Received: 19 September 2024 | Revised: 9 January 2025 | Accepted: 5 February 2025

DOI: 10.1002/pro.70074

RESEARCH ARTICLE

THE PROTEIN SOCIETY WILEY

Refinement and curation of homologous groups facilitated by structure prediction

Richard Dustin Schaeffer[1] | Jimin Pei[1,2,3] | Jing Zhang[2,3] | Qian Cong[1,2,3] | Nick V. Grishin[1,4]

[1]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, Texas, USA

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

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

[4]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Correspondence
Richard Dustin Schaeffer, Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA.
Email: dustin.schaeffer@gmail.com

Funding information
Welch Foundation, Grant/Award Numbers: I-1505, I-2095-20220331; National Institute of General Medical Sciences, Grant/Award Numbers: GM127390, GM147367; National Science Foundation, Grant/Award Numbers: DBI 2224128, MED230034, MED240004; National Institute of Allergy and Infectious Diseases, Grant/Award Number: 1K99AI180984-01

Review Editor: Lynn Kamerlin

Abstract

Domain classification of protein predictions released in the AlphaFold Database (AFDB) has been a recent focus of the Evolutionary Classification of protein Domains (ECOD). Although a primary focus of our recent work has been the partition and assignment of domains from these predictions, we here show how these diverse predictions can be used to examine the reference domain set more closely. Using results from DPAM, our AlphaFold-specific domain parsing algorithm, we examine hierarchical groupings that share significant levels of homologous links, both between groups that were not previously assessed to be definitively homologous and between groups that were not previously observed to share significant homologous links. Combined with manual analysis, these large datasets of structural and sequence similarities allow us to merge homologous groups in multiple cases which we detail within. These domains tend to be families of domains from families that are either small, previously had few experimental representatives, or had unknown function. The exception to this is the chromodomains, a large homologous group which were increased from "possibly homologous" to "definitely homologous" to increase the consistency of ECOD based their strong homologous links to the SH3 domains.

KEYWORDS

domain classification, DPAM, ECOD, protein homology, structure prediction

1 | INTRODUCTION

Proteins can be classified into domains by detecting their structure and sequence similarities. Small differences in how structural and sequence similarity are considered can lead to differences between domain classifications. The Evolutionary Classification of protein Domains (ECOD) is a structural domain classification that has been actively updated for over a decade (Cheng et al. 2014; Schaeffer et al. 2017). ECOD relies on the detection of distant homology using methods such as HHsearch (Soding et al. 2005) and Dali (Holm 2019), allows for fold change between homologous domains, and has a mixed manual/automatic classification method. Automated domain classification can miss functional considerations that might imply homology, or structural nuances not captured by similarity searches. For difficult cases, manual curation has often shown success where automation fails (Cheng et al. 2015). The recent development of highly accurate structure prediction has led to a dramatic shift in the quantity of available structural data (Jumper et al. 2021;

Tunyasuvunakool et al. 2021; Varadi et al. 2022; Varadi et al. 2024). We have published multiple studies of highly focused classification of proteomes of individual species such as human (Schaeffer et al. 2023) or *Vibrio parahaemolyticus* RIMD (Kinch et al. 2023), as well as a set of 48 model organisms (Schaeffer et al. 2024a). In these studies, we principally used a reference set of domains derived from experimental structures to classify domains from predicted protein structures. Conversely, here we demonstrate how these sequence and structure similarity data can be used to identify potential inconsistencies in our reference set and illustrate cases in which these data prompted us to modify our reference set.

Distant homology can be difficult to distinguish from convergent evolution or analogy (Medvedev et al. 2021). Weak structural similarity can be indicative of shared ancestry, but not definitive. ECOD distinguishes between probable (X-group) and definite (H-group) homology for this reason. Domain classifications are incomplete, and reclassification or reconsideration of those groups or domains as time has passed (and more structural data has been determined/predicted) can lead to novel insights or error correction. Using our recently developed Domain Parser for AlphaFold Models (DPAM), we have classified a series of sets of predicted proteins using a reference set of domains derived from experimental structures (Zhang et al. 2022). DPAM partitions and assigns domains to the ECOD reference in two steps, in contrast to our previous method which did it as a single step. Although DPAM assigns a putative domain by the hit with the highest confidence, multiple high-scoring hits for a given domain can be detected. For domain assignments, we use parameter thresholds (e.g., DPAM probability, HHsearch probability of homology) to determine when a homologous link exists. The use of such thresholds is supported by evidence that there is a sharp asymptotic transition above which many thresholds perform similarly (Donald and Shakhnovich 2005). The HHsuite manual suggests that 95% homology is "near certain," but our experience suggests even 90% or greater can be assigned with few false positives (as long as alignment coverage is considered) (Steinegger et al. 2019). Indeed, HH probabilities can be meaningful for identifying leads between 50% and 60%. DPAM probability is a newer measure, but we have found that hits between 0.5 and 1.0 should be considered, with hits above 0.6 often being automatically classified (Schaeffer et al. 2024a). In some cases, reference domains with multiple hits above these thresholds belong to different ECOD homologous groups. Where a region of a predicted model has a potential high-confidence assignment to differing homologous groups, it can reveal potential insights about evolution, regions where the ECOD classification should be amended, or possibly reference domains with boundary problems (i.e., they contain more than one domain or are fragments of other domains). These cases can be resolved by manual curation, considering functional and evolutionary data from literature and additional conserved features that may not be represented well in aligner scores.

Structural similarity between different homologous groups can be observed despite distinct evolutionary origins (Sadreyev et al. 2009). DPAM relies on measures of both structural similarity and sequence similarity to determine an overall confidence of homology. When the DPAM assignment confidence exceeds a threshold value (>0.6), we consider a domain to be assigned to the homologous group of that reference. Although domains are assigned based on a consensus of homologous links among the confident hits, we do not generally analyze those cases where multiple confident assignments are possible. Here we revisit previously assigned cases and examine marginal hits (i.e., confident hits that were not used) as potential evidence of homology and as validation of the classification. We illustrate multiple cases where these assignment data could be used to improve the ECOD classification. In the first case, the preponderance of homologous links between the SH3 domains and the chromo barrel domains (chromodomains), combined with significant literature evidence published since the initial classification, allows us to change the relationship between the SH3 domains and chromodomains from possible to definite homology. We also unify a family of outer membrane proteins and type 3 secretion system components where previously there was insufficient evidence to make a confident classification. We identified a group of winged three-helix bundles where collaboration with Pfam allowed us to make a more consistent homologous group. Overall, this study demonstrates the utility of large-scale structure prediction not only in aiding structural classification but also in reflecting on the reference dataset in a search for inconsistencies, as well as the potential gains from carefully mixing domains from experimental and predicted sources within a classification to gain clarity on ambiguous homologous relationships.

