Nucleic Acids Research, 2026, **54**, D499–D503
https://doi.org/10.1093/nar/gkaf1023
Advance access publication date: 8 October 2025
Database issue

SMART v10: three decades of the protein domain annotation resource

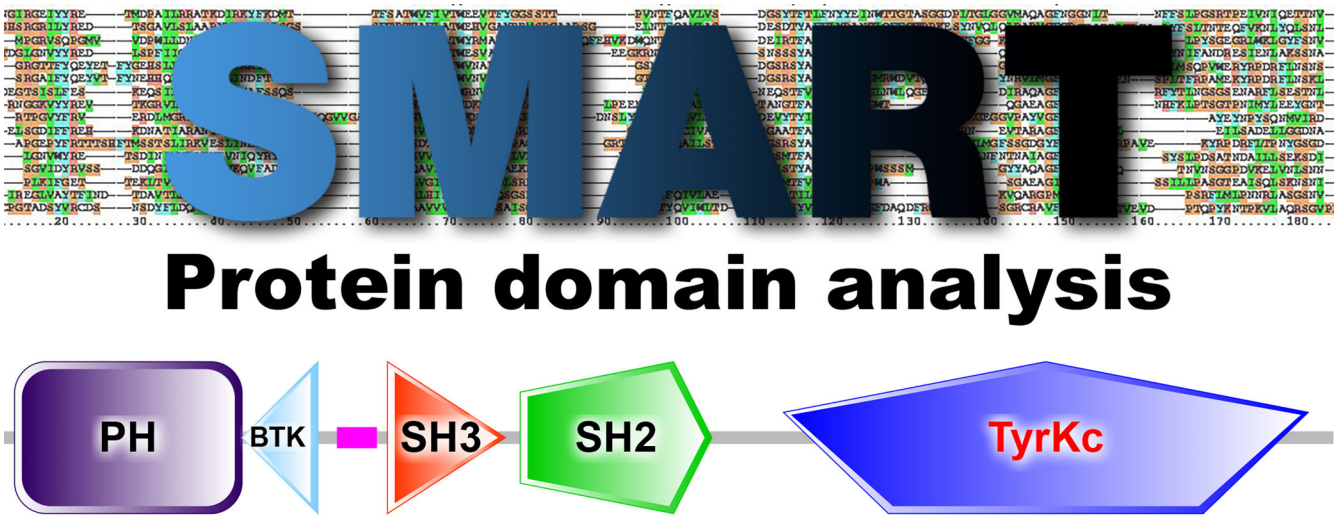
Ivica Letunic[1] **and Peer Bork**[2,3,4,*]

[1]biobyte solutions GmbH, Bothestr 142, 69126 Heidelberg, Germany
[2]EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany
[3]Max Delbrück Centre for Molecular Medicine,13125 Berlin, Germany
[4]Department of Bioinformatics, Biocenter, University of Würzburg, 97070 Würzburg, Germany
[*]To whom correspondence should be addressed. Email: bork@embl.de

Abstract

SMART (Simple Modular Architecture Research Tool, https://smart.embl.de) is a web-based platform for identifying and annotating protein domains and analyzing domain architectures. SMART version 10 features manually curated models for over 1300 protein domains. Approaching its 30th anniversary, SMART's user interface has been redesigned from the ground up, leveraging modern web technologies to enhance intuitiveness and usability. SMART's "Genomic" mode, which annotates proteins from completely sequenced genomes was synchronized with the current release of STRING, and now includes 12 035 species, compared to 5090 in the previous release. Protein and domain annotation pages have been updated with new information sources. Integration with eggNOG provides links to 17.5 million orthologous groups for over 53 million proteins. Additionally, synchronization with the interactive Pathways Explorer version 3 incorporates updated KEGG pathway and orthologous group data, enabling direct visualization on four distinct pathway overview maps.

