Database, 2025, baaf035
DOI: https://doi.org/10.1093/database/baaf035
Database update

GenDiS3 database: census on the prevalence of protein domain superfamilies of known structure in the entire sequence database

Sarthak Joshi[1], Shailendu Mohapatra[2], Dhwani Kumar[1], Adwait Joshi[1], Meenakshi Iyer[1], Ramanathan Sowdhamini[1,2,3,*]

[1]National Centre for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560065, India
[2]Computational Biology, Insitute of Bioinformatics and Applied Biotechnology, Bangalore 560100, India
[3]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

[*]Corresponding author. National Centre for Biological Sciences, Tata Institute of Fundamental Research,GKVK Campus, Bellary Road, Bangalore 560065, India. E-mail: mini@ncbs.res.in

Citation details: Joshi, S., Mohapatra, S., Kumar, D. *et al.* GenDiS3 database: census on the prevalence of protein domain superfamilies of known structure in the entire sequence database. *Database* (2025) Vol. 2025: article ID baaf035; DOI: https://doi.org/10.1093/database/baaf035

Abstract

Despite the vast amount of sequence data available, a significant disparity exists between the number of protein sequences identified and the relatively few structures that have been resolved. This disparity highlights the challenge in structural biology to bridge the gap between sequence information and 3D structural data, and the necessity for robust databases capable of linking distant homologs to known structures. Studies have indicated that there are a limited number of structural folds, despite the vast diversity of proteins. Hence, computational tools can enhance our ability to classify protein sequences, much before their structures are determined or their functions are characterized, thereby bridging the gap between sequence and structural data. GenDiS (Genomic Distribution of Superfamilies) is a repository with information on the genomic distribution of protein domain superfamilies, involving a one-time computational exercise to search for trusted homologs of protein domains of known structures against the vast sequence database. We have updated this database employing advanced bioinformatics tools, including DELTA-BLAST (domain enhanced lookup time accelerated BLAST) for initial detection of hits and HMMSCAN for validation, significantly improving the accuracy of domain identification. Using these tools, over 151 million sequence homologs for 2060 superfamilies [SCOPe (Structural Classification of Proteins extended)] were identified and 116 million out of them were validated as true positives. Through a case study on glycolysis-related enzymes, variations in domain architectures of these enzymes are explored, revealing evolutionary changes and functional diversity among these essential proteins. We present another case, LOG gene, where one can tune in and find significant mutations across the evolutionary lineage. The GenDiS database, GenDiS3, and the associated tools made available at https://caps.ncbs.res.in/gendis3/ offer a powerful resource for researchers in functional annotation and evolutionary studies.

Database URL: https://caps.ncbs.res.in/gendis3/

Introduction

Proteins, the fundamental building blocks of biological functions, are composed of one or more specific functional regions known as domains. Each domain within a protein can be characterized by its distinct structural and functional properties. Structurally, domains are classified by databases such as SCOP (Structural Classification of Proteins) (1,2) and CATH (Class, Architecture, Topology, and Homologous) superfamilies (3,4), which organize protein structures based on hierarchical relationships of folding patterns and architecture. On the sequence level, domains are defined through databases like Pfam (5) and InterPro (6), which use amino acid sequences to identify and group protein domains across different species.

Despite the extensive sequence data that have been accumulated, there remains a significant disparity between the sheer number of protein sequences identified and the relatively few protein structures that have been resolved. This gap underscores a major challenge in structural biology: the need to connect sequence information with 3D structural understanding. Addressing this challenge is critical, as it enhances our comprehension of protein function and interactions. Therefore, there is a pressing need for robust and comprehensive databases that can accurately link distant homologs to known structures, facilitating the prediction and modelling of protein structures from sequence data alone.

Research has shown that there is a limited number of structural folds, despite the immense diversity of protein sequences (7,8). Although the Protein Data Bank (PDB) (9) has experienced substantial growth in the number of deposited structures, the discovery of new folds has not kept pace with this increase. The relatively stable number

Received 20 September 2024; Revised 8 January 2025; Accepted 9 April 2025

of unique folds suggests that protein structural diversity is constrained compared to the exponential growth of sequence data. This observation underscores the importance of computational methods like Hidden Markov models (HMMs) and position-specific scoring matrices (PSSMs) in protein classification. These methods can leverage structural conservation to enhance detection sensitivity, especially when classifying protein domains with low sequence identity among homologous proteins. Basic tools like BLAST (10) often fall short in identifying remote homologs due to subtle sequence similarities. However, advanced methods such as HMMs and PSSMs have demonstrated greater efficacy. HMMs, in particular, are adept at modelling the probabilistic nature of amino acid substitutions and insertions/deletions, providing a powerful means to detect distant relationships within protein families (11).

The previous versions of the GenDiS (Genomic Distribution of Superfamilies) database have identified homologs of SCOP superfamily members, leveraging the PASS2 database (12), which provides curated alignments for SCOP superfamily members. PASS2 is a database that provides structure-based sequence alignments for SCOPe domains that have <40% sequence identity with each other. PASS2 also provides HMMs derived from these alignments. The first version of GenDiS (13) utilized PSI-BLAST (14) and HMMSEARCH (15) for homolog identification. The second version (16,17) utilized CS_BLAST (18) along with PSI-BLAST. CS-BLAST enhances the sensitivity of homolog detection by taking into account the context of amino acid substitutions, thereby providing better detection of distant relationships (19), but is inherently quite greedy. The combination of CS-BLAST and PSI-BLAST, accompanied by strict validation, allowed for a more robust identification process, enabling the detection of homologs with higher sensitivity and specificity (16). Both versions utilized members from the PASS2 database, which provides high-quality alignments of SCOP superfamily members. This database is crucial for ensuring the accuracy and relevance of the homolog identification process during validation, as it offers a reliable reference for structural and functional annotation of protein sequences. The current version of the GenDiS database utilizes DELTA-BLAST (domain enhanced lookup time accelerated BLAST) (20), which integrates domain information from conserved domain databases into the BLAST search process. This allows for more sensitive and accurate detection of distant homologs by leveraging domain-specific information. For validation, the new version employs HMMSCAN, a tool from the HMMER suite (hmmer.org). This combination of DELTA-BLAST for initial detection and HMMSCAN for validation ensures a high level of accuracy and reliability in the identification process. This updated approach builds on the strengths of previous versions by incorporating more sophisticated search and validation tools, thereby improving the comprehensiveness and precision of the GenDiS database.

