Nucleic Acids Research, 2025, gkaf1188
https://doi.org/10.1093/nar/gkaf1188
Database issue

The PROSITE database for protein families, domains, and sites

Christian J. A. Sigrist*, Béatrice A. Cuche, Edouard de Castro, Elisabeth Coudert, Nicole Redaschi, Alan Bridge

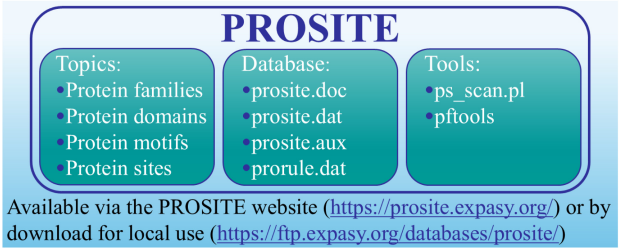
Swiss-Prot Group, Swiss Institute of Bioinformatics (SIB), Centre Médical Universitaire (CMU), 1 rue Michel Servet, CH-1211 Geneva 4, Switzerland

*To whom correspondence should be addressed. Email: christian.sigrist@sib.swiss, prosite@expasy.org

Abstract

PROSITE (https://prosite.expasy.org/) is a database of entries documenting protein domains, families, and functional sites, along with the associated patterns and profiles used to identify them. It is complemented by ProRule, a rule collection that enhances the discriminatory power of these profiles and patterns by providing additional information about amino acids critical for function and/or structure. Together, PROSITE motifs and ProRules are used to annotate domains and features in UniProtKB/Swiss-Prot entries. Since the onset of the COVID-19 pandemic, PROSITE has contributed to SARS-CoV-2 research by leveraging existing tools and by developing new profiles and ProRules for SARS-CoV-2 protein domains. A newly developed profile has also uncovered a link between coregulators of two transcription factor families: POU2F and NF-κB. ProRule has been updated to incorporate the ChEBI ontology to describe chemical ligands and the Rhea reference vocabulary for biochemical reaction annotation. Predicted tridimensional (3D) structures from AlphaFold are now regularly used to define domain boundaries during profile construction. ScanProsite has been enhanced to allow users to visualize motif matches on AlphaFold-predicted structures. In addition, the original pfsearch code has been fully rewritten and optimized to make efficient use of modern multi-core processors, with a new heuristic implemented to further improve performance.

Graphical abstract



Introduction

PROSITE is an annotated collection of motif descriptors used to identify protein domains, sites, and families. It is accessible through its web server at https://prosite.expasy.org/ and its components are available for download from https://ftp.expasy.org/databases/prosite/.

When it was first released in 1989, PROSITE contained 58 documentation entries, described by 60 patterns [1]. At the time, the Swiss-Prot database contained around 10 000 entries, and the Protein Data Bank (PDB) included 365 experimentally determined protein structures (according to PDB Statistics). The Translation of EMBL nucleotide sequence database (TrEMBL), which was created only in 1996, did not yet exist. Initially separate resources, Swiss-Prot and TrEMBL were merged in 2003 to form the UniProt Knowledgebase (UniProtKB), serving respectively as its manually reviewed and automatically annotated sections [2].

In 1999, PROSITE was one of the databases involved in creating InterPro: an integrated documentation resource for protein families, domains, and functional sites. InterPro was developed to rationalize the complementary efforts of individual protein signature database projects [3].

Today, PROSITE (release 2025_03 of 18 June 2025) contains 1956 entries, described by 1311 patterns and 1403 profiles, and is supplemented by 1421 ProRules used to annotate UniProtKB/Swiss-Prot entries. In UniProtKB release 2025_03 (18 June 2025), UniProtKB/Swiss-Prot contains 573 661 entries, while UniProtKB/TrEMBL contains 253 061 697 entries [4]. As of the 10 September 2025 release, PDB [5] includes 241 692 experimentally determined protein structures, and the AlphaFold Protein Structure Database [6] provides 214 683 839 predicted structures (July 2025). PROSITE continues to provide new patterns and profiles to InterPro, which has increased the number of its member databases [7].

Received: September 18, 2025. **Revised:** October 10, 2025. **Accepted:** October 13, 2025

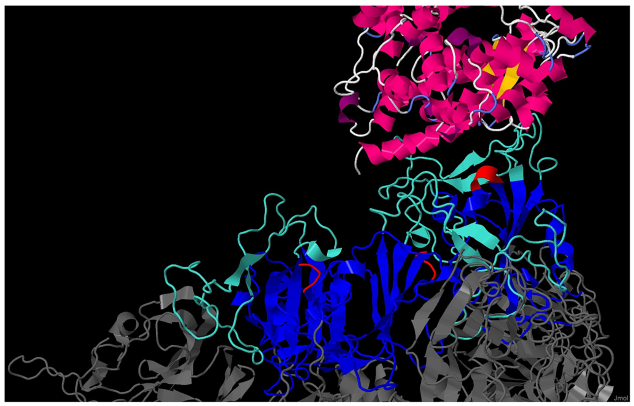


Figure 1. The trimeric SARS-CoV-2 spike protein bound to ACE2. The spike protein is shown in dim grey, the receptor-binding domains (PS51921) in blue, the ACE2-binding regions in turquoise, and the RGD motifs (PS00016) in red. On the top of the figure, the ACE2 protein is shown coloured according to its secondary structures (pink, α-helices; plum, 3_{10}-helices; gold, β-strands; light blue, turns).

