Journal Pre-proofs

Research Article

Bridging the Gap between Sequence and Structure Classifications of Proteins with AlphaFold Models

Jimin Pei, Antonina Andreeva, Sara Chuguransky, Beatriz Lázaro Pinto, Typhaine Paysan-Lafosse, R. Dustin Schaeffer, Alex Bateman, Qian Cong, Nick V. Grishin

PII: S0022-2836(24)00384-X
DOI: https://doi.org/10.1016/j.jmb.2024.168764
Reference: YJMBI 168764

To appear in: *Journal of Molecular Biology*

Received Date: 5 June 2024
Revised Date: 13 August 2024
Accepted Date: 20 August 2024

Please cite this article as: J. Pei, A. Andreeva, S. Chuguransky, B. Lázaro Pinto, T. Paysan-Lafosse, R. Dustin Schaeffer, A. Bateman, Q. Cong, N.V. Grishin, Bridging the Gap between Sequence and Structure Classifications of Proteins with AlphaFold Models, *Journal of Molecular Biology* (2024), doi: https://doi.org/10.1016/j.jmb.2024.168764

Bridging the Gap between Sequence and Structure Classifications of Proteins with AlphaFold Models

Jimin Pei[1,3,4], Antonina Andreeva[2], Sara Chuguransky[2], Beatriz Lázaro Pinto[2], Typhaine Paysan-Lafosse[2], R. Dustin Schaeffer[3], Alex Bateman[2,*], Qian Cong[1,3,4,*] Nick V. Grishin[3,5,*]

[1]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA.

[2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

[3]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA.

[4]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA.

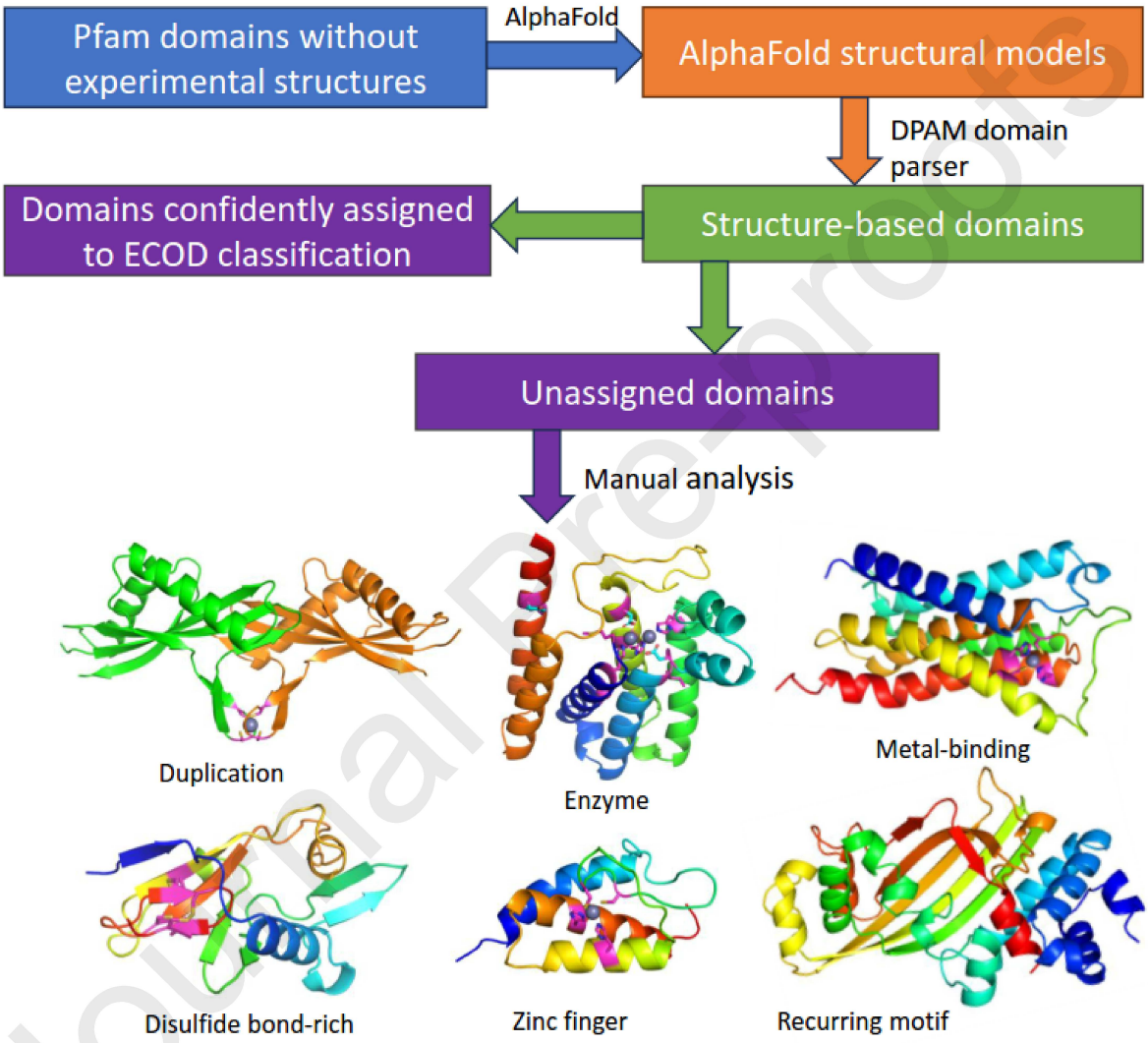
[5]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX, USA.

[*]Corresponding authors:

Alex Bateman: agb@ebi.ac.uk Address: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, United Kingdom.

Qian Cong: qian.cong@utsouthwestern.edu Address: 6001 Forest Park Rd, Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA. Phone: 001-214-645-7401

Nick V. Grishin: grishin@chop.swmed.edu Address: 6001 Forest Park Rd, Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA Phone: 001-214-645-5950



Highlights

- Protein domain classification is important for studying protein evolution and function.
- Highly accurate AlphaFold models offer structural insights into Pfam domains.
- DPAM was able to parse and assign many Pfam domains into ECOD classification.
- Manual inspection of unassigned domains uncovers new folds and remote evolutionary relationships.
- A combined approach to domain classification leads to a better understanding of the protein universe.

Abstract

Classification of protein domains based on homology and structural similarity serves as a fundamental tool to gain biological insights into protein function. Recent advancements in protein structure prediction, exemplified by AlphaFold, have revolutionized the availability of protein structural data. We focus on classifying about 9000 Pfam families into ECOD (Evolutionary Classification of Domains) by using predicted AlphaFold models and the DPAM (Domain Parser for AlphaFold Models) tool. Our results offer insights into their homologous relationships and domain boundaries. More than half of these Pfam families contain DPAM domains that can be confidently assigned to the ECOD hierarchy. Most assigned domains belong to highly populated folds such as Immunoglobulin-like (IgL), Armadillo (ARM), helix-turn-helix (HTH), and Src homology 3 (SH3). A large fraction of DPAM domains, however, cannot be confidently assigned to ECOD homologous groups. These unassigned domains exhibit statistically different characteristics, including shorter average length, fewer secondary structure elements, and more abundant transmembrane segments. They could potentially define novel families remotely related to domains with known structures or novel superfamilies and folds. Manual scrutiny of a subset of these domains revealed an abundance of internal duplications and recurring structural motifs. Exploring sequence and structural features such as disulfide bond patterns, metal-binding sites, and enzyme active sites helped uncover novel structural folds as well as remote evolutionary relationships. By bridging the gap between sequence-based Pfam and structure-based ECOD domain classifications, our study contributes to a more comprehensive understanding of the protein universe by providing structural and functional insights into previously uncharacterized proteins.

Keywords: Protein classification; Pfam; AlphaFold structural modeling; ECOD; DPAM

Introduction

Domain classification of proteins provides the scientific community with a common set of homologous domains. The inference of function based on domain homology is one method for researchers to develop biological insights [1, 2]. These classifications can differ depending on whether they incorporate structural similarity into their methodology [3]. Classifications that rely principally on sequence have access to a broader view of the protein world than those which rely on structure access to a smaller set of proteins where domain boundaries, function, and protein-protein interactions can be more clearly determined. This fundamental difference led to the development of distinct sequence and structural resources such as Pfam [4], CDD [5], SUPERFAMILY [6], SCOP [7], SCOPe [8], CATH [9], and ECOD [10]. However, the recent development of highly accurate structure prediction software such as AlphaFold2 [11] and RoseTTAFold [12] has led to a wealth of structural data for proteins and protein families that have not yet been structurally characterized by experiment. The subsequent release of predicted structures of over 200 million protein structures representing the known protein universe by the AlphaFold Structure Database (AFDB) was a significant inflection point in the aggregation of structural data [13]. Classification of these predicted structures can illuminate areas of the protein universe previously unpopulated in these structural classifications, refine domain boundaries in sequence classifications, and better understand the performance of prediction algorithms on proteins more distant than those commonly used for model training. A recent study of the UniProt sequence space with high-confidence AlphaFold structural models has expanded our understanding of the dark protein universe of unannotated proteins and uncovered new protein families and unusual folds [14].

Structural classifications such as ECOD use structural similarity to detect distant homology. Where sequence similarity is low, alignment of core topology, identification of shared active sites or cofactor binding can provide additional evidence of homology. Traditionally, structural classifications have been limited by the availability of experimental protein structures. However, we have shown that we can use structural predictions to classify previously uncharacterized human proteins [15], fast-evolving proteins in microbes [2], and across multiple whole proteomes [16]. Using DPAM [17], a domain parser that specifically incorporates consideration of AlphaFold measures of inter-residue errors, we can quickly and precisely classify predicted structures. Domain classification of predicted structures can provide clarification about domain boundaries, especially where domains commonly co-occur; they can identify compact globular regions from predictions that contain disordered regions or that are unstructured in the absence of binding partners, and they can suggest the presence of cofactor and metal-binding sites, even when those compounds are not present.

The ECOD classification scheme classifies domains by the confidence degrees of homology. ECOD differs from similar classifications (e.g., SCOP and CATH) in that it recognizes homology between domains of differing topologies. The X-group (or possible homology level) classifies domains where there is some homologous signal, but it is not definitive. X-groups contain homologous groups (H-groups) that cluster domains that are definitively homologous. Some H-groups are made up of multiple topological groups (T-groups) that contain domains that are homologous but differ in topology (e.g., 4- and 5-bladed beta-propellers). Finally, the F-group represents sequence families containing domains with high sequence similarity. The F-group in ECOD is most analogous to a Pfam family. DPAM classifies domains at the T-group level. In the past, domains classified into ECOD were divided into F-groups using a highly modified version of

Pfam dubbed 'ECODf' [18]. This and subsequent works describe a series of efforts to harmonize Pfam and ECOD domain definitions and allow the use of Pfam to directly classify ECOD domains into F-groups.

The Pfam classification encompasses proteins from reference proteomes from UniProtKB [19]. Proteins in Pfam are classified exclusively, each residue in a protein should belong to a single Pfam family. Previous versions of Pfam have classified >50% of the residues in UniProtKB [4]. The fraction of UniProtKB classified by Pfam has remained remarkably stable over time as more proteins have been deposited. Insofar as Pfam classifies disordered and coiled-coil domains in addition to globular domains, it encompasses a broader protein universe than traditionally targeted by structural domain classifications. The Pfam classification is also active and frequently updated, and available through the InterPro [20] website. We consider Pfam to be a good proxy for the protein universe and a useful target for increasing the coverage of the ECOD classification. Similarly, by classifying Pfam domains and identifying differences, we can target instances where Pfam domains could be improved by structural classification or cases where they should be split into multiple distinct domains.

Here we classify the domains of those proteins in Pfam families with no associated experimental structure into ECOD. Using predicted models from AFDB, we attempt automated classification using our DPAM domain parser. Where those classifications succeed, they are added to ECOD. We report on the distribution of these highly duplicated and divergent domains among ECOD homologous groups. Additionally, we discuss several examples where automated classification failed. Specifically, we discuss cases where identifying domain duplications aids in the classification of unassigned domains, domains where the presence of an enzymatic active site can aid in the detection of homology, and cases where metal-binding sites and disulfide bonds can indicate potential homology. We indicate where these examples result in forthcoming additions and changes to Pfam and ECOD and make available those unassigned DPAM domains identified from Pfam families as a test set for future classification.