2 | RESULTS AND DISCUSSION

2.1 | Consideration of confident DPAM hits to reveal links between ECOD homologous groups

We searched AFDB protein structure predictions for similarity to ECOD domains by both sequence and structure (see section 4). Proteins were partitioned into domains by DPAM, and those putative domains were assigned to the ECOD hierarchy by their structure and sequence similarity to reference domains from experimental structures.

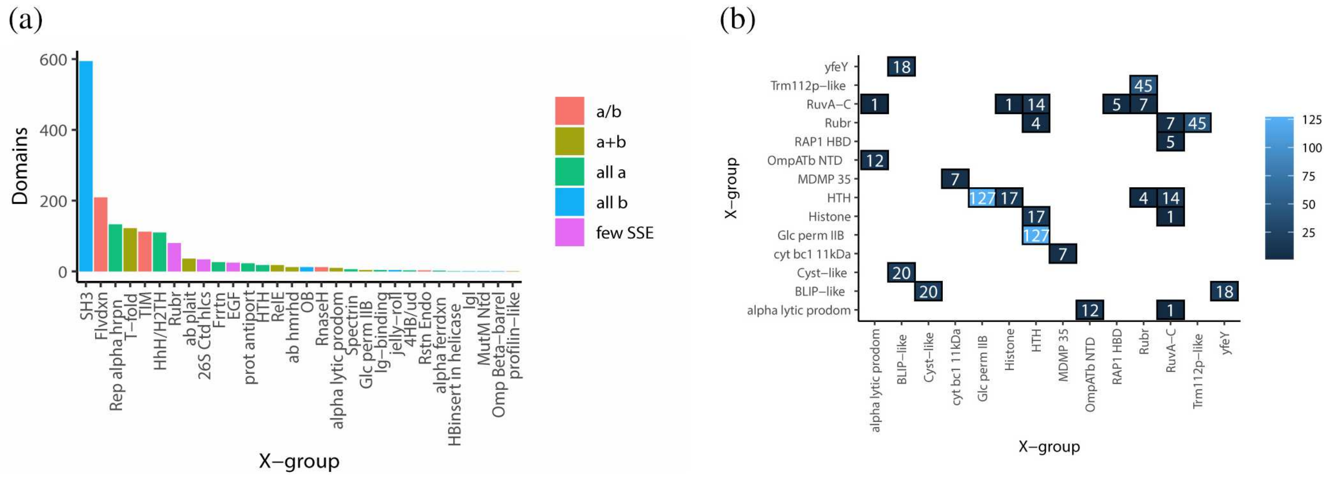

FIGURE 1 Homologous links within and between ECOD potentially homologous (X) groups. (a) X-groups and 48 proteomes AFDB domains with links to multiple H-groups. (b) X-group pairs and the number of AFDB domains with substantive homologous links selected by curators for remediation and/or curation. X-group short names (e.g., SH3, Flvdxn) are indexed in Table S1. The plot is colored by the absolute number of domain pairs detected.

We focused on those domains where the detected homologous relationships (or links) suggest possible confident assignments to multiple X- or H-groups. Homologous links between topology groups or sequence families are expected using measures of distant homology and were not further examined here. We expect most proteins examined to have confident homologous links to a single X- or H-group. Where this is not true, it signals (1) potential errors in the domain partition process, (2) inconsistencies in the reference set, or (3) domain boundary problems where multiple domains have been combined. Because DPAM probability can be less sensitive to alignment coverage, we also use query and reference coverage (>50%) of HHsearch alignments to filter for additional confident hits.

We considered multiple hierarchy levels: links within groups of possible homology (X-groups) and definite homology (H-groups). Among more than 490,000 proteins and their 1.18M putative domains from the AFDB 48 proteomes we analyzed, we found that 23,210 (1.95%) domains had confident links (DPAM probability >0.5 and bidirectional alignment coverage >50%) to more than one ECOD homology group, whereas 11,845 (1.00%) domains had confident links to multiple X-groups. Broadly, these levels of consistency were within the expected bounds from our previous large-scale classifications of PDB structures (Schaeffer et al. 2021). In a consistent classification, we expect that most domains should be homologous to a single ECOD homologous group.

More specifically, we were interested in where these inconsistencies were focused within ECOD. We examined inconsistencies between homologous groups with X-groups (Figure 1a) and inconsistencies that linked disparate X-groups (Figure 1b). In both cases, a single case dominated the results and was selected for further analysis. SH3-like domains (ECOD X: 4) showed significant links (36% of total links found) between the SH3 domains (ECOD H: 4.3) and the chromodomains (ECOD H: 4.8), which when combined with recent literature evidence (Schaeffer et al. 2021) directly leads to the decision to consider the chromodomains as definitively homologous to the SH3 domains (see below). The flavodoxin-like domains (ECOD X: 2007) show some internal mixing (12%) between the Toll/Interleukin receptor (TIR) domains (ECOD H: 2007.9) and the N-deoxyribosyltransferases (ECOD H: 2007.15). At the time these flavodoxin links are not sufficiently populated nor is there sufficient literature evidence to justify further merging of these groups. The repetitive alpha hairpins (ECOD X: 109) are a difficult group to classify due to frequent boundary problems (Schaeffer et al. 2016) and often show mixing between H-groups, seen here as 8% of intra H-group mixing. We also identified homologous links between X-groups (rather than within) nD2 and used those data to identify cases where remediation or reclassification was necessary. The full collection of cross X-group data is presented in Table S2, Supporting Information. The most prolifically linked groups among those we selected were the (1) helix-turn-helix domains and glucose permease IIB-like domains (which contains the eIF1-like H-group), (2) the Trm112p-like domains and the rubredoxin domains, and the (3) cystatin-like domains and the BLIP-like domains. We elaborate on the consequences of this categorization and describe further criteria for its clear identification and the subsequent repair to the data set.

