RepeatsDB in 2025: expanding annotations of Structured Tandem Repeats proteins on AlphaFoldDB

Damiano Clementel, Paula Nazarena Arrías, Soroush Mozzafari, Zarifa Osmanli, Ximena Aixa Castro, Carlo Ferrari, Andrey V Kajava, Silvio C E Tosatto, Alexander Miguel Monzon

OXFORD UNIVERSITY PRESS | Nucleic Acids Research

RepeatsDB in 2025: expanding annotations of Structured Tandem Repeats proteins on AlphaFoldDB

Journal:	*Nucleic Acids Research*
Manuscript ID	NAR-02906-2024
Manuscript Type:	6 Database Issue
Key Words:	Structured Tandem Repeat Proteins, AlphaFoldDB, Protein Data Bank, Protein Tandem Repeat Annotation

DATA AVAILABILITY

Does the manuscript use or report the following? If so, please provide details in a Data Availability statement below and in the manuscript.	
New genome expression or sequencing data (ChIP-seq, RNA-seq…) - Must comply with ENCODE Guidelines. - All datasets must be validated via biological replicates. - Must deposit data in GEO or an equivalent publicly available depository and provide accession numbers, private tokens, reviewer login details and/or private URLs for Referees. - Excluding RNA-Seq, data must be viewable on the UCSC (eukaryotes) or other suitable genome browsers; must provide genome browser session links (even if GEO entries are publicly available). See next box below.	**No**
New genome-wide binding/interaction data - Must be viewable on the UCSC or another suitable genome browser. - Must provide genome browser session link in the Data availability field below.	**No**
Novel nucleic acid sequences - Must deposit in EMBL / GenBank / DDBJ. - Must provide sequence names and accession numbers.	**No**
Illumina-type sequencing data - Must submit data to BioProject/SRA, ArrayExpress or GEO. - Must provide link for reviewers (BioProject/SRA), login details (ArrayExpress) or accession numbers and private tokens (GEO).	**No**
Novel protein sequences - Must deposit UniProt using the interactive tool SPIN. - Must provide sequence names and accession number.	**No**
Novel molecular structures determined by X-ray crystallography, NMR and/or CryoEM/EM - Must deposit to a member site of the Worldwide Protein Data Bank (RCSB PDB, PDBe, PDBj) and provide the accession numbers. - If structures are unreleased (i.e. status HPUB), **MUST** upload: - the validation reports (.pdf) - molecular coordinates (.pdb or .mmcif). - one of the following: • X-ray data (.mtz, .cif) • NMR restraints and chemical shift files (.mr, .tbl or .str) • CryoEM map files (.map).	**No**
Novel molecular models based on SAXS, computational modeling, or other combinations of strategies that are generally not appropriate for deposition in the PDB - Must deposit coordinates and all underlying data in appropriate databases (including but not limited to the Small Angle Scattering Database and PDB-Dev). - Must report on validation of the structure against experimental data (if available)	**No**

or report on statistical validation of the structure by model quality assessment programs. If applicable, these should be uploaded as a Data file.	
Molecular behaviour studies derived from biological NMR spectroscopy data (not necessarily leading to new structures) - Must deposit NMR spectral data, including assigned chemical shifts, coupling constants, relaxation parameters (T1, T2, and NOE values), dipolar couplings, in BMRB.	**No**
Novel nucleic acids structure - Must deposit to NDB (via PDB if possible) and provide accession numbers.	**No**
Structures of nucleosides, nucleotides, other small molecules - Must deposit in the Cambridge Crystallographic Data Centre (CCDC) and provide the structure identifiers.	**No**
Mass spectrometry proteomics - Must deposit to ProteomeXchange consortium and provide Dataset Identifier and reviewer account details. If appropriate, data and corresponding details can also be deposited in the Panorama repository for targeted mass spec assays and workflows.	**No**
Microarray data - Must comply with the MIAME Guidelines - Must deposit the data to GEO or Array Express, and provide accession numbers and private tokens (GEO) or login details (ArrayExpress).	**No**
Quantitative PCR - Must comply with the MIQE Guidelines. - Details should be supplied in Materials and Methods section of manuscript.	**No**
Synthetic nucleic acid oligonucleotides including siRNAs or shRNAs - The manuscript should include controls to rule out off-target effects, such as use of multiple siRNA/shRNAs or inclusion of cDNA rescue data. - Manuscript should provide exact sequences, exact details of chemical modifications at any position, and source of reagent or precise methods for creation. These can be included in the main text or in Supplementary Material.	**No**
Software and source codes - Must deposit in FigShare and provide link to code and/or DOI or upload source code as Data file.	**No**
Gel images, micrographs, graphs, and tables - Optionally, may deposit in a general-purpose repository such as Zenodo or Dryad. If applicable, provide access details.	**No**

REFEREES – you will find data deposition details below
All the data is available from the URL: https://dev.repeatsdb.org

KEY POINTS

(3 bullet points summarizing the manuscript's contribution to the field)

Expanded STRP coverage with a 15-fold increase, including AlphaFoldDB predicted structures.