Results and Discussions

DPAM identifies corresponding structural domains from AlphaFold models for 87% Pfam families without experimental structures

We focused on 9,284 Pfam families without experimental 3D structures (see Methods). AlphaFold models of UniProt proteins in AFDB have significantly increased the fraction of protein families with 3D structure data. Using Pfam's annotation of UniProt proteins, we selected 25,893 representative UniProt proteins with AlphaFold models from these families (3 models per Pfam family), representing 8,631 (93.0%) Pfam families without experimental structures. The 653 protein families without predicted structures consist principally of viral proteins, which were excluded from AFDB. We applied DPAM [17] to partition and assign domains to ECOD classification for the representative proteins, and identified 28,725 DPAM domains (from 7,595 Pfam families representing 87%) mapping to the same region as the Pfam domains in these proteins (Figure 1).

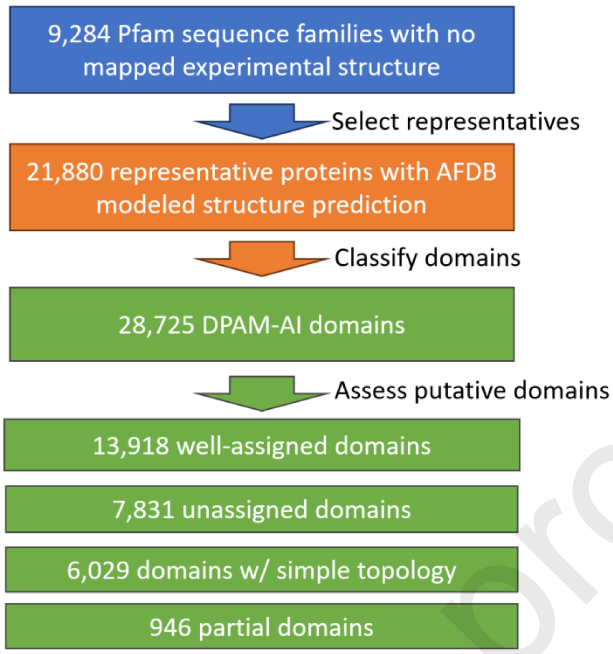


Figure 1. Flowchart of assigning Pfam domains by DPAM-AI

These domains belong to four categories (see Materials and methods): globular domains with high DPAM confidence scores to ECOD entries (well-assigned domains), globular domains with low DPAM confidence scores to ECOD entries (unassigned domains), domains with few secondary structure elements (simple topology domains), and domains aligned to parts of ECOD domains (partial domains). 13,918 (48%) domains were well-assigned domains. 7,831 (27%) DPAM domains were classified as unassigned domains. 6,029 DPAM domains were classified with simple topology. Finally, 946 domains were partial domains. We focus our analysis on the well-assigned domains and unassigned domains.

DPAM confidently classifies numerous well-assigned domains from Pfam families without experimental structures

The 13,918 well-assigned domains belong to 4,940 Pfam families. These domains were predominantly all-α (36%), with the remainder being divided among the all-β (24%), α/β (16%), α+β (19%), and few secondary structure elements (5%) (Fig. 2A). The ECOD reference homologous groups most populated by the high confidence DPAM domains were the immunoglobulin-like (IgL) domains (ECOD H: 11.1), armadillo (ARM) repeats (ECOD H: 109.4), helix-turn-helix (HTH) domains (ECOD H: 101.1), Src homology 3 (SH3) domains (ECOD H: 4.1), and restriction endonuclease-like domains (ECOD H: 2008.1) (Fig 2B). The IgL homologous group belongs to the beta sandwiches architecture and is implicated in protein-protein interactions, specifically in the acquired immune systems of vertebrates [21] and extracellular sensing in bacteria [22, 23]. They are also commonly observed as structural repeats (e.g., titin and fibronectin) [24, 25]. Among the 731 well-assigned IgL domains, 26% came from proteins that have a transmembrane (TM) region (87% of these TM regions were external to the IgL domain) and 44% from a protein that contained a detectable signal peptide (Figure 2C). Of 233 Pfam families that contained at least one IgL domain, 120 (52%) were labeled as “Domains of

Unknown Function“ (DUFs), which signals that their function was not clear at the time the underlying Pfam was initiated. The ARM repeats H-group contains various helical repeats in addition to the canonical ARM repeats, such as the consensus tetratricopeptide repeats (CTPRs) and the HEAT repeats [26, 27]. Among the 707 well-assigned ARM repeats, 19% were from a TM protein and 6% from a potentially secreted or extracellular protein. By contrast, neither the HTH domains [28] nor the SH3 domains [29] contained a significant fraction of proteins that could be transmembrane or extracellular. The SH3 and HTH domains both contained 4% TM proteins, and 6% and 5% proteins with a signal peptide, respectively. This is consistent with the tendency of these homologous groups to be involved in protein or DNA recognition and signaling. In bacteria, SH3 domains are implicated in extracellular peptidoglycan binding [30]. For SH3 and HTH domains, it is more likely that they are encompassed by a larger Pfam domain that has not been structurally characterized and thus not yet further subdivided into globular domains. 29% of DPAM SH3 domains are components of larger Pfam domains (i.e., they are entirely covered by a larger Pfam domain). Of these, 52 (19%) DPAM SH3 domains are from proteins where multiple well-assigned DPAM domains were detected, whereas 26 (9%) are SH3 domains co-occurring with at least one other well-assigned SH3 domain. The remaining DPAM SH3 domains are clearly covered by a single Pfam domain. Similarly, among the DPAM HTH domains, 46% are identified as longer and potentially multidomain Pfam domains. These SH3 and HTH examples illustrate the strength of structural information in identifying potential homologs and clarifying domain boundaries.

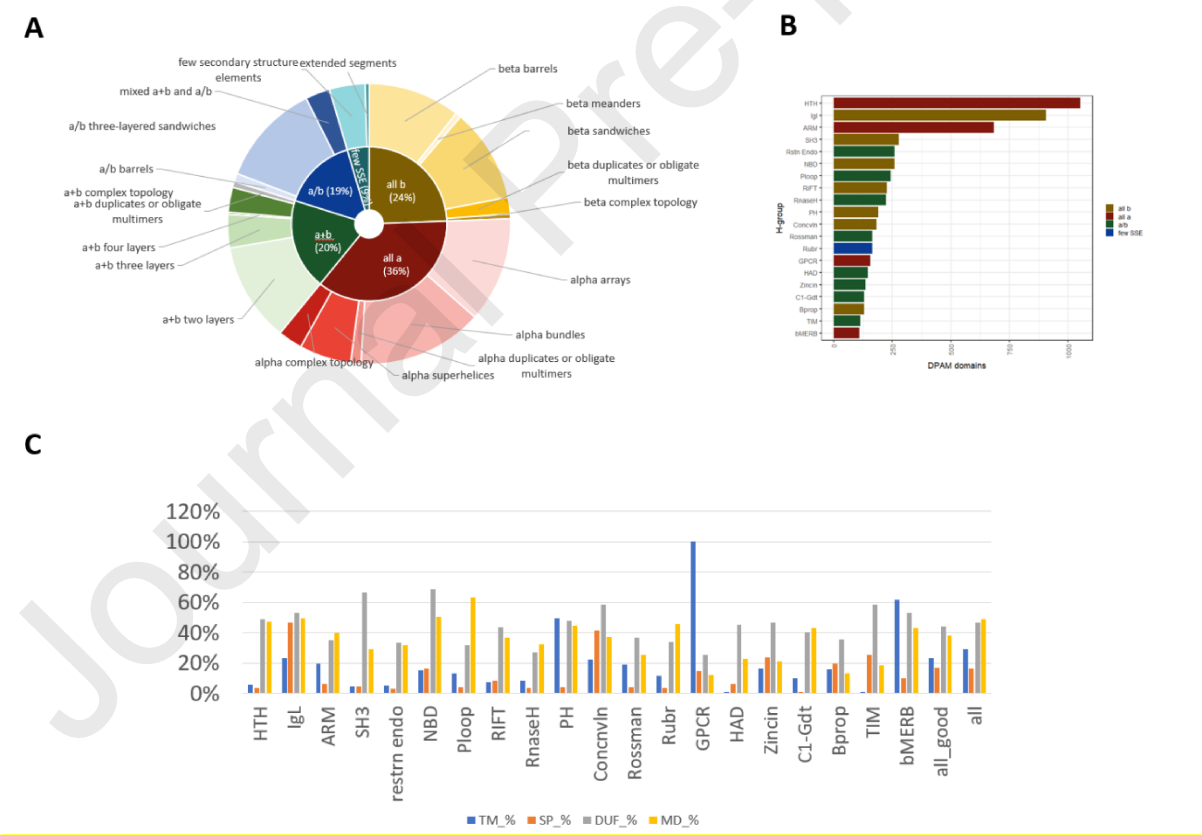


Figure 2. Overall DPAM classification of Pfam families missing from ECOD. A) Well-assigned domains from this set and the classification of their hit domains in an ECOD reference by class (overall secondary structure content) and architecture (specific secondary structure arrangement). B) The top 20 most populated homologous groups of well-assigned domains. C) From the top 20 most populated homologous groups, proportion of domains derived from proteins with transmembrane helices (TM_%) as well as signal

peptides (SP_%). Also, the fraction of proteins belonging to a DUF Pfam family that are mapped to at least one protein within an H-group (DUF_%). Finally, the number of DPAM domains whose mismatch with their mapped Pfam region exceeds 30 residues, indicating the possibility of a multidomain Pfam (MD_%).

Characterization of DPAM domains that cannot be confidently assigned to ECOD classification

Of the 28,562 domains classified by DPAM, 7,799 globular domains (about 27%) could not be confidently assigned to ECOD H-groups and were classified as unassigned domains. These unassigned domains had a breadth of physical properties differing from well-assigned domains. We observed a slight skew towards larger lengths among the well-assigned domains (Figure 3A). On average, the lengths of well-assigned compared to unassigned domains was 172 ± 115 residues and 129 ± 74 residues, respectively. Comparing the number of secondary structure elements (SSEs) between the well-assigned and unassigned domains, we observed an average of 10.2 ± 6.35 SSEs and 7.59 ± 4.94 SSEs respectively, with a slight skew towards smaller numbers of SSEs in unassigned domains (Fig 3B). Between well-assigned and unassigned domains, nearly the same fraction (about 17%) contained signal peptides, regardless of whether that signal peptide was internal to the partitioned domain boundaries (Figure 3C). Generally, signal peptides are not included in ECOD domains and are instead separated into a special architecture in the ECOD domain classification. Unlike signal peptides, there was a noted increase in classified domains with internal transmembrane segments among the unassigned domains: about 14% of well-assigned domains have at least one transmembrane segment compared to about 22% of unassigned domains, suggesting that transmembrane domains are more difficult to assign (Fig 3D). These unassigned domains with transmembrane segments likely require the creation of new transmembrane groups or careful manual analysis to detect distant homology to known transmembrane domains.

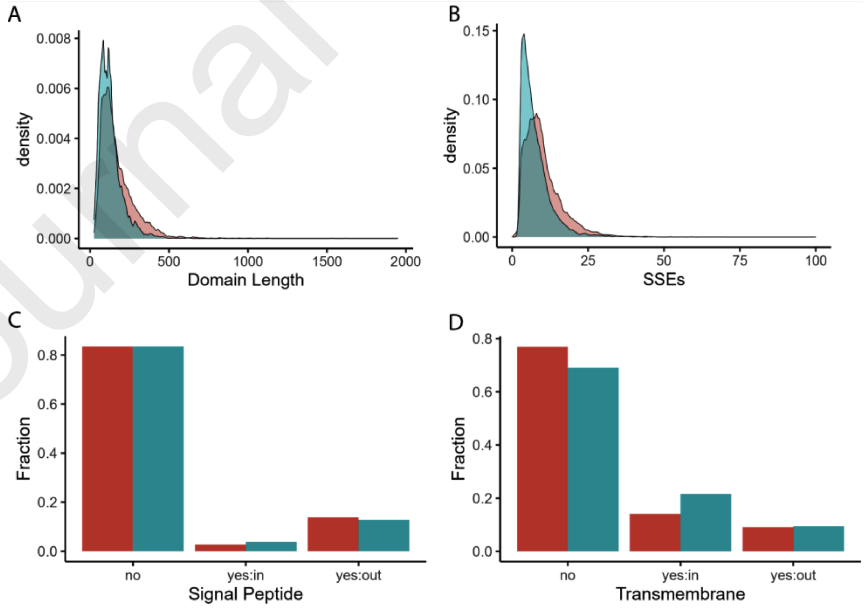


Figure 3. Statistics of well-assigned domains and unassigned domains. A) Comparison of domain length distribution between well-assigned (pink) and unassigned (cyan) domains. B) Comparison of SSE count (helix + strand) between well-assigned (pink) and unassigned (cyan) domains. C) Fraction of well-assigned (pink) and unassigned domains (cyan) with no signal peptide, signal peptide internal to domain boundaries

(yes:in), signal peptide elsewhere in protein (yes:out). D) Fraction of well-assigned (pink) and unassigned (cyan) domains with no transmembrane helix, or with at least one transmembrane helix internal to domain boundaries (yes:in) or external to domain boundaries (yes:out).