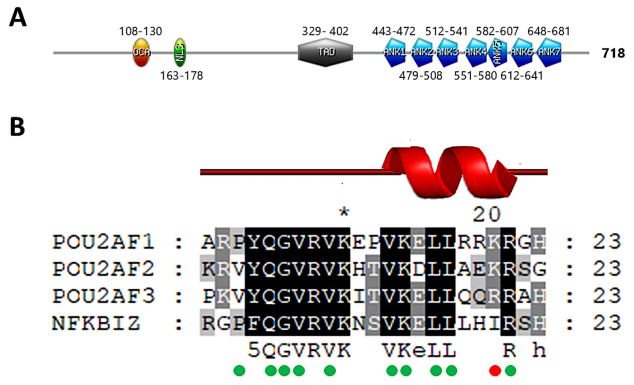


Figure 2. Identification of an OCA motif in IκBζ. (**A**) Domains found in IκBζ: OCA motif (OCA, orange), nuclear localization signal (NLS, green), *trans*-activation domain (TAD, grey), and ankyrin repeats (ANK1-7, blue). (**B**) MSA of the OCA motif of the OCA family members POU2AF1 (OCA-B), POU2AF2 (OCA-T1 or C11orf53), and POU2AF3 (OCA-T2 or COLCA2) and IκBζ. The green circles indicate the conserved residues involved in POU2F1 or DNA binding. The red circle indicates the only residue with a non-conservative substitution in IκBζ. The structure adopted by the OCA motif (PDB: 1CQT) is shown above the alignment.

This paper reviews the evolution of PROSITE over the years to improve the quality of its predictions. It also presents its applications and the updates introduced since our previous publication [8].

Materials and methods

Figures

The 3D structures shown in Figs 1 and 3 were visualized with the JSmol program (https://jmol.sourceforge.net/) used by PROSITE. The secondary structure shown in Fig. 2 was taken from the protein view of 1CQT from PDBsum [9]. The multiple sequence alignment (MSA) shown in Fig. 2 was visualized with the full-featured MSA editor GeneDoc (http://nrbsc.org/gfx/genedoc/).

The PROSITE database

PROSITE is a database of protein families, domains, and sites. It originally consisted of two files: a documentation file (prosite.doc), which compiles all PROSITE entries describing the families, domains, and sites in the database; and a data file (prosite.dat), which contains PROSITE motifs—either patterns or matrices—designed to identify these elements. The prosite.dat file also included, for each motif, a curated match list for the Swiss-Prot section of UniProtKB [1], as well as an additional match list for PDB since 2003 [10]. In 2004, we introduced a new file, prorule.dat, which contains PROSITE rules (ProRules). These rules notably capture the positions of structurally and/or functionally important amino acids, specify the conditions they must fulfil to play their biological roles, and enable the generation of UniProtKB-formatted annotation for the DE, CC, KW, or FT lines [11, 12].

prosite.doc

Documentations include a concise description providing useful biological information about protein families, domains, and/or sites [13]. They generally contain all or part of the following information: the origin of the name, the taxonomic occurrence, examples of matching proteins and their domain architectures, the function and, if relevant, the catalytic mechanism, an example of a 3D structure and its description, the main characteristics of the primary sequence, and some references, which often include alignments that helped develop the motif descriptor(s).

prosite.dat

The motif descriptors used in PROSITE to detect the protein families, domains, or sites described in the documentation are of two types: patterns or profiles [13]. Both are derived from multiple alignments of homologous sequences. This gives these motif descriptors the notable advantage of identifying distant relationships between sequences that would have gone unnoticed based solely on a pairwise sequence alignment. Both patterns and profiles have their own strengths and weaknesses, which define their area of optimum application [13].

Patterns

Originally, PROSITE used patterns—also known as regular expressions—as its first motif descriptors [1]. These motifs, typically spanning 10 to 20 amino acids, often correspond to critical functional regions such as enzyme catalytic sites, prosthetic group attachment sites (e.g. haem, pyridoxal-phosphate, biotin), metal ion binding residues, cysteines forming disulphide bonds, and regions involved in binding molecules (e.g. ADP/ATP, GDP/GTP, calcium, DNA) or other proteins. While regular expressions may seem somewhat 'old-fashioned' today, they remain widely used by the PROSITE community. Users can scan sequences using either existing patterns from the PROSITE database or custom-defined ones. Moreover, regular expressions continue to be frequently used in scientific publications to describe short, conserved sequence regions of interest. Although the syntax of these patterns may vary somewhat from author to author, they can easily be modified using PROSITE's own syntax [13].