Improved user interface, enhanced search tools, and new statistics page for data exploration.

Advanced STRP detection, integrating manual curation and automatic annotations.

RepeatsDB in 2025: expanding annotations of Structured Tandem Repeats proteins on AlphaFoldDB

Damiano Clementel[1,§], Paula Nazarena Arrías[2,§], Soroush Mozzafari[1], Zarifa Osmanli[1], Ximena Aixa Castro[1], **RepeatsDB curators**, Carlo Ferrari[3], Andrey V. Kajava[4], Silvio C. E. Tosatto[1,5,*], Alexander Miguel Monzon[3,*]

[1]Department of Biomedical Sciences, University of Padova, Padova, Italy
[2]Department of Protein Science, KTH Royal Institute of Technology, Stockholm, Sweden
[3]Department of Information Engineering, University of Padova, Padova, Italy
[4]Centre de Recherche en Biologie Cellulaire de Montpellier, CNRS, Universite´ Montpellier, Montpellier, France
[5]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, Italy

To whom correspondence should be addressed. E-mail silvio.tosatto@unipd.it and alexander.monzon@unipd.it; Tel: +39 049 827 6269

§Co-first authors

Abstract

RepeatsDB (URL: https://dev.repeatsdb.org) stands as a key resource for the classification and annotation of Structured Tandem Repeat Proteins (STRPs), incorporating data from both the Protein Data Bank (PDB) and AlphaFoldDB. This latest release features substantial advancements, including annotations for over 34,000 unique protein sequences from more than 2,000 organisms, representing a fifteenfold increase in coverage.

Leveraging state-of-the-art structural alignment tools, RepeatsDB now offers faster and more precise detection of STRPs across both experimental and predicted structures. Key improvements also include a redesigned user interface and enhanced web server, providing an intuitive browsing experience with improved data searchability and accessibility. A new statistics page allows users to explore database metrics based on repeat classifications, while API enhancements support scalability to manage the growing volume of data. These advancements not only refine the understanding of STRPs but also streamline annotation processes, further strengthening RepeatsDB's role in advancing our understanding of STRP functions

Introduction

Tandem Repeat Proteins (TRPs) are a diverse class of proteins characterized by the repetition of specific sequence motifs (1). A subset within TRPs, known as Structured Tandem Repeat Proteins (STRPs), is characterized for preserving specific structural motifs, in the absence of sequence similarity (2). STRPs are a specialized subset of TRPs distinguished not only by their sequence repetition but also by their structural characteristics. Defined by regular secondary structure elements and discernible tertiary structures, STRPs may integrate into larger molecular assemblies. According to Kajava's classification (3), most STRPs fall into Classes III (elongated) and IV (closed). However, they can also be found in other classes such as aggregates, fibrous, and bead-on-string repeats. STRPs play pivotal roles in various biological processes and mechanisms. For example, leucine-rich repeats (LRRs) are crucial components of the extracellular domains of Toll-like receptors, which activate host immune responses (4), while STRPs such as DNA sliding clamps are essential in DNA replication processes (5). Moreover, their ability to interact with numerous and diverse proteins favors their involvement in complex formation mechanisms, contributing to the regulation of expression, transcription, and splicing (6). The growing focus on TRPs in recent research stems from their crucial roles in health-related studies (7, 8) and their utility in the field of protein engineering (9–11).

The first version of RepeatDB (12) was released over a decade ago, becoming the primary repository for STRPs with high-quality annotations of repeat regions, units, and insertions on experimental protein structures according to Kajava's classification. Subsequent releases in 2016 (12) and 2021 (13) have further enhanced these capabilities, extending the manually curated and automatic annotations, and expanding the classification to comprehensively describe STRPs present in the Protein Data Bank (PDB) (14).

The exponential growth in accurate protein structural data in recent years, mostly driven by state-of-the-art Artificial Intelligence (AI) protein structure prediction methods such as AlphaFold (15), RoseTTAFold (16), and ESM Fold (17), has revolutionized the structural bioinformatics community. This surge presents unique challenges in automatically detecting and classifying STRPs with tools like RepeatsDB-Lite (18), TAPO (19), and RAPHAEL (20). Recently developed, STRPsearch (21) is a rapid and accurate method for detecting STRPs from protein structures. This tool enables quick scanning of thousands of structures, facilitating the detection and classification of repeat regions, and identifying units and insertions.

Here, we introduce the updated version of RepeatsDB, accessible at https://dev.repeatsdb.org. This version is designed to incorporate the latest structural data and advances in TRPs, increasing the number of annotated protein sequences fifteen fold compared to the previous version. Extensive community curation efforts have enhanced the

database, identifying specific regions, units, and insertions within PDB experimental structures. Additionally, it features automatic annotations of structural models from SwissProt proteins deposited in AlphaFoldDB (22), using STRPsearch. It includes a complete re-annotation of manually curated annotations in order to guarantee their quality, as well as a revised classification. Furthermore, we have upgraded the website and web server architecture, introducing a dedicated platform for manual annotations of STRPs. This not only allows for a faster review process of manually curated annotations but also improves data accessibility and usability.