Pfam families can be aggregated into larger clans based on a purported single evolutionary origin [31]. Conceptually, clans are similar to ECOD homologous groups (or X-groups). We analyzed the relative fraction of DPAM domains (both well-assigned and unassigned) that arose from Pfam families and whether or not those families already possessed a clan assignment. Pfam v36.0 contains 659 distinct clans, and approximately 43% of Pfam families are assigned to a clan. DPAM domains (well-assigned and unassigned) were classified from proteins belonging to 387 distinct Pfam clans, 358 of which contained at least one well-assigned domain. This result indicates that many proteins lacking structural characterization could nevertheless be classified with other homologous proteins purely based on sequence information. However, 60% of the well-assigned domains (from 3,111 Pfams) lacked a clan classification prior to our structural DPAM classification. Furthermore, a majority (87%) of the unassigned domains in this DPAM domain set also lacked a clan classification. Focusing on these domains lacking a clan classification offers a potential direction both for expansion of Pfam as well as ECOD.

Consistency and discrepancy of Pfam and DPAM assignments

More than half of the Pfam domains contain a single DPAM domain for all three representatives, including 2,335 domains classified as a single well-assigned domain (“W1|W1|W1”), 1,761 domains classified as a single simple topology domain (“S1|S1|S1”), and 908 domains classified as a single “unassigned” domain (“U1|U1|U1”). Some single-domain Pfams are not consistently classified. For example, 343 Pfam domains were classified as “W1|W1|U1” for the three representatives (two classified as well assigned DPAM domains and one as an unassigned domain) and 331 domains were classified as “W1|U1|U1”. Inconsistent assignments among different proteins selected for the same Pfam family could be caused by different levels of sequence divergence among representatives and the use of a hard cutoff value in DPAM assignment that differentiates well-assigned and unassigned domains.

A subset of Pfam domains contain multiple DPAM domains. For example, for 166 Pfam domains, DPAM defined two well-assigned domains for each of the three representatives (“W2|W2|W2”). In addition, for 75 Pfam domains, DPAM defined a well-assigned domain and an unassigned domain (“W1U1|W1U1|W1U1”); and for 25 Pfam domains, DPAM defined two unassigned domains (“U2|U2|U2”). The cases of three or more DPAM domain assignments for a Pfam domain are much fewer. For example, there are 15, 10, 5, and 4 cases of “W3|W3|W3”, “W1U2|W1U2|W1U2”, “W2U1|W2U1|W2U1”, and “U3|U3|U3”, respectively.

Manual inspections of the multi-DPAM assignments for a single Pfam family suggest that most of them correspond to independent folding units. Some examples are shown in Figure 4. The CRF-BP family (PF05428) (Figure 4A) is found in corticotropin-releasing factor binding proteins (CRF-BP) in metazoans. CRF-BP may play inhibitory or activation roles by binding corticotropin-releasing hormone (CRH) and other CRH-like ligands [32]. An AlphaFold model revealed that the CRF-BP Pfam domain region contains two jelly roll-like domains (Figure 4A), both of which are well-assigned domains by DPAM. In Pfam version 38.0, we split the CRF-BP family into the N-terminal CUB-like family (PF05428) and the C-terminal CUB-like family (PF23541). DUF1512 (PF07431) represents a case where one well-assigned domain and one unassigned domain were found in the Pfam domain region (Figure 4B). The C-terminal DPAM domain is assigned to the

ECOD H-group of Peptidyl-tRNA hydrolase-like, which is supported by HHpred hits to known members with structures such as germination protease (pdb:1c8b) [33] and the presence of conserved aspartic acid residues [34]. DUF1512-containing proteins, mainly found in archaea, could possess hydrolase activity. The N-terminal region of the DUF1512 domain is a helical bundle domain with 5 α-helices and is an unassigned domain. In Pfam version 38.0, a new Pfam family (PF23452) is created for the C-terminal domain. Pfam domains from the DUF6688 (PF20394) family contain two DPAM unassigned domains (Figure 4C). The N-terminal domain is a membrane-bound domain with 7 transmembrane segments, while the C-terminal domain is a soluble domain consisting of several α-helices and a β-hairpin. DUF6688-containing proteins, mostly found in bacteria, that do not have known functions. In Pfam version 38.0, a new Pfam family (PF23453) is created for the C-terminal soluble regions of DUF6688 proteins.

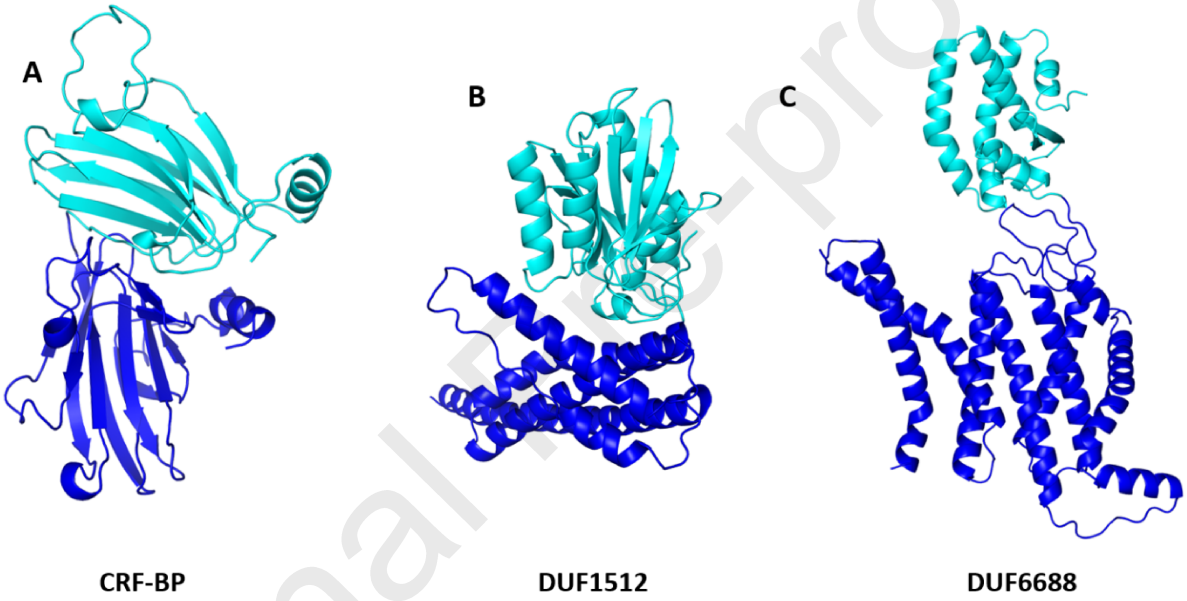


Figure 4. Examples of two DPAM domains parsed from a Pfam domain region. A) Pfam domain CRF-BP contains two well-assigned DPAM domains assigned as jelly roll fold (UniProt accession: I3KE96). B) Pfam domain DUF1512 contains two DPAM domains: an unassigned N-terminal alpha-helical domain and a well-assigned C-terminal domain (UniProt accession: A0A7C4YAY0). C) DUF6688 contains two unassigned DPAM domains: a N-terminal transmembrane domain and a C-terminal soluble domain (UniProt accession: A0A1M6N611). N- and C-terminal domains are colored blue and cyan, respectively.

Manual analysis of unassigned domains

While DPAM-AI assigned a significant portion of domains to ECOD structural categories, many DPAM domains overlapping with Pfam domains remain unassigned. Specifically, 908 Pfam domains contain a single unassigned domain for all three representatives. Some of these unassigned domains may represent new families or groups that do not fit into existing ECOD classifications, while others could still be remotely related to existing ECOD groups. The sensitivity of DPAM assignment is limited in cases lacking strong sequence and structural signals.

We focused on manually analyzing a subset of the 908 Pfam families that contain a single unassigned domain for all three representatives. The goal of our manual analysis is to uncover remote evolutionary relationships through comprehensive consideration of domain structures, functional associations, weak sequence and structural cues, unconventional sequence and structural attributes, and transitive searches for relationships. Approximately 150 unassigned domains underwent manual scrutiny, resulting in around 20% of them being categorized within the ECOD framework. Furthermore, we identified recurring sequence, structural, and evolutionary patterns within these unassigned domains, which are described below.

Internal domain duplications are recurring in unassigned domains

Domain duplication is a key driving force in the evolution of multidomain proteins and contributes to the complexity and dynamics of proteomes. We observed numerous internal domain duplications in unassigned domains. For instance, domain duplications were observed in proteins in the Pfam family DUF898 (PF05987). Proteins in this family possess five repeats of a βααβ (β: β-strand, α: α-helix) unit. The α-helices are mostly hydrophobic, suggesting their localization in the membrane. The beta hairpins formed by the two β-strands in each repeating unit together form a β-barrel covering the top of the α-helices (Figure 5A). Some α-helices (for example, those in the second and fifth units in protein A0A1W9VTP9, Figure 5A) are longer than others and likely form transmembrane helices. In contrast, relatively short α-helices may not fully penetrate the membrane. Duplications of the βααβ units were supported by their structural similarities and the HHpred hits between the units. DUF898 proteins were mainly found in bacteria, including the YjgN inner membrane protein from *Escherichia coli*. The functions of these proteins are unknown. We used the STRING database [35] to investigate possible functional associations for these proteins. For YjgN, some of its predicted associated proteins are involved in colanic acid biosynthesis, such as WcaC and WcaK [36]. It is thus possible that YjgN, with its central pore, is involved in colonic acid transportation. In addition, some proteins containing the Peptidase_M48 domain were found in the gene neighborhood of DUF898 proteins. For example, A1WJ38 (DUF898) and A1WJ39 (Peptidase_M48) are products of neighboring genes from *Verminephrobacter eiseniae*. Some bacterial proteins are fused gene products containing both DUF898 and Peptidase_M48, further suggesting possible functional associations between these two families. We applied AlphaFold to model the complex between A1WJ38 (DUF898) and A1WJ39 (Peptidase_M48) and found strong interactions between the N-terminal domain of the Peptidase_M48-containing protein and the β-barrel of the DUF898 protein. Peptidase_M48-containing proteins play important roles in the maturation of a central component of the lipopolysaccharide (LPS) biogenesis machinery in bacteria [37]. DUF898 proteins may also be involved in LPS biosynthesis based on the predicted association with Peptidase_M48-containing proteins.

Interestingly, HHpred also identified another Pfam family, DUF6693 (PF20403), that contains two units of such repeats (Figure 5B). Individual units of AlphaFold models of DUF6693 proteins structurally resemble those of the DUF898 family proteins. However, proteins in the DUF6693 family adopt an open conformation since there are only two repeating units. It is plausible that DUF6693 proteins form oligomers involving more repeats, which could result in the formation of a β-barrel similar to those observed in DUF898-containing proteins. We used AlphaFold-multimer to investigate possible homo-oligomers formed by a DUF6693 protein (UniProt accession: A0A2T1A990 from *Epibacterium scottomollicae*) and indeed found a homodimer complex formed with high confidence (interdomain interaction score iPTMs > 0.8). The four beta hairpins of the two subunits in the complex form a β-barrel in a similar fashion as the DUF898 proteins. Modeling

a homotrimer complex of the DUF6693 protein did not yield high interaction scores. The functions of DUF6693 proteins remain to be revealed experimentally.