Some PROSITE motif descriptors are too short and/or degenerate to carry biological significance on their own, as they appear in most known protein sequences. They are

marked with the /SKIP-FLAG = TRUE qualifier and are not associated with a match list (see below). By default, they are excluded from ScanProsite analyses; their exclusion needs to be manually deselected. These motifs, some of which predict post-translational modification sites—e.g. N-glycosylation sites, phosphorylation sites, or phosphopantetheine attachment sites—produce matches that are only indicative of a possible function. Independent biological evidence must be considered to confirm the appropriateness of these matches and scans should only be performed against small sets of proteins potentially concerned. Nevertheless, if used appropriately, these patterns can provide very useful information where other methods fail. For example, the N-glycosylation site pattern (https://purl.expasy.org/prosite/signature/PS00001) is used to annotate N-linked (GlcNAc…) asparagines in the extracellular regions of eukaryotic proteins in the UniProtKB/Swiss-Prot database. The accuracy of this annotation process is illustrated by its ability to detect N-glycosylation sites in mammalian interferon-β (IFN-β) that can carry one to five predicted glycosylated asparagines [14]. The pattern correctly predicted the unique human IFN-β N-glycosylation site as well as the three murine sites, all of which were indeed shown to carry N-linked sugars at the predicted positions [14, 15].

Another illustration of the usefulness of such patterns was provided during the COVID-19 pandemic. One of them enabled the identification of an RGD motif (or cell attachment sequence) (https://purl.expasy.org/prosite/signature/PS00016) in the SARS-CoV-2 spike protein (see Fig. 1) [16]. The presence of this motif, potentially capable of interacting with integrins, had gone unnoticed during the initial SARS-CoV-2 genome analysis [17]. Experimental confirmation subsequently showed that this motif, located close to the region involved in angiotensin-converting enzyme 2 (ACE2) binding, can indeed bind integrins [18, 19].

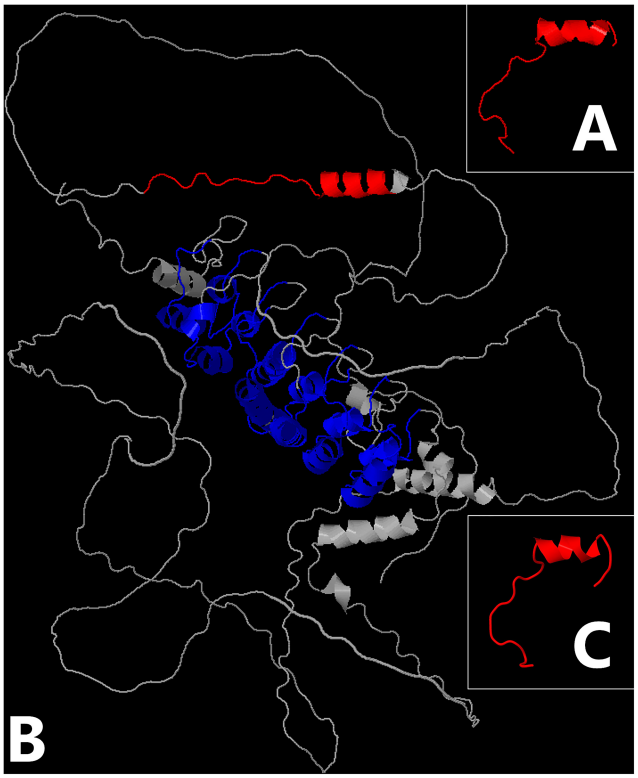


Figure 3. Modelling of the OCA domain. (**A**) The OCA domain of OCA-B (PDB: 1CQT). (**B**) The IκBζ structure predicted by AlphaFold (AF-Q9BYH8-F1-model_v2). The seven ankyrin repeats are represented in blue. (**C**) The human IκBζ OCA domain modelled with SWISS-MODEL using the OCA-B OCA domain (PDB: 1CQT) as a template. The regions shown in red correspond to OCA domains detected by the PROSITE profile.

Generalized profiles

Sensitive position-specific scoring matrix (PSSM) serves as a very useful scoring matrix that contains evolutionary information of protein sequences, which is commonly used to detect distantly related proteins and protein folding patterns [20]. In 1994, PROSITE introduced an extension of PSSMs known as generalized profiles or weight matrices as motif descriptors [21, 22]. Generalized profiles offer the advantage of being less conservative and covering a larger region, enabling models to be built that cover entire proteins or domains. Their enhanced sensitivity enables poorly conserved domains or proteins to be detected.

