Published in final edited form as:
Proteins. 2015 July ; 83(7): 1238–1251. doi:10.1002/prot.24818.

Manual classification strategies in the ECOD database

Hua Cheng[1,3], **Yuxing Liao**[2,3], **R. Dustin Schaeffer**[1,3], and **Nick V. Grishin**[1,2,*]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA

[2]Department of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050

Abstract

ECOD (Evolutionary Classification Of protein Domains) is a comprehensive and up-to-date protein structure classification database. The majority of new structures released from the PDB (Protein Data Bank) every week already have close homologs in the ECOD hierarchy and thus can be reliably partitioned into domains and classified by software without manual intervention. However, those proteins that lack confidently detectable homologs require careful analysis by experts. Although many bioinformatics resources rely on expert curation to some degree, specific examples of how this curation occurs and in what cases it is necessary are not always described. Here, we illustrate the manual classification strategy in ECOD by example, focusing on two major issues in protein classification: domain partitioning and the relationship between homology and similarity scores. Most examples show recently released and manually classified PDB structures. We discuss multi-domain proteins, discordance between sequence and structural similarities, difficulties with assessing homology with scores, and integral membrane proteins homologous to soluble proteins. By timely assimilation of newly available structures into its hierarchy, ECOD strives to provide a most accurate and updated view of the protein structure world as a result of combined computational and expert-driven analysis.

Keywords

protein; sequence; structure; evolution; classification; homology; domain; database

Introduction

Protein classifications organize the protein world in meaningful ways and help to reveal the interplay of protein sequence, structure, function, and evolution. Currently, the most widely used protein structure classifications are SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/

[*]To whom correspondence should be addressed: Nick V. Grishin, Ph.D., Investigator/Professor, Howard Hughes Medical Institute/ Departments of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX, 75390-9050, Voice: (214)645-5952, Fax: (214)645-5948, grishin@chop.swmed.edu.

[3]These authors contributed equally to this work

Supplementary material
Supplementary material includes 6 files: ECOD_examples_FigS1_1feu_HH.pdf, ECOD_examples_FigS2_1feu_DALI.pdf, ECOD_examples_FigS3_3v6p_HH.pdf, ECOD_examples_FigS4_AAY43359dot1_HH.pdf, ECOD_examples_FigS5_3p4g_DALI.pdf, and ECOD_examples_FigS6_3p4g_HH.pdf. Each file is an HHpred or DALI run result.

index.html)[1] and CATH (http://www.cathdb.info/)[2]. CATH is more reliant on automatic methods for classification, whereas SCOP relies more on manual analysis and curation. SCOP and CATH are invaluable resources for studying proteins, but they do not generally include the most recently solved structures in the PDB[3].

We have developed the ECOD (Evolutionary Classification Of protein Domains) database (http://prodata.swmed.edu/ecod/) and published details of its contents and underlying theory elsewhere.[4] Briefly, ECOD strives to recognize broader homology than other domain classifications while retaining family divisions between close homologs. By removing boundaries based on non-evolutionary criteria such as fold or cell location, ECOD can recognize evolutionary relationships between topologically-distinct homologs or between proteins which have recently evolved from soluble to membrane-bound, or vice-versa.

ECOD is a hierarchical classification consisting of five levels: architecture (A), possible homology (X), homology (H), topology (T), and family (F). The entire database is first divided into architectures based on secondary structure element (SSE) composition and overall shape (e.g., alpha bundles and beta sandwiches). Below the architecture level, possible homology (or X-level) groups domains that might be homologous as implied usually by overall structural similarity or fold similarity. Fold similarity itself is not enough to establish homology because it may result from either homology or analogy.[5–7] Below the possible homology level, homology (or H-level) groups domains that are descended from a common ancestor as indicated by significant sequence and/or structure scores, shared functional properties, opinions in literature and in SCOP, etc. Below the homology level, topology (or T-level) groups domains that have similar arrangements of and connections between the SSEs (T-level reflects the observations that homologs can have different topologies[8–10]). Below the topology level, family (or F-level) groups domains that have significant sequence similarity primarily based on Pfam[11].

ECOD employs an automatic software pipeline to classify newly released structures. Proteins that cannot be classified confidently and completely by automated methods are manually curated. The manual classification process involves partitioning the query protein into domains and identifying homologs or possible homologs for each domain. In this process, we rely on scientific literature, sequence and structure similarity comparison programs, popular protein databases (e.g., Pfam, SCOP, and CATH), visual inspection and comparison, as well as our knowledge and experience.

Here, through examples, we illustrate the characteristics of ECOD and explain the curation methodology from the broader perspective of protein classification. In the example of the ANTAR domain, a newly released structure led to the modification of the domain boundaries of previously classified proteins and the establishment of a new homologous group. The multi-domain proteins example demonstrates where ECOD has partitioned the multi-domain proteins in SCOP into individual domains. The cysteine proteinase example illustrates the expansion of this large and diverse superfamily to incorporate many newly released structures that are not yet represented in other databases. The next four examples explore the complex relationship between homology and sequence/structure similarity scores: the homology between MRP and Whirly cannot be detected by sequence methods but can be identified by good structure scores; the homology between TL5/L25 C-terminal domain and YbbR is accompanied by a low structure score but a high sequence score; the 'high sequence or structure score may not imply homology' example shows that high scores do not necessarily imply homology; and the 'low scores given by software do not preclude homology' example demonstrates that low sequence and structure scores do not necessarily mean there is no homology. In the final example, we discuss homology between soluble and intramembrane proteins and explain why ECOD classifies these two kinds of proteins in the same hierarchy.