Materials and methods

Datasets and tools used in sequence search

Datasets

SCOPe (Structural Classification of Proteins extended) is an extended version of the SCOP database for the classification of proteins based on structural relations. Astral compendium (21) provides sequences and coordinates for domains from 7 out of 12 SCOPe classes. All sequences from Astral SCOPe version 2.08, filtered at 40% sequence identity, were used as query sequences to identify potential homologs. The National Center for Biotechnology Information (NCBI) non-redundant (NR) database of protein sequences (version September 2022) (22) was used to search for homologs. For domain identification in query sequences, the searches utilized the conserved domain database (23) (cdd_delta) from NCBI.

Sequence search

The DELTA-BLAST tool was employed for this purpose, using the BLAST executables version 2.13.0 (24). The searches utilized the conserved domain database (cdd_delta) from NCBI with a domain inclusion threshold of 0.05 for RPS-BLAST. An *E*-value threshold of 0.001 was set to consider a hit significant (Fig. 1). The sequence database used for these searches was the NCBI NR database from September 2022.

Validation

The hits identified from the DELTA-BLAST search, starting from individual PASS2 entries as queries, were combined for each superfamily. These sequence hits were validated using HMMs from PASS2 alignment and single-query HMMs built from Astral sequence set at 70% sequence identity. The HMMs from PASS2 versions 2.4 and 2.7 were used for validation. The sequences for building single-query HMMs were obtained from the scop.berkley.edu webserver. Single-query HMMs were built using the hmmbuild program from the HMMER (version 3.3.2) suite. HMMSCAN from the HMMER suite was used for validation with an *E*-value threshold of 0.01. The sequence hits from each superfamily was used as a query set. The HMMs from PASS2.4 and PASS2.7 and single-query HMMs from Astral sequence set at 70% sequence identity were compiled together and used as a target set (Fig. 1). The domain architectures (DAs) for the sequences were also obtained using the domain table output of the HMMSCAN results.

Identification of DAs

The DAs of identified homologs were determined using HMMSCAN, utilizing the PASS2 HMM library along with single-query HMMs. This library consists of PASS2 versions 4 (25) and 7 (26), which correspond to SCOPe versions 1.75 and 2.07, respectively, and single-query HMMs generated from members of SCOPe 2.08. The regions of each sequence that aligned with an HMM were annotated with the domain associated with the superfamily of the HMM. The Pfam DAs for the validated homologs were also identified using HMMSCAN and the Pfam-A HMM library.

Organization of the website

The website for organizing and analysing protein domain superfamilies across the GenDiS database consists of two main components: the user end (front end) and the server end (backend). The front end was developed using HTML, CSS, JavaScript, bootstrap, and jQuery, providing an intuitive and user-friendly interface and a dynamic and responsive website. AJAX was used for asynchronous requests to the server. Data for the backend is stored using MongoDB, a flexible and scalable NoSQL database. The backend logic and webpage rendering are handled by Python Flask,

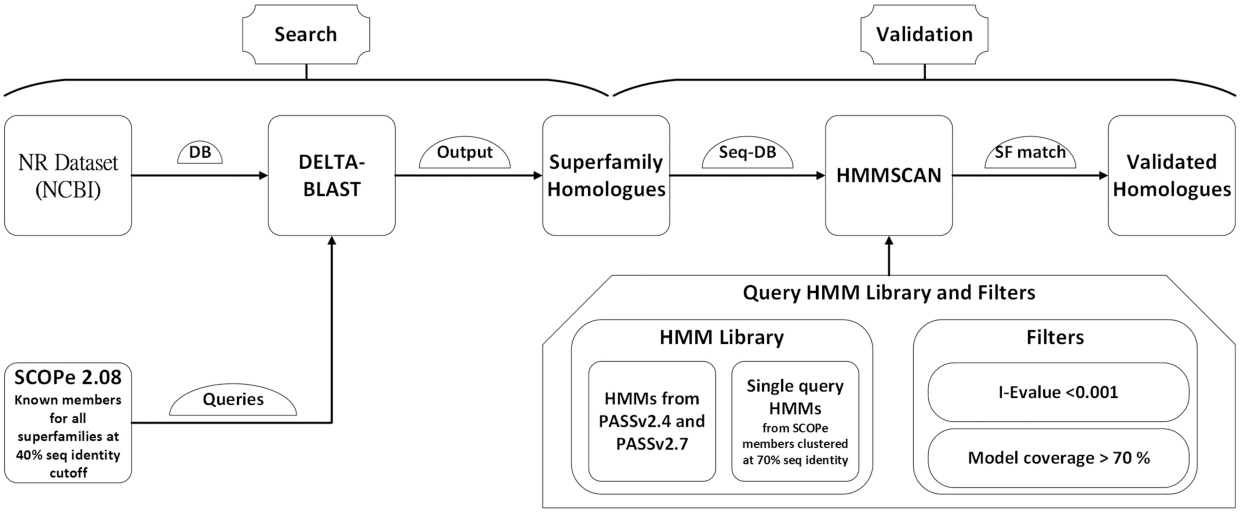

Figure 1. GenDiS3 workflow for search and validation of homologs.

a lightweight web framework. Data collection and storage involved PyMongo for accessing the MongoDB database containing precollected data. Custom Python modules were used for parsing XML files and generating CSV files for visualization purposes (27). The tool for DA prediction utilizes HMMSCAN from HMMER version 3.3.2, while the sequence alignment tools employ Clustal Omega version 1.2.4 (28).

Analysis using case studies

This study utilized a bioinformatics pipeline to investigate the taxonomic distribution and DAs of glycolysis-related enzymes. All KEGG (29) orthologies related to glycolysis were retrieved using the KEGG-API. This provided a comprehensive list of orthologous groups involved in glycolytic pathways. Using the obtained KEGG orthologies, all associated genes were identified through the KEGG-ORGANISMS database. This ensured the collection of genes that are directly linked to the orthologies involved in glycolysis across different organisms. The taxonomic distribution of the identified genes was analysed to understand the prevalence of these genes across various taxa. This involved categorizing the genes according to their taxonomic affiliations using the KEGG-ORGANISMS database. The validated homologs for the identified genes were obtained from the GenDiS 3 database. This step involved searching for homologous sequences using NCBI Accession IDs. The glycolysis enzymes were identified within the GenDiS database. The DA of the identified glycolysis enzymes was fetched from the GenDiS 3 database using HMMSCAN results provided in the GenDiS database. The taxonomic distribution of the DAs was analysed to understand the evolutionary relationships and distribution patterns of the DAs across different taxa.