Over the past few years, we have continued to create new generalized profiles covering structural domains, which are used for annotating UniProtKB/Swiss-Prot entries and for analysing protein sequences for PROSITE users. Some of them have identified domains in new protein families, revealing unexpected relationships between them. Hence, in 2022, we identified a previously unnoticed link between two transcription factor (TF) families, namely POU2F and nuclear factor (NF)-κB (https://purl.expasy.org/prosite/documentation/PDOC52003). The POU2AF (POU2AF1-3) coactivator proteins have been shown to contain an OCA motif, which forms a ternary complex with an octameric DNA sequence and the POU domain of POU2F (POU2F1-3) TFs to control expression of target genes [23, 24]. A generalized profile built from the OCA motif of POU2AF1 detected the presence of the motif with a significant score not only in the paralogous POU2AF2 and POU2AF3 proteins as expected, but also in NF-κB inhibitor ζ (IκBζ) proteins (Fig. 2A) [25]. Reciprocal searches starting from an IκBζ OCA motif produced a similar result: besides IκBζ OCA motifs, the profile retrieved members of the POU2AF family. Finally, before making our model public and using it to annotate the presence of an OCA motif in IκBζ UniProtKB/Swiss-Prot entries, we made some structural comparisons to strengthen our observations. The OCA motif structure of POU2AF1 has been solved and revealed an N-terminal extended polypeptide strand and a C-terminal two-turn α-helix (Figs 2B and 3A) [24]. Although there is no solved structure for IκBζ, AlphaFold [26] predictions show a large N-terminal globally unstructured region, with few secondary structures, sandwiching the C-terminal domain made of ANK repeats (Fig. 3B). Interestingly, the region mapped by the OCA motif profile on the predicted IκBζ structure is located precisely in one of the few secondary structures present in the largely unstructured N-terminal half and shows a structure like the one observed in POU2AF1, namely an unstructured region followed by an α-helix (Fig. 3). Furthermore, the prediction of the start of the α-helix in IκBζ corresponds exactly to the beginning of the α-helix of the OCA motif in the solved structure (Fig. 2B). However, the N-terminal region preceding the α-helix adopts an elongated structure in the IκBζ structure predicted by AlphaFold as compared to the bends, allowing the corresponding region of POU2AF1 to fit in the major groove of the octamer DNA (Fig. 3). To see if the IκBζ OCA motif could adopt a similar structure, SWISS-MODEL

PS51124 **PEPTIDASE_C16** *Peptidase family C16 domain profile* :

1634 - 1898: **score =** 60.973

YYHTTDPSFLGRYMSALNHtkKWKYPQVNGLTSIKWADNNCYLATALLTLQQIELKFNPP
ALQDAYYRARAGEAANFCALILAYCNKTVGELGDVRETMSYLFQHANLDSCKRVLNVVCK
tCGQQQTTLKGVEAVMYMGTLSYEqfkKGVQIPCTCGKQATKYLVQQESPFVMMSAPPAQ
YELKH--GTFTCASEYTGNYQCGHYKhITSKETLYCIDGALLTKSSEYKGPITDVFYKEN
SYTTTIKPVTYKLDGVVCTEIDPKLDN

Predicted features:

DOMAIN	1634	1898	/note="Peptidase C16 "	[condition: none]	
ACT_SITE	1674		/note="For PL1-PRO activity"	[condition: C]	[group: 1]
BINDING	1752		/ligand="Zn(2+)" / ligand_id="ChEBI:CHEBI:29105" / ligand_label="1"	[condition: [CH]]	[group: 2]
ZN_FING	1752	1789	/note="C4-type "	[condition: <117=[C]> and <119=[C]> and <148=[C]> and <150=[C]>]	
BINDING	1755		/ligand="Zn(2+)" / ligand_id="ChEBI:CHEBI:29105" / ligand_label="1"	[condition: [CH]]	[group: 2]
BINDING	1787		/ligand="Zn(2+)" / ligand_id="ChEBI:CHEBI:29105" / ligand_label="1"	[condition: [CH]]	[group: 2]
BINDING	1789		/ligand="Zn(2+)" / ligand_id="ChEBI:CHEBI:29105" / ligand_label="1"	[condition: [CH]]	[group: 2]
ACT_SITE	1835		/note="For PL1-PRO activity"	[condition: H]	[group: 1]
ACT_SITE	1849		/note="For PL1-PRO activity"	[condition: D]	[group: 1]

Figure 4. The matched peptidase family C16 [or papain-like (PL) protease] domain in SARS-CoV-2 replicase polyprotein 1ab (P0DTD1; R1AB_SARS2). The catalytic triads C1674–H1835–D1849 and the four zinc-binding C residues of the C4-type zinc finger are detected by the PROSITE profile (PS51124; PEPTIDASE_C16) and its associated ProRule (PRU00444) and are shown in a table on the ScanProsite Results page. The Zn^{2+} ligand is annotated with its unique ChEBI identifier.

[27] has been used to build an IκBζ OCA motif model using the POU2AF1 OCA motif (PDB: 1CQT) as a template [5, 24]. The modelled structure showed that the IκBζ OCA motif N-terminal region can adopt bends, which should position the conserved residues for interacting both with the octamer DNA major groove and the POU domain (Fig. 3). Among the residues known to interact with the DNA and/or the POU domain of POU2F1, there is only one that is not conserved in the OCA motif of IκBζ, namely the positively charged Lys or Arg at position 20 that makes a hydrogen bond with the POU domain, which is replaced by a hydrophobic aliphatic Ile in IκBζ (Fig. 2B). This substitution would prevent the formation of the hydrogen bond at this position but also extend the hydrophobic surface on the face of the α-helix that binds to a complementary hydrophobic pocket in the POU domain [24]. This increase in the hydrophobic surface could compensate for the loss of the hydrogen bond. As both the primary and secondary structural evidence indicated the presence of a *bona fide* OCA motif in IκBζ, we validated the model. This was made publicly available in PROSITE release 2022_04 on 12 October 2022 and was used to annotate the presence of an OCA motif in the UniProtKB/Swiss-Prot IκBζ entries with the same release date. A year and a half after the OCA motif profile was made publicly available in PROSITE and one year after its integration into InterPro (IPR047571; Release 93.0, 2 March 2023) [7], a paper was published confirming our observations with experimental data [28]. To date, PROSITE remains the only protein domain database to provide a model for the OCA motif.

