Proteins: Structure, Function, and Bioinformatics

WILEY

PROTEINS

RESEARCH ARTICLE OPEN ACCESS

Case Studies of Orphan Domain Reclassification in ECOD by Expert Curation

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

[1]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas, USA | [2]Department of Biophysics, 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: Qian Cong (qian.cong@utsouthwestern.edu) | Nick V. Grishin (grishin@chop.swmed.edu)

Received: 22 January 2025 | **Revised:** 22 April 2025 | **Accepted:** 5 May 2025

Funding: This work was supported by National Science Foundation (DBI 2224128), Welch Foundation (I-2095-20220331 and I-1505), and National Institutes of Health (GM127390 and GM147367).

Keywords: ECOD | isolated domain families | protein classification | protein evolution | remote homology

ABSTRACT

Homology-based protein domain classification is a powerful tool for gaining biological insights into protein function. This classification process has been significantly enhanced by the availability of experimental structures and high-accuracy structural models generated by advanced tools such as AlphaFold. Our Evolutionary Classification of protein Domains (ECOD) database provides a continuously updated and refined domain classification system. Isolated ("orphan") protein domain families, which have a limited distribution in the protein universe, present a unique challenge in this classification process. These families lack clear or identifiable evolutionary relationships with other sequence families. While some isolated domain families may have emerged through de novo evolution, others potentially share common evolutionary origins with existing domain families but represent difficult cases for traditional classification methods. In this study, we conducted a manual analysis of a set of isolated families of small domains in ECOD. By exploring sequence, structural, and functional evidence, we uncovered distant members and likely homologous relationships between different isolated domain families that were previously unrecognized. Our analysis provides valuable insights into the evolution of isolated domain families and has led to improved classification within ECOD. This work enhances our understanding of protein evolution and underscores the importance of continuous refinement in domain classification systems as new data and analytical methods become available.

1 | Introduction

Isolated protein domain families ("orphan" families) do not have clear or identifiable evolutionary relationships with other sequence families. These domain families are intriguing: how do the evolutionary origins of a domain family become obscured? Their origins can be explained by one of two possibilities: (1) independent evolution or (2) distant homology. These two pathways offer different perspectives on how proteins might acquire unique structures and functions. Independent evolution suggests that some protein domains arise without any discernible ancestral domain, essentially evolving de novo. In this case, these proteins are true evolutionary innovations, emerging from non-coding regions of the genome or previously unstructured regions within existing proteins [1–4].

Alternatively, isolated domain families may be homologous to other domain families, but those relationships have been obscured over time due to changes in structure and sequence [3, 5–7]. As evolutionary time passes, the sequence identity between homologous proteins can degrade due to mutations, insertions, deletions, and other genomic events.

These changes can diminish the recognizable sequence or structural features that link them to their ancestral domain family. Isolated domain families can be created through gene duplication followed by divergence [8]. This process allows for proteins to evolve new functions while preserving the gene's original function. Duplicated copies of a gene can not only perform the same function as the original gene but also acquire new functions by accelerated evolution of the new copy [9]. Domain duplication can lead to structural divergence or the emergence of new binding properties, giving rise to proteins with novel functions that distinguish them from their ancestral family. This process can lead to functional diversification within an organism, where the duplicated domains contribute to novel metabolic pathways, regulatory mechanisms, or environmental adaptations [10]. Over long evolutionary timescales or through elevated evolutionary rates, this divergence can obscure the fact that the two domain families are evolutionarily related.

Modern computational methods such as sequence alignment and structural comparison often struggle to detect distant relationships, leaving some protein families seemingly isolated. However, as search algorithms advance and with careful manual analysis of weak sequence and structural signals, these domains may eventually be linked to their evolutionary origins, uncovering hidden kinships [11–14]. This notion of distant, unobserved homology highlights how extensive evolutionary divergence can mask the shared ancestry of proteins that initially appear unrelated.

Both possibilities—independent evolution and distant homology—are important for understanding protein evolution. Independent evolution expands the repertoire of molecular functions by creating entirely new protein structures. Distant homology underscores the vast timescales over which evolution operates, sometimes obscuring connections that were once more apparent. In both cases, isolated protein domains challenge our current understanding of how life evolves at the molecular level, offering fascinating glimpses into the adaptability and creativity of nature's evolutionary processes.

Our domain classification, the Evolutionary Classification of protein Domains (ECOD) [15], was designed specifically with these two possibilities in mind. The hierarchy of ECOD is designed such that similar topology, which can be indicative of homology, is not definitively suggestive of homology. Two of ECOD's top hierarchical levels, the “possible homology” group (X-group) and “definite homology” group (H-group) allow for both distant relationships between domain families and the possibility of independent evolution of topologically similar domains (i.e., convergent evolution). Because the protein universe is not fully structurally characterized and new structural comparison methods are continually being developed, we must allow for isolated domain families with no apparent homology, and that they might be merged with existing homologous groups in the future. The recent development and deployment of highly accurate structural prediction methods [16, 17] has resulted in a wealth of new data, such as those in the AlphaFold database. By leveraging comparative analyses of these predictions, we have been able to revisit and improve the classification of isolated domain families in ECOD.

ECOD X-groups and H-groups exhibit significant variation in their population sizes. Certain domain families, particularly those associated with superfolds [18], are exceptionally populous, encompassing hundreds to thousands of individual domains. In contrast, many other domain families contain only a few structures and form distinct X-groups or H-groups. Identifying evolutionary relationships for these isolated domain families presents a challenge. However, careful analysis of subtle sequence and structural similarities may help uncover distant members of these domain families and establish potential connections with more populous domain families. Small domains are likely candidates for isolated domain families due to the ease with which they are recruited from disordered regions and their limited sequence information and structural constraints. Here, we present an analysis and reclassification of four groups of small domain families in ECOD that unify domains previously thought to be isolated. One of the domain groups involves an isolated small zinc finger domain originally found in bacterial ADA proteins. Distant members of this domain family were detected in a variety of proteins with diverse domain contents. Three cases of domain groups involve small helical domains and represent difficult cases in homology detection due to short domain lengths. Our analysis provides insight into the evolution of a class of LigA-related domains found in the precursors and enzymes of ribosomally synthesized and post-translationally modified peptides, the expansion of the domain group likely related to the HMG-box domains, and the union of a group of STI1-HOP_DP-like domains. The curation performed herein will aid us in our further classification of predicted structures and their domains.

2 | Results and Discussion

2.1 | Strategies for Homology Inference of Isolated Domains in the AlphaFold Era

The ECOD database organizes protein domains into a hierarchical structure that reflects evolutionary relationships, distinguishing between definite homology (H-groups) and possible homology (X-groups). A substantial portion of ECOD consists of small or isolated domain families—so-called “orphan” domains—found in X-groups or H-groups with only a few members. These isolated domains lack clear evolutionary links to other classified domains and often represent some of the most challenging and ambiguous cases in protein classification. The increasing availability of high-confidence predicted structures from AlphaFold, along with more sensitive sequence and structure comparison tools, offers an opportunity to revisit these isolated domains and probe for overlooked relationships.