Results

Using DELTA-BLAST to search for homologs, 15 128 members from 2060 superfamilies were queried against the NCBI NR database (September 2022), which comprises 504 094 943 unique protein sequences. This search yielded 151 622 506 unique hits as potential homologs across all superfamilies (∼30% of the NR) (Table 1). Following validation with HMMSCAN, 116 393 303 of these sequences were confirmed as true positives (∼23% of the NR). The true positive rates across different superfamilies are distributed widely (see Supplementary Tables ST2 and ST3 for a list of superfamilies with the highest and lowest true positive rates, respectively), with 50% of the superfamilies showing a true positive rate of >0.77 and 75% having rates of >0.44 for DELTA-BLAST hits (Fig. 2).

Analysis of DAs

The delineation of DAs, in homologs where there is at least single presence of the central domain of interests, offers a birds-eye view of the variety in biological processes where the central domain can get recruited. This also provides a comprehensive picture of the 'social' or isolated nature of the central domain (30).

The superfamily c.37.1, identified as P-loop containing nucleoside triphosphate hydrolases (SF: 52540), is the most prevalent in the GenDiS database. This also corresponds to one of the superfolds (8). It is associated with the highest number of sequences, with the single DA (52540) linked to ∼2.5 million sequences. Additionally, the double DA (52540 ∼ 52540) is found in 736 000 sequences. The majority of homologs associated with this superfamily are from eukaryotes, with a smaller proportion belonging to viruses.

The second most populated superfamily is c.2.1, known as NAD(P)-binding Rossmann-fold domains (SF: 51735). There are 1.69 million homologs associated with single DA from this superfamily. Interestingly, ∼500,000 sequences from the 51735 superfamily also include the 6-phosphogluconate dehydrogenase (6PGD) C-terminal domain-like (SF: 48179), as a coexisting domain, with the combined architecture as 51735∼48179. This cooccurrence likely reflects a functional relationship, as both domains are involved in dehydrogenase activity where the Rossmann-fold domain binds cofactors like NAD(P)H, and the 6PGD domain catalyses specific metabolic reactions. All the other superfamilies are similarly associated predominantly with eukaryotic sequences.

The third highest populated superfamily is c.67.1, identified as Pyridoxal phosphate-dependent transferases

Table 1. Total number of superfamilies (SFs), number of DELTABLAST hits, true positives after validation, and mean true positive rate for each structural class of SCOPe

Class	Total number of SFs	Total DELTABLAST hits	Total true positives	Mean true positive rate
Mostly alpha (a)	519	33 745 701	20 690 279	0.63694
Mostly beta (b)	374	36 009 271	21 973 707	0.60450
Alpha and beta [a + b] (d)	578	48 232 485	34 602 755	0.68532
Alpha or beta [a/b] (c)	247	67 224 602	53 394 941	0.77172
Small proteins (g)	73	5 315 902	3 726 663	0.66218
Multi-domain proteins (e)	130	5 867 215	3 991 323	0.67400
Membrane proteins (f)	139	6 003 912	4 053 609	0.67860