prorule.dat

The ProRule database contains a set of manually created rules that provide additional biologically meaningful information about domains detected by PROSITE profiles [11, 12]. Its purpose is to provide domain-specific information in the UniProtKB/Swiss-Prot format, using standardized nomenclature and controlled vocabularies. Occasionally, rules make use of patterns. In these cases, the rules do not work independently but are called by another rule triggered by a profile. ProRule uses the UniRule format [29], which is used for all types of rules created to annotate the UniProtKB, including the HAMAP rules [30].

ProRule is extensively used by UniProtKB/Swiss-Prot curators to facilitate the annotation work and to check the consistency of UniProtKB/Swiss-Prot entries. Some features of ProRule, like predicted active and binding sites, posttranslationally modified residues (PTMs), or disulphide bonds, are accessible for external users through our ScanProsite web page (Fig. 4) [31].

Like UniProtKB, ProRule now uses the Chemical Entities of Biological Interest (ChEBI) ontology [32] as its reference vocabulary for annotating biologically relevant ligands and their binding sites [33]. Some ProRules, particularly those describing Ca^{2+} and Zn^{2+} binding, have been revised and used to add and/or correct the annotation of residues involved in this binding in Swiss-Prot entries on a large scale, using the ChEBI identifiers. PROSITE users can also benefit from ligand binding site annotations using stable, unique ChEBI identifiers. If ligand binding sites are detected with ScanProsite, they will appear in the results with stable unique identifiers from the ChEBI ontology (Fig. 4). ProRules can also predict potential biochemical reactions performed by a domain, provided that the active site residues are present. These reactions are described using the same reference vocabulary as UniProtKB: Rhea, an expert-curated knowledgebase of biochemical reactions which uses ChEBI to represent reaction participants

[34]. However, it should be noted that these predicted reactions should be treated with caution, as profiles can detect evolutionarily distant sequences in different species, and the function may have changed during evolution. Slight differences in reactant binding may result in a related but distinct reaction corresponding to a different Rhea entry [35].

prosite.aux

Since release 2024_03 (29 May 2024), the file prosite.dat includes only core PROSITE motif data (i.e. patterns and generalized profiles). Auxiliary information—recalculated at each release based on PROSITE motif matches in UniProtKB/Swiss-Prot and PDB—is now distributed in a separate file: prosite.aux.

This file contains, for each motif not flagged as high probability of occurrence (i.e. those lacking the line CC /SKIP-FLAG = TRUE in prosite.dat—see above), the following information:

- Name and accession (ID and AC lines)
- Numerical Results (NR lines): include the UniProtKB release number, the number of UniProtKB/Swiss-Prot entries in that release, and a summary of relevant motif matches
- Taxonomic range comment (CC /TAXO_RANGE line)
- Cross-references to UniProtKB/Swiss-Prot (DR lines), with a curation-assigned status to each sequence:
  - T – true positive (matched and biologically relevant)
  - P – partial sequence (not matched; assumed to be a true positive if the sequence were complete)
  - N – false negative (not matched but expected to match)
  - ? – unknown (matched; biological relevance unclear)
  - F – false positive (matched but not biologically relevant)
- Cross-references to PDB structures (3D lines)

Separating auxiliary information from prosite.dat improves manageability and contributes to more efficient scanning. The auxiliary data—such as DR and 3D lines—plays no role in the scanning process, as it is ancillary to the motifs (patterns and profiles) themselves. By omitting large match lists from the core motif file, memory usage during scanning is reduced. Moreover, updates to auxiliary data no longer require recompilation of the PROSITE data files, making maintenance easier. Finally, pftools imposes a technical limit on the size of the input motif file, which was another key reason for moving DR and 3D lines to the separate prosite.aux file. This reorganization results in a leaner motif file and may lead to a slight increase in scan speed.

PROSITE availability

PROSITE web interface

PROSITE can be accessed through a web interface at https://prosite.expasy.org/. The homepage allows access to documentations and motif descriptors (for ProRule there are dedicated pages that can be accessed through a tab) and basic scans of a few sequences (max. 10). More sophisticated scans can be performed with the ScanProsite tool accessible through the dedicated tab. ScanProsite notably allows users to upload their own protein database (max. 16 MB) and to perform scans against protein database with PROSITE motifs or their own pattern [8].

Persistent URLs

Each PROSITE documentation entry, motif, and ProRule is assigned a Persistent URL (PURL) via https://purl.archive.org/ [36]. PURLs should be used when referencing specific PROSITE documentations, motifs, or ProRules in publications or data records. A PURL is a stable identifier in the form of a URL that redirects through a resolver to the current location of the resource. This ensures that references remain valid even if the actual URL changes over time.