Graphical abstract



Introduction

Protein domain analysis remains a crucial research tool, simplified by various user-friendly online databases [1–3] providing comprehensive domain annotations. The SMART database [4] combines manually curated hidden Markov models [5, 6] for many domains with a user-friendly web interface, providing robust analysis and visualization capabilities. For nearly 30 years, SMART has been a widely adopted resource among researchers worldwide.

SMART specializes in mobile protein domains, particularly those involved in signaling, extracellular, and regulatory functions. It emphasizes the modular architecture of proteins, making it particularly valuable for studying complex, multidomain proteins and their evolutionary relationships. SMART's hidden Markov models are manually curated, ensuring high-quality annotations with minimal false positives. SMART's web interface is designed for intuitive navigation, offering interactive tools like domain architecture diagrams, phylogenetic distribution views, and detailed annotation tables. These

Received: August 18, 2025. **Revised:** September 19, 2025. **Accepted:** September 19, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

features make it particularly accessible for researchers studying domain evolution and protein function.

Here, we outline the key updates and new features introduced since our last report [7].

New user interface

SMART version 10 introduces a fully redesigned web user interface, rebuilt from the ground up to leverage modern web technologies (Fig. 1). The redesign preserves the original functionality and page layout to ensure a familiar user experience as much as possible.

Protein domain annotation pages now feature a tabbed interface, incorporating a faster, interactive tree display widget for exploring the taxonomic distribution of related proteins.

All protein schematics in list displays, such as domain architecture analysis results, are rendered as dynamically generated SVG images, which seamlessly scale to the users display size, regardless of its resolution. In SMART version 10, these have been extended with several new features and optimized to support latest web standards. Protein size scales are now included in all generated images, allowing easier identification of the positions of individual domains and other features.

The display applet used in the single protein annotation mode has been extended and remains fixed on-screen during page scrolling. Selecting a predicted domain or feature in data tables now automatically highlights its position in the protein schematic and centers it on the screen. This functionality simplifies the identification of relationships among protein features, including those not directly displayed due to threshold cutoffs or overlaps.

If a more fine-grained evaluation of a protein region is required, the viewer allows interactive selection of various parts of the protein sequence independent of the annotated features and their submission to BLAST analysis [8]. This function now supports direct submission of the sequence to the PSI-BLAST [9] server at the European Bioinformatics Institute [10], or to the standard BLAST server at NCBI [11].

Updated genomic protein database

The main underlying protein database in SMART combines the complete UniProt [12] with all stable Ensembl [13] proteomes. In addition to the main protein database, SMART offers a "Genomic" analysis mode that contains only proteins from completely sequenced genomes. Synchronized with the current STRING version 12 [14], it contains >59 million proteins from 12 535 complete genomes (1322 Eukaryota, 10 756 Bacteria, and 457 Archaea), which is a 2.5-fold increase in both the number of proteins and genomes compared to the previous release.

Expanded protein interaction data

With the update of the underlying protein databases, we have also synchronized our protein interaction data with the version 12 of the STRING database [14]. Updated graphical representations of putative interaction partners are now available for >52 million proteins. These interaction network displays in SMART are now interactive and can be zoomed and navigated directly. Individual proteins in the network can be selected, and their corresponding annotation pages in SMART or STRING accessed for further exploration.

Expanded and updated external information sources

SMART version 10 significantly expands its integration of external data sources. Protein orthology data, sourced from the eggNOG 6.0 database [15], now encompass over 53 million proteins across 12 035 species, organized into >17 million orthologous groups. Annotation pages provide a detailed overview of orthologous groups for each protein, including descriptions and taxonomic classifications. Direct links to eggNOG 6.0 enable in-depth exploration of these groups, complete with their associated alignments and phylogenetic trees.