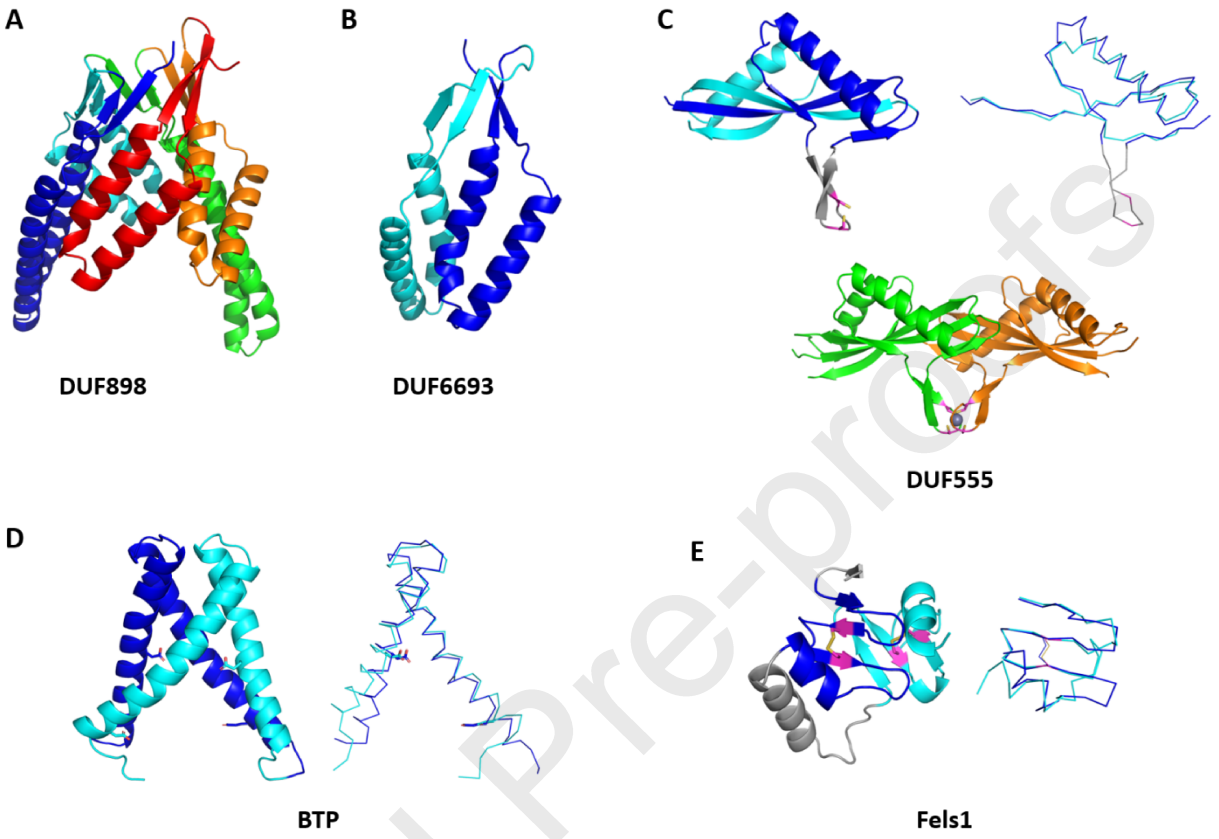


Figure 5. Examples of domain duplications in unassigned domains. A) A DUF898 protein (UniProt accession: A0A1W9VTP9) B) A DUF6693 protein (UniProt accession: A0A2T1A990). C) A DUF555 protein (UniProt accession: A9A9F5). The bottom is a dimer generated by AlphaFold3 with zinc (shown in gray) as an ion ligand. D) A BTB-domain containing protein (UniProt accession: A0A5S9MZ71). E) A Fels1-domain containing protein (UniProt accession: Q6LSA2). For B to E, the N- and C-terminal duplicated units are colored blue and cyan, respectively. For C to E, the right panel shows the superposition of the two duplicated units in lines. Two cysteines in C, conserved negatively charged residues in D, and disulfide bonds in E are shown in sticks.

Another interesting example of an internal duplication was found in proteins from the DUF555 family (PF04475). These proteins are predicted to adopt a fold with a ferredoxin-like topology consisting of a four-stranded twisted antiparallel β-sheet with two α-helices packed on one side of it in an order of βαββαβ (β: β-strand, α: α-helix). Dali searches of DUF555 proteins retrieved multiple matches with significant (Dali Z 5.5-7.8) structural similarity to other known ferredoxin-like fold proteins. Unlike the canonical ferredoxin-like domains, the β-strands of DUF555 proteins are elongated and twisted, and the two α-helices are not parallel and adopt an orthogonal packing (Figure 5C, top right). The N- and C-terminal regions of these proteins share significant sequence and structural similarity (HHpred scores >0.9, RMSD 1.48 Å) suggesting an internal duplication (Figure 5C, top right). These two parts, each corresponding to a non-compact βαβ substructure,

are inverted and interlock in the globular domain. Observed amongst the members of this DUF555 family is a twelve-residue insertion into the second β-strand which forms a β-hairpin pointing out of the β-sheet plane (shown in gray in the top structures of Figure 5C). Similarly positioned insertions in ferredoxin-like domains are not uncommon. For example, one has been previously observed in GhoS protein (pdb:2LLZ), a component of GhoT-GhoS toxin-antitoxin system [38]. Some members of the DUF555 family contain a CXXC sequence motif in this β-hairpin resembling a zinc knuckle observed in various zinc finger domains. The CXXC motif is also found in ferredoxin proteins to bind Fe-S clusters [39] and in thioredoxin proteins involved in redox processes [40]. Two CXXC motifs in zinc ribbon-type domains are involved in chelating zinc [41]. It is thus possible that the DUF555 proteins dimerize via two inserted beta-hairpins to form a zinc ribbon-like finger with the zinc binding ligands contributed by the CXXC motifs from different polypeptide chains. The dimer model of the DUF555 protein generated using AlphaFold3 [42] indeed positioned two CXXC motifs in spatial proximity to chelate a zinc ion, resembling a zinc ribbon-type domain (Figure 5C, bottom structure, ipTM score: 0.91). Alternatively, these proteins can form a higher order assembly similar to the homoheptamers of several uncharacterized archaeal proteins with a ferredoxin-like fold, such as MTH889 protein (pdb:2raq), SSo6206 protein (pdb:2x3d), and O28723_ARCFU (pdb:3bpd). DUF555 proteins were mostly found in archaea. Functional association analysis by the STRING server revealed several proteins linked by gene neighborhood analysis. For example, a DUF555 protein from *Methanolacinia petrolearia* (Mpet_2013) had gene neighborhood links to a nitroreductase family protein (Mpet_2012, score: 0.848) and a sugar kinase (ribokinase) family protein (Mpet_2014, score: 0.844).

Another example of unassigned Pfam domains belong to the BTP (PF05232) family, which represents individual sequence repeats found in tandem and predicted to form a non-compact alpha-alpha helical unit (Figure 5D). This Pfam family of proteins includes transport proteins responsible for the proteobacterial antimicrobial compound efflux (PACE) [43]. A member of this family, chlorhexidine-specific efflux pump AceI, exists in a monomer-homodimer equilibrium [44]. The dimer is probably the functional form of the protein, and the assembly of the dimer is mediated by binding of chlorhexidine and promoted by high pH conditions [44]. HHpred searches identified AlaE (L_Alanine exporter) [45] as a probable remote homolog (probability score = 89.6), which is supported by significant sequence and structural similarity.

Similar to the examples above are the unassigned domains of members of the Pfam family Fels1 (PF05666). Proteins in this family contain a duplication of a βββα unit (three short β-strands and a short α-helix) that are tightly packed against each other to form a compact domain. Each of these units represents a Pfam domain and contains two invariant cysteines, predicted to form a disulfide bond (Figure 5E). Fels1 proteins are annotated as proteins from Fels-1 prophages. These proteins with unknown function are found in various bacterial genomes, including the uncharacterized *E. coli* protein YcgJ. It should be noted that the experimentally characterized YcgJ protein from *Clostridium perfringens* [46, 47] is not evolutionarily related to *E. coli* YcgJ. *C. perfringens* YcgJ does not possess the Fels1 domain and instead has a methyltransferase domain.

Enzyme active sites revealed by AlphaFold models

Several Pfam families incorporate putative, structurally uncharacterized enzymes. AlphaFold models of their members can provide insights into their probable structures, active site architectures, and catalytic mechanisms. One example is the Pfam families from the Frag1-like clan (CL0412), including Frag1/DRAM/Sfk1 (PF10277) and DUF998 (PF06197). DUF998 family

proteins are mainly from bacteria, while Frag1/DRAM/Sfk1 family proteins are mainly distributed in eukaryotes. PGAP2 (Post-GPI Attachment to Proteins 2), previously named FRAG1, is known to play a crucial role in the biosynthesis of glycosylphosphatidylinositol (GPI) anchors [48]. PGAP2 specifically acts in the remodeling and processing of GPI anchored proteins, ensuring their stable expression and attachment to specific membrane components termed lipid-rafts or lipid-microdomains [48]. The GPI anchor serves as a membrane anchor for proteins, allowing them to be localized to the cell surface [49]. In Golgi, the GPI anchored proteins undergo fatty acid remodeling, in which the sn-2-linked unsaturated fatty acid is removed by PGAP3 and replaced with a saturated fatty acid by PGAP2 [50, 51]. It has been proposed that PGAP2 acts as an acyltransferase [48]. Consistent with this annotation, we found a conserved histidine that could serve as an active site residue in the alignment of the PGAP2 family proteins. Other acyltransferases, such as the membrane-bound O-acyltransferases (MBOAT) [52] utilize histidine as an active site residue but do not have structural similarity to DUF998 members or PGAP2. A conserved serine residue is present in the vicinity of the histidine in PGAP2 (Figure 6A). Such a His-Ser diad is conserved in DUF998 proteins and Frag1/DRAM/Sfk1 proteins, including the human proteins PGAP2 (FRAG1), DRAM, TMEM150A/B/C and the yeast protein Sfk1p [53-55].

Another Pfam clan that incorporates several acyltransferase families is the Acyl_transf_3 clan (CL0316), which, despite containing 9 Pfam families, lacks experimental structures except the recently solved structure of human heparan-alpha-glucosaminide N-acetyltransferase (HGSNAT) [56]. Representatives from these families were mostly classified as unassigned domains by DPAM. These proteins possess 8 or more transmembrane segments. We also found that the Pfam family GWT1 (PF06423) (Figure 6B) is their remote homolog and has now been classified into the Acyl_transf_3 clan. The homology of GWT1 to Acyl_transf_3 members was supported by HHpred searches and strong structural similarity between GWT1 and the experimental structure of human HGSNAT. Interestingly, the GWT1 family proteins are also acyltransferases involved in GPI anchor biosynthesis [57]. The structural folds of GWT1 and other Acyl_transf_3 members are different from those from the Frag1-like clan, such as the PGAP2 proteins. We observed several conserved polar residues (D178, R386 and Y391) in *Colletotrichum orbiculare* GWT1 in the vicinity of the proposed active site residue K168 (aligned to H269 in the experimental structure of human HGSNAT) and several conserved hydrophobic residues (F252, L448, F451 and L452) lining along the ligand binding pocket (the modeled ligand shown in spheres in Figure 6B is based on the superposition to human HGSNAT).