Materials and Methods

The key to classifying a newly released PDB (or a query structure) is to identify its homologs in the ECOD hierarchy. As the majority of newly released PDB chains already have close homologs in ECOD, they can be mapped to the hierarchy automatically using a pipeline of scripts described in detail elsewhere[4]. For the remaining query chains, the sequence/structure comparison methods used in the pipeline can only find hits that either have low scores or cover just part of the query (e.g., just one domain in a multi-domain protein). In such cases, the pipeline cannot confidently and completely classify the query and thus passes it to the manual curators. Manual curators then apply multiple considerations to identify homologs for a query structure: 1) the domain assignments attempted by the pipeline, as the pipeline can still suggest correct hits (i.e., homologs) for part of the query (e.g., one domain in a multi-domain protein); 2) scientific literature; 3) results from homology detection tools such as HHsearch[12], DALI[13], and HorA server[14]; 4) other protein classification databases such as Pfam, SCOP, and CATH; and 5) visual inspection and comparison. When a homologous hit with similar topology can be found, the query is classified into the same T-group as the hit; when a homologous hit with different topology can be found, the query is classified in a new T-group but the same H-group as the hit; when only a possibly homologous hit with similar overall structure can be found, the query is classified in a new H-group but the same X-group as the hit; when no possible homologs can be identified, the query is classified in a new X-group by itself. We use a custom web interface and Google Docs to collect and present necessary information for manual analysis as well as to record and share manual classifications and annotations. After a PDB weekly update, the pipeline generates a table containing the newly released structures that it cannot map to ECOD. Two manual curators work on the table consecutively. The first curator goes through the queries rather quickly to resolve easier and less demanding issues, and the second curator studies more challenging cases in greater depth to render final classification decisions. Typically it takes the first curator 1~1.5 days and then the second curator 2~3 days to classify all the queries. Specific details of method used to classify examples discussed in this study are provided with each example below.

Results and Discussion

Recognizing new domain splits and homologous groups: ANTAR domain

Domains are defined in part by their relationship to other protein domains. Discovery of new structures can reveal previously unknown domain boundaries in existing structures. The

ANTAR (AmiR and NasR transcription antitermination regulators) domain is an RNA-binding module found in bacterial response regulatory proteins.[15] In the Pfam database, the ANTAR family has three proteins with solved structures: AmiR (PDB 1QO0, chain D)[16], Rv1626 (PDB 1SD5)[17], and NasR (PDB 4AKK)[18]. All three structures show that ANTAR domain adopts a small 3-helical bundle conformation with characteristic helix-helix packing angles. In AmiR and Rv1626, ANTAR follows an N-terminal Rossmann-like domain, whereas in NasR, it follows an N-terminal helical bundle domain (Fig. 1). SCOP classifies both AmiR and Rv1626 in the ‘Flavodoxin-like’ fold and ‘CheY-like’ superfamily, noting that there is an additional small helical subdomain at the C-terminus. NasR is a relatively new structure and is not yet classified in SCOP or SCOPe[19]. When trying to classify NasR in ECOD, the pipeline found AmiR as the best hit with HHsearch probability 98% but query coverage only 22% and therefore passed this case to the manual curators. We studied this case and realized that ANTAR is an established domain in literature and in Pfam and that it occurs in combination with different domains. Therefore, we created a new ANTAR domain H-group in ECOD to accommodate the ANTAR domains in NasR, AmiR and Rv1626 (the latter two are split from the SCOP entries d1qo0d_ and d1sd5a_). This example demonstrates that it is important to consider the impact of newly released structures on existing members of the classification in order to reflect the most up-to-date understanding of the protein world.

Domain partitioning of multi-domain proteins

SCOP contains a ‘Multi-domain proteins (alpha and beta)’ class that accommodates ‘folds consisting of two or more domains belonging to different classes’. Historically, members of the multi-domain proteins class contained domains belonging to different potential SCOP classes that had never been seen in an independent context, making confident partition difficult. Since their initial classification, homologs to many of these proteins have been discovered.[20] Therefore, by consulting SCOP notes as well as literature, we have split each of the entries in SCOP1.75 ‘Multi-domain proteins’ class and classified the resulting domains in appropriate locations in the ECOD hierarchy. As a result, homologs existing in both the multi-domain class and one of the other classes in SCOP can be joined in a single ECOD H-group, potentially benefiting the training and testing of bioinformatics algorithms.

For example, Toprim (topoisomerase-primase) domain is present in various proteins that function in DNA/RNA metabolism.[21] Toprim adopts a Rossmann-like fold consisting of repeating beta-alpha units with conserved acidic residues clustered at the C-terminal end of the parallel four-stranded beta-sheet.[21,22] Toprim domains typically occur in DNA/RNA-manipulating enzymes (topoisomerases type IA and type II, DnaG-type primases, and OLD family nucleases) and likely use the conserved acidic residues to coordinate catalytically active metal ions; however, RecR Toprim has lost some of the conserved acidic residues and functions instead as a protein-protein interaction module.[21–23] In SCOP, Toprim is present in five folds (‘Prokaryotic type I DNA topoisomerase’, ‘Type II DNA topoisomerase’, ‘DNA topoisomerase IV, alpha subunit’, ‘DNA primase core’, and ‘Recombination protein RecR’) in ‘Multi-domain proteins’ class and in one fold (‘Toprim domain’) in ‘Alpha and beta proteins (a/b)’ class. Fig. 2 depicts one representative SCOP entry from each of these six folds and shows how the multi-domain entries are split in ECOD. This splitting enables

ECOD to gather all of the Toprim domains from their different contexts into one homologous group.