Biological pathway integration has been substantially enhanced, with synchronization to the interactive Pathways Explorer version 3 (iPath) [16]. Pathway data, available for nearly 40 million proteins, now include links to 423 KEGG pathways and 15 758 KEGG orthologous groups (KOs) [17]. Users can visualize these pathways and KOs interactively on four iPath3 overview maps: metabolic pathways, biosynthesis of secondary metabolites, biosynthesis of antibiotics, and microbial metabolism in diverse environments.

Updated architecture SMART and taxonomic tree data export

SMART's domain architecture analysis enables users to identify proteins with specific domain combinations. These can also be generated using combinations of GO (Gene Ontology) terms [18] associated with protein domains and restricted to selected taxonomic classes. In addition to the standard SMART protein schematic visualization, these data can also be exported into FASTA files or represented as annotated phylogenetic trees.

The backend for domain architecture queries has been completely overhauled, achieving 10- to 100-fold faster performance. Analysis results, including datasets exceeding 50 000 proteins, now load quickly and display efficiently in modern web browsers. These results are visualized as a taxonomic tree structure, and SMART protein schematics are displayed directly in a floating popup dialog for any selected tree clade (Fig. 2). This allows simple exploration of large datasets, and quick comparison of protein domain architectures across taxonomic clades.

Phylogenetic tree export functionality has been enhanced and aligned with the iTOL version 6 [19]. Phylogenetic trees and their associated protein domain datasets can be downloaded or visualized directly in iTOL, and now also include helpful popup information with detailed taxonomic information for all clades and species. Furthermore, the taxonomic database used for the tree generation was synchronized with the current NCBI taxonomy release [20].

Database and web server optimizations

The backend of SMART is a relational database management system powered by PostgreSQL. It stores annotations for all SMART domains, protein annotations and sequences, taxonomy information, and precalculated analyses for the entire UniProt [12], Ensembl [13], and STRING [14] proteomes. In addition to the predictions of all SMART and Pfam [21] domains, this includes various protein intrinsic features, like signal peptides, transmembrane, and coiled coil regions.