2.2 | Chromodomains are homologous to SH3 barrel domains

ECOD v292 contains distinct homologous groups for the SH3 domains (H: 4.3) and the chromodomains

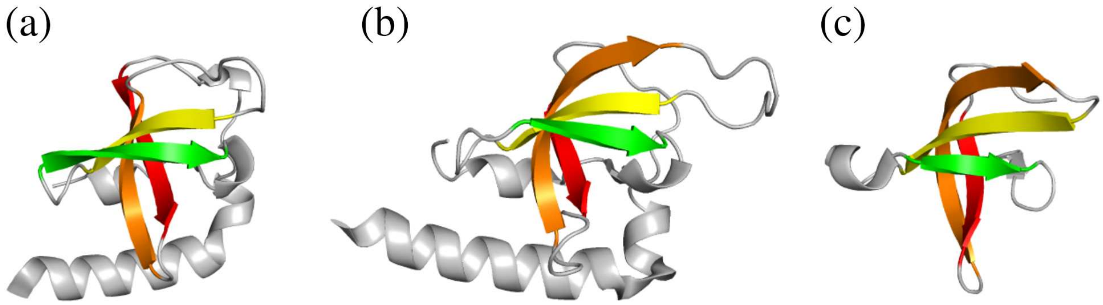

FIGURE 2 Homology between an SH3 domain and a chromodomain using an ECOD domain from predicted structures. (a) An SH3 domain from an uncharacterized *T. trichuria* protein (ECOD 48p: A0A077Z1I8_F1_nD2) with canonical SH3 strands colored individually (S1: red, S2: orange, S3: yellow, S4: green). (b) Structure of *B. mori* Tudor ECOD SH3 domain (ECOD PDB: e5vqhA1), the representative domain used in initial assignment. (c) Structure of *A. thaliana* Tudor-knot chromodomain (ECOD PDB: e4pl6B1), domain with confident but unused assignment to differing homologous groups.

(H: 4.8). In the pilot version, the ECOD SH3 H-group was formed from domains from 13 SCOP superfamilies, most from the SH3 fold (SCOP: b.34) but also from two superfamilies in the Sm-like fold (SCOP: b.38). At the time, the chromodomain superfamily (SCOP: b.34.13) was considered not definitively homologous and formed a distinct homologous group (Andreeva et al. 2020). The canonical function of chromodomains is binding nucleic acids and methylated histones, which contributes to their ability to bind and remodel chromatin (Eissenberg 2012). This function is exemplified by Pfam/ECOD families such as the Chromo (PF00385, ECOD F: 4.8.1.1), Chromo_2 (PF18704, ECOD F: 4.8.1.9), and Chromo_shadow (PF01393, ECOD F: 4.8.1.2) domains. The MBT family (PF02820, ECOD F: 4.8.1.4) is classified within the Chromo domain-like homologous group, but has a protein–protein interaction function, rather than nucleic acid binding. ECOD, having recently standardized its sequence family classification against Pfam, can also evaluate the distribution of our sequence families in homologous groups versus Pfam families in their Clans classification (Schaeffer et al. 2024b). In ECOD's case, all but one family (ComK, PF06338, ECOD F: 4.8.1.13) are classified in the SH3 (CL0010) Pfam clan. Additionally, although a single Tudor domain sequence family (Tudor-knot) is classified as a chromo-like domain (PF11717, ECOD F: 4.8.1.6), numerous Tudor-domain families (such as the canonical Tudor domain, PF00567, ECOD F: 4.1.1.9, and the PTM7/DIR17-like Tudor domain, PF21743, 4.1.1.141) are classified within the SH3-like homologous group. Accordingly, although the canonical chromo-like and SH3 domain functions are nucleic-acid and protein-motif recognition, respectively, there are examples of either function occurring in both the SH3-like and chromo domain-like homologous groups. Subsequent sequence and structure analysis further substantiated the common ancestry of canonical SH3 folds and chromodomains through an ancestral zinc-ribbon fold (Kaur et al. 2018).

We found repeated transitive links to SH3 domains among 20% of ECOD representative (F70) chromodomains. Figure 2 illustrates an example of such a transitive link: AFDB domain A0A077Z1I8_F1_nD2 (Figure 2a) contains the canonical SH3 barrel strands and was assigned with high confidence (DPAM probability = 0.99) to a domain (ECOD: e5vqhA1) in the SH3 ECOD homologous groups (Figure 2b). This SH3 domain also was linked with lower confidence (DPAM probability = 0.76) to a domain (ECOD: e4pl6B1) in the Chromodomain ECOD H-group (Figure 2c).

2.3 | Enumerating the probable homology between beta-lactamase inhibitor domains and cystatin-like domains

The cystatin-like domains have a conserved structural motif consisting of a five-stranded antiparallel β-sheet and a single α-helix. This motif is principally found in the cystatin family of proteins but can be observed elsewhere (Grzonka et al. 2001; Quimby et al. 2001). These domains are commonly found in proteins with inhibitory functions, particularly against various proteases (Dubin 2005). Domains with this structural motif are grouped in the cystatin-like X-group (ECOD X: 243). Within this X-group, the cystatin sequence family (PF0037) and other protease inhibitor families such as the PePSY (PF03413), SQAPI (PF16845), and YPEB_PepSY1-2 (PF14620) were grouped in the cystatin/monellin homologous group (ECOD H: 243.1). Among our AFDB classification data, we found significant hits (DPAM probability >0.97) linking the BLIP-like (beta-lactamase inhibitor-like) X-group (ECOD X: 809) and the cystatin-like homologous group. For example, a periplasmic protein (UNP: Q0P802) from *Campylobacter jejuni* found a hit with DPAM probability 0.99 to a domain classified in the BLIP-like X-group (e3db7A4; Figure 3a) and with DPAM probability 0.95 to a domain

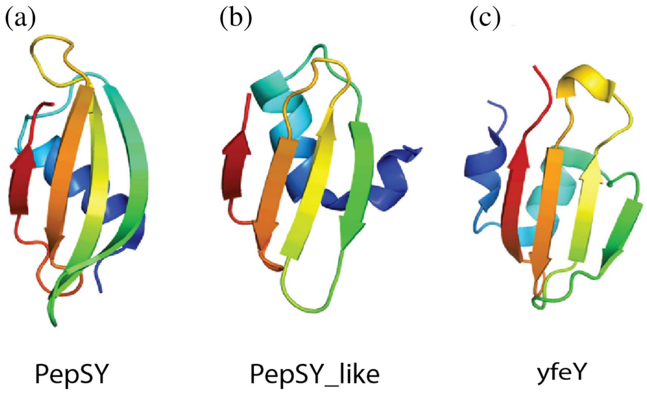

FIGURE 3 Similarities between the BLIP-like and cystatin-like X-groups in ECOD. (a) A PepsSY domain (e2gu3A1) from the cystatin-like X-group. (b) A PepSY-like domain (e3db7A4) from the BLIP-like X-group. (c) A CAP_assoc_N domain (e4ifaA1) from the yfeY-like X-group.