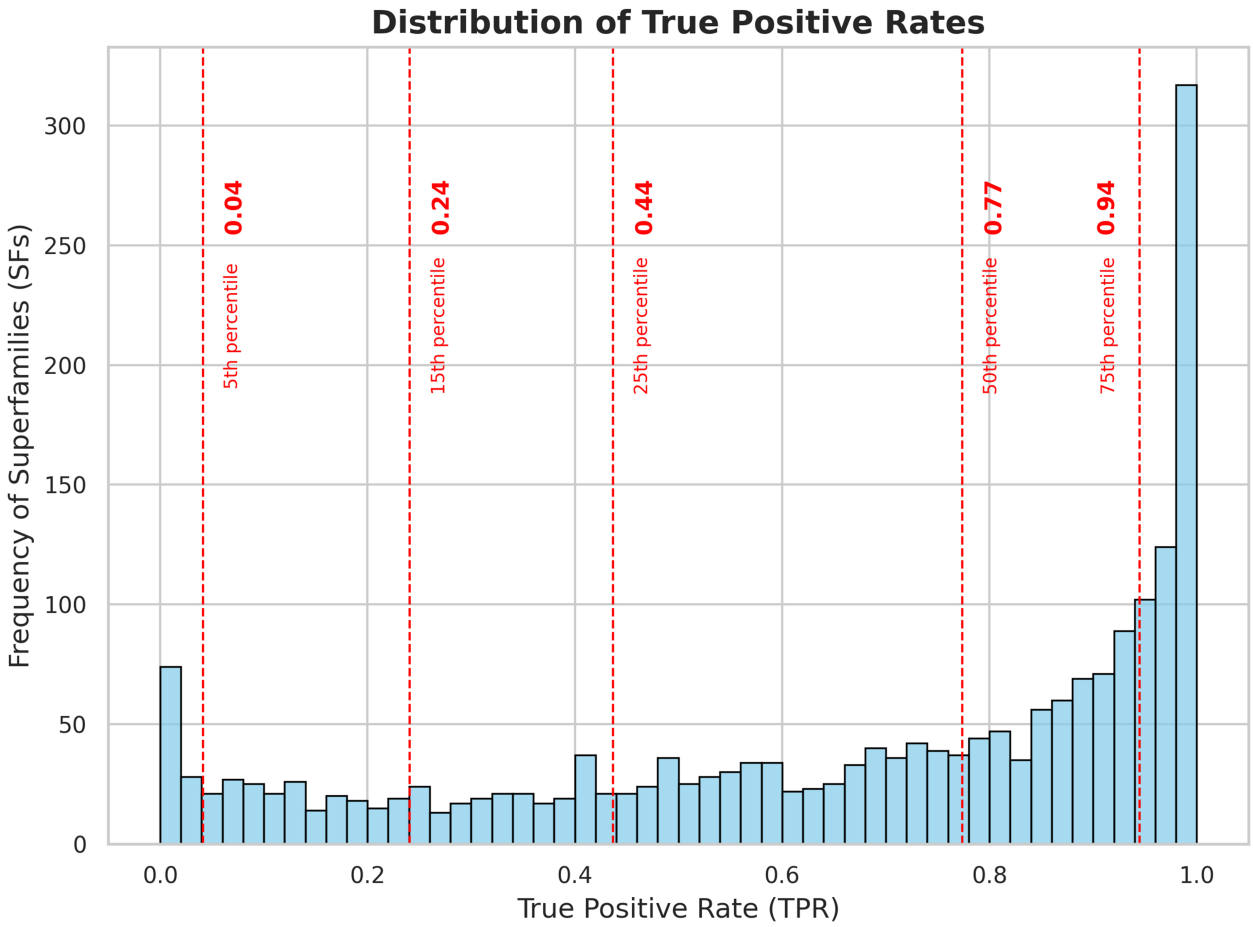

Figure 2. Frequency distribution of superfamilies with respect to true positive rates of the search strategy. The dashed lines indicate the 5th, 15th, 25th, 50th, and 75th percentiles of the data. From the plot, we can observe that 50% of the superfamilies have a true positive rate of >0.77 for DELTA-BLAST and 75% of the superfamilies have a true positive rate of >0.44 for DELTA-BLAST.

(SF: 53383) (also known as PLP-dependent transferases). The single DA of this superfamily is associated with 1.6 million sequences.

Among the 10 most prevalent domain architectures of homologs, 9 consist of single domains, suggesting that single DAs are more common overall.

Indeed, the participation in multi-DAs is not uniform for all domains. Only a few domains have a higher tendency to form multi-DAs. A graph theoretical approach to the study of DAs in GenDis3.0, where the nodes were individual domains and the edges represent the coexistence of domains in a gene, revealed that the average degree of this graph was about 69. But some domains have a higher degree, implying that they are more likely to form multi-DAs. Domains from superfamily c.37.1, P-loop containing nucleoside triphosphate hydrolases, have the highest degree of 1244 along with being the most populated superfamily.

Table 2 lists the SCOPe codes and corresponding superfamily names of protein domains that show the highest degree of coexistence with other domains in multi-DAs. The degree of coexistence is indicated by the number of unique domain combinations in which each domain is found. The data highlight domains such as P-loop containing nucleoside triphosphate hydrolases (c.37.1), NAD(P)-binding Rossmann-fold

Table 2. Domains with the highest degree of coexistence in multi-DAs

SCOPe code	Superfamily name	Degree
c.37.1	P-loop containing nucleoside triphosphate hydrolases	1244
c.2.1	NAD(P)-binding Rossmann-fold domains	919
d.144.1	PK-like	911
c.69.1	alpha/beta-Hydrolases	735
a.4.5	Winged helix' DNA-binding domain	728
c.47.1	Thioredoxin-like	714
b.40.4	Nucleic acid-binding proteins	670
d.58.7	RNA-binding domain, RBD, aka RNA recognition motif	652
b.69.4	WD40 repeat-like	641
d.211.1	Ankyrin repeat	629
c.66.1	S-adenosyl-L-methionine-dependent methyltransferases	620
c.108.1	HAD-like	598
a.118.8	TPR-like	596
g.44.1	RING/U-box	592
a.4.1	Homeodomain-like	581
c.1.8	(Trans)glycosidases	574
c.67.1	PLP-dependent transferases	559
g.37.1	beta-beta-alpha zinc fingers	549
a.39.1	EF-hand	546
c.55.1	Actin-like ATPase domain	529
c.55.3	Ribonuclease H-like	510
c.3.1	FAD/NAD(P)-binding domain	497
d.3.1	Cysteine proteinases	493
b.1.1	Immunoglobulin	490
c.10.2	L domain-like	489

domains (c.2.1), and protein kinase-like (PK-like) (d.144.1) as the most frequently cooccurring, suggesting their widespread involvement in diverse biological functions.

Database browsing and visualization

The updated GenDiS3 database offers hierarchical browsing of superfamily hits through the SCOPe classification, including folds and classes. This feature facilitates user navigation and discovery. For each superfamily, users can view SCOPe and Pfam DAs of the homologs, along with the frequency of each DA within the superfamily. This functionality provides valuable insights into the structural diversity within superfamilies.

Taxonomic distribution and visualization

An interactive sunburst chart visualizes the taxonomic distribution of superfamily hits, offering an intuitive way to explore the evolutionary spread of the identified homologs. This visualization aids in understanding the taxonomic breadth and evolutionary context of the superfamilies (Fig. 3a).

Tools and capabilities

DA prediction tools

This tool allows users to predict the DA of their provided sequences. Users can choose to obtain either SCOPe DA or Pfam DA. This capability is crucial for researchers looking to annotate new sequences or verify the domain structures of known proteins, facilitating functional annotation and evolutionary studies.

Align two genomes of a given superfamily

This tool aligns superfamily hits of a specific superfamily from two different genomes. It helps in comparative genomics studies by allowing researchers to examine the evolutionary conservation and divergence of superfamilies between two species. This tool is particularly useful for identifying conserved domains and understanding the functional implications of superfamily distribution across different organisms.

Align input sequence with genome of a given superfamily

Similar to the previous tool, this one aligns a user-provided sequence with the genome of a given superfamily. It enables researchers to investigate how a specific sequence aligns with known superfamily members within a genome, aiding in the identification of homologous regions and potential functional elements within the genome. This tool is beneficial for pinpointing evolutionary relationships and functional annotations of user-specific sequences.

Blast search superfamily homologs

This tool allows users to search for similar sequences within all the homologs of a superfamily using a query sequence. The tool uses blastp executable from NCBI. The tool will help researchers to look for genes within superfamilies using reference sequences.

Case study 1: the glycolysis pathway in GenDiS

Glycolysis is a fundamental metabolic pathway found in nearly all living organisms, from bacteria to humans. It involves a diverse set of enzymes with varying structures and functions. In this study, we explored the taxonomic distribution of various SCOPe domains within glycolysis pathway enzymes using the GenDiS database. To identify these enzymes, we mapped known glycolysis enzymes to their respective KEGG orthology (KO) identifiers via KEGG. These KO identifiers were then used to query GenDiS for homologs containing the relevant SCOPe domains (Supplementary Fig. S1). Some notable observations are discussed in the net section.