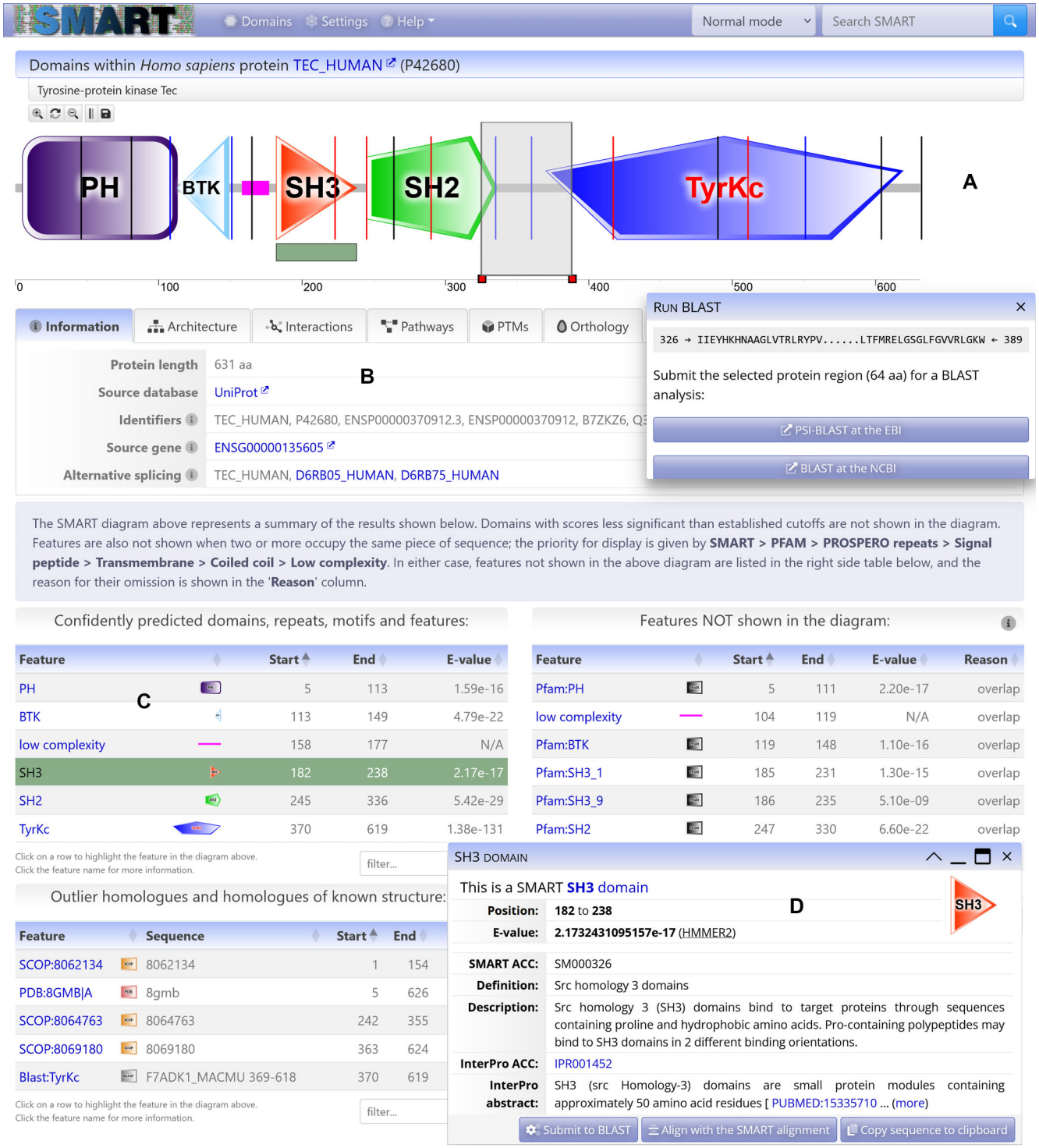


Figure 1. SMART annotation page for protein TEC_HUMAN. (**A**) Protein schematic representations are displayed using an interactive SVG (scalable vector graphics) applet. Schematics are zoomable without quality loss and can be saved as vector (SVG) images. Using the interactive scale, any protein region can be selected and submitted for further BLAST (Basic Local Alignment Search Tool) analysis. (**B**) Tabbed interface collects various sources of external information about the protein analyzed. (**C**) Data tables listing the domains and other features detected in the protein. Each table can be individually searched or sorted on any column. Selecting a feature will highlight it in the protein schematic, while scrolling the display to make it visible on screen. (**D**) Movable and resizable popup dialog displays the most important bits of information for any selected feature, with links to complete annotation.