To uncover remote homologies among these orphan domains, we applied an integrative strategy that combines information from multiple sources: sequence similarity, structural similarity, conserved sequence and structural motifs, and functional associations. Starting with a representative ECOD domain, we performed sensitive sequence-based searches using HHpred [19] against the PDB database [20], the Pfam database [21], and selected organismal proteomes. An HHpred probability score above 95% was used as a criterion for a highly likely homology relationship [19], and any hits with probability scores above 30% were also inspected [19]. We evaluated candidate relationships

based on alignment statistics, the presence of conserved motifs and functional context, such as domain architecture or common functions in related pathways. Transitive searches were performed by initiating new HHpred searches with candidate hits as queries, which could expand homology to other domain families beyond the initial search results or increase the confidence of original weak hits. For example, if a distant domain (A) weakly matches the original query (Q), but both A and Q strongly match a shared intermediate (B), this transitive connection strengthens the evidence for a relationship between Q and A.

We complemented this sequence-based analysis with structure-based searches using DaliLite [22] and Foldseek [23], which were particularly useful for divergent domains where sequence similarity was limited. Experimental structures or AlphaFold models of found hits were used as input to these structural searches. AlphaFold models aid in distant homology detection by making high-quality structural information available for protein domain families that lack experimental structures. AlphaFold models played a central role not only in enabling reliable structure comparisons but also in validating the presence of conserved core regions or structural motifs and clarifying domain boundaries. DaliLite was used to compare either experimental structures or AlphaFold models against the PDB database and to perform pairwise structural alignments to evaluate potential homology. Structural similarity was assessed using Dali Z-scores, with scores above 2 considered indicative of significant fold-level similarity [24]. Foldseek offers a much faster alternative for structure-based searches, enabling rapid comparisons of query structures against large repositories of experimental structures and AlphaFold models. Its speed and scalability make it particularly useful for exploring the distribution of a domain across the protein universe, including its presence in diverse proteins and taxonomic lineages.

When strong evidence from these sources converged, we assigned domains to the same H-group; when the evidence was suggestive but less conclusive, we unified them at the X-group level. This integrative approach, combining complementary data types, provides a framework for refining domain relationships in ECOD and better capturing the evolutionary connections among previously isolated domain families. Below, we present four cases of distant homology inference involving isolated domains. In each case, sensitive sequence and structural comparisons allowed us to establish homology relationships between distantly related domains. The use of AlphaFold models further enabled the expansion of ECOD classification to include proteins and domain families lacking experimental structures.

2.2 | Remote Homologs of the N-Terminal Zinc Finger Domain of the DNA Repair Protein ADA

ADA (alkyladenine DNA glycosylase) is a bacterial DNA-binding protein involved in the repair of alkylated DNA damage [25]. It plays a key role in the adaptive response to alkylating agents, which modify the DNA by adding alkyl groups to nucleotides, leading to mutagenesis or cell death if not repaired [25]. ADA recognizes and binds to alkylated DNA and removes alkyl groups from specific damaged bases such as methylated guanine (O^6-methylguanine) and methylated thymine (O^4-methylthymine). ADA also functions as a transcriptional activator in the adaptive response to alkylating agents [26]. Once activated by alkylation, the ADA protein induces the expression of genes involved in DNA repair to cope with alkylating damage.

The ADA protein consists of four globular domains. The N-terminal domain (Pfam: Ada_Zn_binding; PF02805) has four conserved cysteines forming a zinc-binding site. One of these cysteines (Cys69 in *Escherichia coli* ADA) is also a methyl group acceptor [27]. The N-terminal domain undergoes a conformational change upon methylation of this cysteine, which enables it to bind specifically to the promoter regions of genes involved in the response to alkylation damage. This binding activates transcription through interactions with the σ^{70} subunit of RNA polymerase, leading to an increase in the production of Ada protein.

Experimental structures revealed (e.g., PDB: 1u8b) [28] that the domain following the N-terminal zinc-binding domain of ADA is an HTH (helix-turn-helix) domain. We found that several experimental ADA structures include the N-terminal zinc-binding domain and part of the HTH domain that is disordered (e.g., PDB: 1eyf, the partial HTH domain shown in gray in Figure 1A). ECOD previously assigned the full structures as the ADA N-terminal domains. Such assignments were modified in the new ECOD classification by separating the ADA N-terminal domain from the disordered partial HTH domain. The ADA N-terminal domain features a four-stranded β-sheet surrounded by several short α-helices. Two zinc-binding cysteines are from the loop after the second core β-strand. The short C-terminal α-helix contains two other zinc-binding cysteines. The positioning of this α-helix is similar to that found in Treble clef zinc fingers [29]. However, unlike most Treble clef zinc fingers, the ADA N-terminal domain lacks a β-hairpin between the two "CXXC" motifs.

We conducted a sequence similarity search of the ADA N-terminal domain and found an HHpred hit with moderate probability score (82.4%) to the C-terminal domain of the 50S ribosomal protein L4 from *Cutibacterium acnes* (RplD, UniProt accession (UNP): Q6A6M7) in the PDB database (PDB: 8cvm, chain e). The HHpred alignment covers most of the query with a sequence identity of 21%. While this domain is disordered in the CryoEM structure, its AlphaFold model (Figure 1B) exhibits the same fold as the ADA N-terminal domain. A DaliLite comparison between the ADA N-terminal domain (PDB: 1eyf) and the C-terminal domain of RplD yields a *Z*-score of 4.1 over 49 aligned residues. HHpred searches against bacterial genomes revealed that domains (called ADA_N-like) showing remote homology to the ADA N-terminal domain are also present in other Gram-positive bacteria. AlphaFold models of two such domains are shown in Figure 1C (C-terminal domain of Cwp27, with an HHpred probability score of 98.1% and a Dali *Z*-score of 3.9 to the ADA N-terminal domain) and Figure 1D (C-terminal domain of ArfC, with an HHpred probability score of 94.1% and a Dali *Z*-score of 3.7 to the ADA N-terminal domain). Two proteins with this domain were found in *Mycobacterium tuberculosis*. One is the uncharacterized membrane protein ArfC (UNP: A1KH33) with an N-terminal transmembrane helix. ADA-N-like domains can co-occur with other domains, such as LGFP repeat (e.g., the *M. tuberculosis* protein UNP: O07219),