classified in the cystatin-like X-group (e4exrA4; Figure 3b). Sequence families in the BLIP-like X-group such as BLIP (PF07467) and PepSY-like (PF11396) also function as enzyme inhibitors. In fact, all sequence families in the BLIP-like X-group, including SmpA_OmlA (PF04335), DUF3862 (PF12978), DUF4309 (PF14172), BLIP (PF07467), and PepSY-like (PF11396), are classified in the same Pfam clan PepSY (CL0320) that includes other cystatin-like families (Mistry et al. 2021). We also identified strong links through our AFDB classification data between the BLIP-like X-group and the yfeY-like X-group (ECOD X: 6043): an uncharacterized protein from *Staphylococcus aureus* (UNP: Q2FWX2) has significant DPAM hits (with DPAM probabilities >0.9) to ECOD domains in both the BLIP-like X-group (e.g., e1jtgB2) and the yfeY-like X-group (e.g., e4h0aB4). The yfeY-like X-group includes several families such as CAP_assoc_N (PF14505), DUF4309 (PF14172), and DUF1131 (PF06572) that showed significant sequence similarities (with >90% HH probability scores) to BLIP domains by HHpred searches. Based on these similarities, we unified existing sequence families and domains in the BLIP-like and yfeY-like X-groups under the Cystatin/monellin homologous group in ECOD v292.

2.4 | Similarity between domains in eukaryotic initiation factors 1, 5, and 2B

We found a strong link between the eIF1-like H-group (ECOD H: 306.3) and the eIF-5_eIF-2B family (ECOD F: 101.1.28, PFAM: PF01873) in the HTH homologous group (ECOD H: 101.1). The eIF1-like H-group, classified in the Glucose permease domain IIB-like X-group, contains domains from proteins such as eukaryotic translation initiation factor 1 (UNP: P41567), eukaryotic translation initiation factor 2D (UNP: P41214), and mitochondrial large subunit ribosomal protein L49 (UNP: Q13405). Domains from the eIF-5_eIF-2B family, classified in the HTH X-group, are found in eukaryotic translation initiation factor 5 (UNP: P55010) and eukaryotic translation initiation factor 2B (UNP: P20042). These domains share a similar fold with two α-helices and four β-strands in the order of ββαββα (Figure 4a). This fold is related to the ferredoxin-like fold (βαββαβ) by circular permutation. Significant sequence similarities were found by HHpred (Soding et al. 2005) between eIF1 proteins and the eIF-5_eIF-2B family proteins. For example, the human eIF1 protein (UNP: P41567) was found with a probability score of 95% by using an eIF2B domain (e1neeA1) as query, while no hits to classical HTH domains were found. Compared to the eIF1 proteins, the eIF-5_eIF-2B family proteins have two α-helices between the second and third core β-strands (Figure 4b). These two helices have a similar intra-helical angle as the HTH motif, which likely resulted in the HTH classification. Using FoldSeek, we found that the most structurally similar domains to eIF-5_eIF-2B family proteins are domains from eIF1 proteins and not HTH domains (van Kempen et al. 2024). Based on these observations, we moved the eIF-5_eIF-2B F-group from the HTH X-group to the eIF1-like H-group in the Glucose permease domain IIB-like X-group in ECOD v292.

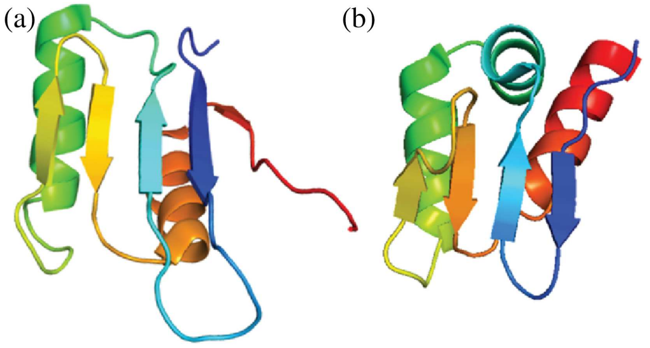

FIGURE 4 Homology between domains in eukaryotic translation initiation factors. (a) Circularly permuted ferredoxin-like fold of eukaryotic translation initiation factor 1 (eIF1, ECOD: e1d1rA1). (b) Structurally similar domain of eukaryotic translation initiation factor 2beta (eIF2B, ECOD: e1neeA1).

2.5 | Common domain topology of outer membrane proteins and type 3 secretion systems

Several families, including BON (PF04972), BON_like (PF21923), CdsD_PD2 (PF22598), Yop-YscD_ppl_1st (PF16693), Yop-YscD_ppl_2nd (PF21937), and Yop-YscD_ppl_3rd (PF21934) were classified in the X-group of amino-terminal domains of OmpATb (ECOD X: 3261). These domains are often found in components of bacterial type III secretion systems. We found links between these families and families from the Alpha-lytic protease prodomain-like X-group (ECOD X:

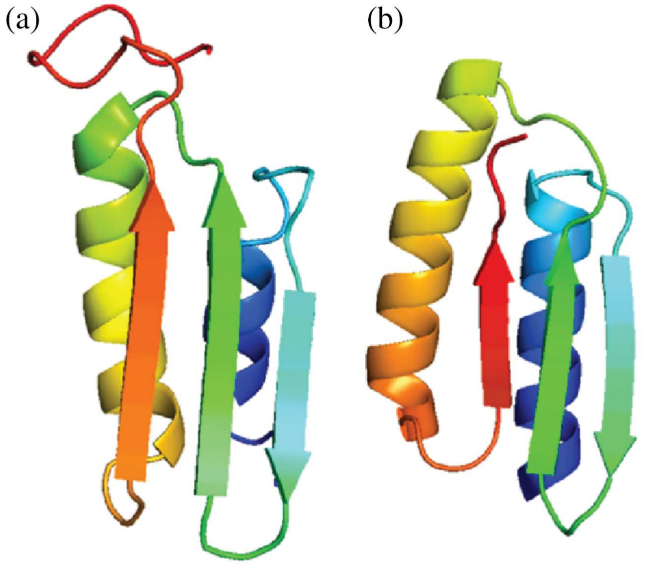

FIGURE 5 Homology between components of type 3 secretion systems and OMP proteins. (a) A domain (e2l26A3) from the BON family in the amino-terminal domain of OmpATb homologous group. (b) A domain (e6rwxU1) from the PrgH in the "Ring-building motif I in type III secretion system" homologous group.