Progress and new features

Database content

RepeatsDB aims to establish itself as the fundamental repository for the annotation and classification of STRPs, sourced from primary structural data databases, such as the PDB for experimental structures and AlphaFoldDB for structural models. Additionally, continuous efforts in the community focus on identifying and characterizing novel tandem repeat proteins, as well as accurately defining repeat protein families. Recently, RepeatsDB has been included as an external resource in the InterPro (23), providing repeat region annotations and enhancing interoperability between these databases. This collaboration also contributes to refining and characterizing repeat protein family boundaries within Pfam at the sequence level (24) and to better defining repeat units from a structural perspective.

In this new release, the number of entries in RepeatsDB has increased more than fifteen times in terms of protein sequences with repeat annotations. A total of 34,319 unique protein sequences, each identified by different UniProtKB identifiers, now feature at least one repeat region detected in PDB and/or AlphaFoldDB (Table 1). By providing online tools for the annotation of STRPs, community curation efforts have been directed towards the review of more than 15,000 experimentally observed structures from the PDB. Using the new algorithm, STRPsearch (21), over 30,000 repeated regions were automatically extracted from AlphaFoldDB/Swiss-Prot. STRPsearch utilizes a Tri-Unit Library composed of manually curated structures from RepeatsDB and employs the fast and accurate structural alignment method, FoldSeek (25). This tool enables rapid and precise detection and classification of repeat regions, units, and insertions within protein structures.

Source	Annotation	Proteins	Single-chain	Regions	Units

	type		structures		
Protein Data Bank	*Reviewed*	4,092	15,141	16,282	106,408
AlphaFoldDB	*Predicted*	30,220	30,220	33,207	250,416

Table 1: RepeatsDB in 2025 data. “Proteins” corresponds to the number of different UniProtKB identifiers mapping to a single PDB chain.

Classification Schema

This updated version of RepeatsDB retains the 4-level classification scheme introduced in version 3.0, which is based on Kajava’s class assignment (3) and aims at integrating and maintaining coherence with the Pfam database. While no new topologies have been introduced in this update, two existing topologies, namely the “Beta trefoil” and the “Alpha/beta trefoil”, have been merged into 'trefoil' based on structural clustering.

Several adjustments have been made at the third level, 'Fold', to accommodate new data and refine the definition of RepeatsDB 'Fold'. For instance, two new folds have been created within the 'Alpha beads' topology to include 'Spectrin repeats' and 'FF-repeats'. Additionally, two folds have been established in the 'Beta solenoid' topology to account for solenoid 'handedness', which refers to the direction in which the chain winds around the molecular axis (26).

Significant progress has also been made in 'Clan' assignment, largely facilitated by clustering the STRPs based on structural similarity at three hierarchical levels of single unit, tri-unit, and region, by DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm of scikit-learn (27). The resulting clusters were analyzed and compared in relation to their corresponding Pfam annotations of their members. This method improved the refinement of clan classifications by enabling the merging of existing clans or their division into new ones. For example, a total of 11 new clans have been assigned across the seven folds of the 'Propeller' topology.

Data generation and updates

RepeatsDB Biocuration tool

A new feature of this version of RepeatsDB is the dedicated Biocuration tool, developed to manually curate and verify STRPsearch predictions or perform *de novo* annotations on protein structures. The RepeatsDB Biocuration tool offers an easy-to-use user interface that allows

biocurators to log in and access STRPsearch predictions stored in the staging database (see the implementation section), or to annotate a protein structure from scratch. It uses the ORCID authentication service to link annotations to a unique researcher identifier. Biocurators can log into the Biocuration tool using their ORCID identifiers, gaining access to a personal dashboard where they can create, edit, and submit annotations to RepeatsDB. All these annotations remain on hold until an expert curator reviews and approves them for publication in the next periodic release.

Biocurators can manually define or adjust the boundaries of repeat regions, units, and insertions, as well as assign the appropriate classifications. To facilitate this, the tool incorporates the same components (structure and sequence viewer) used throughout the database interface. Both are updated to reflect changes in units, regions, and insertions annotated through the main table, located in the bottom right side of the page. Most importantly, each repeat region is associated with a specific branch of the classification hierarchy (Class → Topology → Fold → Clan), and each unit and insertion is linked to a specific repeat region. Once the curation is complete, the user submits it for review.

Data curation on experimental structures

The current version of RepeatsDB includes 15,141 manually curated STRPs based on experimental structures from the PDB, covering a total of 4,092 distinct UniProt protein sequences. This represents a nearly threefold increase compared to version 3.0, published in 2022. This achievement reflects a substantial human effort dedicated to ensuring the highest quality and continuous expansion of our reviewed entries. To achieve this goal, RepeatsDB was a key resource in the REFRACT project, a Marie Skłodowska-Curie Research and Innovation Staff Exchange (RISE) Horizon 2020 consortium funded by the European Union (URL: http://refract-rise.eu). This project involved institutions from Europe and Latin America, focusing on understanding tandem repeat proteins and establishing a standardized framework for their classification and annotation. Throughout the project, many seconded staff members from Latin American institutions made significant contributions to the curation of the RepeatsDB database.