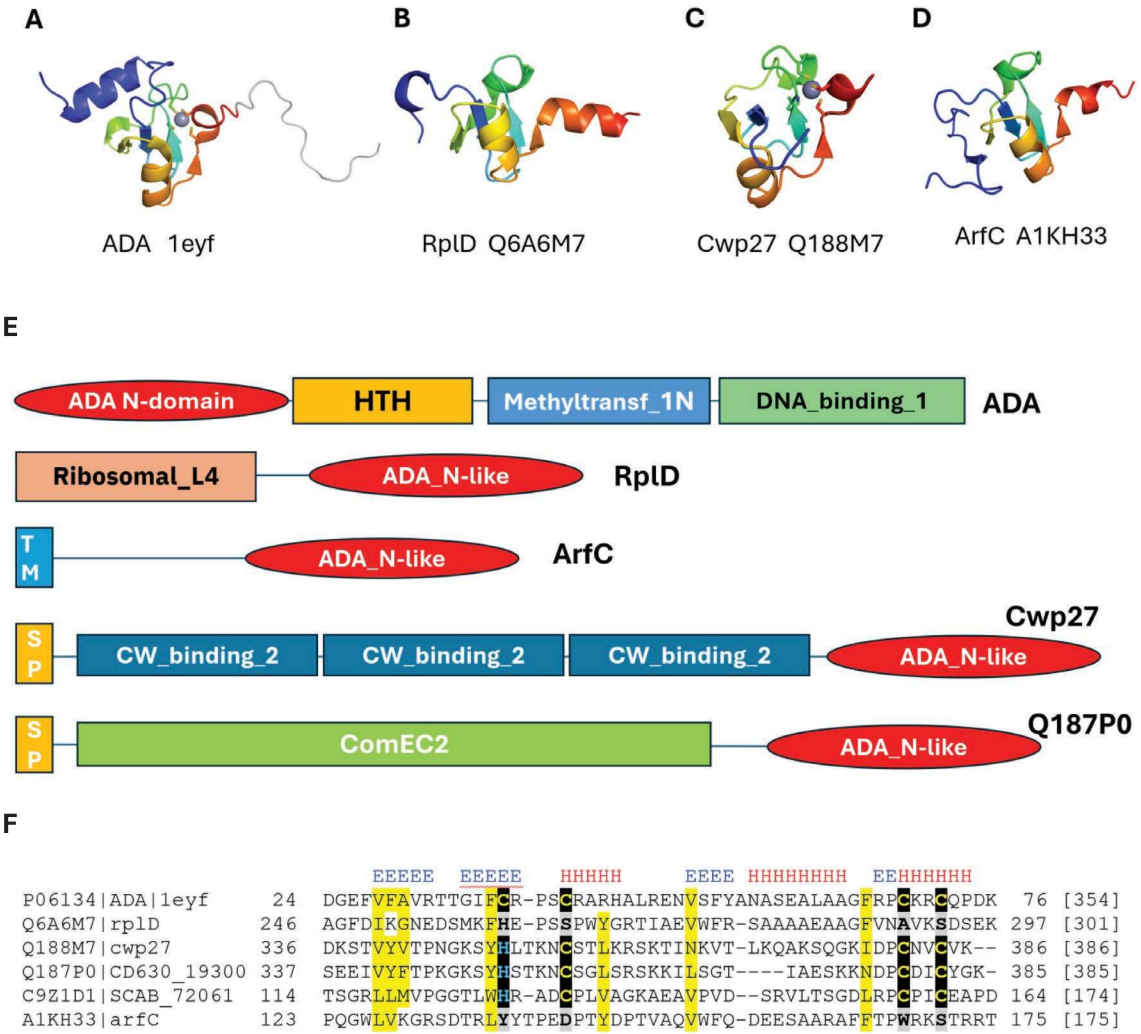


FIGURE 1 | Structure, domain architecture, and alignment of ADA N-terminal domain and ADA_N-like domains. (A) The structure of *Escherichia coli* ADA (PDB: 1eyf). The ADA N-terminal domain is colored in rainbow. The disordered partial HTH domain is colored gray. (B) AlphaFold model of the RplD C-terminal domain. (C) AlphaFold model of the C-terminal domain of Cwp27. (D) AlphaFold model of the ArfC C-terminal domain. (E) Domain architecture of selected proteins that contain the ADA_N-like zinc finger domains. Boxes labeled "TM" and "SP" indicate predicted transmembrane segment and signal peptide, respectively. (F) Multiple sequence alignment of ADA_N-like domains in different proteins. Helices (red) and strands (blue) are marked (H: helix; E: strand). Conserved hydrophobic residues are shaded in yellow. Zinc-binding residues are shaded in black. The start and end positions of the domains are shown, and the protein lengths are shown in brackets. Proteins are identified by their UniProt accession numbers followed by gene or genomic location names.

ComEC (e.g., a ComEC/Rec2 family competence protein from *Clostridioides difficile*, UNP: Q187P0) and CW_binding_2 (e.g., UNP: Q188M7 from *C. difficile*) (Figure 1E). Both *C. difficile* proteins with the ADA_N-like domains possess predicted signal peptides (Figure 1E), suggesting that they are secreted. While this domain lies in the N-terminus of ADA, it is mostly located at the C-terminal ends in other proteins (Figure 1E). Some of these domains, such as those in ribosomal protein L4 and ArfC, have deteriorated zinc-binding sites (Figure 1B,D,F). The functions of ADA_N-like domains remain to be experimentally determined. Like their counterpart in ADA, they could function as sensors of alkylated DNA.

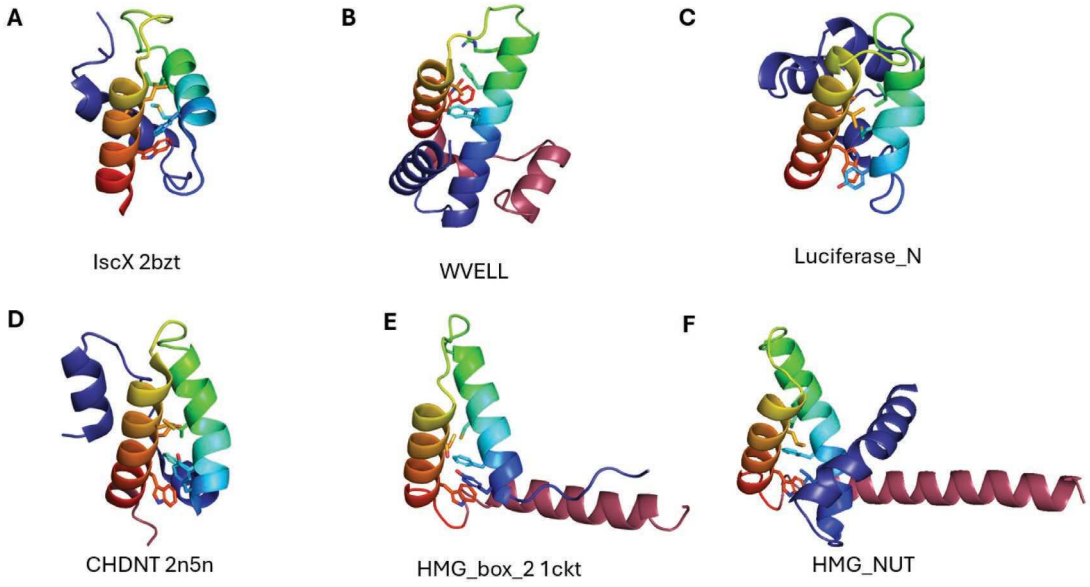
In the updated ECOD classification, we expanded the original X-group of "ADA DNA repair protein, N-terminal domain (N-ada 10)" to include newly identified ADA_N-like domains, including the C-terminal domain of ribosomal protein RplD and four domains from *M. tuberculosis* and *C. difficile* (Figure S1). To reflect this broader classification, we renamed the X-group to "ADA_N-like domains." These newly added domains are considered well-supported homologs based on strong sequence and structural similarities, and the conservation of zinc-binding sites in some cases. We also revised the domain boundaries of the ECOD representative e1eyfA1 from residues 1–90 to 1–76, excluding the partial sequence of the adjacent HTH domain.

2.3 | Expansion of the HMG-Box X-Group

IscX is a small protein involved in the assembly of iron–sulfur (Fe-S) clusters, a process essential for various metabolic activities in Gram-negative bacteria such as *E. coli* [25]. It acts as a molecular adaptor that binds to the cysteine desulfurase IscS, which plays a central role in Fe-S cluster assembly. IscX domains (Pfam:

Fe-S_assembly; PF04384) adopt the fold of a three-helical bundle and were classified in its own H-group within the HTH X-group. HHpred search results suggested that IscX could be remotely related to the WVELL family of proteins from Gram-positive bacteria, such as the YfhJ protein from *Bacillus subtilis* (HHpred probability score: 82.4%, sequence identity: 13%, alignment coverage: > 90% of the query). This family of proteins is characterized by a conserved "WVELL" motif (Pfam: WVELL; PF14043). Interestingly, *E. coli* IscX was previously named YfhJ [30]. The similarity between *E. coli* IscX and *B. subtilis* YfhJ occurs mostly in the helical hairpin formed by the last two α-helices of IscX (Figure 2), where they adopt similar patterns of hydrophobic positions, including a conserved tryptophan in the last α-helix (Figure 2). However, the overall structural similarity between *E. coli* IscX and *B. subtilis* YfhJ is low (Dali *Z*-score below 2).