327), including those from the "Ring-building motif I in type III secretion system" homologous group (ECOD H: 327.13), which contains families such as YscJ_FliF (PF01514), PrgH (PF09480), and SpoIIIAH (PF12685), and the "Ring-building motif II in type III secretion system" homologous group (ECOD H: 327.16), which contains families such as Secretin (PF00263) and Secretin_N (PF03958). Domains in these families are structurally similar and adopt the fold of alpha-lytic protease prodomain, consisting of two α-helices and a β-sheet of three β-strands in the order of αββαβ where the two α-helices are packed on the same side of the β-sheet (Figure 5). Due to their sequence and structural similarities and related functions in bacterial secretion systems, we merged these three H-groups as a single H-group "OmpATb and ring-building motifs in type III secretion systems" (ECOD H: 327.22) in the Alpha-lytic protease prodomain-like X-group and removed the amino-terminal domain of OmpATb X-group.

2.6 | Trm112p-like domains are possibly homologous to rubredoxin domains due to their common ligand binding modes

We identified a connection between the Trm112p-like homologous group (ECOD H: 4294.10) and homologous groups belonging to the Rubredoxin-like X-group (ECOD X: 375), which contains sequence families of the zinc beta ribbon fold (Krishna et al. 2003). The Trm112p-like H-group has four families: Trm112p (PF03966), RlmA_N (PF21302), Rieske (PF00355),

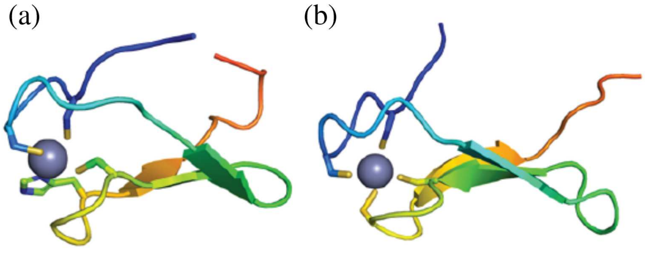

FIGURE 6 Common zinc-binding sites in Trm112p and rubredoxin-like domains support homology between previously distinct groups. (a) A domain (e1p91A1) in the Trm112p-like H-group with bound zinc (gray) in a zinc-ribbon motif. (b) A rubredoxin-like domain (e1pftA1) with canonical zinc ribbon zinc-binding motif.

and Rieske_2 (PF13806). Domains from both the Trim112p (Heurgue-Hamard et al. 2006) and RlmA_N (Das et al. 2004) families are characterized by a zinc beta-ribbon fold and bind zinc. The Trm112p family domains are found in proteins from all domains of life, including eukaryotic proteins that function as activators of methyltransferases involved in protein synthesis and RNA modification (Liger et al. 2011). The RlmA_N domains are principally bacterial and are the N-terminal nucleic acid-binding domains of the 23S rRNA (guanine(745)-N(1))-methyltransferase RlmA from *Escherichia coli* and similar proteins. Domains from the Rieske and Rieske_2 families (Schmidt and Shaw 2001) also adopt the zinc beta ribbon fold but bind Fe-S clusters and are broadly distributed across archaea, eukaryotes, and bacteria. In addition to the initial DPAM results, the possible homology of Trm112p-like families to other zinc beta ribbon families was supported by further HHpred searches. For example, significant hits (HH probability score >95%) to zinc beta ribbon families (e.g., Pfam families A2L_zn_ribbon and YjdM_Zn_Ribbon) were found by using the ECOD domain e2hf1A1 in the Trim112-like H-group as the query (Figure 6). We moved the Trm112p-like H-group to the X-group of Rubredoxin-like and removed the entry of the Trm112p-like X-group.

2.7 | RAP-1 C-terminal domains are assigned to the RuvA-C homologous group

We found homologous links between the RAP1 C-terminal domain (RCT, ECOD: e3k6gA1) in the "Repressor activator protein 1 (RAP1) helical bundle domain" X-group (ECOD X: 3764) and domains (e.g., e1aipD1) from the "RuvA-C, UBA, CRAL/TRIO-N, HBS1" homologous group (ECOD H: 103.1). The RCT X-group has a single family Rap1_C (PF11626) containing C-terminal domains from DNA-binding protein RAP1 (e.g., e4bjtA2). These domains have a common left-handed three alpha-helical bundle topology.

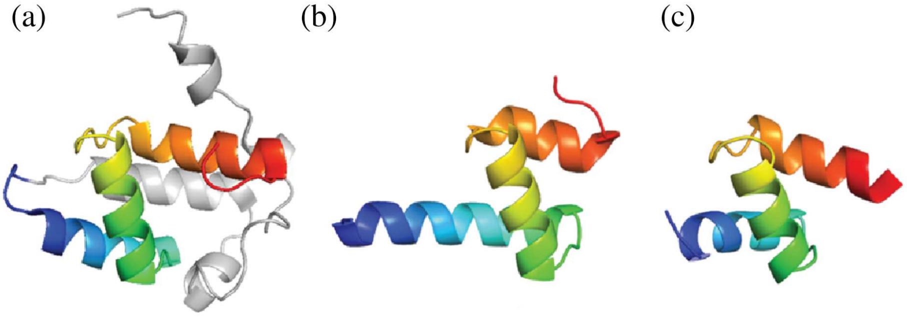

FIGURE 7 Common helical topology among RAP1 C-terminal domains and RuvA helical bundles. (a) RAP-1 C-terminal domain (RCT, e4bjtA2) shares a common helical topology and is homologous to domains from the RuvA-like homologous group. (b) e3k2gA1. (c) A RuvA-like domain from Elongation Factor TS (EF_TS, e1aipD1).

Rap1_C domains also have several N-terminal alpha-helices that are packed against the core of the three helical bundle (Figure 7a). This family was defined based on two separate experimental structures of the RAP1 RCT, one from human (PDB: 4BJT, Figure 7b) and the other from *Saccharomyces cerevisiae* (PDB: 3K6G). These manually curated domains anchored the remaining 19 non-representative domains from this family. The Pfam sequence classification of these domains reveals principally eukaryotic origins. This structural similarity is supported by moderate evidence of homology by HHpred probability. For example, by using the Rap1_C domain e4bjtA1 as the query, a hit to a RuvA-C domain (PDB: 1oai) was found with a probability score of 87.61% (Figure 7c). We merged the domains of the RCT X-group to a new family (Rap1_C) in the "RuvA-C, UBA, CRAL/TRIO-N, HBS1" H-group.