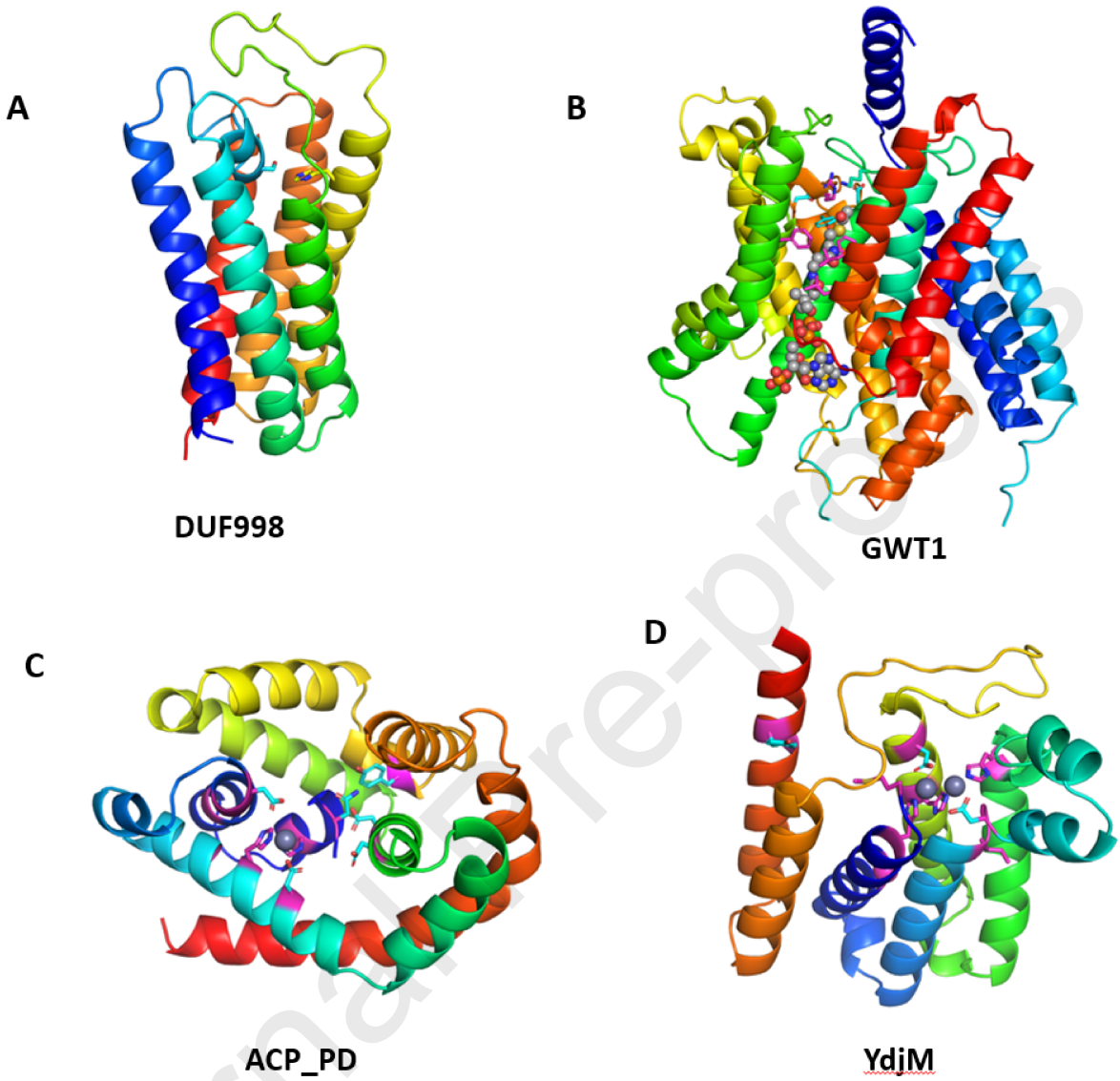


Figure 6. Examples of putative enzymes in unassigned domains. A) DUF998 protein with potential acyltransferase activity (UniProt accession: A0A542YTA8). B) A GWT1 protein with potential acyltransferase activity (UniProt accession: N4UQC3). C) An ACP-PD-domain containing protein, modeled by AlphaFold3 with a zinc ion, is a potential ACP hydrolase (UniProt accession: A0A3N0WVJ5). D) A YdjM protein with potential metal-dependent hydrolase activity (UniProt accession: Q8Y6J4) modeled by AlphaFold3 with a zinc ion. Sidechains of conserved residues are shown in sticks.

AcpH is an acyl carrier protein (ACP) hydrolase that catalyzes the conversion of holo-ACP to apo-ACP by hydrolytic cleavage of the phosphopantetheine prosthetic group from ACP [58]. AcpH belongs to the Pfam family ACP_PD (PF04336) that DPAM classified as an unassigned domain. The secondary structural contents of AcpH, consisting of mainly α-helices, is quite different from the phosphopantetheinyl hydrolase PptH from *Mycobacterium tuberculosis*, which adopts a beta/alpha fold of metallo-dependent phosphatases classified in the H-group of Carbon-nitrogen hydrolase in ECOD [59]. AcpH, a metal dependent enzyme, has been proposed to be a non-canonical member of the HD family of phosphatases and phosphodiesterases [60]. This proposal was mainly based on the primarily helical contents of the AcpH protein and the presence of

multiple conserved histidines and acidic residues. The AlphaFold model of AcpH confirmed the mainly helical content of the protein. However, it displayed a different fold from the HD family of phosphatases and phosphodiesterases (Figure 6C). In fact, AcpH appears to adopt a novel structural fold based on manual analysis that consulted the Dali and FoldSeek searches. The N-terminus of the protein forms a helical hairpin consisting of two short α-helices that lie in the middle of the structure and are surrounded by other α-helices (Figure 6C). The conserved histidine and aspartic acid residues in the first and fourth α-helices appear to form the metal binding site, which is supported by the AlphaFold3 model (Figure 6C). Another aspartic acid from the second α-helix could act as the catalytic residue, like those in many metalloproteases.

Another interesting Pfam family with potential enzymatic activity is YdjM (PF04307) [61]. It has been classified in the same Pfam clan (PhosC_NucP1, CL0368) as two soluble Pfam families: S1-P1_nuclease (S1/P1 Nuclease, PF02265) [62] and Zn_dep_PLPC (Zinc dependent phospholipase C, PF00882) [63]. However, the YdjM family and two other families in the same clan, DUF2227 and DUF4184, are transmembrane proteins. Other cases of transmembrane proteins sharing the same evolutionary origin as soluble proteins have been described before [64, 65]. The homology among these proteins was supported by conserved metal-binding residues such as histidines and acidic residues (Figure 6D). In contrast to the aforementioned Pfam families (S1-P1_nuclease and Zn_dep_PLPC) which contain a three-metal binding site, members of the YdjM family have not retained all residues that constitute the metal-binding site and are likely to bind only two instead of three metal ions, which is supported by its AlphaFold3 model (Figure 6D). Further HHpred searches identified four more Pfam families that are remote homologs of YdjM. These Pfam families (DUF6122, DUF1286, DUF4260, and DUF3307) consist of proteins with unknown functions. Proteins in these families are likely transmembrane metal-binding enzymes with hydrolase activities.

Disulfide bond-containing protein families

About 10 manually analyzed unassigned domains contain two or more disulfide bonds. Some of them have restricted phylogenetic distribution. For example, the DB module (PF01682) appears to be restricted in metazoans, the CX module (PF01705) and the TRA-1_regulated (PF02343) family proteins are mainly found in nematodes, and the BURP (PF03181) family proteins are plant specific. The functions of these disulfide bond-containing domains are largely unknown. Careful inspection of weak sequence and structural similarity search results and comparison of disulfide bond connection patterns helped reveal remote homology of some unassigned domains to known structures, as exemplified by the DB module and CX module described below.

The DB module is an α-helical domain (Figure 7A) distributed in various metazoan lineages. Weak HHpred hits and structural similarity indicate that the DB module is remotely related to the cysteine-rich helical bundle domains found in the human protein RECK [66]. They indeed have the same pattern of disulfide bond connections as revealed by the AlphaFold models of the DB module and the experimental structure of a cysteine-rich domain in RECK (Figure 7A). The cysteine-rich domain of RECK was also found to be remotely related to another protein, Her-1 from *C. elegans* [66, 67]. The cores of these structures are composed of 4 α-helices (Figure 7A). These proteins all contain the “CC” motif in the first core α-helix. The first cysteine in the CC motif forms a disulfide bond with a cysteine in the last α-helix. Both the DB module and Her-1 contain internal duplication as they have two units of the four helical domain. The second domain in the DB module is structurally divergent, as the last two α-helices merged to form one kinked long α-helix (Figure 7A).

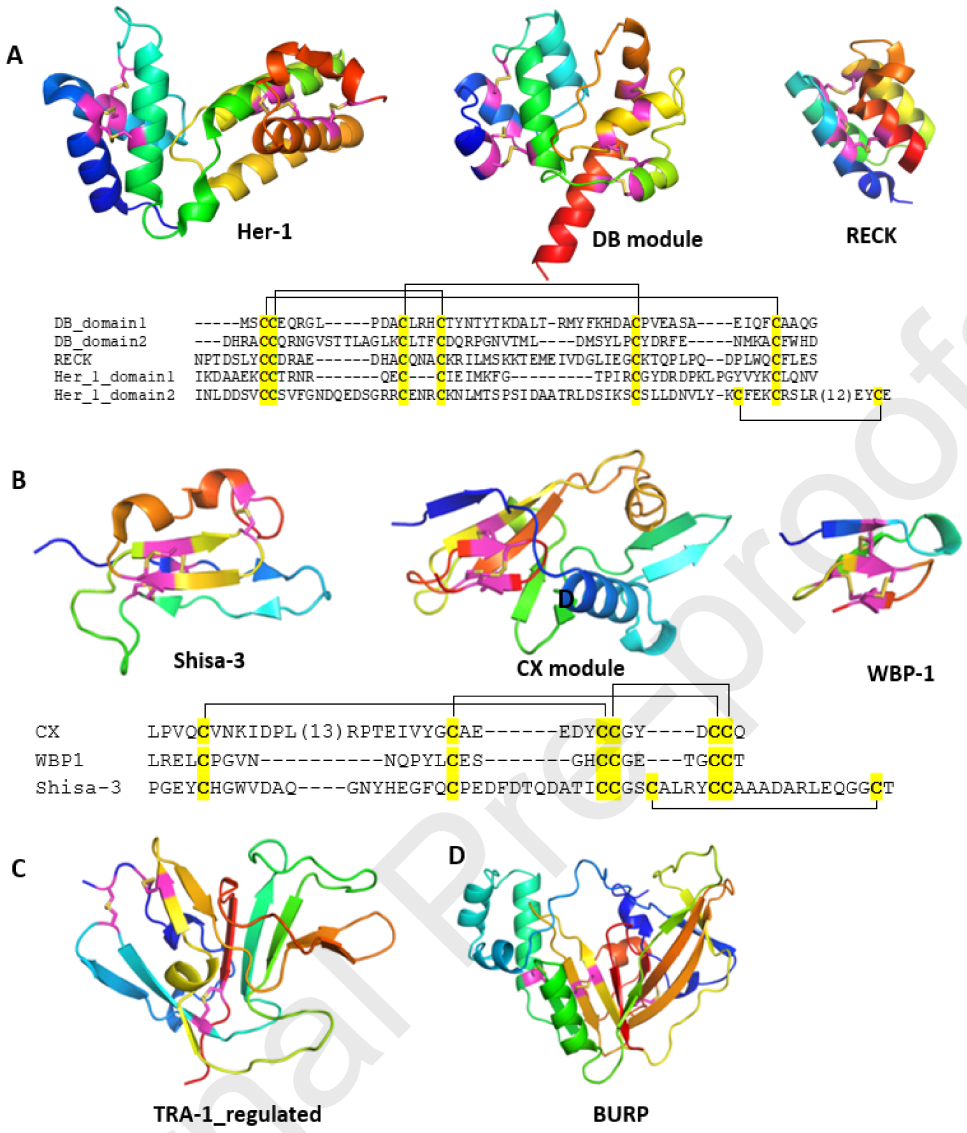


Figure 7. Examples of disulfide bond rich unassigned domains. A) Structures and alignment of Her-1, DB module and RECK. B) Structures and alignment of Shisa3, CX module and WBP-1. C) AlphaFold model of a TRA-1_regulated domain. D) AlphaFold model of a BURP domain.

The CX module is mainly made up of β-hairpins (Figure 7B). It has 6 conserved cysteines forming three disulfide bonds. The CX module was proposed to be distantly related to the Shisa domain [68]. Together with a number of other domains such as WBP-1, TMEM92 and CYYR1, they have been classified as a large group of cysteine-rich domains called STMC6 that are often found in single-transmembrane proteins [68]. These proteins all contain two “CC” motifs, which, together with two N-terminal cysteines, contribute to three disulfide bonds in two β-hairpins. Experimental structure of mouse Shisa-3 [69] and AlphaFold models of other STMC6 domains such as CX and WBP-1 confirmed their structural similarities with the same disulfide bond patterns (Figure 7B). The CX module contains the cysteine-rich domain at the C-terminus. It also possesses other structural elements such as α-helices and β-hairpins that are N-terminal to the cysteine-rich domain. The Shisa domain has an additional disulfide bond formed by one cysteine in between

the two “CC” motifs and a C-terminal cysteine [69]. The AlphaFold model of the cysteine-rich domain in WBP-1 [70] appears to have a minimal fold with less than 30 residues (Figure 7B).