By examining the results of transitive HHpred searches, we found that this helical hairpin motif is also present in the N-terminal domains of luciferases (Pfam: Luciferase_N, PF05295) and several families of the HMG-box (High Mobility Group box) domain. The Luciferase_N family was found to be a weak hit (HHpred probability score: 58.4%) by using IscX as the query. The Luciferase_N hit has a low sequence identity (12%) and covers the two core helical segments (residues 30–60) of the query IscX with the conserved tryptophan aligned. By using a domain from the Luciferase_N family as the query, the HMG-box domain (Pfam: CHDNT, PF08073) [31] in the human chromodomain-helicase-DNA-binding protein 3 (CHD3) was found as a weak hit (HHpred probability score: 54.2%; Dali *Z*-score: 2.8). The HMG-box is a structural DNA-binding domain present in various eukaryotic proteins [32]. The HMG-box enables proteins to bind to and bend DNA, playing a role in regulating DNA-dependent processes like transcription, replication, and DNA repair. The presence of the helical hairpin with the conserved tryptophan residue suggests likely homologous relationships between the HMG-box domains and bacterial domains such as IscX and WVELL. In addition, we found that the NUT family proteins [33] contain a divergent HMG-box domain without experimental structures that has not been described previously (Figure 2F). The HMG-box domain in human NUT1 (UNP: Q86Y26) has a Dali *Z*-score of 6.5 to an HMG-box domain with structure (1ckt, Figure 2E) and a Dali *Z*-score of 3.2 to the WVELL domain (Figure 2B). We designate this new family of HMG-box domains as HMG_NUT.



G

```
                                    HHHHHHHHHHHH                HHHHHHHHHHHHH
P0C0L9|IscX|Fe-S_assembly|2bzt     23 DPKTVRFT-DMHQWICDLEDFD-DDPQASNEKILEAILLVWLDEAE  66 [66]
O31578|YfhJ|WVELL                  26 TWVELLWS-DFEATYAKA------GHAYQGEKMTEKIVTQWIENYG  64 [89]
Q5UAZ6|Luciferase|Luciferase_N     37 YWTRDEYERGMQDDIISQLELF--RSKPAGKLQVARLRAAWQGAQQ  80 [1237]
Q14839|CHD4|CHDNT|2n5n            172 YRTLTNYK-AFSQFVRPLIAAK--NPKIAVSKMMMVLGAKWREFST 214 [1912]
P63159|Hmgb1|HMG_box_2|1ckt        10 RGKMSSYA-FFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSA  54 [215]
Q86Y26|NUTM1|HMG_NUT              231 SPDTEALS-CFLIPVLRSLARL--KPTMTLEEGLPLAVQEWEHTSN 273 [1132]
```

FIGURE 2 | A structural motif common to a set of HMG-box-like domains. (A) IscX from *Escherichia coli* (PDB: 2bzt). (B) AlphaFold model of a WVELL domain from *Bacillus subtilis* YfhJ (UNP: O31578). (C) AlphaFold model of a Luciferase_N domain (UNP: Q5UAZ6). (D) An HMG-box domain from the human protein CHD4 with Pfam family CHDNT (PDB: 2n5n). (E) An HMG-box domain from the mouse protein Hmgb1 with the Pfam family HMG_box_2 (PDB: 1ckt). (F) AlphaFold model of a remote HMG box domain (HMG_NUT) from the human protein NUTM1 (UNP: Q86Y26). The two core α-helices are colored in rainbow. The additional N- and C-terminal regions are colored in dark blue and dark red, respectively. Sidechains of hydrophobic positions are shown. (G) Multiple sequence alignment of two core α-helices for domains in the above structures. These domains are annotated by UniProt accession, protein name, domain name and PDB code (if available). Hydrophobic positions are shaded in yellow. The conserved tryptophan residues are highlighted in bold. The start and end positions of the domains are shown, and the protein lengths are shown in brackets. α-helices are marked by red "H" letters.

In the updated ECOD classification, we expanded the HMG-box X-group to incorporate newly identified potential homologs (Figure S2) and renamed the X-group to "HMG-box-like." The IscX H-group was transferred from the HTH X-group to the HMG-box-like X-group. Additionally, we established two new H-groups within the HMG-box-like X-group that currently lack experimental structures: the WVELL H-group, which includes the Pfam family WVELL, and the Luciferase_N H-group, which includes the Pfam family Luciferase_N. We also added a representative of the newly identified HMG_NUT family, which lacks an experimental structure, to the existing HMG-box H-group. Due to the limited sequence and structural similarities—primarily confined to a shared helical hairpin motif—the IscX, WVELL, Luciferase_N, and HMG-box domains were classified into separate H-groups. Future improvements in homology detection sensitivity, along with the growth of sequence databases, may enable the establishment of reliable evolutionary relationships among these H-groups.

2.4 | Evolutionarily Related Domains in LigA and Various RiPP Precursors and Enzymes

Aromatic-ring-opening dioxygenase LigAB is an enzyme complex that plays a crucial role in the degradation of lignin-derived aromatic compounds, particularly in bacteria [34]. This enzyme is involved in the cleavage of aromatic rings, a key step in breaking down aromatic hydrocarbons and lignin-derived molecules into simpler forms that can be further metabolized. The catalytic subunit of this complex is LigB, while LigA is a regulatory subunit.

We found significant similarities between LigA (PDB: 1b4u) [35] and several domains in known structures involved in the generation and modification of ribosomally synthesized and post-translationally modified peptides (RiPPs). A LigA-like domain is found in the ophMA protein (PDB: 5n0o) from fungi (HHpred probability score: 98.7%, sequence identity: 24%, Dali *Z*-score: 5.9), which contains a peptide N-methyltransferase domain at the N-terminus and the omphalotin core peptide at the C-terminus [36]. The gene encoding this protein is part of a gene cluster involved in the biosynthesis of omphalotin A, a highly methylated cyclic dodecapeptide with nematicidal activity. Using LigA (PDB: 1b4u) as the query, a LigA-like domain was also found in the N-terminal region of TglI (PDB: 8hi7) [37] (HHpred probability score: 96.2%, sequence identity: 2%, Dali *Z*-score less than 2), which is a regulatory subunit of the TglHI enzyme complex required for the biosynthesis of 3-thiaglutamate, a RiPP generated by the plant pathogen *Pseudomonas syringae*. The region corresponding to the three N-terminal helices of LigA also found HHpred hits to regions in the alpha chains of nitrile hydratases (e.g., PDB: 4ob0; Pfam: PF02979, NHase_alpha; HHpred probability score: 75.5%, sequence identity: 16%, Dali *Z*-score: 2.5), another family of enzymes involved in RiPP biosynthesis [38, 39].