2.8 | Reclassification of a helical domain of MCM4 from histone-like to helix-turn-helix

We found strong links between the winged helix domain of DNA replication licensing factor MCM4 (MCM4_WHD, ECOD: e5v8f45) in the histone-like X-group (ECOD: 148) and a domain (e1cf7A1) in the HTH X-group. The "minichromosome maintenance proteins" (MCM) are required for the initiation of eukaryotic DNA replication and elongation (Georgescu et al. 2017). The C-terminal region of MCM domains contains both a histone-like lid domain in addition to a winged-helix HTH domain (Figure 8a). At the time of classification, this domain lacked a sequence family and was assigned to the histone-like X-group, likely due to boundary contamination from the N-terminal domain leading to misclassification by the automated classifier. Later, the MCM4_WHD sequence family was generated by Pfam, partly informed by ECOD domains lacking a sequence family (such as this MCM4 domain) and was classified into their HTH clan and identified as a winged helix HTH domain. Subsequently, these (and other) domains were automatically assigned to the MCM4_WHD family in the histone-related homologous group. However, it adopts the fold of a winged HTH that contains a three-helical bundle with a C-terminal beta-hairpin (i.e., the wing). The link of this domain to HTH domains is also supported by strong HHpred scores to HTH domains instead of domains with the histone-like fold. The misclassification of this domain in the Histone-like X-group is probably due to its presence in the C-terminus of AAA+ ATPase subunit, since many C-terminal domains of AAA+ ATPases adopt the histone-like fold and are classified in the Histone-like X-group. Manual inspection also revealed that domains from the family of MCM5_C (Figure 8b) in the T-group of AAA+ ATPase lid domain in the H-group of Histone-related are in fact HTH domains and do not have a histone-like fold (Figure 8c). We thus moved the two families MCM4_WHD and MCM5_C from the X-group of Histone-like to the X-group of HTH. Domain classifications are based on incomplete information, in this case, the accumulation of additional data, changes to our sequence family classification, and the consideration of mass prediction data allowed us to identify a mistake in a previously small family of domains. The winged-helix domain in the MCM proteins alter configurations upon ssDNA binding, leading to a series of ordered and disordered structures in these regions. These types of dynamics are still challenging for structural domain classifications to model and represent correctly and are an area of potential future development.

2.9 | Generation of a unified CHCH/CX9X homologous group

The conservation of disulfide bonds can be a strong homologous signal, even in the absence of other sequence conservation. Coiled-coil domains are difficult to classify because of their compositional bias, as well as the difficulty in constraining the domain boundaries (Mistry et al. 2013). CHCH domains are disulfide-bonded coiled-coil domains classified principally based on their function and disulfide-bonding patterns. Strong sequence homology between AFDB domains to ECOD reference domains in multiple X-groups unified CHCH domains into a single homologous group.

We found strong similarities between a domain (e5xtiAR1) from the "Cytochrome bc1 complex 11 kDa

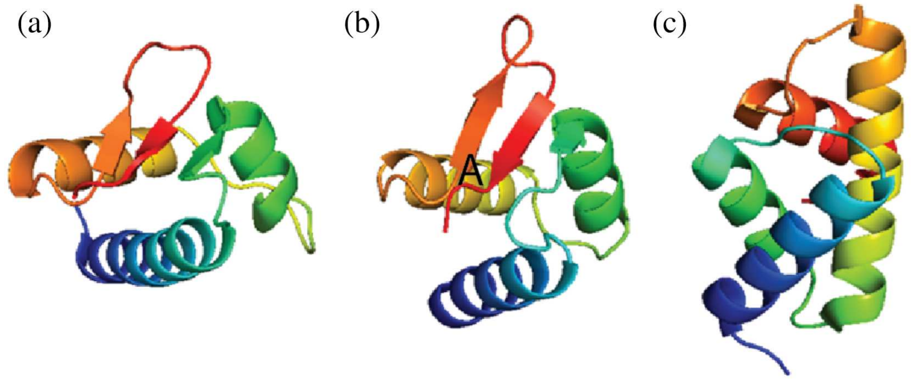

FIGURE 8 MCM4_WHD and MCM5_C are families of winged HTH domains involved in DNA replicase activity and do not share a common topology with AAA+ helicase lid domains with a histone-like fold. (a) A winged helix HTH (e5v8f45) classified the MCM4_WHD family with the characteristic beta-sheet "wing" C-terminal to the helix-turn-helix motif. (b) AlphaFold model of the MCM5_C domain from the human MCM5 protein (UNP: P33992). (c). An AAA_lid_4 domain (e1in4A2) with a histone-like fold in the histone-like X-group, lacking the wing.

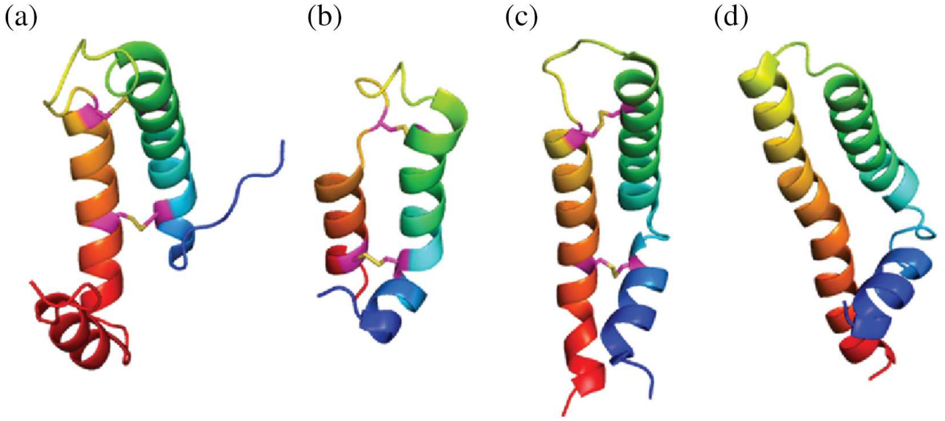

FIGURE 9 Common topologies, conserved disulfide bond patterns, and functions among multiple mitochondrial cytochromes establish a common homologous group of CX9C-related domains. (a) A CX9C domain (e4ytwA1) from the "Mitochondrial distribution and morphology protein 35" X-group. (b) A CX9C domain (e2lqlA2) from the "p8-MTCP1-related" X-group. (c) A CX9C domain (e1ppjH1) from the "Cytochrome bc1 complex 11 kDA protein-like" X-group. (d) A DUF465 domain (e1zhcA1) from the "Cytochrome bc1 complex 11 kDA protein-like" X-group.

protein-like" X-group (ECOD X: 5071; Figure 9a) and several domains from the "Mitochondrial distribution and morphology protein 35" X-group (ECOD X: 3981; Figure 9b) and a domain (e2lqlA2) from the of "p8-MTCP1-related" X-group (ECOD X: 568; Figure 9c). Manual examination of these domains suggests that they are evolutionarily related and belong to the CX9C superfamily (Longen et al. 2009) characterized by two disulfide bonds formed between two alpha-helices.