For example:

- The documentation PDOC52003 can be referenced using: https://purl.expasy.org/prosite/documentation/PDOC52003
- The motif PS52003 uses: https://purl.expasy.org/prosite/signature/PS52003
- The ProRule PRU00444 is accessed via: https://purl.expasy.org/prosite/rule/PRU00444

ScanProsite versus UniProtKB/TrEMBL

Redundant sequences do not significantly add to the information content. Their presence slows down and compromises the output of our services, and they are costly to process and perform repetitive computations on. Since the end of 2021, we have restricted scans against the UniProtKB/TrEMBL database to its reference proteomes, i.e. one representative from each cluster of proteomes that are grouped by their overall sequence similarity [4]. Scans using UniProtKB identifiers (IDs) or accession numbers (ACs) are also limited to UniProtKB/Swiss-Prot and reference proteomes entries. However, scans against sequences not belonging to this group are permitted, provided they are supplied in FASTA format.

AlphaFold

Since domains correspond to protein units that fold independently, structural information is essential for defining domain boundaries. For many years, PROSITE has used the PDB database [5] to build more precise models with correct domain boundaries. ScanProsite also allows scans against the PDB database and enables visualization of matches on protein structures. For UniProtKB entries associated with a PDB entry, a link to the associated structure is available on the results page, enabling users to view the match on the structure [10]. Unfortunately, PDB, which relies on experimental data, only contains structural information for a limited part of the protein universe. The AlphaFold Protein Structure Database (AlphaFold DB), however, is a vast digital library of predicted protein structures containing over 214 million entries, which greatly increases the size of available structures [6]. AlphaFold uses deep learning models trained on evolutionary relationships and physical constraints to predict a protein's 3D structure from its amino acid sequence [26]. Given the high quality of these predicted models, we use them to define the boundaries of our profiles. If a precalculated AlphaFold model exists for a UniProtKB entry, ScanProsite now offers the option of visualizing PROSITE hits on the predicted structure. This enables users to evaluate the quality of PROSITE hits in two ways. Firstly, the hit mapped on the AlphaFold predicted structure should resemble the expected domain structure as described in the corresponding PROSITE documentation and

PDB match list. If it is completely different, the match is most probably a false positive. Second, partial similarity could indicate an incorrect boundary assessment, and this can be adjusted according to the predicted structure.

PROSITE for download

The files making up the PROSITE database (prosite.doc, prosite.dat, prosite.aux, and prorule.dat) as well as files containing additional information can be downloaded at https://ftp.expasy.org/databases/prosite/. The tools (ps_scan.pl, pfsearch, pfscan, psa2msa) needed to use them are also available at the same URL in the ps_scan folder as tarballs containing precompiled versions of the tools for various operating systems.

A suite of tools to build and search generalized profiles can be accessed at the following URLs: https://github.com/sib-swiss/pftools3 or https://doi.org/10.5281/zenodo.17360684 [37]. It contains both the original FORTRAN 77 pftools (release 2.3) and the new programme: pftoolsV3. pfsearch2.3 is considered as the standard for PROSITE match lists but, as announced in our previous paper [8], there is now a new version, pfsearchV3, that has been rewritten and optimized to efficiently use modern multi-core processors, and a heuristic has been implemented for further speed enhancements [38]. Except for circular profiles, which require the standard programme pfsearch2.3, most profiles are compatible with both versions of pfsearch. When used with pfsearchV3, circular profiles do not cause the program to crash, but they do not circularize anymore and therefore produce matches of a single unit. ps_scan.pl, PROSITE's standard analysis tool for patterns and profiles [39], has been adapted to accept both pfsearch2.3 and pfsearchV3. For large-scale scans, we recommend that users who have installed PROSITE locally use pfsearchV3 to reduce the time required. For a limited number of sequences, it is preferable to use pfsearch2.3 to obtain the same reference results as PROSITE and to have the circular profile functionality enabled.

Conclusion

Since our last article for the NAR database issue [8], PROSITE has continued to develop useful profiles for Swiss-Prot annotation and external users. During the COVID-19 shutdown, we focused particularly on the protein domains present in the SARS-CoV-2 virus in order to offer researchers broad coverage in terms of modules that truly correspond to structural and functional units. An old PROSITE pattern has also proven useful for identifying a new mode of interaction between the virus and the cell surface [16].

Among the new profiles developed, one of them made it possible to identify a previously unidentified link between the POU2F and NF-κB families of TFs. The profile developed for the OCA motif identified in the POU2AF1-3 coactivators of POU2F revealed that the OCA motif was also present in IκBζ, an inhibitor of the NF-κB TF. This discovery allowed PROSITE to annotate the presence of an OCA motif in UniProtKB/Swiss-Prot IκBζ entries more than a year before its presence in these proteins was confirmed experimentally [28].

Thanks to the close collaboration between UniProtKB/Swiss-Prot and PROSITE annotators, the development of new profiles should also be fruitful and lead to new discoveries in the future. To meet the needs of UniProtKB/Swiss-Prot annotators, the ProRules associated with PROSITE profiles and patterns have adopted the ChEBI ontology and the Rhea reference vocabulary to ensure consistent annotation of UniProtKB/Swiss-Prot entries [33].