Expanding and subdividing homologous groups: the cysteine proteinases

The release of new structures can join together sequence families that were previously unknown to be related. The SCOP superfamily ‘Cysteine proteinases’ comprises a large number of homologous proteins and includes not only proteases/peptidases such as papain (Enzyme Nomenclature EC 3.4.22.2, http://www.chem.qmul.ac.uk/iubmb/enzyme/) but also transglutaminases such as coagulation factor XIII (EC 2.3.2.13) and acetyltransferases such as arylamine N-acetyltransferase (EC 2.3.1.118). As noted in SCOP and in the literature,[24–28] the structural core of this diverse superfamily consists of an alpha helix packed against a three-stranded antiparallel beta-sheet, and the catalytic triad is composed of a cysteine (Cys) residing on the N-terminus of the alpha helix and a histidine (His) and a polar residue (usually asparagine (Asn) or aspartate (Asp)) residing on the beta-sheet (Fig. 3). In ECOD, the cysteine proteinases superfamily (H-group ‘Cysteine proteinases’) has been expanded to include multiple new members that have recently been released in the PDB but have not yet been classified in SCOP, SCOPe, or CATH. This large H-group is subdivided into multiple F-groups based on sequence similarity. While the majority of these F-groups correspond to Pfam families, some F-groups are not yet included in Pfam. Below we discuss two F-groups whose identities as members of the cysteine protease superfamily are established in the literature but are not yet recorded in other databases. By cataloging these sequence families in the cysteine proteinases H-group, ECOD strives to provide researchers a comprehensive and up-to-date view of this diverse superfamily.

The ECOD F-group ‘DUF4285’ corresponds to ‘Domain of unknown function (DUF4285)’, which is a functionally uncharacterized Pfam-A family that is not assigned to any Pfam clan. This F-group has two manually classified representatives and multiple automatically classified non-representatives, and at the time of writing, none of these structures are recorded in SCOP, SCOPe, or CATH yet. When one manual representative, *Salmonella typhimurium* Tae4 (PDB 4hff, chain A), is used as a query in sequence searches, the Pfam website (http://pfam.xfam.org/) finds DUF4285 with E-value 6.3e-28, whereas the CDD website (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)[29,30] finds DUF4285 with E-value 2.09e-39. Although uncharacterized or unrepresented in popular databases, published studies[31–33] show that Tae4-related proteins are peptidoglycan amidase effectors secreted by a type VI secretion system and are related to the CHAP and NlpC/P60 families[26,34]. In Pfam, CHAP and NlpC/P60 are members of the ‘Peptidase_CA’ clan that collects papain-like peptidases, whereas in SCOP, they are classified in the ‘Cysteine proteinases’ superfamily. Indeed, when compared to papain (Fig. 3), Tae4 shows a similar arrangement of the core secondary structural elements (SSEs) and the catalytic triad. Therefore, ECOD classifies Tae4-related proteins in the cysteine proteinases H-group.

The ECOD F-group ‘Cycle inhibiting factor (Cif)’ has two manual representatives and multiple non-representatives, and at the time of writing, none of these structures are in SCOP, SCOPe, or CATH. When one manual representative, *Photorhabdus luminescens* Cif (PDB 3gqj, chain A), is used as a query, neither Pfam nor CDD identifies a matching family.

Though unrepresented in popular databases, recent works[35–38] demonstrate that Cifs are papain-like deamidases utilized by bacteria as type III secreted effectors to modulate host cell function and that, by converting a specific glutamine to glutamate in ubiquitin and ubiquitin homolog NEDD8, Cifs impair ubiquitin-dependent protein degradation pathway and cause cell cycle arrest. As shown in Fig. 3, Cif shares the characteristic structural core and catalytic triad with papain. Thus, ECOD classifies Cifs in the cysteine proteinases H-group.

Homology with undetectable sequence similarity but pronounced structure similarity

Detection of structural similarity is often necessary for identifying evolutionary relationships between distant homologs. The mitochondrial RNA binding protein complex consists of two homologous proteins, MRP1 and MRP2, which bind to guide RNAs and are essential for kinetoplastid RNA editing in trypanosomatids.[39,40] Although sequence homology detection methods, such as PSI-BLAST[41] and HHsearch[12], fail to detect any other homolog of MRP, the crystal structures of MRP1 and MRP2 exhibit remarkable structural similarity to the Whirly family of single-stranded DNA (ssDNA) binding proteins in plant (Fig. 4A and 4B).[42] A Dali alignment between MRP1 (PDB 2GIA) and WHY1 (PDB 1L3A) has a Z-score of 13.8 and a RMSD of 2.4 Å over 125 residues (Fig. 4C). We note that the structure prediction server I-TASSER[43] identified WHY2 (PDB 3N1H, ranked 2nd) and WHY1 (PDB 4KOO, ranked 9th) with normalized Z-scores of 0.73 and 0.53 respectively in the top 10 templates used for threading when MRP2 sequence is provided as input and close templates in PDB sequences are excluded. Whirly proteins bind to ssDNA functioning in transcription regulation and DNA double-strand break repair[44–46] and also can bind to plastid RNA in chloroplast RNA metabolism[47].[48] In addition to the structural similarity of the protomer, MRP and Whirly proteins both form tetramers that superimpose well with a RMSD of 3.7 Å over 248 residues (Fig. 4D).[42,44] MRP complex is a heterotetramer with two MRP1 and two MRP2,[42] while Whirly proteins form a homotetramer[44,45] and are suggested to further assemble into a 24-mer[49]. They also both bind to nucleic acids on the same surface in a sequence-independent fashion (Fig. 4D). However, distinct binding mechanisms are adopted. For MRP, binding is dominated by the electrostatic interaction between the positively charged surface of MRP and the phosphate groups of the guide RNA.[42] Whirly proteins mainly use hydrophobic interactions of nucleobases and the compensation of few sequence-specific interactions is observed in structures with different ssDNAs.[45] The homology relationship is also recorded in SCOP as they are classified in the same superfamily. The MRP and Whirly families represent two highly diverged branches that are distributed in animals plus trypanosomatids and plants, respectively. They likely originated from a duplication event as shown by other homologous families such as human transcription cofactor PC4[50], which is a homodimer of two ββββα units, and Pur-α whose bacterial homodimer structure (PDB 3N8B)[51] and the duplicated form in *Drosophila* (PDB 3K44)[52] are solved. The results of this divergent evolution are reflected in the distinct sequence profiles of MRP and Whirly (Fig. 4C), posing a difficult challenge for sequence homology detection methods.