Two disulfide bond-containing Pfam domains, TRA-1_regulated (PF02343) [71] and BURP (PF03181) [72] are classified as unassigned DPAM domains. These two domains are mainly made up of β-strands (Figure 7C and 7D). They do not show significant structural similarity to known structures and could be considered as new folds.

Metal-binding proteins and zinc fingers

We identified several putative metal-binding proteins based on the vicinity of conserved residues indicating metal-binding sites. One such family is NrsF (PF06532) (negative regulator of sigma F), which includes anti-sigma factors such as CC3252 from *Caulobacter crescentus* and OsrA (blr3039) from *Bradyrhizobium japonicum* [73, 74]. There is evidence that OsrA binds an ECF-type sigma factor (EcfF) in vivo. Two conserved cysteine residues (129 and 179) in OsrA are indispensable for its function [74]. This family of proteins contains 6 transmembrane α-helices (Figure 8A). The two conserved cysteines and two conserved histidines are close to each other and could serve as a metal binding site, as suggested in the AlphaFold3 model (Figure 8A). Metal binding sites have been identified in other sigma factors, such as RsrA from *Streptomyces coelicolor* [75]. Another potential metal-binding protein family is Phage_Orf51 (PF06194), in which three conserved histidines in the vicinity of each other could form a metal-binding site, as suggested by the AlphaFold3 model (Figure 8B).

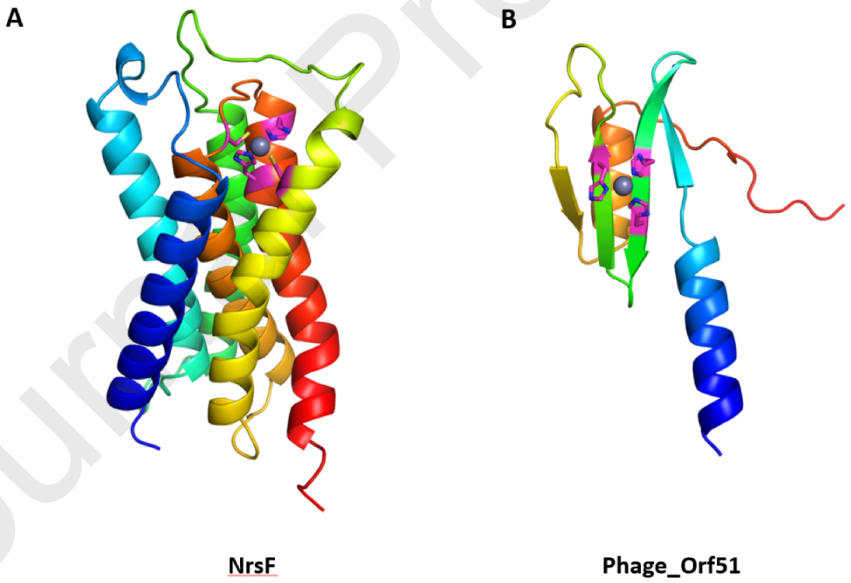


Figure 8. Examples of putative metal-binding domains. A) AlphaFold3 model of a NrsF protein (UniProt accession: A0A7Y6UGL6) chelating a zinc ion. B) AlphaFold3 model of a Phage_Orf51 protein (UniProt accession: Q2FYA6) chelating a zinc ion. Sidechains of putative metal binding residues are shown in sticks.

Manual inspection revealed several zinc finger families in the dataset of unassigned DPAM domains. One of them, zf-XS (PF03470), possesses a domain with three α-helices (Figure 9A). The long loop region between the first two α-helices possesses two conserved cysteines. They, together with two conserved histidine residues from the second α-helix and the third α-helix, form

a potential zinc binding site. This zinc finger appears to be structurally distinct from known classified zinc fingers such as zinc ribbon finger and C2H2 zinc finger [41].

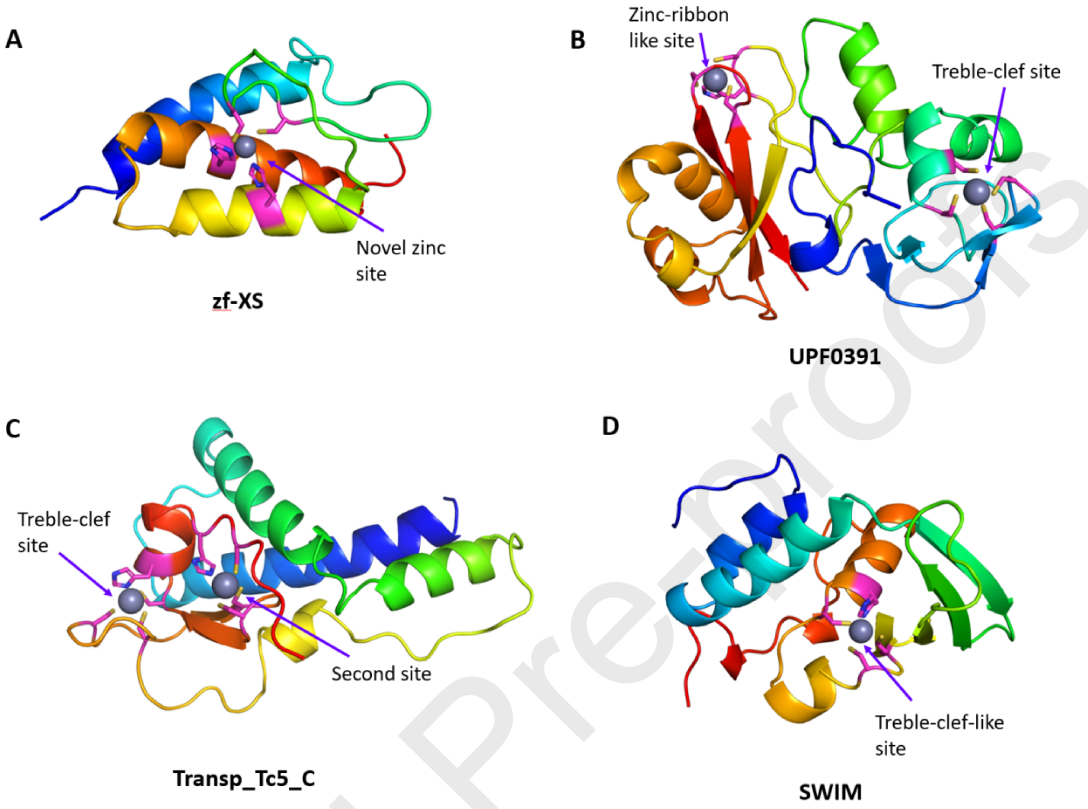


Figure 9. Examples of putative zinc finger containing domains. A) AlphaFold3 model of a zf-XS domain (UniProt accession: Q9SBW2). B) AlphaFold3 model of a UPF0391 domain (UniProt accession: A0A2A7WYE7). C) AlphaFold3 model of a Transp_Tc5_C domain (UniProt accession: A0A1D1UGS4). D) AlphaFold3 model of a SWIM domain (UniProt accession: Q9SZL8). Sidechains of putative metal binding residues are shown in sticks. Putative zinc binding sites and their types are indicated by arrows.

Some zinc fingers classified as unassigned by DPAM are structurally similar to existing zinc fingers. For example, UPF0167 (PF03691) contains an N-terminal treble-clef zinc finger where the zinc-coordinating residues (cysteines) are conserved in all members of the family (Figure 9B). Some members of this family contain a second metal-binding site at the C-terminus that resembles the zinc ribbon finger (Figure 9B). Compared to the classical zinc ribbon domains with two CxxC sequence motifs in the zinc knuckles, these proteins have a CXXC and a HC motif that contribute to the C-terminal zinc binding site. The latter motif is not strictly conserved: some family members have a histidine replaced by cysteine, while others lack this motif entirely. This lack of strict conservation of this C-terminal zinc finger domain suggests that some members of this family could have lost the zinc-binding ability and have become functionally more divergent. The UPF0167 proteins are mostly found in bacteria, including the protein CbrC (also known as YieJ) from *E. coli*. Genes encoding CrbC and CrbB, an inner membrane protein, constitute a regulon of the creBC two-component system that regulates gene expression in response to growth in minimal media [76]. CbrC is predicted to form an operon with the known Cre regulon gene, YieI (CbrB) [77]. Expression of CbrC is under control of the CreBC two-component system and was

shown to be essential for colicin E2 tolerance [78]. A recent study has demonstrated the ability of the wild-type CbrC protein to bind zinc with a measured metal to protein ratio close to two, confirming the presence of two zinc binding sites [79].

The unassigned domain corresponding to Pfam family Transc_Tc5_C (PF04236) also contains a treble clef type zinc finger and has evolved a second zinc binding site (Figure 9C). Proteins with this domain are mainly found in the Tc5 family of transposable elements in metazoans [80]. The SWIM family zinc fingers (PF04434) with a CxCxnCxH motif are present in various prokaryotic and eukaryotic proteins [81]. AlphaFold models revealed their partial structural similarity to a treble clef zinc finger, which contains a C-terminal α-helix with two zinc binding residues (Figure 9D). However, the two N-terminal zinc binding residues of the SWIM zinc finger are located in a short loop region after a four stranded β-sheet (Figure 9D), compared to their typical knuckle location in treble clef zinc fingers [41].

Quality Check: Recurring structural motifs in unassigned domains challenge structure-based classification due to possible convergence

Structural similarity, while an important criterion for inferring homology, is not always sufficient on its own to conclusively determine evolutionary relationships between proteins. Recurring structural motifs [82] such as β-meanders and α-helical bundles appear to be abundant in the set of proteins with unassigned domains. Structures with these motifs could appear through independent invention. In many cases of unassigned domains, structural similarities due to common structural motifs are insufficient to infer homology to known structures in the absence of significant sequence similarities.

Some proteins containing a DUF240 (PF03086) domain (Figure 10A) also have a C-terminal DUF237 (PF03072) domain, and these two domains are predicted to be separated by a long helix. Each of these domains folds into a curved, elongated meander β-sheet with α-helices packed on the concave side of it. These two domains share significant structural similarity, suggesting that they are related by duplication. We observed that domains from another Pfam family, Lipoprotein_3 (PF00938), have a very similar arrangement of secondary structural elements (Figure 10B). AlphaFold models of the Lipoprotein_3 family proteins superposed well with the models of the DUF240 and DUF237 domains (RMSD ranging from 1.9 to 2.9 Å). Based on this significant similarity and the conservation of specific features such as packing of the α-helices and co-location of predicted β-bulges, these three protein families are now grouped into a new clan, CL0835. While no significant sequence hits of DUF237/DUF240 and Lipoprotein_3-containing proteins were found by HHpred to proteins with known structures, structure comparison searches with Dali hit to structures (Z scores ~5) containing an extended meandering β-sheet. One such structure is an actin capping protein from *Plasmodium berghei*, shown in Figure 10C (pdb: 7a0h). The arrangement of α-helices in this experimental structure is quite different from DUF240, DUF237, and Lipoprotein_3, and its β-sheet is relatively flat and extended. Despite finding no other structural hits, the β-sheet curvature of DUF240/DUF237/Lipoprotein_3 is reminiscent of the one observed in Lipopolysaccharide binding protein (LBP), bactericidal/permeability- increasing protein (BPI) and related proteins members of the Aha1/BPI superfamily (pdb entries such as 5tod, 4m4d and 1bp1) (Figure 10D). In addition to the similarly curved β-sheet meander packed with helices, these proteins share in common a number of aromatic hydrophobic residues located at the concave side of the sheet and β-bulges at the edge. Members of the Aha1/BPI superfamily (Pfam clan: CL0648) are either involved in lipid transfer or binding of lipopolysaccharides and lipopeptides. Most of the Lipoprotein_3 family

members contain N-terminal lipoprotein attachment motifs, suggesting that they may be associated with the membrane and similarly bind lipids. Taken together, the common structural and functional features point to a possible distant evolutionary link between these protein domains.