The overall curation process was divided into two phases. The first phase involved a thorough review of all curated data from RepeatsDB version 3.0 to check inconsistencies and apply uniform criteria to all reviewed entries. This initial phase was carried out by expert curators

with specialized knowledge in the biology and structural aspects of STRPs. The second phase consisted of running STRPsearch on the PDB to target the curation of those PDB entries that map to UniProt protein sequences not yet covered in RepeatsDB. The new version ensures that there is at least one curated experimental protein structure per protein sequence mapping to the PDB.

AlphaFoldDB entries

In this release, RepeatsDB included potential STRPs detected on AlphaFold structural models deposited in AlphaFoldDB. To accomplish this, STRPsearch was used to analyze the AlphaFold models of 542,378 protein structures in the SwissProt database. The parameters of STRPsearch were set to identify STRPs with an E-value threshold below 10^{-5}, while all other parameters were left at their default values. This analysis resulted in the identification of 30,220 unique proteins from 2,598 different organisms, containing 33,207 repeat regions and 250,416 repeat units. Figure 1 displays the top 20 organisms based on the number of predicted STRPs, along with the different topologies represented within the total STRPs for each organism.

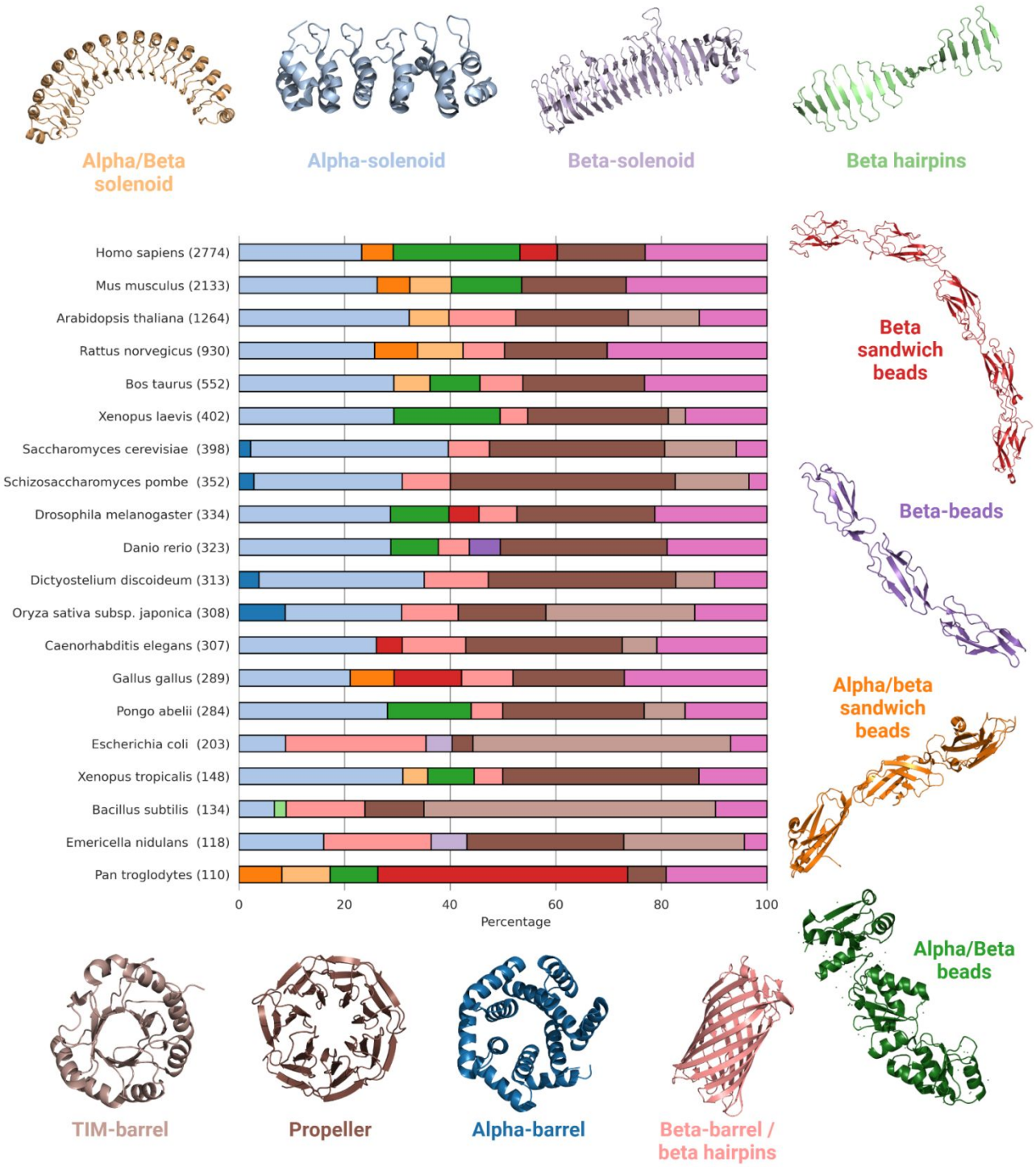


Figure 1: Proportion of STRP topologies identified in the top 20 most frequently occurring organisms in AlphaFoldDB/SwissProt.