Homology with low structure similarity but significant sequence similarity

The CTC (catabolite-controlled) family of proteins is widespread in bacteria and includes both constitutive ribosome components (e.g., *Escherichia coli* ribosomal protein L25 and *Thermus thermophilus* ribosomal protein TL5/L25) and temporary ribosome components that are only produced under special circumstances such as stress (e.g., *Bacillus subtilis* CTC).[53] Although *E. coli* L25 has only a single domain that binds 5S ribosomal RNA (PDB 1dfu),[54] *T. thermophilus* TL5/L25 consists of two domains: an N-terminal domain that is homologous to *E. coli* L25 and a C-terminal beta-sandwich domain that is suggested to have a novel fold (PDB 1feu)[55]. Although TL5/L25 N- and C-terminal domains have distinct structures, SCOP classifies the entire *T. thermophilus* TL5/L25 (SCOP domain d1feua_) protein in the same family as *E. coli* L25 in the fold 'Ribosomal protein L25-like' with the note 'contains additional all-beta (sub)domain in the C-terminal extension'. CATH does excise the C-terminal domain to form a homologous superfamily 'Ribosomal protein L25-like Domain 2' (Homologous superfamily: 2.170.120.20). In the *T. thermophilus* ribosome, TL5/L25 C-terminal domain interacts with both the ribosomal protein L16 and the 23S rRNA and is suggested to contribute to the stability of the local conformation.[5356]

YbbR domains are present in a variety of bacteria lineages.[57] In *Bacillus subtilis*, the YbbR protein consists of four repeating YbbR domains and is in the same operon as a diadenylyl cyclase (DAC) that synthesizes cyclic di-AMP (c-di-AMP). By interacting and stimulating the enzymatic activity of DAC, YbbR modulates the level of c-di-AMP, which is a second messenger regulating important physiological processes.[57,58] The structures of the first and the fourth YbbR domains in the *Desulfitobacterium hafniense* Y51 YbbR-like protein have been recently reported as part of the structural genomics effort, and interestingly, the authors note that the characteristic fold of the YbbR domains is only shared by TL5/L25 C-terminal domains.[59] Indeed, as shown in Fig. 5A, both TL5/L25 C-terminal domain and YbbR domain appear as elongated and twisted beta-sandwiches with the peptide chains traversing between the beta-sheets in the same way, although the DALI Z-score between them is low. When TL5/L25 C-terminal domain (PDB 1feu, chain A, residue 94–185) is used as a query, HHpred[12] finds YbbR domains immediately after other L25 proteins with high probabilities. For example, the first (PDB 3lyw, chain A) and the fourth (PDB 2l3u, chain A) YbbR domains in *Desulfitobacterium hafniense* Y51 YbbR-like protein have probabilities 96.6% and 95.1%, respectively, and in both cases, the alignments cover almost the whole lengths of the query and the hit and include conserved matching positions (Supporting Information Fig. S1). Based on their shared peculiar fold and statistically significant sequence similarity, ECOD classifies TL5/L25 C-terminal domain and YbbR domain as remote homologs and puts them in the same H-group. At the time of writing, it appears that YbbR family members are not yet classified in other databases such as SCOP, SCOPe, and CATH.

When TL5/L25 C-terminal domain (PDB 1feu, chain A, residue 94–185) is used as a query, DALI[13] first finds other L25 proteins, then the YbbR domains, then many DNA-directed RNA polymerase subunit alpha (Supporting Information Fig. S2). For example, *E. coli* RNA polymerase alpha subunit (PDB 1bdf, chain C)[60] has Z-score 3.1, RMSD 2.9 Å, and aligned length 67. The aligned part corresponds to SCOP domain d1bdfc2, which belongs to the fold 'Insert subdomain of RNA polymerase alpha subunit'. Although SCOP does not mention any similarity between TL5/L25 C-terminal domain and RNA polymerase alpha subunit insert domain, CATH considers them to be structurally similar and classifies them as two homologous superfamilies (Homologous superamilies: 2.170.120.12 and 2.170.120.20, respectively) in the same topology group (Topology: 2.170.120). Fig. 5B shows TL5/L25 C-terminal domain (left) as well as two RNA polymerase alpha subunit insert domains from the same SCOP family (middle, *E. coli* protein; right, yeast protein[61]). Although DALI Z-score is low, TL5/L25 C-terminal domain and RNA polymerase alpha subunit insert domain do share similar topologies, and *T. thermophilus* TL5/L25 C-terminal domain and *E. coli* RNA polymerase alpha subunit insert domain even have corresponding loops that have nearly identical conformations (indicated by red arrows), though this loop conformation is not preserved in the yeast protein. TL5/L25 C-terminal domain and RNA polymerase alpha subunit insert domain do have two structural differences: 1) the second major secondary structural element (SSE) is a beta-strand in TL5/L25 C-terminal domain but an alpha-helix in RNA polymerase alpha subunit insert domain; 2) RNA polymerase alpha subunit insert domain has one additional strand at the C-terminus which sits in the middle of one of the beta-sheets. Based on these observations, ECOD classifies TL5/L25 C-terminal domain and RNA polymerase alpha subunit insert domain as possible homologs in different H-groups but the same X-group.