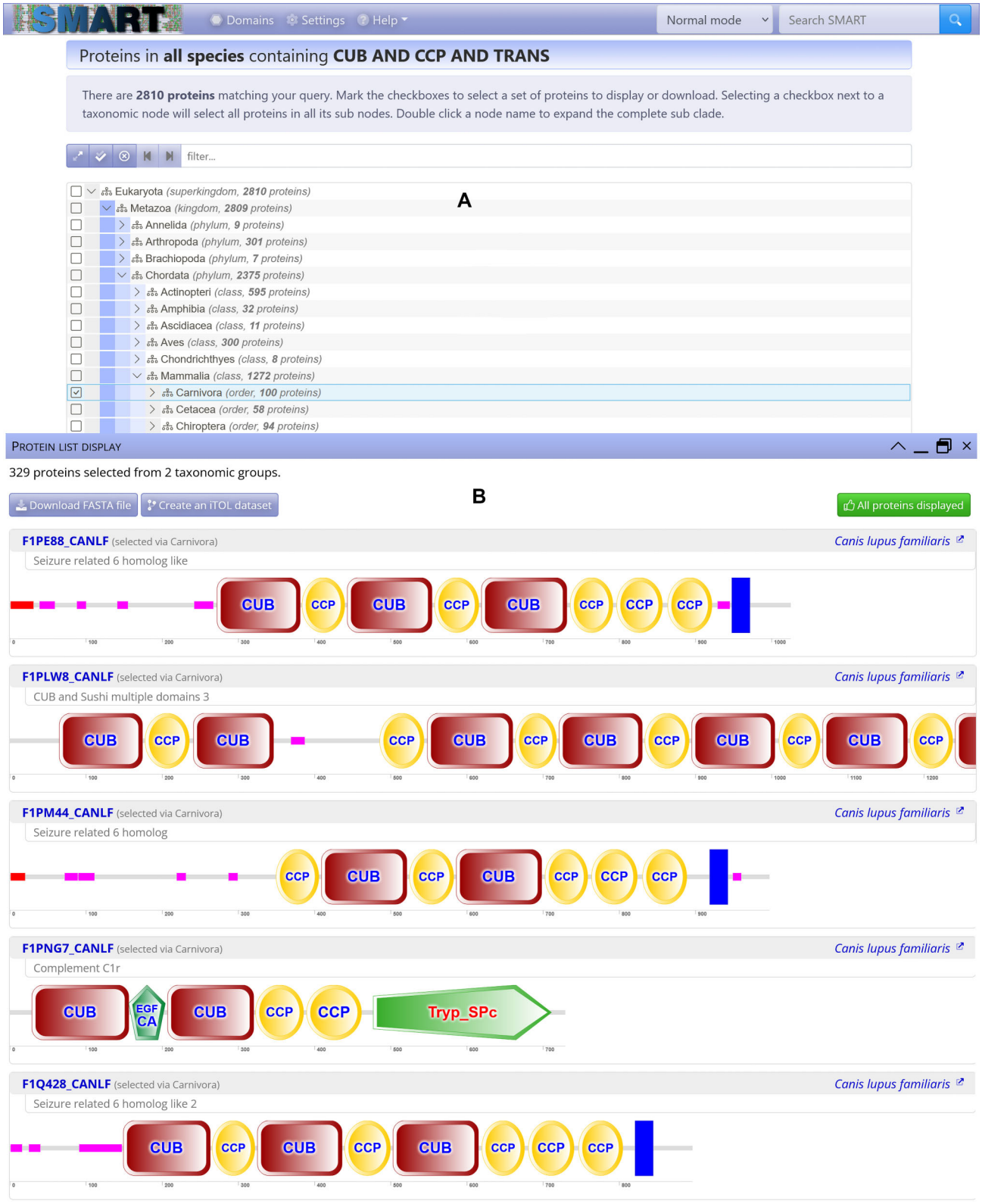


Figure 2. SMART's new domain architecture analysis page. (**A**) All proteins containing the domains matching user's query are displayed in a taxonomic tree browser. User can select any combination of taxonomic clades or individual proteins. (**B**) Proteins selected by the user are visualized in a dialog window directly in the results page, allowing simpler comparisons of domain architectures between across taxonomic clades. Proteins selected can also be downloaded as a FASTA file or visualized in iTOL (Interactive Tree of Life) [19].

In SMART version 10, we now also store the precalculated results for the detection of remote homologues and homologues of protein structures. These are generated via regular BLAST searches against the latest PDB [22], SCOP [23], and the sequences of all detected SMART domains.

Due to constant growth of annotated features, we are regularly restructuring our backend databases, and optimizing various parts of the server code to ensure a satisfactory user experience. Additionally, we have upgraded the server hardware supporting sequence annotation searches and database queries, incorporating expanded RAM and CPU capacity. These enhancements significantly improve processing speeds for user-submitted proteins and reduce response times.

Acknowledgements

Author contributions: Ivica Letunic (Conceptualization [equal], Software [lead], Visualization [lead], Writing—original draft [lead]) and Peer Bork (Conceptualization [equal], Writing—review & editing [equal]).