Web site

RepeatsDB data is now accessible through both the website and a dedicated server API. Key improvements in this latest RepeatsDB release include: i) enhancements to the API, which now offer improved performance with pagination and a RESTful architecture, ensuring scalability to handle the growing number of entries; ii) a redesigned entry page that more

effectively highlights repeat regions, associated Pfam domains, external annotations from InterPro, and structural classifications; iii) a new statistics page that allows users to explore database metrics based on repeat classifications; and iv) more streamlined browse and search interface that improves data searchability and accessibility, making it easier for users to find specific entries.

Browsing and searching data

In the new release of RepeatsDB, the main way of filtering a specific subset of annotations is through the search page. It presents a search form at the top, which is visible, hence usable, despite the user scrolling down the page. The forms allow filters for annotated protein structure, chain and source database. Recognized source databases are currently the PDB and the AlphaFoldDB. Moreover, it allows one to search for the type of annotation: whether this is a manually curated or an automatically predicted one by clicking on the top buttons, namely “reviewed” and “predicted”. Those work as checkboxes. Moreover, the interface allows searching for region classification and Pfam domain. In both those fields more than one value can be specified by clicking on the add button, or by pressing enter while focusing the input field. Each field performs a logical *or* operation: in case at least one searched term is found in an annotation, it is returned. Otherwise, it is discarded from the results.

Another way of accessing the search page is through the home page or the search box on the right corner of the upper navigation bar. In both cases a text form is presented: if the searched text matches exactly a UniProtKB identifier or a PDB structure and chain identifier, it will automatically attempt to redirect the user to the associated annotation. Otherwise, it will redirect the user to the search page, where the form will be partially compiled with the previously entered information.

It is worth mentioning that the classification tree on the home page is clickable. Thus, by clicking on one of its nodes the user will be redirected to the search page. In this case, the search form will be configured to search only for regions in the selected classification branch.

Entry page

The entry page starts with a summary of essential information, including the positions of repeated regions, their classifications, the corresponding UniProt, Pfam, and InterPro accessions with cross-links and boundaries for each annotation, as well as the Gene Ontology terms. The next section features an interactive structure viewer, accompanied by detailed information for each region displayed in table form. In the table, rows representing regions are colored violet, repeated units appear in alternating red and blue patterns, and insertions are

highlighted in yellow. The table details the start and end positions of repeat regions, units, and insertions, as well as the corresponding region classification or parent region for units and insertions.

The next section of the page includes a protein feature viewer, which allows users to view different sequence features from multiple sources in a single visualization. The 'RepeatDB annotations' track displays repeat regions, units, and insertions (if present) mapped on the sequence from the source structure, either PDB or AlphaFoldDB. The other two tracks display external features. For PDB structures, UniProt coverage is shown in blue; for AlphaFold models, InterPro coverage is shown in blue. Pfam coverage is displayed in green for both. These tracks can be expanded into sub-tracks for visualizing the accession numbers. All tracks are zoomable using the mouse wheel or by selecting a specific region.

In the bottom section, users can select a region using the violet buttons to access detailed sequence and structural analyses of its units. Below this, a breakdown of all classifications annotated for the entry and a brief description are provided. Multiple sequence alignment is conducted using Clustal Omega software with default parameters. The alignment is displayed in a sequence viewer, where residues are colored based on their physicochemical properties using the Jalview Zappo color scheme. Additionally, a sequence similarity matrix is provided to visualize the similarities between each pair of unit sequences. Multiple structural alignments are performed using TM-align (28) with default parameters. The displayed results include the superimposed unit structures and a matrix displaying TM-scores on the upper diagonal and Root Mean Square Deviation (RMSD) on the lower between units. Additionally, a sequence viewer shows the sequence alignment derived from the structural alignment. All this information can be easily downloaded in several formats by clicking the blue buttons next to the corresponding titles.

Server and API

RepeatsDB features a novel architecture that integrates data storage and distribution capabilities. The architecture is divided into two partitions: staging and production. The staging partition handles biocuration activities, allowing annotations to be created, updated, and deleted by users. These annotations contain only the minimal information required to describe themselves. The production partition manages consolidated and enriched data, which has been processed using external resources and third-party software. Data flows into the production partition through a specific update pipeline executed before each public release of the database. This pipeline performs structural and multiple sequence alignments among units, calculates TM-Score and RMSD matrices, and maps regions annotated on PDB sequences to their corresponding UniProt sequences using SIFTS (29).