High sequence or structure score may not imply homology

Common methods of detecting homology by sequence or structure can report false positive results. TAL (transcription activator-like) effectors are secreted by plant pathogens and bind host DNA via TAL repeats.[62] Each TAL repeat consists of two helices and interacts with one nucleotide in the target DNA, and multiple repeats tandemly pack together to form a left-handed alpha-alpha superhelix.[63,64] When manually curating ECOD, we observed that when the TAL repeats in TAL effector dHax3 (PDB 3v6p, chain A) is used as query, HHpred[12] lists the tetratricopeptide repeat (TPR) domain of kinesin light chain 1 (PDB 3nf1, chain A) even before other TAL effectors such as PthXo1 (PDB 3ugm, chain A) (Supporting Information Fig. S3). As shown in Fig. 6A and supplementary Fig. S3, although HHpred gives dHax3 TAL repeats and kinesin light chain 1 TPR very good scores, the alignment is of low quality: sequence similarity is low, several long gaps are present, and no motifs are identifiable (as short stretches of stronger similarity among weakly similar regions). Interestingly, dHax3 is an artificially modified protein,[63,65] and when the natural protein Hax3 (GenBank: AAY43359.1, 96% identical with dHax3) is used as a query, HHpred does not find any TPR proteins with significant scores (Supporting Information Fig. S4). When submitted to DALI[13], dHax3 TAL repeats and kinesin light chain 1 TPR only have Z-score 1.5, indicating that they are not structurally similar. Most importantly, these two kinds of repeat domains have opposite handedness: TAL repeats assume a left-handed packing between the two-helical bundle units, whereas TPR assumes a right-handed packing (Fig. 6A)[64]. Due to the low alignment quality, the different search results when modified and natural sequences are used as queries, the poor DALI Z-score, and the opposite handedness, we think that TAL repeats and TPR are not homologous and that the high HHpred scores between dHax3 TAL repeats and kinesin light chain 1 TPR are a false positive result.

Structure comparison programs may also identify false positive hits. For instance, when the antifreeze protein (AFP) (PDB 3p4g, chain B)[66] is used as a query, the DALI server[67] finds the hypothetical protein YDCK (PDB 2f9c, chain A) with a high Z-score of 13.9 and an unusually large root mean square deviation (RMSD) of 10.7 Å (Fig. 6B and Supporting Information Fig. S5). A close comparison of the two structures explains the large RMSD: although they are both beta-helices, AFP is right-handed and YDCK is left-handed. As noted by Li *et al.*[68], the DALI algorithm compares distance matrices and can ignore the handedness of a structure. Due to the opposite handedness, we believe that AFP and YDCK are not homologous and that the high DALI Z-score is a false positive. We note that when AFP is used as a query to search the SCOP1.75 database, HHpred only identifies proteins in the 'Single-stranded right-handed beta-helix' fold with the correct handedness as significant hits (Supporting Information Fig. S6).

As shown by the above examples, both sequence-based and structure-based comparison programs can generate false positives. Thus, when inferring homology, it is helpful to not rely solely on one score from a single program but rather to consider information from multiple diverse sources such as scores from different programs, alignment qualities, and visual inspection of the structures in question.

Low scores given by software do not preclude homology

Homology between small proteins can be hard for sequence or structure comparison methods to detect. The Crustacean Hyperglycemic Hormone (CHH)-like superfamily of peptides are widespread in ecdysozoans and regulate a variety of physiological processes. Members of this superfamily include CHH, moult-inhibiting hormone (MIH), gonad/vitellogenesis-inhibiting hormone (GIH/VIH), and ion transport peptide (ITP), and multiple sequence alignments reveal six absolutely conserved cysteines forming three disulfide bonds.[69,70] Currently, SCOP1.75[1] and SCOPe2.04[19] 'Crustacean CHH/MIH/GIH neurohormone' fold contains only one protein, Kuruma prawn MIH (PDB 1j0t)[71]. A recent work[72] has expanded CHH-like superfamily to include many venom peptides from spiders and other species, and one of these peptides turns out to have a solved structure (spider *Tegenaria agrestis* toxin TaITX1, PDB 2KSL). In addition, HorA server[14] identifies prawn MIH as the first hit of the newly released structure of the neurotoxin Kappa-scoloptoxin-Ssm1a or k-Ssm1a from centipede *Scolopendra subspinipes* (PDB 2m35)[73]. As shown in Fig. 7A, spider toxin TaITX1, prawn MIH, and centipede neurotoxin k-Ssm1a share similar overall folds of a small helical bundle stabilized by three disulfide bonds. A manual superposition reveals that three helices are structurally equivalent (colored in blue, green, and yellow, respectively). MIH has an additional C-terminal helix, whereas TaITX1 has an additional N-terminal helix. Importantly, the disulfide bond between the 2nd and the 4th cysteines linking the blue and the green helices and the disulfide bond between the 3rd and the 6th cysteines linking the blue and the yellow helices both superimpose closely in all three structures. However, the disulfide bond between the 1st and the 5th cysteines linking the N-terminal region and the green helix do not superimpose well: this disulfide is much closer to the other two disulfides in k-Ssm1a than in TaITX1 or MIH due to insertions/deletions in their sequences. The DALI[13] Z-scores and HHsearch[12] probabilities between these three small proteins are quite low, but the DALI and HHsearch alignments largely agree with each other (Fig. 7B). In addition, when used as queries in DALI database searches, TaITX1 and k-Ssm1a both find MIH as the first hit. Based on the shared overall folds, matching disulfide bond patterns, and the agreement between sequence and structure alignment algorithms, ECOD has expanded the CHH-like superfamily to include TaITX1 and k-Ssm1a. As shown in this example, remote homology can be difficult for sequence and structure comparison programs to detect due to low levels of similarity reflected by low scores, and thus it usually takes careful manual analysis to identify such distant evolutionary relationship.

Homology between water-soluble and intramembrane proteins

Soluble and intramembrane proteins have distinct biophysical properties. While SCOP has a separate class called ‘Membrane and cell surface proteins and peptides’, the CATH hierarchy does not have a special place for membrane proteins.[74] ECOD has chosen not to discriminate between soluble and membrane proteins but instead to classify them in one hierarchy due to two observations: 1) soluble and membrane proteins can be homologous; and 2) some proteins can transform from a soluble state to a transmembrane state, as discussed below.