Seeing the knowns: phosphofructokinase

The KEGG pathway for glycolysis includes two orthologies for phosphofructokinase: 6-phosphofructokinase (6-PFK; K24182) and ADP-dependent phosphofructokinase (ADP-PFK; K00918). The 6-PFK is primarily observed in bacteria, while ADP-PFK is predominantly present in archaea (Supplementary Fig. S2). This observation aligns with expected patterns, as bacteria typically utilize ATP-dependent 6-PFK in classical glycolysis pathways (31). In contrast, archaea, often adapted to extreme environments, employ ADP-PFK to maintain metabolic flexibility (32).

Hexokinase

Hexokinase is observed to be present in all kingdoms, emphasizing its essential role in glycolysis. The most frequently found SCOPe DAs in these proteins are two- and four-domain repeats of 53067 (actin-like ATPase domain superfamily containing the hexokinase family). All kingdoms, except animals, typically exhibit a two tandem repeat DA, whereas animals display a four tandem repeat DA (Supplementary Fig. S3).

Enolase

Enolase is present across all kingdoms, demonstrating its fundamental role in glycolysis. The most common DA identified was 54826∼51604 (Enolase NTD∼CTD) (Supplementary Fig. S4). Additionally, it was notable that some enolase proteins also coexisted with the Helix Loop Helix (HLH) DNA-binding domain (47459), pointing to alpha-enolases that are nuclear and found to regulate gene expression (33). This also exemplifies the analysis of DAs could reveal the potential multi-functional role of central domains.

The analysis of glycolysis enzymes across different kingdoms revealed distinct patterns in their taxonomic distribution and DAs (Fig. 4). Phosphofructokinase showed a clear divergence, with 6-PFK predominantly found in bacteria and ADP-PFK mainly in archaea, reflecting their metabolic adaptations. Enolase was universally present across all kingdoms, with the most common DA being Enolase NTD∼CTD, and some proteins also exhibited the HLH DNA-binding domain, suggesting additional functional roles. Hexokinase, also present in all kingdoms, displayed varying DAs: two tandem repeats in most kingdoms and four tandem repeats in animals, highlighting a possible evolutionary expansion in animals. These findings show the evolutionary diversity and specialization of glycolytic enzymes, driven by the metabolic needs and environmental adaptations of different organisms.

Case study 2: LOG gene implicated in the formation of prickles

Prickles, small, sharp outgrowths on the stems and leaves of many plants, serve as a natural defence mechanism against herbivores. While various morphological features and underlying genetic factors contributing to prickle formation have been identified, recent studies have highlighted the role of the LOG (Lonely Guy) gene as a key player in this process (34). The LOG gene codes for cytokinin riboside 5′-monophosphate phosphoribohydrolase (CRMP) enzyme. It was initially known for its role in cytokinin biosynthesis and has been implicated in the formation of prickles across a wide range of plant species, including economically important crops such as eggplant and rose. Satterlee *et al.* showed that mutations in the LOG gene can result in a prickle-less phenotype in *Solanum* sp. Some of these mutations are mapped onto the structure of CRMP from *Arabidopsis thaliana* (Fig. 5). Interestingly, even the intronic mutations are mapped on the peripheral helices, away from the central beta-sheet and the active site region.

We know that CRMP (encoded by the LOG gene) belongs to the MCP/YpsA-like superfamily as classified by SCOPe. We searched for all the homologs of the LOG gene in the MCP/YpsA-like superfamily using the built-in blastp search tool of the GenDis3.0 database, with LOG genes from *A. thaliana* and *O. sativa* (from UniProt) as queries. From this search, we found 5812 homologs for the LOG gene protein belonging to the Viridiplantae kingdom. Over a thousand of these are unannotated hypothetical proteins, and this pipeline can be used to annotate these as LOG genes.

Enhancing homolog detection accuracy

The use of DELTA-BLAST and HMMSCAN in identifying homologs has significantly improved the accuracy and reliability of the GenDiS database. DELTA-BLAST's integration of domain-specific information allows for more sensitive searches, while HMMSCAN provides robust validation by comparing identified sequences against well-curated profile HMM databases. The results show that this approach yields a higher true positive rate across various structural classes, particularly for superfamilies with conserved DAs. These findings suggest that the updated GenDiS database is a powerful resource for the classification and annotation of protein domains, which is crucial for both functional and evolutionary studies. One of the primary challenges addressed in this study is the exponential growth of protein sequence data, which far outpaces the determination of their corresponding structures. The updated GenDiS database helps mitigate this issue by providing a more efficient framework for linking sequences to known structures. By enhancing the accuracy of DA predictions and facilitating the identification of distant homologs, the GenDiS database plays a critical role in expanding our understanding of protein structures and their functional implications, even in the absence of experimental structure determination.

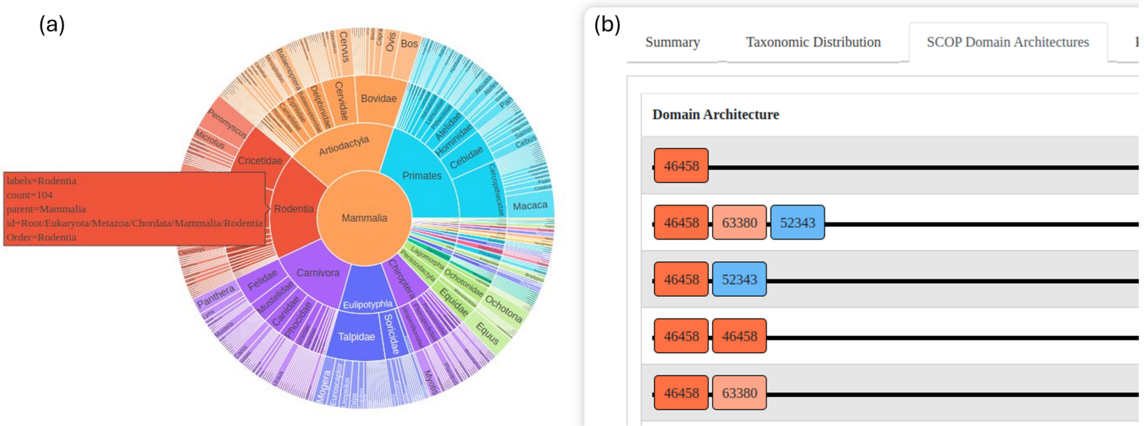

Figure 3. The GenDiS3 web interface for visualizing taxonomic distribution and DAs. (a) Sunburst chart (interactive) showing the taxonomic distribution of a superfamily. (b) SCOP DAs of a superfamily.