The "Cytochrome bc1 complex 11 kDa protein-like" X-group (ECOD X: 5071; Figure 9c) was composed of five representative domains spanning three sequence families, each occupying a separate homologous group. These three families included the Ubiquinol-cytochrome C reductase hinge protein (UCR_hinge, PF02320) domains (Iwata et al. 1998), the NDUFS5 domains (PF10200), and the DUF465 domains (PF04325). Each domain was an alpha bundle, and the UCR_hinge and NDUFS5 domains are characterized by conserved C9XC motifs. The DUF465 domains lacked these disulfide bonds and did not share the predominance of participation in mitochondrial electron transport chains. The UCR_hinge ECOD family contains 266 domains from both experimental and predicted structures. These domains are often found in single-domain proteins and are sometimes C-terminal to an intrinsically disordered region. They commonly mediate electron transport between cytochrome c1 to cytochrome c. This domain contains a pair of disulfide bonds separated by 9 residues (i.e., a C9XC motif). The NDUFS5 family contains 182 domains: 166 non-representative experimental domains and 16 non-representative domains from AFDB predictions. It is represented by a single domain (ECOD: e5lnkl1) from a cryoEM structure of the ovine respiratory complex (PDB: 5lnk). This domain is 100 residues long and is a component of the NADH:ubiquinone oxidoreductase complex I (Loeffen et al. 1999). The precise function of this subunit is unknown. Members of the associated Pfam sequence family (PF10200) are included in the "CH domain" Pfam clan (CL0351). The DUF465 family contains 182 domains, principally from predicted structures. DUF465 is represented by a single experimental structure of a *Helicobacter pylori* hypothetical protein (HP1242) with unknown function. The Pfam DUF family containing this sequence (DUF465) belongs to no clan and also reports no known function, although it is noted this domain is commonly found C-terminal to kinesin domains. There are no indications that these domains participate as components of the electron transport chain. The original manual classification of this ECOD domain was likely based on structural similarity to other alpha hairpins in this X-group.

These proteins are mostly found in mitochondria and are involved in various protein–protein interactions. The p8-MTCP1-related X-group contains the greatest number of families of the CX9C superfamily. The p8-MTCP1-related X-group (ECOD X: 568) contains a single homologous group comprising 11 associated sequence families. These families include the CHCH (PF06747), CX9C (PF16860), MTCP1 (PF08991), COX17 (PF05051), and GCK domains (PF07802). Many of the sequence families in this group are characterized by two coiled-coil domains bound by two pairs of conserved cysteines and with proposed function in the electron transport chain: 8 of 11 belong to the Pfam "CHCH" clan (CL0351). We thus merged these X-groups into a single H-group (ECOD H: 568, "CHCH/CX9C-like domains") for all families with the signature cysteines. One exception is the DUF465 homologous group (ECOD H: 5071.3) from the original "Cytochrome bc1 complex 11 kDa protein-like" X-group. This H-group contains a single family DUF465 with a single ECOD domain (ECOD: e1zhcA1). Proteins in the DUF465 family do not have the signature cysteines found in CX9C proteins (Figure 9d). They adopt a fold consisting of three alpha-helices similar to CX9C domains in the "Cytochrome bc1 complex 11 kDa protein-like" X-group. Consequently, DUF465 is retained as a separate homologous group.

3 | CONCLUSIONS

The deployment of accurate structure prediction on a large scale has resulted in a bewildering surge of new structural data. Although the principal focus has rightly been the analysis of the domains and quality of these predictions, what we have attempted here is to show how this diverse data can be used to identify and address inconsistencies in our previous classification of experimental data in groups of domains that might have been undersampled or otherwise able to avoid close scrutiny. Domain classifications rely on a series of decisions based on incomplete data using methods which change over time. As such, periodic checks of consistency and revisiting prior curation decisions are a necessary component of classification maintenance. Here we have attempted to show how ambiguous reference domains can be more clearly addressed using large datasets based on structure prediction results. Ideally, the convergence of structure and sequence classifications in the future will allow us to shift resources from classifying new proteins to reconciling inconsistencies and determining methods and schemas for classifying even more distant homologous domain relationships.

4 | METHODS

4.1 | DPAM assignment of predicted structures of proteins in the AFDB

We have previously described the domain classification of proteins from the 48 whole proteomes deposited in the AFDB (Schaeffer et al. 2024a). Briefly, these proteins were partitioned using a combination of secondary structure measures, interresidue distances, measures of interdomain prediction confidence (predicted aligned error or PAE), homology to reference domains by HHsearch profile-profile hits, and structure similarity by DALI. Putative partitioned domains were assigned to an ECOD reference domain by a neural network trained against a reference set of ECOD structural domains (ECOD v285).

DPAM domains are assigned status based on their alignment and score parameters. Well-assigned domains have multiple secondary structure elements, strong DPAM probability to their assigned ECOD hierarchical group, and the alignments (DALI or HHsearch) used to generate the database hits cover the majority of the reference domain (i.e., the putative domain is not a fragment or partial domain compared to the reference domain). For this analysis, we considered only those well-assigned domains. DPAM generates both HHsearch and DALI scores where possible during the generation of domain boundaries and assignments. Both scores were considered along with the overall DPAM probability. In total, we considered over 213M individual domain-domain comparisons between AFDB domains and ECOD reference domains. 53,225 AFDB domains were found to have possible homologous links to multiple ECOD homologous groups. The full list of these domains is presented in Table S2 and is the basis for which domains were chosen for manual curation.

Following initial data generation, DPAM intermediate files were loaded to a PostgreSQL database for exploratory data analysis. R/Rstudio was used for data analysis, graphs and plots were generated using ggplot2. Protein structure images were generated using PyMol with cartoon representations of ECOD domain PDB files retrieved from ECOD PDB and AFDB (http://prodata.swmed.edu/).

AUTHOR CONTRIBUTIONS

Richard Dustin Schaeffer: Conceptualization; writing – original draft; writing – review and editing; methodology; software; data curation; investigation; funding acquisition; supervision. **Jimin Pei:** Writing – original draft; writing – review and editing; investigation; methodology. **Jing Zhang:** Methodology; software; data curation. **Qian Cong:** Data curation; software; methodology. **Nick V. Grishin:** Funding acquisition; supervision; conceptualization; project administration.