Terpenoids (a.k.a. isoprenoids) are built from 5-carbon isoprene units and have myriad chemical structures and manifold physiological functions.[75,76] Class I terpenoid synthases are a group of homologous enzymes that produce various terpenoid molecules; for example, farnesyl diphosphate synthase (FPS) synthesizes the linear 15-carbon farnesyl diphosphate (FPP), and pentalenene synthase cyclizes farnesyl diphosphate to make the cyclic 15-carbon pentalenene.[75–78] Within the broad group of class I terpenoid synthases, *trans*-prenyltransferases (a.k.a. *trans*-isoprenyl diphosphate synthases) form a conserved subgroup and catalyze the magnesium ion (Mg^{2+})-dependent transfer of prenyl chains of various lengths from an isoprenyl diphosphate (prenyl donor) to isopentenyl diphosphate (prenyl acceptor).[75–77,79] As a prototype of *trans*-prenyltransferases, FPS is a soluble enzyme with a central hydrophobic cavity to accommodate the isoprenyl chains of its substrates and products.[77,80,81]

The UbiA family comprises intramembrane prenyltransferases that synthesize a variety of biomolecules, and the structures of two family members have recently been reported.[82,83] Interestingly, the authors noted that the intramembrane UbiA and the soluble *trans*-prenyltransferases such as FPS are probably related. As discussed in the original reports[82,83] and also shown in Fig. 8, UbiA and FPS exhibit similarities in the following aspects. 1) Both UbiA and FPS adopt a helical bundle fold that probably results from the duplication and fusion of a 4-helix bundle. DALI superimposes UbiA (PDB 4od5, chain A) and FPS (PDB 1rqi, chain A) with a significant Z-score of 14.6 and RMSD 3.8 Å on 220 aligned residues. 2) Both UbiA and FPS present two conserved aspartate-rich motifs in corresponding places, i.e., between the 2nd and the 3rd helices of each 4-helix bundle. 3) Both UbiA and FPS catalyze the Mg^{2+}-dependent prenyl chain transfer from an isoprenyl diphosphate donor to a prenyl acceptor (UbiA uses an aromatic acceptor while FPS uses a linear acceptor). 4) UbiA and FPS show active site resemblance (i.e., the aspartate-rich motifs coordinate the Mg^{2+} ions that in turn bind the diphosphate moiety of the prenyl donor) and may exploit similar catalytic mechanisms. Based on these structural and functional similarities, ECOD classifies UbiA as homologous to FPS and other class I terpenoid synthases (same H-group). In addition to the class I terpenoid synthases superfamily, the type II phosphatidic acid phosphatases (PAP2) homologous superfamily also has both soluble and transmembrane members.[84,85]

Pore-forming proteins (PFPs) are produced by diverse organisms and participate in various physiological processes such as pathogenesis, immunity, and apoptosis.[86] PFPs are synthesized as soluble proteins, but through conformational change and often oligomerization, they manage to penetrate membranes and form pores.[87–89] Based on the secondary structural elements that make up the transmembrane pore, PFPs can be divided into two categories: alpha-PFPs and beta-PFPs. Alpha-PFPs such as pore-forming colicins and Bcl2 family apoptosis regulators use helices to traverse the membrane, whereas beta-PFPs such as membrane attach complex/perforin (MACPF) and cholesterol-dependent cytolysin (CDC) domains form transmembrane beta barrels.[87–89]The remarkable ability of PFPs to transform from soluble state to transmembrane state blurs the boundary between soluble and transmembrane proteins.

As discussed above, intramembrane and soluble proteins can be homologs, and pore-forming proteins can convert from soluble state to transmembrane state. Therefore, ECOD does not have a special category for membrane proteins but instead classifies membrane and soluble proteins in the same hierarchy.

In summary, the key issue in protein evolutionary classification is identifying homologs for the query. As the automated pipeline is usually able to detect close homologs with significant scores and high coverage, the main task for manual curators is to find remote homologs. Various lines of evidence could support homology: significant sequence or structure comparison scores, shared conserved motifs, common functional properties, common cofactor binding modes, similar disulfide patterns, shared structural features such as an unusual left-handed connection or a rare fold, similar oligomerization modes, similar domain organizations, et al.[5,90] As indicated by the above examples, argument for remote homology is often based on multiple lines of evidence. Therefore, manual curators need to integrate information from various sources (sequence and structure comparison results from programs, knowledge and insights in the literature and other databases, and observations from visual inspection) and use their experience to decide if there is adequate evidence for homology.

Conclusions

A first-principles definition of domains and their classifications is hardly possible because the true evolutionary history of all protein domains is not known and in some cases may be unknowable. We evaluate the homologous relationships between proteins and partition them into domains using our manual expertise, the results from well-known sequence and structure comparison programs, and the knowledge recorded by others in literature. The large number of known proteins as well as the increase in the rate of discovery of new proteins implies that a completely curated manual classification is impossible. Similarly, limits on the ability of automated methods to determine homology at large evolutionary distances, as well as problems with domain detection at the boundaries of our theoretical definition (e.g., small motifs that are not always domains, proteins with internal repeats, intrinsically disordered proteins), prevent the creation of a fully automated classification. Nonetheless, we expect that as the protein space becomes more covered, the burden of manual curation will decrease. Additionally, as more advanced structure and sequence comparison tools are being developed, we expect that detection of less conserved region, regions of complex topology, and clear delineation of proteins with internal repeats will become more automatic. The complexity of biology limits our ability to derive entirely consistent rules of protein classification and thus construct a fully-automated classification. Because of these complicating factors, many of the most successful protein classifications incorporate both automated and manual classification elements in their methodology. However, the precise types of cases that necessitate manual curation are often not fully described. Here we have attempted to reveal the intricacies of these cases in a number of examples that are demonstrative of the types of conflicts that can arise in an automated system, where new proteins can alter or necessitate changes to existing proteins and domains in the classification. Our protein classification, ECOD, attempts to minimize the differences between the known tree and the complete tree by incorporating as many proteins as possible through frequent updates. Although ECOD F-groups are primarily derived from Pfam families, ongoing works focus on refining domain boundaries of incorporated families based on structural evidence and generating new families that were previously unrepresented. We aim to further increase the coverage of our classification by incorporating those sequences whose structures are not known in the near future with improved F-group profile hidden Markov models, a strategy that has also been adopted by other protein classifications such as CATH[91].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Lisa Kinch and Jimin Pei for helpful discussions. This work was supported in part by National Institutes of Health (GM094575 to NVG) and the Welch Foundation (I-1505 to NVG).