Conflict of interest

None declared.

Funding

This work was supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK), the Federal Ministry of Research, Technology and Space (W-de.NBI-014), and by EMBL. Funding to pay the Open Access publication charges for this article was provided by EMBL.

Data availability

SMART profiles and alignments are freely available for academic users via EMBLem (https://www.embl-em.de). Commercial licenses and support are available via biobyte solutions GmbH (https://www.biobyte.de).

References

1. 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
2. Wang J, Chitsaz F, Derbyshire MK *et al.* The conserved domain database in 2023. *Nucleic Acids Res* 2023;**51**:D384–8. https://doi.org/10.1093/nar/gkac1096
3. Sigrist CJ, 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
4. Schultz J, Milpetz F, Bork P *et al.* SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA* 1998;**95**:5857–64. https://doi.org/10.1073/pnas.95.11.5857
5. Krogh A, Brown M, Mian IS *et al.* Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 1994;**235**:1501–31. https://doi.org/10.1006/jmbi.1994.1104
6. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;**7**:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
7. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res* 2021;**49**:D458–60. https://doi.org/10.1093/nar/gkaa937
8. Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
9. Altschul SF, Madden TL, Schaffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402. https://doi.org/10.1093/nar/25.17.3389
10. Madeira F, Madhusoodanan N, Lee J *et al.* The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res* 2024;**52**:W521–5. https://doi.org/10.1093/nar/gkae241
11. Johnson M, Zaretskaya I, Raytselis Y *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res* 2008;**36**:W5–9. https://doi.org/10.1093/nar/gkn201
12. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 2025;**53**:D609–17.
13. Dyer SC, Austine-Orimoloye O, Azov AG *et al.* Ensembl 2025. *Nucleic Acids Res* 2025;**53**:D948–57. https://doi.org/10.1093/nar/gkae1071
14. Szklarczyk D, Nastou K, Koutrouli M *et al.* The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res* 2025;**53**:D730–7. https://doi.org/10.1093/nar/gkae1113
15. Hernandez-Plaza A, Szklarczyk D, Botas J *et al.* eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res* 2023;**51**:D389–94. https://doi.org/10.1093/nar/gkac1022
16. Darzi Y, Letunic I, Bork P *et al.* iPath3.0: interactive pathways explorer v3. *Nucleic Acids Res* 2018;**46**:W510–3. https://doi.org/10.1093/nar/gky299
17. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30. https://doi.org/10.1093/nar/28.1.27
18. Huntley RP, Sawford T, Mutowo-Meullenet P *et al.* The GOA database: Gene Ontology Annotation updates for 2015. *Nucleic Acids Res* 2015;**43**:D1057–63. https://doi.org/10.1093/nar/gku1113
19. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* 2024;**52**:W78–82. https://doi.org/10.1093/nar/gkae268
20. Cox E, Tsuchiya MTN, Ciufo S *et al.* NCBI Taxonomy: enhanced access via NCBI Datasets. *Nucleic Acids Res* 2025;**53**:D1711–5. https://doi.org/10.1093/nar/gkae967
21. Poretsky E, Andorf CM, Sen TZ. Structural variability of Pfam domains based on Alphafold2 predictions. *Proteins: Structure, Function, and Bioinformatics* 2025. https://doi.org/10.1002/prot.70021
22. Choudhary P, Kunnakkattu IR, Nair S *et al.* PDBe tools for an in-depth analysis of small molecules in the Protein Data Bank. *Protein Sci* 2025;**34**:e70084. https://doi.org/10.1002/pro.70084
23. Andreeva A, Kulesha E, Gough J *et al.* The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res* 2020;**48**:D376–82. https://doi.org/10.1093/nar/gkz1064

Received: August 18, 2025. **Revised:** September 19, 2025. **Accepted:** September 19, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.