Annotations are described by the Annotation Graph library. Although based on a NoSQL database, the library adopts a graph-like data structure where nodes represent generic biological annotations. This architecture allows the staging partition to remain online without shutting down the library implementation when releasing a new production version. Moreover, it decouples bio-curation efforts, ensuring that individual users are unaware of others' activities. A limited group of trusted users, referred to as reviewers, selects annotations suitable for the next release and flags them as reviewed. Once an annotation (node) is reviewed, it becomes immutable; to modify or delete its content, a child node must be created. This approach ensures that specific snapshots of the staging partition can always be reproduced.

The Annotation Graph library is included with RepeatsDB. Its implementation has three levels: the front end, the back end, and the database. The front end is based on the Angular framework. The back end is developed using NestJS over the Express framework. The database is a MongoDB instance. In RepeatsDB, there are two different front-end applications: (1) RepeatsDB Bio-Curation Tool: handles staging data, allowing entries in the database to be created, modified, or deleted; (2) RepeatsDB User Interface: allows users to browse through production data. Both applications communicate with the same back end, which in turn communicates with two different databases within the same MongoDB instance, implementing staging and production partitioning. To achieve this, ngx-mol-viewers package (URL: https://biocomputingup.github.io/ngx-mol-viewers/) was developed to provide a collection of modular and reusable components easily integrated into any Angular-based web application. This package supports the sequence, structure, and feature viewers used in the RepeatsDB UI and Bio-Curation tool.

The Swagger documentation page for the API (URL: https://dev.repeatsdb.org/api/) provides a comprehensive overview of the available endpoints, operations, and data models for interacting with the RepeatsDB API. The page is structured in a user-friendly manner, with a navigation bar on the left-hand side that allows you to browse through different sections of the documentation. The main content area displays detailed information about each API endpoint, including its path, HTTP methods (such as GET, POST, PUT, DELETE), and a brief description of its purpose.

Conclusions and future work

Understanding STRPs is crucial due to their extensive roles in biological systems and their potential for biotechnological applications. STRPs are fundamental not only to numerous cellular functions but also represent a rich domain for therapeutic interventions and biomaterial

design. Their unique structural features and dynamic roles in cell signaling, molecular recognition, and self-assembly processes make them a focal point for advanced research in protein science. RepeatsDB serves as the primary repository for annotating and classifying these proteins. This new version significantly advances the curation and classification of STRPs, facilitating integration with broader structural protein data repositories. The combination of manual and automated curation processes, supported by innovative tools such as STRPsearch and the Bio-Curation tool, has enabled a comprehensive update and expansion of the database. RepeatsDB was the core resource of the REFRACT consortium and is crucial for the efforts pursued by the COST Action “Machine Learning for Non-Globular Proteins” (ML4NGP).

Looking ahead, RepeatsDB plans to integrate with APICURON (30), aiming to recognize and reward the efforts of biocurators. This integration will foster a more engaged and active community around STRP data curation, which is particularly crucial as we extend our efforts to include AlphaFold models from TrEMBL. Additionally, we will continue to enhance our classification schemas and refine repeat definitions in collaboration with the InterPro consortium. These initiatives will ensure that RepeatsDB remains a leading resource in the detection and classification of STRPs, providing essential support for researchers worldwide.

Data Availability

All the data is available from the URL: https://repeatsdb.org.

Acknowledgements

We acknowledge ELIXIR-IIB (elixir-italy.org), the Italian Node of the European ELIXIR infrastructure (elixir-europe.org), for supporting the development and maintenance of RepeatsDB. The authors would like to express their gratitude to all members of the REFRACT consortium (refract-rise.eu) for their invaluable contributions to advancing the study of protein tandem repeats. Special thanks go to Miguel Andrade for his insightful discussions and constructive feedback on our database. We also extend our appreciation to Gustavo Parisi, Antonio Lagares, and Layla Martinez Hirsh for their support in curation efforts and for promoting this important field throughout Latin America. We would also like to acknowledge Alex Bateman and Sara Chuguransky from the InterPro consortium for their joint efforts in improving protein repeat annotations in Pfam and RepeatsDB. We thank Jianchen Lu for her initial help in data curation and Matina Bevilacqua for her support in database updates and

visualization. We express our gratitude to Ivan Mičetić for his contributions in updating the IT infrastructure of the RepeatsDB server and his support in the server backend.

Funding

ELIXIR, the research infrastructure for life-science data. This work was supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 823886 (H2020 MSCA-RISE "REFRACT") as well as European Union - NextGenerationEU through "Italiadomani-PNRR" project ElixirxNextGenIT [IR0000010] to SCET. This publication is partially based upon work from COST Action ML4NGP (CA21160), supported by COST (European Cooperation in Science and Technology). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Funding for open access charge: COST Action ML4NGP (CA21160).