Several Pfam families without known structures were also found, including Frankia_peptide (Pfam: PF14407, ribosomally synthesized peptide from bacteria such as Frankia) (HHpred probability score: 99.1%, sequence identity: 11%) [40], Nif11 (Pfam: PF07862) (HHpred probability score: 90.7%, sequence identity: 21%) [39], ComX (Pfam: PF05952) (HHpred probability score: 89.3%, sequence identity: 16%) [41], and DurB (Pfam: PF19398) (HHpred probability score: 82.2%, sequence identity: 7%) [42]. These hit alignments with various lengths all cover a core segment (residues 37–65) with three α-helices in the query LigA domain. Proteins with these domains are precursors of various RiPPs [43]. The LigA-like domains are in the leader peptide regions of these precursors. Homology of these domains to LigA was supported by their AlphaFold structural models (Figure 3E–H). The most conserved parts of these homologs are the first two α-helices connected by a tight turn with an angle of about 60°. The third α-helix can differ in length, and in the case of ComX, it is deteriorated into a loop. These domains also exhibit structural differences in the region after the three core α-helices (shown in dark red in Figure 3A–H). Most contain one or more C-terminal α-helices with differing orientations, except for the NHase_alpha domain, which has two C-terminal β-strands (dark red in Figure 1D) after the three core α-helices. The NHase_alpha domain possesses a duplication of the LigA-like domain (one copy is rainbow colored and the other copy shown in gray in Figure 1D).

In the updated ECOD classification, we unified LigA and LigA-like domains into a single H-group within the revised X-group "LigA-like domain" (formerly named "LigA subunit of an aromatic-ring-opening dioxygenase LigAB"; see Figure S3). This updated X-group now includes the LigA-like domain from TglI (PDB: 8hi7, chain A, residues 1–59), which had not been previously classified in ECOD. We also corrected the classification of the LigA-like domain in 5n0o (e5n0oA3), which was previously misassigned to the X-group "all-alpha NTP pyrophosphatases." It has now been moved to the "LigA-like domain" X-group with revised domain boundaries (residues 315–378). Additionally, the former X-group "Nitrile hydratase alpha chain" was removed, and the NHase_alpha family was moved into the "LigA-like domain" X-group. This updated X-group also includes four Pfam families lacking experimental structures—Frankia_peptide, Nif11, ComX, and DurB. Evolutionary relationships among these domains are supported by strong sequence and structural similarities, as well as functional associations, as many of them are found in RiPP precursors or RiPP biosynthesis enzymes.

2.5 | A Group of α-Helical Domains Related to STI1_HOP-DP

The STI1-HOP_DP domain (Pfam: PF17830) was found in a variety of proteins such as Sti1, HOP, and Tic40 [44–46]. As cochaperones associated with chaperones such as Hsp70 and Hsp90, they are adapter proteins capable of transferring client proteins between chaperones. The STI1-HOP_DP domains are characterized by several α-helices connected by short turns of about 90°, often characterized by the [DN]P signature (prolines are shown in magenta in Figure 4 structures). Two STI1-HOP_DP domains (DP1 and DP2) are present in the STI1 protein (PDB: 2llv and 2llw) [44]. However, they have been placed in different X-groups in ECOD (2llw in the HTH X-group and 2llv in the X-group of "DP domain"), likely due to the high level of structural divergence (Dali *Z*-score less than 2). DP1 (Figure 4A) adopts a more closed conformation compared to DP2 (Figure 4B). Despite this structural difference, the two

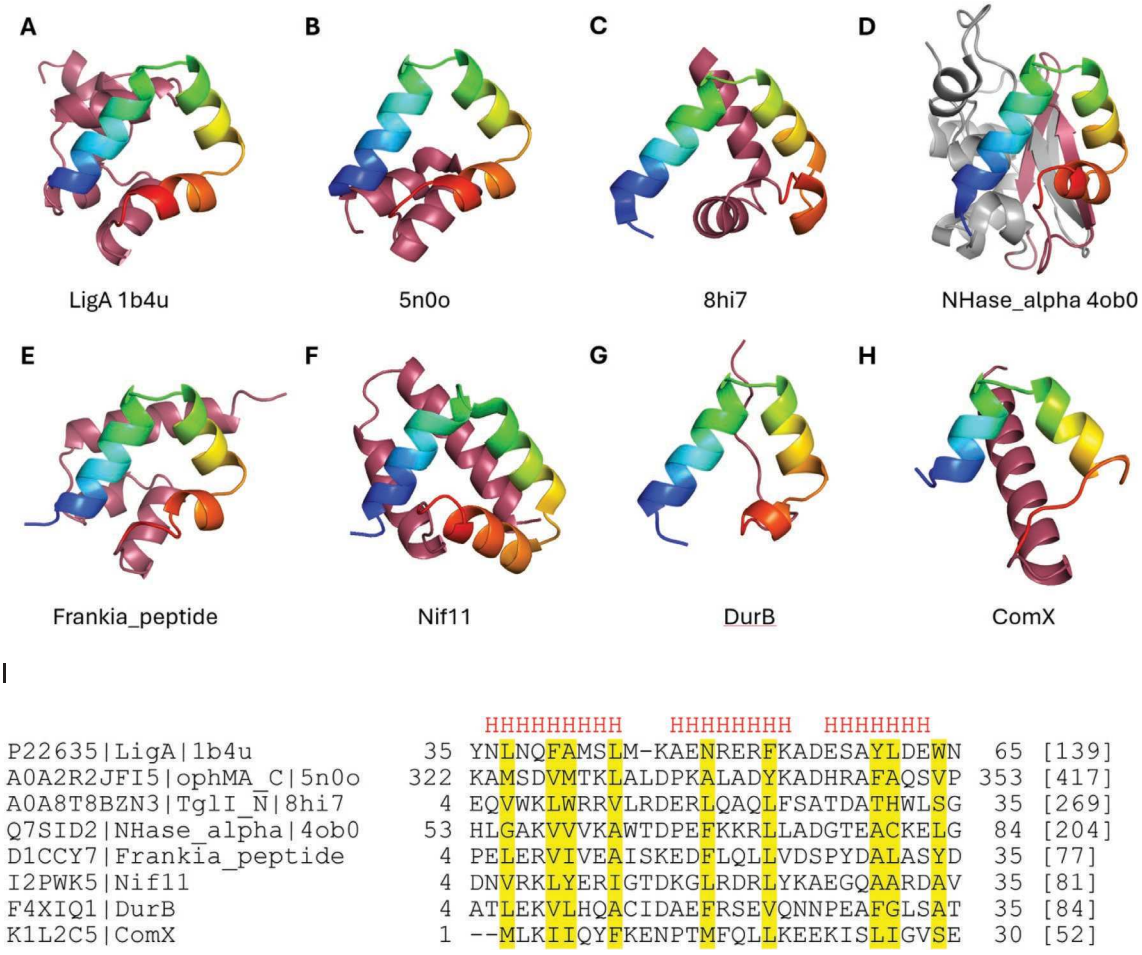


FIGURE 3 | A structural motif common to a set of domains related to LigA. (A) The LigA subunit of the aromatic-ring-opening dioxygenase complex (PDB: 1b4u). (B) The C-terminal domain of a peptide N-methyltransferase (PDB: 5n0o). (C) The N-terminal domain of a RiPP recognition protein that functions as a regulatory subunit of the TglHI enzyme complex (PDB: 8hi7). (D) The alpha chain of a nitrile hydratase (PDB: 4ob0). (E) A protein with the Frankia_peptide domain. (F) A protein with the Nif11 domain. (G) A protein with the DurB domain. (H) A protein with the ComX domain. The top four domains (A–D) have known experimental structures. The bottom four domains (E–H) do not have experimental structures and AlphaFold models are shown for them. The common structural motif with three core α-helices is colored in rainbow. The additional C-terminal α-helices and β-strands are colored in dark red. The duplicated second domain in the NHase_alpha structure is shown in gray. (I) Multiple sequence alignment of the eight structures. Proteins are annotated by UniProt accession, domain name, and PDB code (if available). The start and end positions of the domains are shown, and the protein lengths are shown in brackets. α-helices are marked by red "H" letters.

domains share significant sequence similarity (HHpred probability score: 98.8%, sequence identity: 23%, with alignment covering over 95% of the query), suggesting they may have arisen from a domain duplication event.