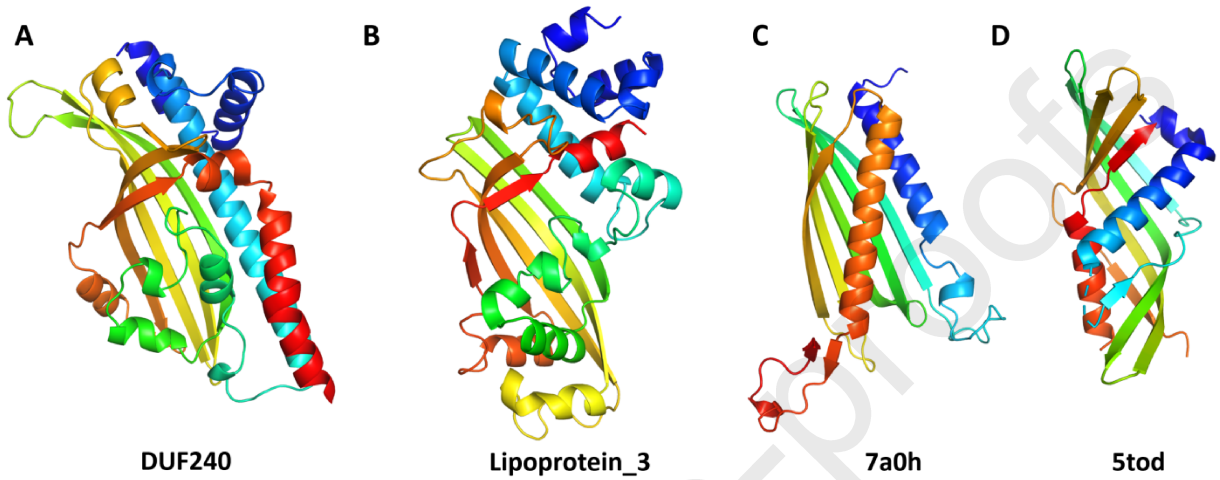


Figure 10. Examples of common structural motifs. A) AlphaFold model of a DUF240 domain (UniProt accession: A0A0H3DPT9). B) AlphaFold model of a protein (UniProt accession: A0A0H3DK99) containing the Lipoprotein_3 domain. C) A top Dali hit to DUF240 domain with known structure (PDB: 7aoh) shares the long β meandering sheet as a common structural motif. D) A member of the Aha1/BPI superfamily, TMEM24 (PDB: 5tod), shares a very similar curvature of the β-sheet meander.

Conclusions

Protein domain classification provides a systematic approach for organizing the vast diversity of proteins into meaningful categories, facilitating the study of their evolutionary and structural relationships. Categorizing domains based on sequence similarity and common structural features helps infer their functions and derive biological insights. Recent advancements in computational methods, such as AlphaFold, have revolutionized the field of protein structure prediction, making it more accessible and accurate than ever before. This has opened up new avenues for understanding domain evolution within the protein universe.

In our study, we leveraged AlphaFold's predictions to enhance domain classification using the tool DPAM-AI, which incorporates various sequence and structural measures derived from AlphaFold models in domain parsing and assignment. We focused on a large set of Pfam domains from Pfam families lacking experimental structures. By applying DPAM-AI, we were able to precisely classify a large fraction of the predicted models of Pfam domains and map them onto the existing ECOD classification hierarchy. This integration between Pfam and ECOD not only improves classification accuracy and coverage but also provides valuable functional insights, enriching our understanding of protein structure and function.

The structural data obtained from AlphaFold models proved invaluable in refining domain boundaries within Pfam regions and even suggesting potential new domain splits. However, while many domains could be confidently assigned to existing classification frameworks, a significant fraction of Pfam domains without experimental structures remained unassigned by the automatic DPAM-AI method. These unassigned domains presented challenges in understanding their evolutionary relationships, often exhibiting distinct characteristics in terms of their lengths, secondary structural contents, and transmembrane contents when compared to well-assigned domains.

Nevertheless, through manual analysis, we uncovered intriguing structural patterns within these unassigned domains, such as internal duplications and recurring structural motifs. We observed that specific sequence signals, such as conserved enzyme active sites, metal binding residues, and disulfide bond patterns, were instrumental in detecting weak homologous relationships. These findings underscore the importance of considering both structural and sequence information in domain assignment and highlight areas for improving automatic classification methods.

In summary, our study demonstrates that the structural classification of protein domains by using AlphaFold models enriches our understanding of the structural and functional diversity of the protein world and the complexity of biological systems. By integrating advanced computational tools with manual analysis, we can continue to unravel the mysteries of protein evolution and function, paving the way for future discoveries in molecular biology and biotechnology. Future research could focus on enhancing domain parsing and assignment by integrating the results from various domain prediction tools and applying these methods to classify the dark protein universe that includes proteins currently lacking domain annotations.

Materials and methods

Domain classification of AlphaFold-predicted structures using DPAM

We used the Pfam seed alignments and the Pfam annotations of UniProt proteins from Pfam release 36 [4] available on the Pfam FTP (https://ftp.ebi.ac.uk/pub/databases/Pfam/). We downloaded the sequences for all PDB entries (in Aug 2023) and merged identical sequences. We searched every non-redundant PDB sequence against the Pfam Hidden Markov Models (HMM, Pfam-A.hmm) using hmmscan from HMMER [38]. We ranked the Pfam hits for each PDB query and selected the best hits covering different regions of the query: we included a hit into the set of best hits if 50% of query residues it aligned to were not covered by existing best hits. 11,511 (55.4%) of the 20,795 Pfam families were among the best hits of PDB entries and were excluded from this study. We thus focused on 9,284 Pfam families that cannot be mapped to PDB structures.

For each Pfam family, we identified proteins containing domains of this family and having AlphaFold models in AFDB [83]. We selected three (if available) representative AlphaFold models for each family, and we favored proteins in the Pfam seed alignment and those with few inserted or deleted residues compared to the Pfam HMMs. We used DPAM to predict domains from each selected AlphaFold model. The details of the DPAM algorithm are described elsewhere [17]. Briefly, using the predicted aligned errors (PAE) distributed with the AlphaFold predictions, regions that appear disordered or as linkers are excluded. DPAM predicts the probability for a

pair of residues to belong to the same domain by their distance in the 3D structure, PAE between them, and whether this pair was aligned to the same ECOD domain based on sequence (HHsearch) and structure (DALI) searches. These probabilities were then used to cluster 5-residue segments into domains.

DPAM then utilizes a Neural Network [84] to classify the parsed domains in DPAM partitions and assign domains to ECOD homologous groups. We consider those globular domains with a DPAM confidence score above 0.85 to an ECOD entry and with a significant number of secondary structural elements (SSEs) as well-assigned domains. There are three categories of domains and regions that cannot be confidently assigned by DPAM to an ECOD homologous group. 1) “Unassigned" domains are globular domains with a significant number of SSEs that cannot be confidently assigned to a single homologous group (DPAM confidence score < 0.85). Often, these domains (referred to as unassigned domains) are sufficiently distant from known ECOD domains that additional expert considerations are needed, such as cofactor binding or functional inference from literature, to make an assignment. Frequently, these domains are candidates to expand the reference set by initiating a new homologous group, either with some partial homology signal to a known X-group (i.e., a new ECOD H-group within an existing X-group) or as an entirely new X-group. 2) “Simple topology” domains are composed of one or two SSEs and may contain intrinsically disordered or poorly predicted regions. Frequently, these domains have significant stability provided by non-SSEs (e.g., disulfide bonds, cofactors or metal binding sites). These regions may also be components of larger protein complexes and lack a globular structure outside of this broader context. 3) “Partial” domains have confident homology to a much longer reference ECOD domain. These domains are usually the result of errors either in the query protein (i.e., genome annotation) or due to inconsistency within the reference set with respect to repeat and duplication (i.e., a query domain hitting a homologous domain duplication incorrectly annotated as a single domain).

Manual analysis of unassigned Pfam domains

We manually analyzed a subset (about 150) of 908 Pfam domains where a single unassigned DPAM domain was found in all three representative proteins. We looked into sequence and structural similarity search results. To facilitate analysis, sequence conservation was calculated by AL2CO [85] and mapped to structures with the top conserved residues and disulfide-bond forming cysteines highlighted in PyMOL. We inspected the HHpred [86] results against PDB [87] and Pfam databases and paid attention to conserved motifs among weak hits. Structural similarity searches were conducted by DaliLite [88], and for some proteins also by the FoldSeek server [89]. Functional associations were carried out by the STRING web server [35]. Similar searches were also carried out for any Pfam domain that was found to be homologous to the domain under manual analysis. Such a transitive search strategy sometimes helps uncover the evolutionary relationships of multiple Pfam domains without experimental structures.

Supplementary Materials

The lists of Pfam families with the results of DPAM assignments are available at: https://conglab.swmed.edu/pfam_web/.

Funding

The study is supported by grants from the National Institute of General Medical Sciences of the National Institutes of Health GM127390 (to N.V.G.), GM147367 (to R.D.S), the Welch Foundation I-1505 (to N.V.G.), the National Science Foundation DBI 2224128 (to N.V.G.). Q.C. is a Southwestern Medical Foundation-endowed scholar. This research is partly supported by grant I-2095-20220331 to Q.C. from Welch Foundation. This work was supported by the UKRI Biotechnology and Biological Sciences Research Council [BB/X012492/1, BB/X018660/1] and EMBL core funds.

References

[1] Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014;10:e1003926.

[2] Kinch LN, Schaeffer RD, Zhang J, Cong Q, Orth K, Grishin N. Insights into virulence: structure classification of the Vibrio parahaemolyticus RIMD mobilome. mSystems. 2023;8:e0079623.

[3] Zhang Y, Chandonia JM, Ding C, Holbrook SR. Comparative mapping of sequence-based and structure-based protein domains. BMC Bioinformatics. 2005;6:77.

[4] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: The protein families database in 2021. Nucleic acids research. 2021;49:D412-D9.

[5] Wang J, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu S, et al. The conserved domain database in 2023. Nucleic Acids Res. 2023;51:D384-D8.

[6] Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res. 2002;30:268-72.

[7] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247:536-40.

[8] Chandonia JM, Fox NK, Brenner SE. SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. Nucleic Acids Res. 2019;47:D475-D81.

[9] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res. 2021;49:D266-D73.

[10] Schaeffer RD, Liao Y, Cheng H, Grishin NV. ECOD: new developments in the evolutionary classification of domains. Nucleic Acids Res. 2017;45:D296-D302.

[11] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021.

[12] Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, et al. Computed structures of core eukaryotic protein complexes. Science. 2021;374:eabm4805.

[13] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50:D439-D44.

[14] Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, et al. Uncovering new families and folds in the natural protein universe. Nature. 2023;622:646-53.

[15] Schaeffer RD, Zhang J, Kinch LN, Pei J, Cong Q, Grishin NV. Classification of domains in predicted structures of the human proteome. Proc Natl Acad Sci U S A. 2023;120:e2214069120.

[16] Schaeffer RD, Zhang J, Medvedev KE, Kinch LN, Cong Q, Grishin NV. ECOD domain classification of 48 whole proteomes from AlphaFold Structure Database using DPAM2. PLoS Comput Biol. 2024;20:e1011586.

[17] Zhang J, Schaeffer RD, Durham J, Cong Q, Grishin NV. DPAM: A domain parser for AlphaFold models. Protein Sci. 2023;32:e4548.

[18] Liao Y, Schaeffer RD, Pei J, Grishin NV. A Sequence Family Database Built on ECOD Structural Domains. Bioinformatics. 2018.

[19] UniProt C. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51:D523-D31.

[20] Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res. 2023;51:D418-D27.

[21] Litman GW, Rast JP, Fugmann SD. The origins of vertebrate adaptive immunity. Nat Rev Immunol. 2010;10:543-53.

[22] Bodelon G, Palomino C, Fernandez LA. Immunoglobulin domains in Escherichia coli and other enterobacteria: from pathogenesis to applications in antibody technologies. FEMS Microbiol Rev. 2013;37:204-50.

[23] Chatterjee S, Basak AJ, Nair AV, Duraivelan K, Samanta D. Immunoglobulin-fold containing bacterial adhesins: molecular and structural perspectives in host tissue colonization and infection. FEMS Microbiol Lett. 2021;368.