Abbreviations

PDB	Protein Data Bank
SCOP	Structural Classification of Proteins
CATH	Class, Architecture, Topology, Homology
SCOPe	Structural Classification of Proteins – extended
SSE	secondary structure element
RMSD	root mean square deviation

References

1. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247(4):536–540. [PubMed: 7723011]
2. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure. 1997; 5(8):1093–1108. [PubMed: 9309224]
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic acids research. 2000; 28(1):235–242. [PubMed: 10592235]
4. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014; 10(12):e1003926. [PubMed: 25474468]
5. Murzin AG. How far divergent evolution goes in proteins. Current opinion in structural biology. 1998; 8(3):380–387. [PubMed: 9666335]
6. Orengo CA, Sillitoe I, Reeves G, Pearl FM. Review: what can structural classifications reveal about protein evolution? J Struct Biol. 2001; 134(2–3):145–165. [PubMed: 11551176]
7. Krishna SS, Grishin NV. Structurally analogous proteins do exist! Structure. 2004; 12(7):1125–1127. [PubMed: 15242587]
8. Coles M, Hulko M, Djuranovic S, Truffault V, Koretke K, Martin J, Lupas AN. Common evolutionary origin of swapped-hairpin and double-psi beta barrels. Structure. 2006; 14(10):1489–1498. [PubMed: 17027498]
9. Chaudhuri I, Soding J, Lupas AN. Evolution of the beta-propeller fold. Proteins. 2008; 71(2):795–803. [PubMed: 17979191]
10. Grishin NV. KH domain: one motif, two folds. Nucleic acids research. 2001; 29(3):638–643. [PubMed: 11160884]
11. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. Nucleic acids research. 2014; 42:D222–D230. (Database issue). [PubMed: 24288371]
12. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005; 21(7): 951–960. [PubMed: 15531603]
13. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol. 1993; 233(1):123–138. [PubMed: 8377180]
14. Kim BH, Cheng H, Grishin NV. HorA web server to infer homology between proteins using sequence and structural similarity. Nucleic acids research. 2009; 37:W532–W538. (Web Server issue). [PubMed: 19417074]
15. Shu CJ, Zhulin IB. ANTAR: an RNA-binding domain in transcription antitermination regulatory proteins. Trends in biochemical sciences. 2002; 27(1):3–5. [PubMed: 11796212]
16. O'Hara BP, Norman RA, Wan PT, Roe SM, Barrett TE, Drew RE, Pearl LH. Crystal structure and induction mechanism of AmiC-AmiR: a ligand-regulated transcription antitermination complex. EMBO J. 1999; 18(19):5175–5186. [PubMed: 10508151]
17. Morth JP, Feng V, Perry LJ, Svergun DI, Tucker PA. The crystal and solution structure of a putative transcriptional antiterminator from Mycobacterium tuberculosis. Structure. 2004; 12(9): 1595–1605. [PubMed: 15341725]
18. Boudes M, Lazar N, Graille M, Durand D, Gaidenko TA, Stewart V, van Tilbeurgh H. The structure of the NasR transcription antiterminator reveals a one-component system with a NIT nitrate receptor coupled to an ANTAR RNA-binding effector. Mol Microbiol. 2012; 85(3):431–444. [PubMed: 22690729]
19. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic acids research. 2014; 42:D304–D309. (Database issue). [PubMed: 24304899]
20. Hubbard TJ, Murzin AG, Brenner SE, Chothia C. SCOP: a structural classification of proteins database. Nucleic acids research. 1997; 25(1):236–239. [PubMed: 9016544]

21. Aravind L, Leipe DD, Koonin EV. Toprim--a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Nucleic acids research. 1998; 26(18):4205–4213. [PubMed: 9722641]

22. Yang W. Topoisomerases and site-specific recombinases: similarities in structure and mechanism. Crit Rev Biochem Mol Biol. 2010; 45(6):520–534. [PubMed: 21087076]

23. Honda M, Inoue J, Yoshimasu M, Ito Y, Shibata T, Mikawa T. Identification of the RecR Toprim domain as the binding site for both RecF and RecO. A role of RecR in RecFOR assembly at double-stranded DNA-single-stranded DNA junctions. J Biol Chem. 2006; 281(27):18549–18559. [PubMed: 16675461]

24. Berti PJ, Storer AC. Alignment/phylogeny of the papain superfamily of cysteine proteases. J Mol Biol. 1995; 246(2):273–283. [PubMed: 7869379]

25. Makarova KS, Aravind L, Koonin EV. A superfamily of archaeal, bacterial, and eukaryotic proteins homologous to animal transglutaminases. Protein science : a publication of the Protein Society. 1999; 8(8):1714–1719. [PubMed: 10452618]

26. Anantharaman V, Aravind L. Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. Genome Biol. 2003; 4(2):R11. [PubMed: 12620121]

27. Bromme D. Papain-like cysteine proteases. Curr Protoc Protein Sci. 2001; Chapter 21(Unit 21):22.

28. Sinclair JC, Sandy J, Delgoda R, Sim E, Noble ME. Structure of arylamine N-acetyltransferase reveals a catalytic triad. Nature structural biology. 2000; 7(7):560–564. [PubMed: 10876241]

29. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic acids research. 2011; 39:D225–D229. (Database issue). [PubMed: 21109532]

30. Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. Nucleic acids research. 2004; 32:W327–W331. (Web Server issue). [PubMed: 15215404]

31. Zhang H, Gao ZQ, Wang WJ, Liu GF, Xu JH, Su XD, Dong YH. Structure of the type VI effector-immunity complex (Tae4-Tai4) provides novel insights into the inhibition mechanism of the effector by its immunity protein. J Biol Chem. 2013; 288(8):5928–5939. [PubMed: 23288853]