The newly developed PROSITE motifs continue to be integrated into InterPro [7], of which PROSITE is one of the founding members, and the PROSITE documentation is also largely used for InterPro entry descriptions.

pfsearch2.3 is still the PROSITE reference tool for scanning profiles, but a faster version has been developed, pfsearchV3, which is used in particular for PROSITE scans performed with InterProScan. Both programmes, as well as all the tools needed to install PROSITE locally, are available on our FTP site. The matchlists for UniProtKB/Swiss-Prot and PDB, previously stored with the patterns and profiles in prosite.dat, are now stored separately in a new file, prosite.aux.

Some new features have also been implemented on our website. Our documentation entries, motifs, and ProRules have been supplied with a PURL to ensure that references remain valid even after actual URL changes. When the primary structure of a PDB entry was matched by a PROSITE motif, it was already possible to visualize the match on the tertiary structure. As AlphaFold offers pre-calculated 3D structure predictions for most UniProtKB entries, the output of ScanProsite has been modified to allow users to visualize profile and pattern matches on these 3D models as well.

Acknowledgements

We would like to thank:

- the University of Geneva for hosting the SIB's Swiss-Prot group, of which PROSITE is a member,
- Nadine Gruaz-Gumovski, the Swiss-Prot curator who asked for the creation of a PROSITE profile corresponding to the OCA motif for the annotation of POU2AF family members,
- Ivo Pedruzzi and Salvo Paesano for their help in formatting the manuscript,
- the reviewers for their helpful comments, suggestions, and careful testing of ScanProsite. We will consider their suggestions when planning future developments.

Author contributions: Christian J. A. Sigrist (Conceptualization [lead], Data curation [lead], Methodology [equal], Resources [equal], Validation [equal], Writing—original draft [lead]), Béatrice A. Cuche (Resources [supporting], Software [supporting], Visualization [lead], Writing—review & editing [equal]), Edouard de Castro (Formal analysis [equal], Resources [equal], Software [equal], Writing—review & editing [equal]), Elisabeth Coudert (Supervision [equal], Writing—review & editing [equal]), Nicole Redaschi (Supervision [equal]), and Alan Bridge (Funding acquisition [lead], Project administration [lead], Supervision [lead], Writing—review & editing [lead])

Conflict of interest

None declared.

Funding

PROSITE is supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI). Funding to pay the Open Access publication charges for this article was provided by Swiss Federal Government, State Secretariat for Education, Research and Innovation (SERI).

Data availability

All data are available from the main website (https://prosite.expasy.org/) and can be downloaded from the FTP site (https://ftp.expasy.org/databases/prosite/). The ScanProsite tool can be accessed programmatically through a RESTful web service. Clients send HTTP GET or POST requests, and results are returned directly as plain data in formats such as txt, xml and json. The endpoint is https://prosite.expasy.org/cgi-bin/prosite/scanprosite/PSScan.cgi. For details see https://prosite.expasy.org/scanprosite/scanprosite_doc.html#rest. The pftools are available at https://github.com/sib-swiss/pftools3 or https://doi.org/10.5281/zenodo.17360684.