ACKNOWLEDGMENTS

The study is supported by grants from the National Institute of General Medical Sciences of the National Institutes of Health GM147367 (to R.D.S.) and GM127390 (to N.V.G.), the Welch Foundation I-1505 (to N.V.G.) and I-2095-20220331 (to Q.C.), the National Science Foundation DBI 2224128 (to N.V.G.). We acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources. 1K99AI180984-01 from the National Institute of Allergy and Infectious Disease (to J.Z.). We acknowledge DeltaGPU resources hosted by National Center For Supercomputing Applications at Illinois. The allocation was awarded by NSF ACCESS program through MED240004 and MED230034.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in ECOD at http://prodata.swmed.edu/ecod, reference number 22.

ORCID

Richard Dustin Schaeffer https://orcid.org/0000-0001-6502-1425
Jimin Pei https://orcid.org/0000-0002-3505-9665
Jing Zhang https://orcid.org/0000-0003-4190-3065
Qian Cong https://orcid.org/0000-0002-8909-0414

REFERENCES

Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 2020;48(D1):D376–82.

Cheng H, Liao Y, Schaeffer RD, Grishin NV. Manual classification strategies in the ECOD database. Proteins. 2015;83(7):1238–51.

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(12):e1003926.

Das K, Acton T, Chiang Y, Shih L, Arnold E, Montelione GT. Crystal structure of RlmAI: implications for understanding the 23S rRNA G745/G748-methylation at the macrolide antibiotic-binding site. Proc Natl Acad Sci U S A. 2004;101(12):4041–6.

Donald JE, Shakhnovich EI. Determining functional specificity from protein sequences. Bioinformatics. 2005;21(11):2629–35.

Dubin G. Proteinaceous cysteine protease inhibitors. Cell Mol Life Sci. 2005;62(6):653–69.

Eissenberg JC. Structural biology of the chromodomain: form and function. Gene. 2012;496(2):69–78.

Georgescu R, Yuan Z, Bai L, de Luna Almeida Santos R, Sun J, Zhang D, et al. Structure of eukaryotic CMG helicase at a replication fork and implications to replisome architecture and origin initiation. Proc Natl Acad Sci U S A. 2017;114(5):E697–706.

Grzonka Z, Jankowska E, Kasprzykowski F, Kasprzykowska R, Lankiewicz L, Wiczk W, et al. Structural studies of cysteine proteases and their inhibitors. Acta Biochim Pol. 2001;48(1):1–20.

Heurgue-Hamard V, Graille M, Scrima N, Ulryck N, Champ S, van Tilbeurgh H, et al. The zinc finger protein Ynr046w is plurifunctional and a component of the eRF1 methyltransferase in yeast. J Biol Chem. 2006;281(47):36140–8.

Holm L. Benchmarking fold detection by DaliLite v.5. Bioinformatics. 2019;35(24):5326–7.

Iwata S, Lee JW, Okada K, Lee JK, Iwata M, Rasmussen B, et al. Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. Science. 1998;281(5373):64–71.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.

Kaur G, Iyer LM, Subramanian S, Aravind L. Evolutionary convergence and divergence in archaeal chromosomal proteins and Chromo-like domains from bacteria and eukaryotes. Sci Rep. 2018;8(1):6196.

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(6):e0079623.

Krishna SS, Majumdar I, Grishin NV. Structural classification of zinc fingers: survey and summary. Nucleic Acids Res. 2003;31(2):532–50.

Liger D, Mora L, Lazar N, Figaro S, Henri J, Scrima N, et al. Mechanism of activation of methyltransferases involved in translation by the Trm112 "hub" protein. Nucleic Acids Res. 2011;39(14):6249–59.

Loeffen J, Smeets R, Smeitink J, Triepels R, Sengers R, Trijbels F, et al. The human NADH: ubiquinone oxidoreductase NDUFS5 (15 kDa) subunit: cDNA cloning, chromosomal localization, tissue distribution and the absence of mutations in isolated complex I-deficient patients. J Inherit Metab Dis. 1999;22(1):19–28.

Longen S, Bien M, Bihlmaier K, Kloeppel C, Kauff F, Hammermeister M, et al. Systematic analysis of the twin cx(9)c protein family. J Mol Biol. 2009;393(2):356–68.

Medvedev KE, Kinch LN, Dustin Schaeffer R, Pei J, Grishin NV. A fifth of the protein world: Rossmann-like proteins as an evolutionarily successful structural unit. J Mol Biol. 2021;433(4):166788.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41(12):e121.

Quimby BB, Leung SW, Bayliss R, Harreman MT, Thirumala G, Stewart M, et al. Functional analysis of the hydrophobic patch on nuclear transport factor 2 involved in interactions with the nuclear pore in vivo. J Biol Chem. 2001;276(42):38820–9.

Sadreyev RI, Kim BH, Grishin NV. Discrete-continuous duality of protein structure space. Curr Opin Struct Biol. 2009;19(3):321–8.

Schaeffer RD, Kinch LN, Liao Y, Grishin NV. Classification of proteins with shared motifs and internal repeats in the ECOD database. Protein Sci. 2016;25(7):1188–203.

Schaeffer RD, Kinch LN, Pei J, Medvedev KE, Grishin NV. Completeness and consistency in structural domain classifications. ACS Omega. 2021;6(24):15698–707.

Schaeffer RD, Liao Y, Cheng H, Grishin NV. ECOD: new developments in the evolutionary classification of domains. Nucleic Acids Res. 2017;45(D1):D296–302.

Schaeffer RD, Medvedev KE, Andreeva A, Chuguransky SR, Pinto BL, Zhang J, et al. ECOD: integrating classifications of protein domains from experimental and predicted structures. Nucleic Acids Res. 2024b;53:D411–8.

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(12):e2214069120.

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. 2024a;20(2):e1011586.

Schmidt CL, Shaw L. A comprehensive phylogenetic analysis of Rieske and Rieske-type iron-sulfur proteins. J Bioenerg Biomembr. 2001;33(1):9–26.

Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005;33:W244–8.

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019;20(1):473.

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590–6.

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(2):243–6.

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(D1):D439–44.

Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 2024;52(D1):D368–75.

Zhang J, Schaeffer RD, Durham J, Cong Q, Grishin NV, et al. DPAM: a domain parser for AlphaFold models. Protein Sci. 2022;32: e4548.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Schaeffer RD, Pei J, Zhang J, Cong Q, Grishin NV. Refinement and curation of homologous groups facilitated by structure prediction. Protein Science. 2025;34(3): e70074. https://doi.org/10.1002/pro.70074