References

1. Kajava,A.V. and Tosatto,S.C.E. (2018) Editorial for special issue "Proteins with tandem repeats: sequences, structures and functions". *J. Struct. Biol.*, **201**, 86–87.
2. Monzon,A.M., Arrías,P.N., Elofsson,A., Mier,P., Andrade-Navarro,M.A., Bevilacqua,M., Clementel,D., Bateman,A., Hirsh,L., Fornasari,M.S., *et al.* (2023) A STRP-ed definition of Structured Tandem Repeats in Proteins. *J. Struct. Biol.*, **215**, 108023.
3. Kajava,A.V. (2012) Tandem repeats in proteins: From sequence to structure. *J. Struct. Biol.*, **179**, 279–288.
4. Leulier,F. and Lemaitre,B. (2008) Toll-like receptors--taking an evolutionary approach. *Nat. Rev. Genet.*, **9**, 165–178.
5. Arrías,P.N., Monzon,A.M., Clementel,D., Mozaffari,S., Piovesan,D., Kajava,A.V. and Tosatto,S.C.E. (2023) The repetitive structure of DNA clamps: An overlooked protein tandem repeat. *J. Struct. Biol.*, **215**, 108001.
6. Mac Donagh,J., Marchesini,A., Spiga,A., Fallico,M.J., Arrías,P.N., Monzon,A.M., Vagiona,A.-C., Gonçalves-Kulik,M., Mier,P. and Andrade-Navarro,M.A. (2024) Structured Tandem Repeats in Protein Interactions. *Int. J. Mol. Sci.*, **25**, 2994.
7. Fournier,D., Palidwor,G.A., Shcherbinin,S., Szengel,A., Schaefer,M.H., Perez-Iratxeta,C. and Andrade-Navarro,M.A. (2013) Functional and Genomic Analyses of Alpha-Solenoid Proteins. *PLoS ONE*, **8**, e79894.
8. de Wit,J., Hong,W., Luo,L. and Ghosh,A. (2011) Role of leucine-rich repeat proteins in the development and function of neural circuits. *Annu. Rev. Cell Dev. Biol.*, **27**, 697–729.
9. Höcker,B. (2014) Design of proteins from smaller fragments — learning from evolution. *Curr. Opin. Struct. Biol.*, **27**, 56–62.
10. Brunette,T.J., Parmeggiani,F., Huang,P.-S., Bhabha,G., Ekiert,D.C., Tsutakawa,S.E., Hura,G.L., Tainer,J.A. and Baker,D. (2015) Exploring the repeat protein universe through computational protein design. *Nature*, **528**, 580–584.
11. Wu,K., Bai,H., Chang,Y.-T., Redler,R., McNally,K.E., Sheffler,W., Brunette,T.J., Hicks,D.R., Morgan,T.E., Stevens,T.J., *et al.* (2023) De novo design of modular peptide-binding proteins by superhelical matching. *Nature*, **616**, 581–589.