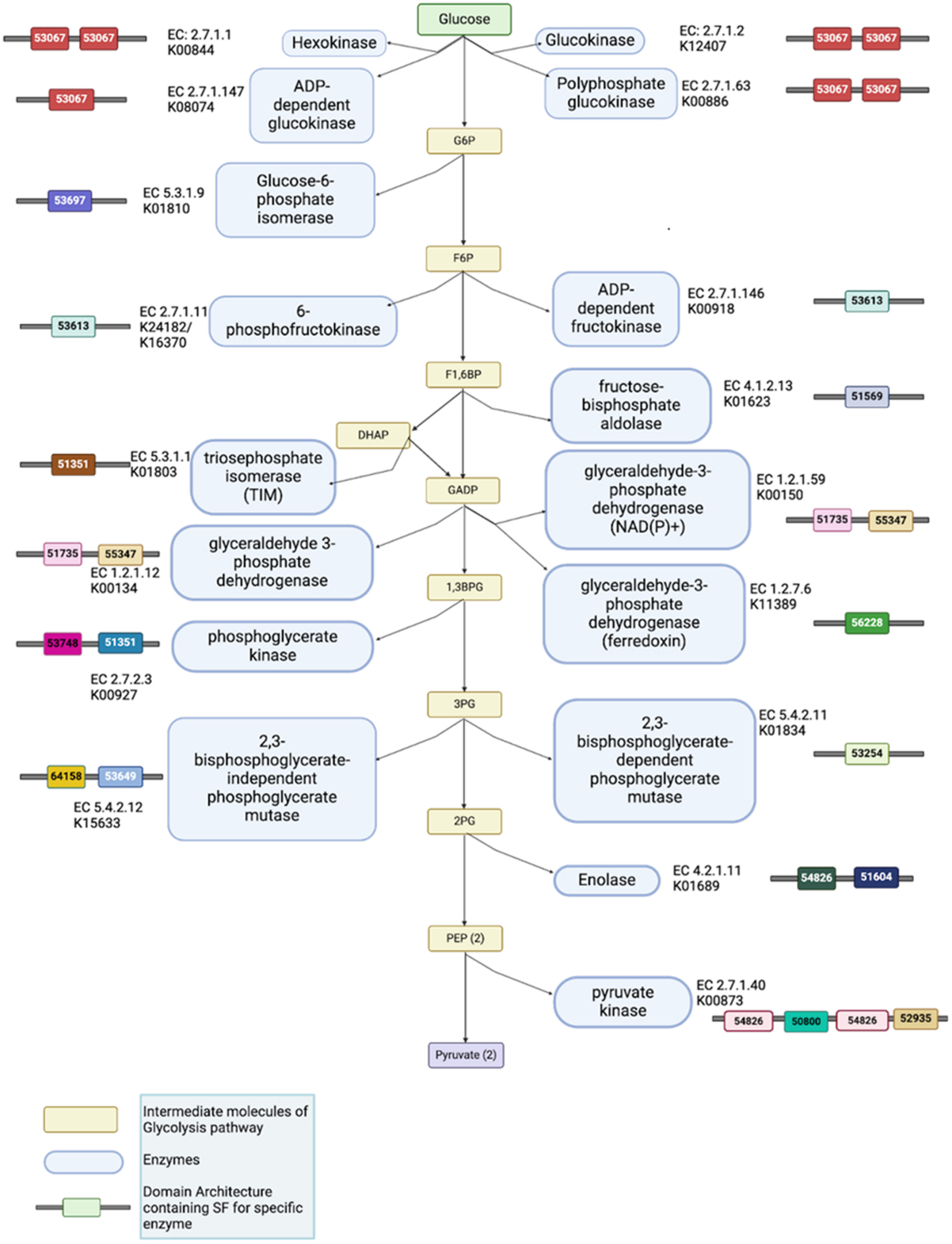

Figure 4. The glycolysis pathway. Enzymes are represented by light blue boxes; yellow boxes correspond to intermediate molecules within the pathway. Colourful boxes flanking the pathway represent the DA of each enzyme where each SCOP domain is represented by the superfamily code. SF 53067: c.55.1 (actin-like ATPase domain), SF 53697: c.80.1 (SIS domain), SF 53613: c.72.1 (ribokinase-like), SF 51569: c.1.10 (aldolase), SF 51351: c.1.1 [triosephosphate isomerase (TIM)], SF 51735: c.2.1 (NAD(P)-binding Rossmann-fold domains), SF 55347: d.81.1 (glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain), SF 53748: c.86.1 (phosphoglycerate kinase), SF 56228: d.152.1 (aldehyde ferredoxin oxidoreductase, N-terminal domain), SF 64158: c.105.1 (2,3-bisphosphoglycerate-independent phosphoglycerate mutase, substrate-binding domain), SF 53649: c.76.1 (alkaline phosphatase-like), SF 53254: c.60.1 (phosphoglycerate mutase-like), SF 54826: d.54.1 (enolase N-terminal domain-like), SF 51604: c.1.11 (enolase C-terminal domain-like), SF 50800: b.58.1 (PK beta-barrel domain-like), SF 52935: c.49.1 (PK C-terminal domain-like).

