

Bucher P, Karplus K, Moeri N, Hofmann K. 1996. A flexible motif search technique based on generalized profiles. *Comput Chem*, 20:3–24.

Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. 2002. PROSITE, a documented database using patterns and profiles as motif identification tools. *Brief Bioinform*, 3. Forthcoming.