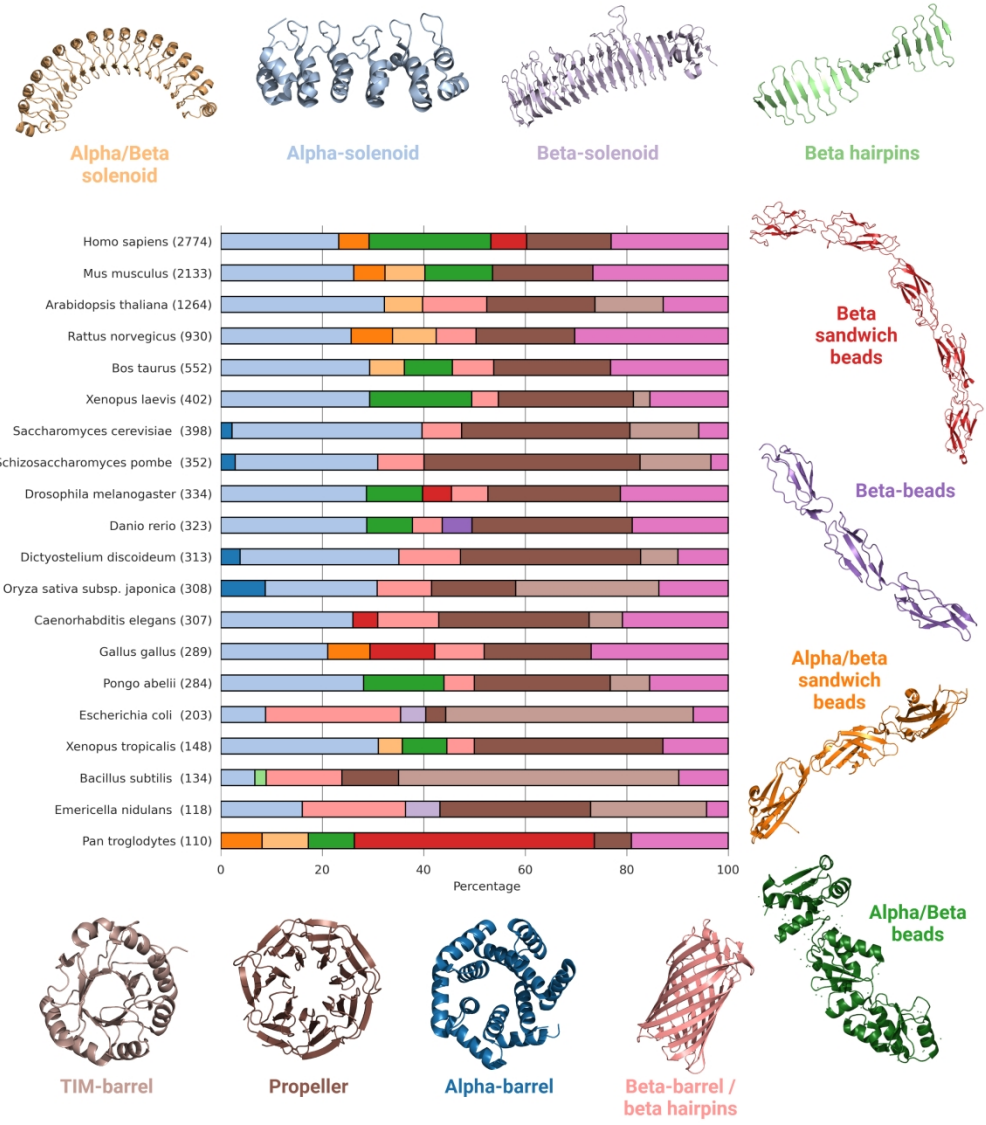
12. Di Domenico,T., Potenza,E., Walsh,I., Parra,R.G., Giollo,M., Minervini,G., Piovesan,D., Ihsan,A., Ferrari,C., Kajava,A.V., *et al.* (2014) RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Res.*, **42**, D352-357.
13. Paladin,L., Bevilacqua,M., Errigo,S., Piovesan,D., Mičetić,I., Necci,M., Monzon,A.M., Fabre,M.L., Lopez,J.L., Nilsson,J.F., *et al.* (2021) RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Res.*, **49**, D452–D457.
14. PDBe-KB consortium (2022) PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.*, **50**, D534–D542.
15. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
16. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D., *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
17. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
18. Hirsh,L., Paladin,L., Piovesan,D. and Tosatto,S.C.E. (2018) RepeatsDB-lite: a web server for unit annotation of tandem repeat proteins. *Nucleic Acids Res.*, **46**, W402–W407.
19. Do Viet,P., Roche,D.B. and Kajava,A.V. (2015) TAPO: A combined method for the identification of tandem repeats in protein structures. *FEBS Lett.*, **589**, 2611–2619.
20. Walsh,I., Sirocco,F.G., Minervini,G., Di Domenico,T., Ferrari,C. and Tosatto,S.C.E. (2012) RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics*, **28**, 3257–3264.
21. Mozaffari,S., Arrías,P.N., Clementel,D., Piovesan,D., Ferrari,C., Tosatto,S.C.E. and Monzon,A.M. (2024) STRPsearch: fast detection of structured tandem repeat proteins. 10.1101/2024.07.10.602726.
22. Varadi,M., Bertoni,D., Magana,P., Paramval,U., Pidruchna,I., Radhakrishnan,M., Tsenkov,M., Nair,S., Mirdita,M., Yeo,J., *et al.* (2024) AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.*, **52**, D368–D375.
23. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L., *et al.* (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
24. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
25. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (2024) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, **42**, 243–246.
26. Kajava,A.V. and Steven,A.C. (2006) Beta-rolls, beta-helices, and other beta-solenoid proteins. *Adv. Protein Chem.*, **73**, 55–96.
27. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
28. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
29. Velankar,S., Dana,J.M., Jacobsen,J., van Ginkel,G., Gane,P.J., Luo,J., Oldfield,T.J., O'Donovan,C., Martin,M.-J. and Kleywegt,G.J. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.*, **41**, D483-489.
30. Hatos,A., Quaglia,F., Piovesan,D. and Tosatto,S.C.E. (2021) APICURON: a database to

credit and acknowledge the work of biocurators. *Database*, **2021**, baab019.

RepeatsDB curators

Estefanía Lorena Borucki[5], Maia Cabrera[6], Patricio Chinestrad[6], Ian Czarnowski[6], Jose Francisco Lombardo[7], Pablo Lorenzano Menna[6], Ezequiel Gerardo Mogro[8], Carla Luciana Padilla Franzotti[9], Julia Yamila Santillan[9]

[5]Laboratorio de Biotransformaciones y Química de Ácidos Nucleicos, Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Argentina
[6]Laboratorio de Farmacología Molecular, Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Argentina
[7]Instituto de Investigaciones Bioquímicas de La Plata (INIBIOLP), Facultad de Ciencias Médicas, Universidad Nacional de La Plata (UNLP)-Consejo Nacional de Investigaciones Científicas Y Técnicas (CONICET), La Plata, Argentina.
[8]Instituto de Biotecnología y Biología Molecular (IBBM), CONICET, CCT-La Plata, Universidad Nacional de La Plata (UNLP), Argentina.
[9]Department of Science and Technology, National University of Quilmes-CONICET, Bernal, Argentina



Proportion of STRP topologies identified in the top 20 most frequently occurring organisms in AlphaFoldDB/SwissProt predictions.

645x774mm (118 x 118 DPI)