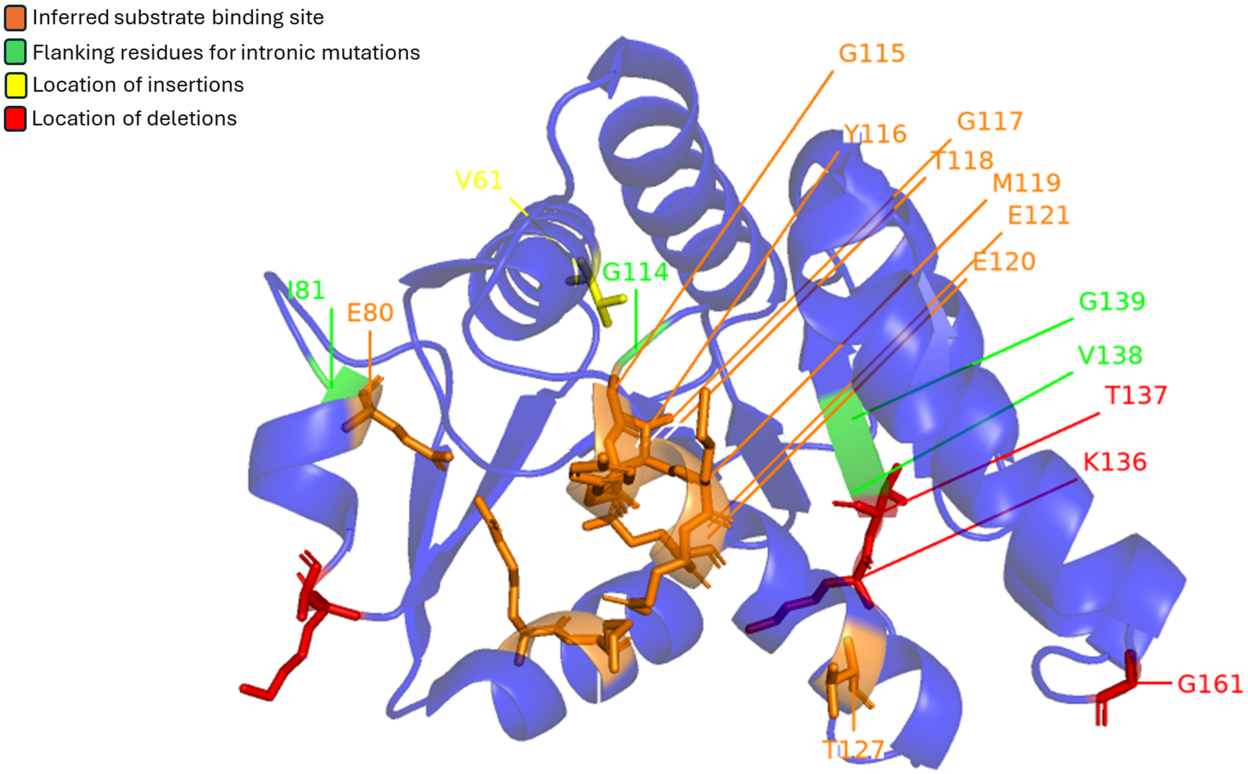

Figure 5. Mutations found in LOG genes from the *Solanum* genus mapped on CRMP (LOG8) from *A. thaliana* (PDB: 1YDH). Red represents location of deletions, yellow represents location of insertions, and green marks the flanking residues for intronic mutations. The orange area marks the inferred substrate binding site.

Understanding evolutionary relationships and structural diversity

The taxonomic analysis of glycolysis-related enzymes, including phosphofructokinase, revealed patterns consistent with established knowledge. For example, ATP-dependent phosphofructokinase is well documented in bacteria, while ADP-PFK is more commonly found in archaea. These findings align with previous research and reflect the metabolic adaptations of these organisms to their environments. Similarly, the variation in hexokinase DAs across kingdoms shows the evolutionary pressures that have shaped these enzymes to fulfil essential metabolic roles. These insights provide a deeper understanding of how protein DAs evolve in response to functional demands and environmental conditions.

Complexity of multi-DAs

The study also delved into the complexity of multi-DAs, which are prevalent in many proteins and critical for their multifunctionality. The analysis revealed that whereas single DAs are common, certain domains, particularly those from the c.37.1 superfamily, have a high propensity to form multi-DAs. This finding highlights the importance of such 'social' domains (30) in facilitating diverse biological functions. Understanding the tendencies of specific domains to participate in multi-DAs can provide valuable insights into their evolutionary history and functional roles.

Conclusion

The study demonstrates the importance of advanced computational methodologies in enhancing protein domain classification and mutations, which provide valuable insights into the evolutionary trends of important proteins. While the updated GenDiS database represents a significant advancement in protein domain classification and evolutionary studies, several challenges remain. Despite the rapid and robust nature of DELTA-BLAST, the mere runs on such sequence searches require months (see GenDis3.0 about-timeframe webpage) (Supplementary Table ST1). The scalability of the database will acquire prime importance since protein sequence data continue to grow and how the computational resources respond to this ever-growing primary data will be crucial for maintaining its utility. Future updates will need to focus on optimizing data management, avoidance of validation steps, and processing capabilities to handle this influx efficiently.

This study has demonstrated the effectiveness of advanced bioinformatics tools in enhancing the classification and annotation of protein domains within the GenDiS database. By addressing key challenges in homolog detection, sequence data growth, and DA complexity, the updated database offers a valuable resource for understanding the structural and functional diversity of proteins. Continued development and expansion of the GenDiS database will be essential for keeping pace with the rapid advancements in genomics and structural biology, ultimately contributing to a deeper

understanding of protein function and evolution. It also offers a powerful resource for researchers in functional annotation and evolution of protein domain superfamilies.

Supplementary data

Supplementary data is available at *Database* online.

Conflict of interest: None declared.

Funding

R.S. acknowledges funding and support provided by the JC Bose Fellowship (JBR/2021/000006) from the Science and Engineering Research Board, India, and the Bioinformatics Centre Grant funded by the Department of Biotechnology, India (BT/PR40187/BTIS/137/9/2021). R.S. would also like to thank the Institute of Bioinformatics and Applied Biotechnology for the funding through her Mazumdar-Shaw Chair in Computational Biology (IBAB/MSCB/182/2022).

Data availability

All domain-superfamily homologues are accessible via the GenDiS-3 web portal at https://caps.ncbs.res.in/gendis3/. Homology searches were performed against the NCBI NR protein database (release September 2022; https://ftp.ncbi.nlm.nih.gov/blast/db/). HMM profiles from PASS2 can be downloaded from https://caps.ncbs.res.in/pass2.8/. SCOPe v2.08 member sequences are available for download from the SCOPe website (https://scop.berkeley.edu/downloads/). Raw DELTA-BLAST and HMMSCAN output files and any intermediate datasets can be obtained from the corresponding author upon reasonable request.

References

1. Sayers EW, Bolton EE, Brister JR. *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2022;**50**:D20–6. https://doi.org/10.1093/nar/gkab1112
2. Yoon B-J. Hidden Markov models and their applications in biological sequence analysis. *Curr Genomics* 2009;**10**:402–15. https://doi.org/10.2174/138920209789177575
3. Biegert A, Söding J. Sequence context-specific profiles for homology searching. *Proc Natl Acad Sci* 2009;**106**:3770–75. https://doi.org/10.1073/pnas.0810767106
4. Mohapatra SS, Sowdhamini R. *The Development of GenDiS V3.0: A Comprehensive Web Application for Organizing and Analyzing Protein Domain Superfamilies across Genomes*. Bengaluru: Bangalore University, 2023
5. Iyer MS, Bhargava K, Pavalam M. *et al*. GenDiS database update with improved approach and features to recognize homologous sequences of protein domain superfamilies. *Database* 2019;**2019**:baz042. https://doi.org/10.1093/database/baz042
6. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;**7**:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
7. Siebers B, Schönheit P. Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. *Curr Opin Microbiol* 2005;**8**:695–705. https://doi.org/10.1016/j.mib.2005.10.014
8. Camacho C, Coulouris G, Avagyan V. *et al*. BLAST+: architecture and applications. *BMC Bioinf* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421
9. Berman HM. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42. https://doi.org/10.1093/nar/28.1.235
10. Chothia C. One thousand families for the molecular biologist. *Nature* 1992;**357**:543–44. https://doi.org/10.1038/357543a0
11. Murzin AG, Brenner SE, Hubbard T. *et al*. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;**247**:536–40. https://doi.org/10.1016/S0022-2836(05)80134-2
12. 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
13. Wang W, Wang L, Endoh A. *et al*. Identification of α-enolase as a nuclear DNA-binding protein in the zona fasciculata but not the zona reticularis of the human adrenal cortex. *J Endocrinol* 2005;**184**:85–94. https://doi.org/10.1677/joe.1.05909
14. Pugalenthi G, Bhaduri A, Sowdhamini R. GenDiS: genomic distribution of protein structural domain superfamilies. *Nucleic Acids Res* 2005;**33**:D252–D255. https://doi.org/10.1093/nar/gki087
15. Chandonia J-M, Guan L, Lin S. *et al*. SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Res* 2022;**50**:D553–9. https://doi.org/10.1093/nar/gkab1054
16. Orengo CA, Jones DT, Thornton JM. Protein superfamilles and domain superfolds. *Nature* 1994;**372**:631–34. https://doi.org/10.1038/372631a0
17. Sievers F, Wilm A, Dineen D. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011;**7**:539. https://doi.org/10.1038/msb.2011.75
18. Chandonia J-M. The ASTRAL compendium in 2004. *Nucleic Acids Res* 2004;**32**:189D–192. https://doi.org/10.1093/nar/gkh034
19. Gandhimathi A, Nair AG, Sowdhamini R. PASS2 version 4: an update to the database of structure-based sequence alignments of structural domain superfamilies. *Nucleic Acids Res* 2012;**40**:D531–4. https://doi.org/10.1093/nar/gkr1096
20. Paysan-Lafosse T, Blum M, Chuguransky S. *et al*. InterPro in 2022. *Nucleic Acids Res* 2023;**51**:D418–27. https://doi.org/10.1093/nar/gkac993
21. Satterlee JW, Alonso D, Gramazio P. *et al*. Convergent evolution of plant prickles by repeated gene co-option over deep time. *Science* 2024;**385**:eado1663. https://doi.org/10.1126/science.ado1663
22. Sillitoe I, Bordin N, Dawson N. *et al*. CATH: increased structural coverage of functional space. *Nucleic Acids Res* 2021;**49**:D266–73. https://doi.org/10.1093/nar/gkaa1079
23. Kanehisa M, Furumichi M, Sato Y. *et al*. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* 2023;**51**:D587–92. https://doi.org/10.1093/nar/gkac963
24. 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
25. Joshi AG, Raghavender US, Sowdhamini R. Improved performance of sequence search approaches in remote homology detection. *F1000Research* 2014;**2**:93. https://doi.org/10.12688/f1000research.2-93.v2
26. Mistry J, Chuguransky S, Williams L. *et al*. Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9. https://doi.org/10.1093/nar/gkaa913
27. Bhattacharyya T, Nayak S, Goswami S. *et al*. PASS2.7: a database containing structure-based sequence alignments and associated features of protein domain superfamilies from SCOPe. *Database* 2022;**2022**:baac025. https://doi.org/10.1093/database/baac025
28. Boratyn GM, Schäffer AA, Agarwala R. *et al*. Domain enhanced lookup time accelerated BLAST. *Biol Direct* 2012;**7**:12. https://doi.org/10.1186/1745-6150-7-12
29. Altschul S. 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
30. Bhattacharyya T, Nayak S, Goswami S. *et al*. PASS2.7: a database containing structure-based sequence alignments and associated features of protein domain superfamilies from SCOPe. *Database* 2022;**2022**:baac025.

31. Orengo C, Michie A, Jones S. *et al*. CATH – a hierarchic classification of protein domain structures. *Structure* 1997;**5**:1093–109. https://doi.org/10.1016/S0969-2126(97)00260-8
32. Kim BH, Gadd GM. Prokaryotic metabolism and physiology. In: *Prokaryotic Metabolism and Physiology*. Cambridge: Cambridge University Press, 2019, 1–4.
33. Iyer MS, Joshi AG, Sowdhamini R. Genome-wide survey of remote homologues for protein domain superfamilies of known structure reveals unequal distribution across structural classes. *Mol Omics* 2018;**14**:266–80. https://doi.org/10.1039/C8MO00008E
34. Mohanty S, Purwar M, Srinivasan N. *et al*. Tethering preferences of domain families co-occurring in multi-domain proteins. *Mol Biosyst* 2013;**9**:1708. https://doi.org/10.1039/c3mb25481j

Database, 2025, 00, baaf035, DOI: https://doi.org/10.1093/database/baaf035, Database update