Through transitive HHpred searches, we identified several additional protein domain families potentially related to the STI1-HOP_DP domains. For example, the Pfam family ZNHIT3_C (PF21373) was identified using the DP2 domain (PDB: 2llw) as a query (HHpred probability score: 95.1%, sequence identity: 13%). The Pfam family Hit1_C (PF18268) was detected by using a ZNHIT3_C member (PDB: 5l85) as the query (HHpred probability score: 94.1%, sequence identity: 9%). Some of these domains were placed in two isolated H-groups in the HTH X-group: the H-group of "XPC-binding domain" contains some STI1-HOP_DP domains (e.g., PDB: 2llw) and domains belonging to the XPC-binding domain family (Pfam: PF09280; e.g., PDB: 2f4m), and the H-group of "Helical domain in DNA-Damage-Inducible 2 (Ddi2)" contains domains (PDB: 5kes and 5k57, not classified in Pfam) from DNA damage-inducible proteins. Other domain families homologous to STI1-HOP_DP were placed in several X-groups. The X-group of "DP domain" contains Pfam families ZNHIT3_C (Pfam: PF21373; PDB: 5l85) (Figure 4C) and STI1-HOP_DP (PDB: 2llv), the X-group of "C-terminal domain of Hit1" contains the Hit1_C family (Pfam: PF18268; PDB: 2mjf) (Figure 4D), and the X-group of "Helical box domain of E3 ubiquitin-protein ligase HECW1" contains the HECW1_helix family (Pfam: PF18436; PDB: 3l4h) (Figure 4E). The HECW1_helix domain (Figure 4E) shows the greatest sequence and structural similarity to the first STI1_HOP-DP domain of the STI1 protein (PDB: 2llv, Figure 4A), with an HHpred probability score of 79%, sequence identity of 8% and a Dali *Z*-score of 4.9. In contrast, the ZNHIT3_C domain and the Hit1_C domain (Figure 4C,D) adopt a more open conformation, closely resembling the DP2 domain of STI1 (PDB: 2llw, Dali *Z*-scores: 4.9 and 3.3) rather than the DP1 domain of STI1 (PDB: 2llv, Dali *Z*-scores: 2.4 and 2.5). Both ZNHIT3_C and Hit1_C domains bind their substrates (depicted in gray in Figure 4C,D) comparably.

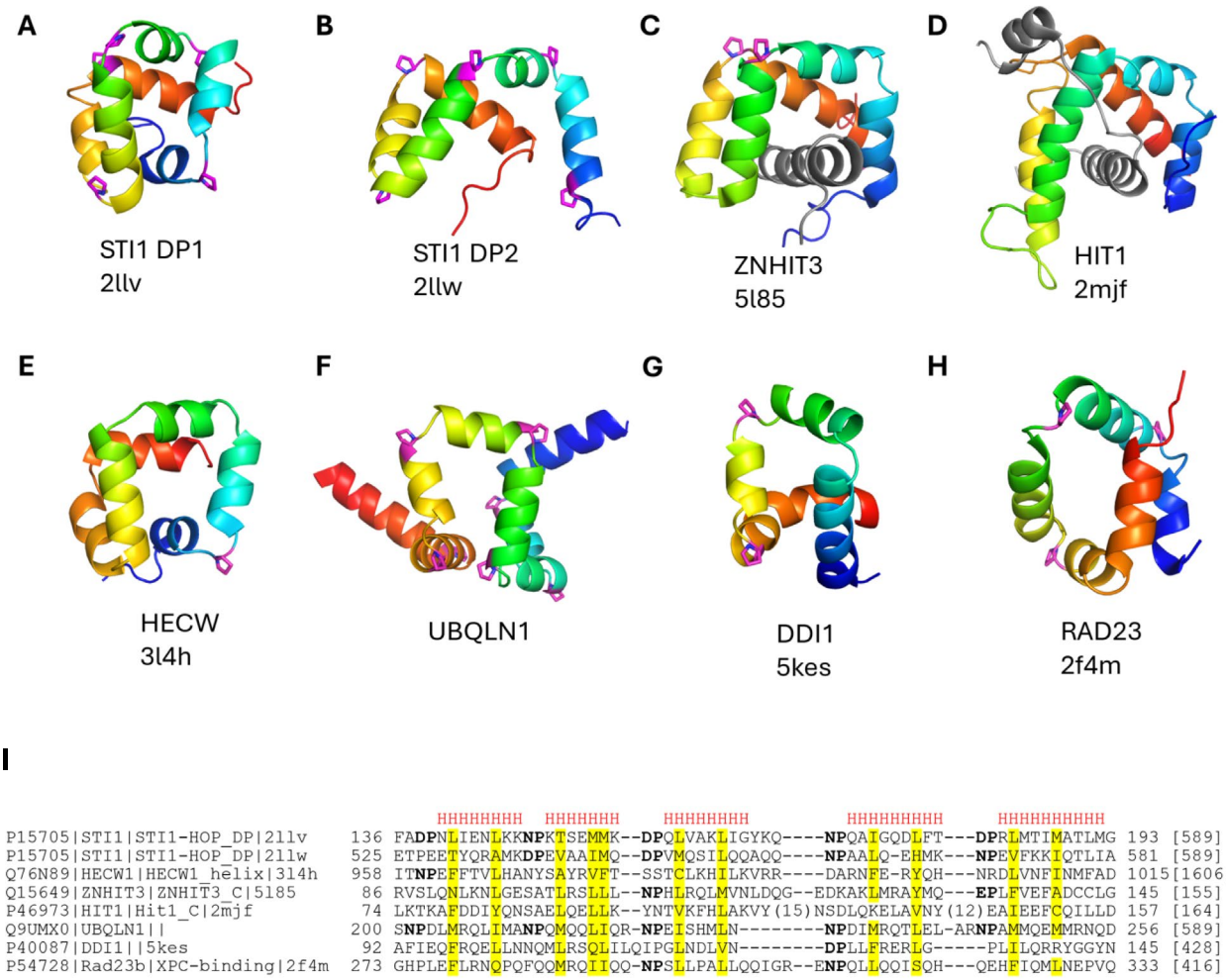


FIGURE 4 | A structural motif common to a set of domains related to STI1-HOP_DP. (A) The first STI1-HOP_DP domain (DP1) of yeast protein STI1. (B) The second STI1-HOP_DP domain (DP2) of yeast protein STI1. (C) The ZNHIT3_C domain in human Zinc finger HIT domain-containing protein 3 (ZNHIT3). (D) The Hit1_C domain in yeast protein HIT1. (E) The HECW1_helix domain in human E3 ubiquitin-protein ligase HECW1. (F) A STI1-HOP_DP-like domain in yeast DNA damage-inducible protein 1 (DDI1). (G) A STI1-HOP_DP-like domain in human protein Ubiquilin-1 (UBQLN1). (H) The XPC-binding domain in mouse protein Rad23b. Experimental structures are shown for these domains, which are annotated by the protein name and the four-letter PDB code except for Ubiquilin-1, for which the AlphaFold model is shown. These structures are rainbow-colored. Prolines in tight turns are colored in magenta. For ZNHIT3 and HIT1, their binding partners are shown in gray. (I) Multiple sequence alignment of the shown structures. The UniProt accession, protein name, Pfam domain name (if available), and PDB code (if available) are separated by vertical bars. Hydrophobic positions are colored in yellow. [NDE]P motifs in turns are highlighted in bold. The start and end positions are shown, and the protein lengths are shown in brackets. α-helices are marked by red "H" letters.

Divergent STI1-HOP_DP domains were also found in human ubiquilin proteins [47] (Figure 4F). A Ddi domain (PDB: 5kes, Figure 4G) as a query identified the XPC-binding domain (PDB: 2f4m) with a high confidence (HHpred probability score: 95.7%, sequence identity: 26%) and the DP2 domain (PDB: 2llw) as a weaker hit (HHpred probability score: 43.9%, sequence identity: 16%). The XPC-binding domains (Pfam: PF09280; PDB: 2f4m), found in RAD23 proteins, were identified as weak hits to STI1-HOP_DP domains. For example, an XPC-binding domain (PDB: 1x3z) was found as a weak hit (HHpred probability: 37.2%, sequence similarity: 21%) by using the DP2 domain (PDB: 2llw) as the query. The XPC-binding domains adopt a different overall topology (Figure 4H) compared to other STI1-HOP_DP-like domains. In the XPC-binding domains, the orientation of the last two α-helices relative to other helices is right-handed, in contrast to the left-handed arrangement observed in other STI1-HOP_DP-like domains.

In the previous ECOD classification, the domains described above with experimental structures were distributed across four separate X-groups (Figure S4). In the updated classification, we unified these domains into a single X-group named "DP domain," which contains two H-groups (Figure S4). We removed the X-groups "C-terminal domain of Hit1" and "Helical box domain of E3 ubiquitin-protein ligase HECW1," transferring their associated families into the "DP domain" H-group. We also removed the H-group "Helical domain in DNA-Damage-Inducible 2 (Ddi2)" from the HTH X-group and reassigned these domains to a new H-group named "XPC-binding domain and DDI helical domain" within the "DP domain" X-group. Additionally, we incorporated the ubiquilin DP domain into the "DP domain" H-group and removed the H-group "XPC-binding domain" from the HTH X-group. The STI1-HOP_DP family from this H-group was merged with the STI1-HOP_DP family already in the "DP domain" X-group while the "XPC-binding" family was reassigned

to the new H-group "XPC-binding domain and DDI helical domain." Although XPC-binding and DDI helical domains show high sequence similarity to each other (HHpred probability scores >95%), they exhibit only weak similarity to other DP domain-related families. Due to differences in overall topology, they are classified into two distinct T-groups within the same H-group.

3 | Materials and Methods

3.1 | Manual Analysis of Isolated Domains in ECOD

We manually analyzed a set of isolated ECOD X-groups and H-groups within the X-group of HTH. We inspected the HHpred [19] results against PDB [20] and Pfam [21] databases and examined conserved motifs among weak hits. For some queries, HHpred searches against the human and various bacterial proteomes were also conducted. Structural similarity searches were conducted by DaliLite [22], and for some proteins also by the Foldseek server [23]. Functional associations were analyzed by using the STRING web server [48]. HHpred and PSI-BLAST searches were also performed on found Pfam domains without experimental structures. Such a transitive search strategy sometimes helped uncover evolutionary relationships of multiple Pfam domains without experimental structures.

Author Contributions

Jimin Pei: investigation, formal analysis, writing – original draft, data curation, conceptualization, methodology. **R. Dustin Schaeffer:** formal analysis, investigation, writing – review and editing, methodology. **Qian Cong:** conceptualization, methodology, writing – review and editing, supervision. **Nick V. Grishin:** writing – review and editing, conceptualization, methodology, supervision.

Acknowledgments

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.), and 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 the Welch Foundation. The authors thank Dr. Lisa Kinch for helpful discussions.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. S. Light, W. Basile, and A. Elofsson, "Orphans and New Gene Origination, a Structural and Evolutionary Perspective," *Current Opinion in Structural Biology* 26 (2014): 73–83.

2. N. Vakirlis and A. Kupczok, "Large-Scale Investigation of Species-Specific Orphan Genes in the Human Gut Microbiome Elucidates Their Evolutionary Origins," *Genome Research* 34 (2024): 888–903.

3. S. Lu, J. Zhang, X. Lian, et al., "A Hidden Human Proteome Encoded by 'Non-Coding' Genes," *Nucleic Acids Research* 47 (2019): 8111–8125.

4. J. Durairaj, A. M. Waterhouse, T. Mets, et al., "Uncovering New Families and Folds in the Natural Protein Universe," *Nature* 622 (2023): 646–653.

5. C. A. Orengo and J. M. Thornton, "Protein Families and Their Evolution—A Structural Perspective," *Annual Review of Biochemistry* 74 (2005): 867–900.

6. D. Zhang, L. M. Iyer, A. M. Burroughs, and L. Aravind, "Resilience of Biochemical Activity in Protein Domains in the Face of Structural Divergence," *Current Opinion in Structural Biology* 26 (2014): 92–103.

7. C. P. Ponting and R. R. Russell, "The Natural History of Protein Domains," *Annual Review of Biophysics and Biomolecular Structure* 31 (2002): 45–71.

8. J. S. Taylor and J. Raes, "Duplication and Divergence: The Evolution of New Genes and Old Ideas," *Annual Review of Genetics* 38 (2004): 615–643.

9. G. C. Conant and A. Wagner, "Asymmetric Sequence Divergence of Duplicate Genes," *Genome Research* 13 (2003): 2052–2058.

10. P. W. Holland, F. Marlétaz, I. Maeso, T. L. Dunwell, and J. Paps, "New Genes From Old: Asymmetric Divergence of Gene Duplicates and the Evolution of Development," *Philosophical Transactions of the Royal Society, B: Biological Sciences* 372 (2017): 20150480.

11. I. Melvin, J. Weston, W. S. Noble, and C. Leslie, "Detecting Remote Evolutionary Relationships Among Proteins by Large-Scale Semantic Embedding," *PLoS Computational Biology* 7 (2011): e1001047.

12. J. Pei, A. Andreeva, S. Chuguransky, et al., "Bridging the Gap Between Sequence and Structure Classifications of Proteins With AlphaFold Models," *Journal of Molecular Biology* 436 (2024): 168764.

13. M. Kilinc, K. Jia, and R. L. Jernigan, "Improved Global Protein Homolog Detection With Major Gains in Function Identification," *National Academy of Sciences of the United States of America* 120 (2023): e2211823120.

14. R. Mudgal, S. Sandhya, N. Chandra, and N. Srinivasan, "De-DUFing the DUFs: Deciphering Distant Evolutionary Relationships of Domains of Unknown Function Using Sensitive Homology Detection Methods," *Biology Direct* 10 (2015): 1–23.

15. R. D. Schaeffer, Y. Liao, H. Cheng, and N. V. Grishin, "ECOD: New Developments in the Evolutionary Classification of Domains," *Nucleic Acids Research* 45 (2017): D296–D302.

16. J. Jumper, R. Evans, A. Pritzel, et al., "Highly Accurate Protein Structure Prediction With AlphaFold," *Nature* 596 (2021): 583–589.

17. K. Tunyasuvunakool, J. Adler, Z. Wu, et al., "Highly Accurate Protein Structure Prediction for the Human Proteome," *Nature* 596 (2021): 590–596.

18. C. A. Orengo, D. T. Jones, and J. M. Thornton, "Protein Superfamilies and Domain Superfolds," *Nature* 372 (1994): 631–634.

19. F. Gabler, S. Z. Nam, S. Till, et al., "Protein Sequence Analysis Using the MPI Bioinformatics Toolkit," *Current Protocols in Bioinformatics* 72 (2020): e108.

20. A. Kouranov, L. Xie, J. de la Cruz, et al., "The RCSB PDB Information Portal for Structural Genomics," *Nucleic Acids Research* 34 (2006): D302–D305.

21. J. Mistry, S. Chuguransky, L. Williams, et al., "Pfam: The Protein Families Database in 2021," *Nucleic Acids Research* 49 (2021): D412–D419.

22. L. Holm and J. Park, "DaliLite Workbench for Protein Structure Comparison," *Bioinformatics* 16 (2000): 566–567.

23. M. van Kempen, S. S. Kim, C. Tumescheit, et al., "Fast and Accurate Protein Structure Search With Foldseek," *Nature Biotechnology* 42 (2024): 243–246.

24. L. Holm, S. Kääriäinen, P. Rosenström, and A. Schenkel, "Searching Protein Structure Databases With DaliLite v. 3," *Bioinformatics* 24 (2008): 2780–2781.

25. J. H. Kim, J. R. Bothe, R. O. Frederick, J. C. Holder, and J. L. Markley, "Role of IscX in Iron–Sulfur Cluster Biogenesis in *Escherichia coli*," *Journal of the American Chemical Society* 136 (2014): 7933–7942.

26. K. Sakumi, K. Igarashi, M. Sekiguchi, and A. Ishihama, "The Ada Protein Is a Class I Transcription Factor of *Escherichia coli*," *Journal of Bacteriology* 175 (1993): 2455–2457.

27. Y. Lin, V. Dötsch, T. Wintner, et al., "Structural Basis for the Functional Switch of the *E. coli* Ada Protein," *Biochemistry* 40 (2001): 4261–4271.

28. C. He, J.-C. Hus, L. J. Sun, et al., "A Methylation-Dependent Electrostatic Switch Controls DNA Repair and Transcriptional Activation by *E. coli* Ada," *Molecular Cell* 20 (2005): 117–129.

29. S. S. Krishna, I. Majumdar, and N. V. Grishin, "Structural Classification of Zinc Fingers: Survey and Summary," *Nucleic Acids Research* 31 (2003): 532–550.

30. C. Pastore, S. Adinolfi, M. A. Huynen, et al., "YfhJ, a Molecular Adaptor in Iron-Sulfur Cluster Formation or a Frataxin-Like Protein?," *Structure* 14 (2006): 857–867.

31. E. Staub, P. Fiziev, A. Rosenthal, and B. Hinzmann, "Insights Into the Evolution of the Nucleolus by an Analysis of Its Protein Domain Repertoire," *BioEssays* 26 (2004): 567–581.

32. M. Štros, D. Launholt, and K. D. Grasser, "The HMG-Box: A Versatile Protein Domain Occurring in a Wide Variety of DNA-Binding Proteins," *Cellular and Molecular Life Sciences* 64 (2007): 2590–2606.

33. C. A. French, C. Ramirez, J. Kolmakova, et al., "BRD–NUT Oncoproteins: A Family of Closely Related Nuclear Proteins That Block Epithelial Differentiation and Maintain the Growth of Carcinoma Cells," *Oncogene* 27 (2008): 2237–2242.

34. K. P. Barry and E. A. Taylor, "Characterizing the Promiscuity of LigAB, a Lignin Catabolite Degrading Extradiol Dioxygenase From *Sphingomonas paucimobilis* SYK-6," *Biochemistry* 52 (2013): 6724–6736.

35. K. Sugimoto, T. Senda, H. Aoshima, E. Masai, M. Fukuda, and Y. Mitsui, "Crystal Structure of an Aromatic Ring Opening Dioxygenase LigAB, a Protocatechuate 4, 5-Dioxygenase, Under Aerobic Conditions," *Structure* 7 (1999): 953–965.

36. H. Song, N. S. Van Der Velden, S. L. Shiran, et al., "A Molecular Mechanism for the Enzymatic Methylation of Nitrogen Atoms Within Peptide Bonds," *Science Advances* 4, no. 8 (2018): eaat2720, https://doi.org/10.1126/sciadv.aat2720.

37. Y. Zheng, X. Xu, X. Fu, et al., "Structures of the Holoenzyme TglHI Required for 3-Thiaglutamate Biosynthesis," *Structure* 31 (2023): 1220–1232.e5.

38. M. Montalbán-López, T. A. Scott, S. Ramesh, et al., "New Developments in RiPP Discovery, Enzymology and Engineering," *Natural Product Reports* 38, no. 1 (2021): 130–239, https://doi.org/10.1039/d0np00027b.

39. D. H. Haft, M. K. Basu, and D. A. Mitchell, "Expansion of Ribosomally Produced Natural Products: A Nitrile Hydratase-and Nif11-Related Precursor Family," *BMC Biology* 8 (2010): 1–15.

40. L. M. Iyer, S. Abhiman, A. M. Burroughs, and L. Aravind, "Amidoligases With ATP-Grasp, Glutamine Synthetase-Like and Acetyltransferase-Like Domains: Synthesis of Novel Metabolites and Peptide Modifications of Proteins," *Molecular BioSystems* 5 (2009): 1636–1660.

41. M. Ansaldi, D. Marolt, T. Stebe, I. Mandic-Mulec, and D. Dubnau, "Specific Activation of the Bacillus Quorum-Sensing Systems by Isoprenylated Pheromone Variants," *Molecular Microbiology* 44 (2002): 1561–1573.

42. A. Fredenhagen, G. Fendrich, F. Märki, et al., "Duramycins B and C, Two New Lanthionine Containing Antibiotics as Inhibitors of Phospholipase A2 Structural Revision of Duramycin and Cinnamycin," *Journal of Antibiotics* 43, no. 11 (1990): 1403–1412, https://doi.org/10.7164/antibiotics.43.1403.

43. P. G. Arnison, M. J. Bibb, G. Bierbaum, et al., "Ribosomally Synthesized and Post-Translationally Modified Peptide Natural Products: Overview and Recommendations for a Universal Nomenclature," *Natural Product Reports* 30 (2013): 108–160.

44. A. B. Schmid, S. Lagleder, M. A. Gräwert, et al., "The Architecture of Functional Modules in the Hsp90 Co-Chaperone Sti1/hop," *EMBO Journal* 31 (2012): 1506–1517.

45. N. S. Silva, D. E. Bertolino-Reis, P. R. Dores-Silva, et al., "Structural Studies of the Hsp70/Hsp90 Organizing Protein of Plasmodium Falciparum and Its Modulation of Hsp70 and Hsp90 ATPase Activities," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1868 (2020): 140282.

46. Y.-F. Kao, Y.-C. Lou, Y.-H. Yeh, C.-D. Hsiao, and C. Chen, "Solution Structure of the C-Terminal NP-Repeat Domain of Tic40, a Co-Chaperone During Protein Import Into Chloroplasts," *Journal of Biochemistry* 152 (2012): 443–451.

47. Z. Kurlawala, P. P. Shah, C. Shah, and L. J. Beverly, "The STI and UBA Domains of UBQLN1 Are Critical Determinants of Substrate Interaction and Proteostasis," *Journal of Cellular Biochemistry* 118 (2017): 2261–2270.

48. D. Szklarczyk, J. H. Morris, H. Cook, et al., "The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible," *Nucleic Acids Research* 45 (2017): D362–D368.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.