References

1. Bairoch A. Prosite: a dictionary of sites and patterns in proteins. *Nucleic Acids Res* 1991;19:2241–5. https://doi.org/10.1093/nar/19.suppl.2241
2. Apweiler R, Bairoch A, Wu CH *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004;32:D115–9. https://doi.org/10.1093/nar/gkh131
3. Apweiler R, Attwood TK, Bairoch A *et al.* InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 2000;16:1145–50. https://doi.org/10.1093/bioinformatics/16.12.1145
4. Consortium UP. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 2025;53:D609–17. https://doi.org/10.1093/nar/gkae1010
5. Burley SK, Bhikadiya C, Bi C *et al.* RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res* 2023;51:D488–508. https://doi.org/10.1093/nar/gkac1077
6. Varadi M, Bertoni D, Magana P *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;52:D368–75. https://doi.org/10.1093/nar/gkad1011
7. Blum M, Andreeva A, Florentino LC *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res* 2025;53:D444–56. https://doi.org/10.1093/nar/gkae1082
8. Sigrist CJA, de Castro E, Cerutti L *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res* 2013;41:D344–7. https://doi.org/10.1093/nar/gks1067
9. Laskowski RA, Jabłońska J, Pravda L *et al.* PDBsum: structural summaries of PDB entries. *Protein Sci* 2018;27:129–34. https://doi.org/10.1002/pro.3289
10. Hulo N, Sigrist CJA, Le Saux V *et al.* Recent improvements to the PROSITE database. *Nucleic Acids Res* 2004;32:D134–7. https://doi.org/10.1093/nar/gkh044
11. Sigrist CJA, De Castro E, Langendijk-Genevaux PS *et al.* ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics* 2005;21:4060–6. https://doi.org/10.1093/bioinformatics/bti614
12. Hulo N, Bairoch A, Bulliard V *et al.* The PROSITE database. *Nucleic Acids Res* 2006;34:D227–30. https://doi.org/10.1093/nar/gkj063
13. Sigrist CJA, Cerutti L, Hulo N *et al.* PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinformatics* 2002;3:265–74. https://doi.org/10.1093/bib/3.3.265
14. Sommereyns C, Michiels T. N-glycosylation of murine IFN-beta in a putative receptor-binding region. *J Interferon Cytokine Res* 2006;26:406–13. https://doi.org/10.1089/jir.2006.26.406
15. Karpusas M, Nolte M, Benton CB *et al.* The crystal structure of human interferon beta at 2.2-A resolution. *Proc Natl Acad Sci USA* 1997;94:11813–8. https://doi.org/10.1073/pnas.94.22.11813
16. Sigrist CJA, Bridge A, Mercier LP. A potential role for integrins in host cell entry by SARS-CoV-2. *Antiviral Res* 2020;177:104759. https://doi.org/10.1016/j.antiviral.2020.104759
17. Zhou P, Yang X-L, Wang X-G *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;579:270–3. https://doi.org/10.1038/s41586-020-2012-7
18. Bugatti A, Filippini F, Bardelli M *et al.* SARS-CoV-2 infects human ACE2-negative endothelial cells through an αvβ3 integrin-mediated endocytosis even in the presence of vaccine-elicited neutralizing antibodies. *Viruses* 2022;14:705. https://doi.org/10.3390/v14040705
19. Bugatti A, Filippini F, Messali S *et al.* The D405N mutation in the spike protein of SARS-CoV-2 Omicron BA.5 inhibits spike/integrins interaction and viral infection of human lung microvascular endothelial cells. *Viruses* 2023;15:332. https://doi.org/10.3390/v15020332
20. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 1987;84:4355–8. https://doi.org/10.1073/pnas.84.13.4355
21. Bucher P, Bairoch A. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc Int Conf Intell Syst Mol Biol* 1994;2:53–61.
22. Bairoch A, Bucher P, Hofmann K. The PROSITE database, its status in 1995. *Nucleic Acids Res* 1996;24:189–96. https://doi.org/10.1093/nar/24.1.189
23. Wu XS, He X-Y, Ipsaro JJ *et al.* OCA-T1 and OCA-T2 are coactivators of POU2F3 in the tuft cell lineage. *Nature* 2022;607:169–75. https://doi.org/10.1038/s41586-022-04842-7
24. Chasman D, Cepek K, Sharp PA *et al.* Crystal structure of an OCA-B peptide bound to an Oct-1 POU domain/octamer DNA complex: specific recognition of a protein–DNA interface. *Genes Dev* 1999;13:2650–7. https://doi.org/10.1101/gad.13.20.2650
25. Totzke G, Essmann F, Pohlmann S *et al.* A novel member of the IkappaB family, human IkappaB-zeta, inhibits transactivation of p65 and its DNA binding. *J Biol Chem* 2006;281:12645–54. https://doi.org/10.1074/jbc.M511956200
26. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2
27. Waterhouse A, Bertoni M, Bienert S *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018;46:W296–303. https://doi.org/10.1093/nar/gky427
28. Alpsoy A, Wu XS, Pal S *et al.* IκBζ is a dual-use coactivator of NF-κB and POU transcription factors. *Mol Cell* 2024;84:1149–57. https://doi.org/10.1016/j.molcel.2024.01.007
29. MacDougall A, Volynkin V, Saidi R *et al.* UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics* 2020;36:4643–8. https://doi.org/10.1093/bioinformatics/btaa485
30. Pedruzzi I, Rivoire C, Auchincloss AH *et al.* HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res* 2015;43:D1064–70. https://doi.org/10.1093/nar/gku1002

31. de Castro E, Sigrist CJA, Gattiker A *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 2006;34:W362–5. https://doi.org/10.1093/nar/gkl124
32. Hastings J, Owen G, Dekker A *et al.* ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;44:D1214–9. https://doi.org/10.1093/nar/gkv1031
33. Coudert E, Gehant S, de Castro E *et al.* Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* 2023;39:btac793. https://doi.org/10.1093/bioinformatics/btac793
34. Bansal P, Morgat A, Axelsen KB *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res* 2022;50:D693–700. https://doi.org/10.1093/nar/gkab1016
35. Junker A, Fischer J, Sichhart Y *et al.* Evolution of the key alkaloid enzyme putrescine N-methyltransferase from spermidine synthase. *Front Plant Sci* 2013;4:260. https://doi.org/10.3389/fpls.2013.00260
36. Ducut E, Liu F, Fontelo P. An update on Uniform Resource Locator (URL) decay in MEDLINE abstracts and measures for its mitigation. *BMC Med Inform Decis Mak* 2008;8:23. https://doi.org/10.1186/1472-6947-8-23
37. Lüthy R, Xenarios I, Bucher P. Improving the sensitivity of the sequence profile method. *Protein Sci* 1994;3:139–46.
38. Schuepbach T, Pagni M, Bridge A *et al.* pfsearchV3: a code acceleration and heuristic to search PROSITE profiles. *Bioinformatics* 2013;29:1215–7. https://doi.org/10.1093/bioinformatics/btt129
39. Gattiker A, Gasteiger E, Bairoch A. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics* 2002;1:107–8.

Received: September 18, 2025. **Revised:** October 10, 2025. **Accepted:** October 13, 2025