[24] Zacharchenko T, von Castelmur E, Rigden DJ, Mayans O. Structural advances on titin: towards an atomic understanding of multi-domain functions in myofilament mechanics and scaffolding. Biochem Soc Trans. 2015;43:850-5.

[25] Schwarzbauer JE, DeSimone DW. Fibronectins, their fibrillogenesis, and in vivo functions. Cold Spring Harb Perspect Biol. 2011;3.

[26] Perez-Riba A, Itzhaki LS. The tetratricopeptide-repeat motif is a versatile platform that enables diverse modes of molecular recognition. Curr Opin Struct Biol. 2019;54:43-9.

[27] Andrade MA, Bork P. HEAT repeats in the Huntington's disease protein. Nat Genet. 1995;11:115-6.

[28] Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS microbiology reviews. 2005;29:231-62.

[29] Mayer BJ. SH3 domains: complexity in moderation. Journal of cell science. 2001;114:1253-63.

[30] Tay PKR, Lim PY, Ow DS. A SH3_5 Cell Anchoring Domain for Non-recombinant Surface Display on Lactic Acid Bacteria. Front Bioeng Biotechnol. 2020;8:614498.

[31] Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res. 2010;38:D211-22.

[32] Ketchesin KD, Stinnett GS, Seasholtz AF. Corticotropin-releasing hormone-binding protein and stress: from invertebrates to humans. Stress. 2017;20:449-64.

[33] Ponnuraj K, Rowland S, Nessi C, Setlow P, Jedrzejas MJ. Crystal structure of a novel germination protease from spores of Bacillus megaterium: structural arrangement and zymogen activation. J Mol Biol. 2000;300:1-10.

[34] Carroll TM, Setlow P. Site-directed mutagenesis and structural studies suggest that the germination protease, GPR, in spores of Bacillus species is an atypical aspartic acid protease. J Bacteriol. 2005;187:7119-25.

[35] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45:D362-D8.

[36] Scott PM, Erickson KM, Troutman JM. Identification of the functional roles of six key proteins in the biosynthesis of Enterobacteriaceae colanic acid. Biochemistry. 2019;58:1818-30.

[37] Bryant JA, Cadby IT, Chong Z-S, Boelter G, Sevastsyanovich YR, Morris FC, et al. Structure-function characterization of the conserved regulatory mechanism of the Escherichia coli M48 metalloprotease BepA. Journal of Bacteriology. 2020;203:10.1128/jb. 00434-20.

[38] Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic acids research. 2018;46:W200-W4.

[39] Krishna SS, Sadreyev RI, Grishin NV. A tale of two ferredoxins: sequence similarity and structural differences. BMC structural biology. 2006;6:1-7.

[40] Schultz LW, Chivers PT, Raines RT. The CXXC motif: crystal structure of an active-site variant of Escherichia coli thioredoxin. Acta Crystallographica Section D: Biological Crystallography. 1999;55:1533-8.

[41] Krishna SS, Majumdar I, Grishin NV. Structural classification of zinc fingers: survey and summary. Nucleic acids research. 2003;31:532-50.

[42] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024.

[43] Hassan KA, Liu Q, Elbourne LD, Ahmad I, Sharples D, Naidu V, et al. Pacing across the membrane: the novel PACE family of efflux pumps is widespread in Gram-negative pathogens. Research in Microbiology. 2018;169:450-4.

[44] Bolla JR, Howes AC, Fiorentino F, Robinson CV. Assembly and regulation of the chlorhexidine-specific efflux pump AceI. Proc Natl Acad Sci U S A. 2020;117:17011-8.

[45] Hori H, Yoneyama H, Tobe R, Ando T, Isogai E, Katsumata R. Inducible L-alanine exporter encoded by the novel gene ygaW (alaE) in Escherichia coli. Applied and environmental microbiology. 2011;77:4027-34.

[46] Ohtani K, Takamura H, Yaguchi H, Hayashi H, Shimizu T. Genetic analysis of the ycgJ-metB-cysK-ygaG operon negatively regulated by the VirR/VirS system in Clostridium perfringens. Microbiol Immunol. 2000;44:525-8.

[47] Ohtani K, Hayashi H, Shimizu T. The luxS gene is involved in cell-cell signalling for toxin production in Clostridium perfringens. Mol Microbiol. 2002;44:171-9.

[48] Tashima Y, Taguchi R, Murata C, Ashida H, Kinoshita T, Maeda Y. PGAP2 is essential for correct processing and stable expression of GPI-anchored proteins. Molecular biology of the cell. 2006;17:1410-20.

[49] Kinoshita T. Biosynthesis and biology of mammalian GPI-anchored proteins. Open biology. 2020;10:190290.

[50] Maeda Y, Tashima Y, Houjou T, Fujita M, Yoko-o T, Jigami Y, et al. Fatty acid remodeling of GPI-anchored proteins is required for their raft association. Mol Biol Cell. 2007;18:1497-506.

[51] Tashima Y, Taguchi R, Murata C, Ashida H, Kinoshita T, Maeda Y. PGAP2 is essential for correct processing and stable expression of GPI-anchored proteins. Mol Biol Cell. 2006;17:1410-20.

[52] Coupland CE, Ansell TB, Sansom MS, Siebold C. Rocking the MBOAT: Structural insights into the membrane bound O-acyltransferase family. Current Opinion in Structural Biology. 2023;80:102589.

[53] Crighton D, Wilkinson S, O'Prey J, Syed N, Smith P, Harrison PR, et al. DRAM, a p53-induced modulator of autophagy, is critical for apoptosis. Cell. 2006;126:121-34.

[54] Audhya A, Emr SD. Stt4 PI 4-kinase localizes to the plasma membrane and functions in the Pkc1-mediated MAP kinase cascade. Developmental cell. 2002;2:593-605.

[55] Barthet VJ, Ryan KM. DRAMs and autophagy: a family affair. Autophagy Reports. 2022;1:170-4.

[56] Navratna V, Kumar A, Mosalaganti S. Structure of the human heparan-alpha-glucosaminide N-acetyltransferase (HGSNAT). bioRxiv. 2023.

[57] Umemura M, Okamoto M, Nakayama K-i, Sagane K, Tsukahara K, Hata K, Jigami Y. GWT1 gene is required for inositol acylation of glycosylphosphatidylinositol anchors in yeast. Journal of Biological Chemistry. 2003;278:23639-47.

[58] Thomas J, Cronan JE. The enigmatic acyl carrier protein phosphodiesterase of Escherichia coli: genetic and enzymological characterization. Journal of Biological Chemistry. 2005;280:34675-83.

[59] Mosior J, Bourland R, Soma S, Nathan C, Sacchettini J. Structural insights into phosphopantetheinyl hydrolase PptH from Mycobacterium tuberculosis. Protein Science. 2020;29:744-57.

[60] Thomas J, Rigden DJ, Cronan JE. Acyl carrier protein phosphodiesterase (AcpH) of Escherichia coli is a non-canonical member of the HD phosphatase/phosphodiesterase family. Biochemistry. 2007;46:129-36.

[61] Zhang Y, Lin K. A phylogenomic analysis of Escherichia coli/Shigella group: implications of genomic features associated with pathogenicity and ecological adaptation. BMC evolutionary biology. 2012;12:1-12.

[62] Romier C, Dominguez R, Lahm A, Dahl O, Suck D. Recognition of single-stranded DNA by nuclease P1: High resolution crystal structures of complexes with substrate analogs. Proteins: Structure, Function, and Bioinformatics. 1998;32:414-24.

[63] Hough E, Hansen LK, Birknes B, Jynge K, Hansen S, Hordvik A, et al. High-resolution (1.5 Å) crystal structure of phospholipase C from Bacillus cereus. Nature. 1989;338:357-60.

[64] Neuwald AF. An unexpected structural relationship between integral membrane phosphatases and soluble haloperoxidases. Protein science. 1997;6:1764-7.

[65] Goblirsch BR, Wiener MC. Ste24: an integral membrane protein zinc metalloprotease with provocative structure and emergent biology. Journal of molecular biology. 2020;432:5079-90.

[66] Chang T-H, Hsieh F-L, Smallwood PM, Gabelli SB, Nathans J. Structure of the RECK CC domain, an evolutionary anomaly. Proceedings of the National Academy of Sciences. 2020;117:15104-11.

[67] Hamaoka BY, Dann III CE, Geisbrecht BV, Leahy DJ. Crystal structure of Caenorhabditis elegans HER-1 and characterization of the interaction between HER-1 and TRA-2A. Proceedings of the National Academy of Sciences. 2004;101:11673-8.

[68] Pei J, Grishin NV. Unexpected diversity in Shisa-like proteins suggests the importance of their roles as transmembrane adaptors. Cellular signalling. 2012;24:758-69.

[69] McCoy AJ, Oeffner RD, Wrobel AG, Ojala JR, Tryggvason K, Lohkamp B, Read RJ. Ab initio solution of macromolecular crystal structures without direct methods. Proceedings of the National Academy of Sciences. 2017;114:3637-41.

[70] Sudol M, Chen HI, Bougeret C, Einbond A, Bork P. Characterization of a novel protein-binding module—the WW domain. FEBS letters. 1995;369:67-71.

[71] Thoemke K, Yi W, Ross JM, Kim S, Reinke V, Zarkower D. Genome-wide analysis of sex-enriched gene expression during C. elegans larval development. Developmental biology. 2005;284:500-8.

[72] Hattori J, Boutilier K, Campagne ML, Miki B. A conserved BURP domain defines a novel group of plant proteins with unusual primary structures. Molecular and General Genetics MGG. 1998;259:424-8.

[73] Kohler C, Lourenço RF, Avelar GM, Gomes SL. Extracytoplasmic function (ECF) sigma factor σ F is involved in Caulobacter crescentus response to heavy metal stress. BMC microbiology. 2012;12:1-14.

[74] Masloboeva N, Reutimann L, Stiefel P, Follador R, Leimer N, Hennecke H, et al. Reactive oxygen species-inducible ECF σ factors of Bradyrhizobium japonicum. 2012.

[75] Zdanowski K, Doughty P, Jakimowicz P, O'Hara L, Buttner MJ, Paget MS, Kleanthous C. Assignment of the zinc ligands in RsrA, a redox-sensing ZAS protein from Streptomyces coelicolor. Biochemistry. 2006;45:8294-300.

[76] Avison MB, Horton RE, Walsh TR, Bennett PM. Escherichia coli CreBC is a global regulator of gene expression that responds to growth in minimal media. Journal of Biological Chemistry. 2001;276:26955-61.

[77] Zhou L, Lei XH, Bochner BR, Wanner BL. Phenotype microarray analysis of Escherichia coli K-12 mutants with deletions of all two-component systems. J Bacteriol. 2003;185:4956-72.

[78] Cariss SJ, Constantinidou C, Patel MD, Takebayashi Y, Hobman JL, Penn CW, Avison MB. YieJ (CbrC) mediates CreBC-dependent colicin E2 tolerance in Escherichia coli. J Bacteriol. 2010;192:3329-36.

[79] Cheng Y, Wang H, Xu H, Liu Y, Ma B, Chen X, et al. Co-evolution-based prediction of metal-binding sites in proteomes by machine learning. Nat Chem Biol. 2023;19:548-55.

[80] Collins JJ, Anderson P. The Tc5 family of transposable elements in Caenorhabditis elegans. Genetics. 1994;137:771-81.

[81] Makarova KS, Aravind L, Koonin EV. SWIM, a novel Zn-chelating domain present in bacteria, archaea and eukaryotes. Trends in biochemical sciences. 2002;27:384-6.

[82] Orengo C, Flores T, Jones D, Taylor W, Thornton J. Recurring structural motifs in proteins with different functions. Current biology. 1993;3:131-9.

[83] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research. 2022;50:D439-D44.

[84] Schaeffer RD, Zhang J, Kinch LN, Pei J, Cong Q, Grishin NV. Classification of domains in predicted structures of the human proteome. Proceedings of the National Academy of Sciences. 2023;120:e2214069120.

[85] Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics. 2001;17:700-12.

[86] Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Soding J, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinformatics. 2020;72:e108.

[87] Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM. The RCSB PDB information portal for structural genomics. Nucleic acids research. 2006;34:D302-D5.

[88] Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics. 2000;16:566-7.

[89] van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42:243-6.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for *[Journal name]* and was not involved in the editorial review or the decision to publish this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: