D458–D460 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 26 October 2020*
doi: 10.1093/nar/gkaa937

SMART: recent updates, new developments and status in 2020

Ivica Letunic [1], Supriya Khedkar [2] 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, Berlin, Germany and [4] Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

Received September 08, 2020; Revised October 04, 2020; Editorial Decision October 05, 2020; Accepted October 06, 2020

ABSTRACT

SMART (Simple Modular Architecture Research Tool) is a web resource (https://smart.embl.de) for the identification and annotation of protein domains and the analysis of protein domain architectures. SMART version 9 contains manually curated models for more than 1300 protein domains, with a topical set of 68 new models added since our last update article (1). All the new models are for diverse recombinase families and subfamilies and as a set they provide a comprehensive overview of mobile element recombinases namely transposase, integrase, relaxase, resolvase, cas1 casposase and Xer like cellular recombinase. Further updates include the synchronization of the underlying protein databases with UniProt (2), Ensembl (3) and STRING (4), greatly increasing the total number of annotated domains and other protein features available in architecture analysis mode. Furthermore, SMART's vector-based protein display engine has been extended and updated to use the latest web technologies and the domain architecture analysis components have been optimized to handle the increased number of protein features available.

INTRODUCTION

Protein domain databases remain important annotation and research tools. The SMART database (5) integrates manually curated hidden Markov models (6,7) for many domains with a powerful web-based interface offering various analysis and visualization tools. Almost 25 years since its inception, it remains popular and widely used by scientists worldwide. Here we summarize the major changes and new features that have been introduced since our last report (1).

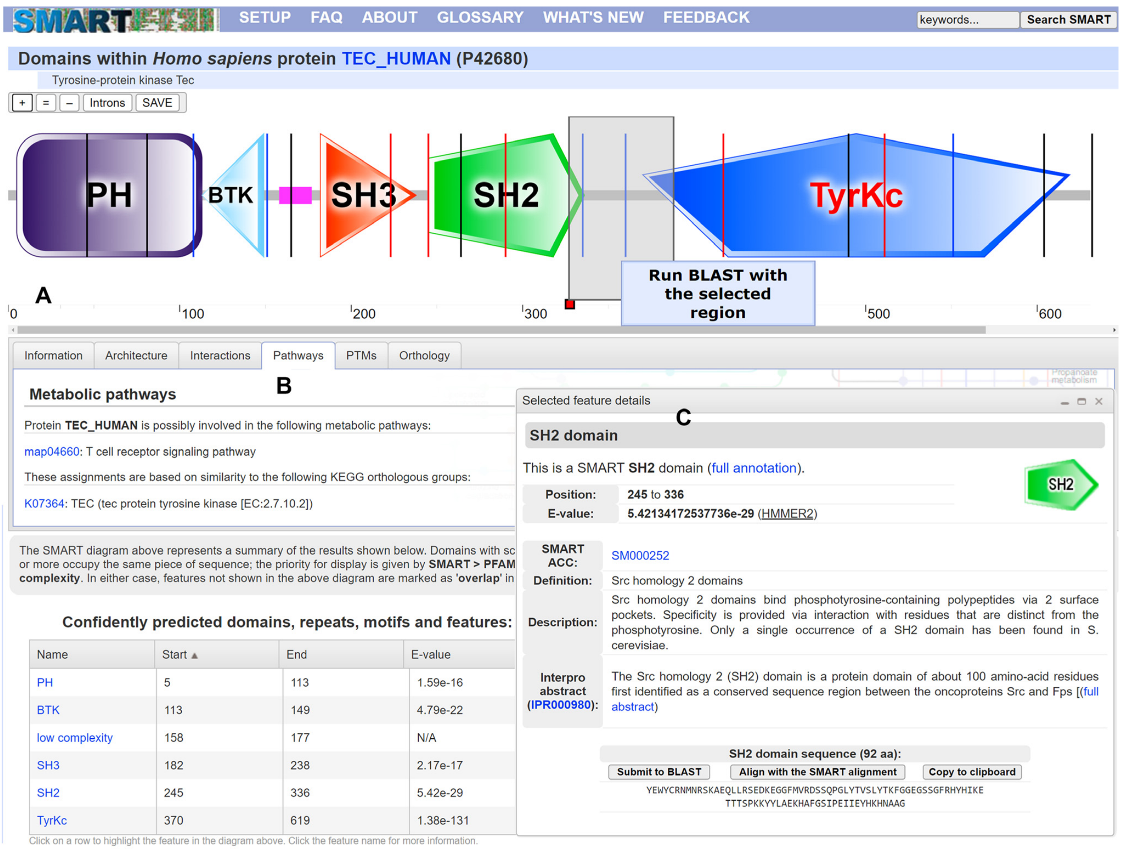
TOPICAL EXPANSION OF DOMAIN COVERAGE

SMART was never intended to be exhaustive, and was initially focused on mobile domains, which occur in various contexts while retaining similar function. Nevertheless, it continues to gradually expand its domain coverage with each new release. The current version introduces a topical expansion geared towards a complete coverage of known prokaryotic recombinases, which can be searched together for a report on mobile genetic elements in bacteria and archaebacteria. For this purpose, we bring together a comprehensive set of 68 domains that cover five major recombinase families namely transposase—DDE (8), relaxase—HUH (9), casposase—cas1 (10), (resolvase-) serine and (integrase-) tyrosine recombinase (11). Thirty-one models for DDE subfamilies, 10 for HUH and 1 for tyrosine recombinase were obtained from Pfam (12), a set of 20 tyrosine recombinases sub-families from (13) and six sub-families three belonging to serine, two to HUH and one to Cas1 were newly developed. Based on the association of most of the recombinase sub-families with specific mobile element types or cellular functions, these 68 domains together can be used as seeds for the identification and discrimination of diverse mobile element types namely transposable elements, phages, integrons, conjugative elements (plasmids and Integrative Conjugative Elements—ICE) and casposons. All together, these new domains are present in close to 1.1 million proteins from the current SMART non-redundant database (Figure 1).

UPDATED PROTEIN DATABASES

The main underlying protein database in SMART combines the complete Uniprot (2) with all stable Ensembl (3) proteomes. The current release contains more than 137 million proteins (a 2.7-fold increase compared to the previous release) from around 537 thousand species, subspecies and strains (a 1.2-fold increase). To reduce the high redundancy that is inherently present in these databases, SMART uses a per-species protein clustering procedure. All the proteins are initially separated into species-specific databases. Each of these databases is clustered separately using the CD-HIT algorithm (14) with a 96% identity cutoff. Longest members of each cluster are used as 'representatives', and are the only proteins included in the database, together with all the

[*] To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; Email: bork@embl.de



Domains within *Homo sapiens* protein TEC_HUMAN (P42680)

Tyrosine-protein kinase Tec

Metabolic pathways

Protein **TEC_HUMAN** is possibly involved in the following metabolic pathways:

map04660: T cell receptor signaling pathway

These assignments are based on similarity to the following KEGG orthologous groups:

K07364: TEC (tec protein tyrosine kinase [EC:2.7.10.2])

The SMART diagram above represents a summary of the results shown below. Domains with sc or more occupy the same piece of sequence; the priority for display is given by **SMART > PFAM** **complexity**. In either case, features not shown in the above diagram are marked as **'overlap'** in

Confidently predicted domains, repeats, motifs and features:

Name	Start ▲	End	E-value
PH	5	113	1.59e-16
BTK	113	149	4.79e-22
low complexity	158	177	N/A
SH3	182	238	2.17e-17
SH2	245	336	5.42e-29
TyrKc	370	619	1.38e-131

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

SH2 domain

This is a SMART **SH2** domain (full annotation).

Position:	**245** to **336**
E-value:	**5.42134172537736e-29** (HMMER2)
SMART ACC:	SM000252
Definition:	Src homology 2 domains
Description:	Src homology 2 domains bind phosphotyrosine-containing polypeptides via 2 surface pockets. Specificity is provided via interaction with residues that are distinct from the phosphotyrosine. Only a single occurrence of a SH2 domain has been found in S. cerevisiae.
Interpro abstract (IPR000980):	The Src homology 2 (SH2) domain is a protein domain of about 100 amino-acid residues first identified as a conserved sequence region between the oncoproteins Src and Fps [(full abstract)

SH2 domain sequence (92 aa):

```
YEWYCRNMNRSKAEQLLRSEDKEGGFMVRDSSQPGLYTVSLYTKFGGEGSSGFRHYHIKE
TTTSPKKYYLAEKHAFGSIPEIIEYHKHNAAG
```

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. Protein features selected in various data tables are dynamically highlighted directly in the viewer. Using the interactive scale, any protein region can be selected and submitted for further BLAST analysis. (**B**) Tabbed interface collects various sources of external information about the protein analyzed. (**C**) Movable and resizable popup dialog displays the most important bits of information for any selected feature, with links to complete annotation.

singletons. This procedure significantly improves the results of all domain architecture queries and brings the domain counts to lower levels, comparable to the genomic mode database described below. The clustering procedure created 9.8 million multi protein clusters containing a total of 25 million proteins.

In addition to the regular protein database described above, SMART offers a 'genomic' analysis mode which contains only proteins from completely sequenced genomes. Synchronized with the current STRING version 11 (4), it currently contains >24 million proteins from 5090 complete genomes (477 *Eukaryota*, 4445 *Bacteria* and 168 *Archaea*), which is a 2.5-fold increase in both the number of proteins and genomes.

UPDATED PROTEIN VISUALIZATION ENGINE

Since the previous SMART release (version 8), all SMART protein schematics ('bubblograms') are generated in vector graphics formats (Figure 1). Protein schematics in various list displays (such as domain architecture analysis results) are displayed as dynamically generated inline SVG (Scalable Vector Graphics) images, which seamlessly scale to the users display size, regardless of its resolution. In SMART version 9, we have optimized all image generation routines to support latest web standards.

Version 9 brings a new version of the display applet used in the single protein annotation mode. It uses interactive SVG for the visualization and user interaction, offering the widest browser support. Protein schematics can be zoomed in to any level without quality loss, and saved as vector graphics for easy inclusion in user documents or publications.

Like the previous versions, the new SMART protein viewer interactively ties various parts of the annotation page. Selecting a predicted domain or other feature in any of the data tables will automatically highlight its position in the protein. Since many predicted features are not directly displayed in the protein schematic (mostly due to overlaps), this function simplifies the visual identification of relations among different protein features.

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.

Detailed information about any detected protein feature can be displayed in streamlined floating popup dialogs, enhancing the user experience and lowering the need to navigate across different web pages. A condensed version of domain annotation pages is included in the dialogs, with optional links to the complete annotation. In addition, several convenience functions are included, allowing users to copy the underlying amino acid sequence to their clipboard, or to submit the feature subsequence for further BLAST analysis.

EXPANDED PROTEIN INTERACTION DATA

With the update of the underlying protein databases, we have also synchronized our protein interaction data with the version 11 of the STRING database (4). Updated graphical representations of putative interaction partners are now available for >21 million proteins, with direct links to the corresponding network display pages in STRING where users can explore the interaction networks in detail.

UPDATED TAXONOMIC TREE DATA EXPORT

Domain architecture analysis functions in SMART allow users to simply access proteins containing combinations of particular domains. These can be also generated using combinations of GO terms associated to protein domains, and restricted to various taxonomic classes. In addition to the standard SMART protein schematic visualization, these data can also be exported into FASTA files or phylogenetic trees. The phylogenetic tree export has been updated and made compatible with the version 5 of the Interactive Tree of Life (iTOL) (15), with which these trees and their associated protein domain datasets are visualized, and can also be further annotated. Furthermore, the taxonomic database used for the tree generation was synchronized with the current NCBI taxonomy release (16).

DATABASE AND WEB SERVER OPTIMIZATIONS

The backend of SMART is a relational database management system (RDBMS), powered by the PostgreSQL engine, which stores the annotation of all SMART domains, protein annotation and sequences, taxonomy information and the pre-calculated protein analyses for the entire Uniprot (2), Ensembl (3) and STRING (4) proteomes. In addition to the predictions of all SMART and Pfam (12) domains, this includes various protein intrinsic features, like signal peptides, transmembrane and coiled coil regions. Due to constant growth of the number of annotated features, we are regularly restructuring our backend databases, and optimizing various parts of the server code in order to make the user experience satisfactory. Additionally, the server hardware that powers the sequence annotation searches and database queries has been replaced and significantly expanded with additional RAM and CPUs, greatly increasing the processing speed of user submitted proteins, and lowering the overall response times.

DATA AVAILABILITY

SMART data is freely available to academic users through EMBLem (www.embl-em.de).

FUNDING

This work has been partially supported by the German Network for Bioinformatics Infrastructure (de.NBI) and by EMBL. SK was supported by the JumpAR grant (FKZ 01Kl1706) funded by the Federal Ministry of Education and Research, Germany within the JPIAMR framework and the European Research Council grant MicrobioS (ERC-AdG-669830). Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Letunic,I. and Bork,P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.*, **46**, D493–D496.
2. UniProt, C. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
3. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
4. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
5. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 5857–5864.
6. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
7. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
8. Hickman,A.B. and Dyda,F. (2015) Mechanisms of DNA transposition. *Microbiol. Spectr.*, **3**, MDNA3–0034–2014.
9. Chandler,M., de la Cruz,F., Dyda,F., Hickman,A.B., Moncalian,G. and Ton-Hoang,B. (2013) Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat. Rev. Microbiol.*, **11**, 525–538.
10. Krupovic,M., Makarova,K.S., Forterre,P., Prangishvili,D. and Koonin,E.V. (2014) Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol.*, **12**, 36.
11. Grindley,N.D., Whiteson,K.L. and Rice,P.A. (2006) Mechanisms of site-specific recombination. *Annu. Rev. Biochem.*, **75**, 567–605.
12. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
13. Smyshlyaev,G., Barabas,O. and Bateman,A. (2019) Sequence analysis allows functional annotation of tyrosine recombinases in prokaryotic genomes. bioRxiv doi: https://doi.org/10.1101/542381, 06 February 2019, preprint: not peer reviewed.
14. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
15. Letunic,I. and Bork,P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, **47**, W256–W259.
16. Schoch,C.L., Ciufo,S., Domrachev,M., Hotton,C.L., Kannan,S., Khovanskaya,R., Leipe,D., McVeigh,R., O'Neill,K., Robbertse,B. *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**, baaa062.