32. Srikannathasan V, English G, Bui NK, Trunk K, O'Rourke PE, Rao VA, Vollmer W, Coulthurst SJ, Hunter WN. Structural basis for type VI secreted peptidoglycan DL-endopeptidase function, specificity and neutralization in Serratia marcescens. Acta Crystallogr D Biol Crystallogr. 2013; 69(Pt 12):2468–2482. [PubMed: 24311588]

33. Russell AB, Singh P, Brittnacher M, Bui NK, Hood RD, Carl MA, Agnello DM, Schwarz S, Goodlett DR, Vollmer W, Mougous JD. A widespread bacterial type VI secretion effector superfamily identified using a heuristic approach. Cell Host Microbe. 2012; 11(5):538–549. [PubMed: 22607806]

34. Bateman A, Rawlings ND. The CHAP domain: a large family of amidases including GSP amidase and peptidoglycan hydrolases. Trends in biochemical sciences. 2003; 28(5):234–237. [PubMed: 12765834]

35. Crow A, Race PR, Jubelin G, Varela Chavez C, Escoubas JM, Oswald E, Banfield MJ. Crystal structures of Cif from bacterial pathogens Photorhabdus luminescens and Burkholderia pseudomallei. Plos One. 2009; 4(5):e5582. [PubMed: 19440549]

36. Crow A, Hughes RK, Taieb F, Oswald E, Banfield MJ. The molecular basis of ubiquitin-like protein NEDD8 deamidation by the bacterial effector protein Cif. Proc Natl Acad Sci U S A. 2012; 109(27):E1830–E1838. [PubMed: 22691497]

37. Cui J, Yao Q, Li S, Ding X, Lu Q, Mao H, Liu L, Zheng N, Chen S, Shao F. Glutamine deamidation and dysfunction of ubiquitin/NEDD8 induced by a bacterial effector family. Science. 2010; 329(5996):1215–1218. [PubMed: 20688984]

38. Jubelin G, Chavez CV, Taieb F, Banfield MJ, Samba-Louaka A, Nobe R, Nougayrede JP, Zumbihl R, Givaudan A, Escoubas JM, Oswald E. Cycle inhibiting factors (CIFs) are a growing family of functional cyclomodulins present in invertebrate and mammal bacterial pathogens. Plos One. 2009; 4(3):e4855. [PubMed: 19308257]

39. Aphasizhev R, Aphasizheva I, Nelson RE, Simpson L. A 100-kD complex of two RNA-binding proteins from mitochondria of Leishmania tarentolae catalyzes RNA annealing and interacts with several RNA editing components. Rna. 2003; 9(1):62–76. [PubMed: 12554877]

40. Vondruskova E, van den Burg J, Zikova A, Ernst NL, Stuart K, Benne R, Lukes J. RNA interference analyses suggest a transcript-specific regulatory role for mitochondrial RNA-binding proteins MRP1 and MRP2 in RNA editing and other RNA processing in Trypanosoma brucei. J Biol Chem. 2005; 280(4):2429–2438. [PubMed: 15504736]

41. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997; 25(17):3389–3402. [PubMed: 9254694]

42. Schumacher MA, Karamooz E, Zikova A, Trantirek L, Lukes J. Crystal structures of T. brucei MRP1/MRP2 guide-RNA binding complex reveal RNA matchmaking mechanism. Cell. 2006; 126(4):701–711. [PubMed: 16923390]

43. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols. 2010; 5(4):725–738. [PubMed: 20360767]

44. Desveaux D, Allard J, Brisson N, Sygusch J. A new family of plant transcription factors displays a novel ssDNA-binding surface. Nature structural biology. 2002; 9(7):512–517. [PubMed: 12080340]

45. Cappadocia L, Marechal A, Parent JS, Lepage E, Sygusch J, Brisson N. Crystal Structures of DNA-Whirly Complexes and Their Role in Arabidopsis Organelle Genome Repair. Plant Cell. 2010; 22(6):1849–1867. [PubMed: 20551348]

46. Desveaux D, Subramaniam R, Despres C, Mess JN, Levesque C, Fobert PR, Dangl JL, Brisson N. A "whirly" transcription factor is required for salicylic acid-dependent disease resistance in Arabidopsis. Dev Cell. 2004; 6(2):229–240. [PubMed: 14960277]

47. Prikryl J, Watkins KP, Friso G, van Wijk KJ, Barkan A. A member of the Whirly family is a multifunctional RNA- and DNA-binding protein that is essential for chloroplast biogenesis. Nucleic acids research. 2008; 36(16):5152–5165. [PubMed: 18676978]

48. Krause K, Kilbienski I, Mulisch M, Rodiger A, Schafer A, Krupinska K. DNA-binding proteins of the Whirly family in Arabidopsis thaliana are targeted to the organelles. Febs Lett. 2005; 579(17): 3707–3712. [PubMed: 15967440]

49. Cappadocia L, Parent JS, Zampini E, Lepage E, Sygusch J, Brisson N. A conserved lysine residue of plant Whirly proteins is necessary for higher order protein assembly and protection against DNA damage. Nucleic acids research. 2012; 40(1):258–269. [PubMed: 21911368]

50. Brandsen J, Werten S, van der Vliet PC, Meisterernst M, Kroon J, Gros P. C-terminal domain of transcription cofactor PC4 reveals dimeric ssDNA binding site. Nature structural biology. 1997; 4(11):900–903. [PubMed: 9360603]

51. Graebsch A, Roche S, Kostrewa D, Soding J, Niessing D. Of Bits and Bugs - On the Use of Bioinformatics and a Bacterial Crystal Structure to Solve a Eukaryotic Repeat-Protein Structure. Plos One. 2010; 5(10)

52. Graebsch A, Roche S, Niessing D. X-ray structure of Pur-alpha reveals a Whirly-like fold and an unusual nucleic-acid binding surface. Proc Natl Acad Sci U S A. 2009; 106(44):18521–18526. [PubMed: 19846792]

53. Gongadze GM, Korepanov AP, Korobeinikova AV, Garber MB. Bacterial 5S rRNA-binding proteins of the CTC family. Biochemistry (Mosc). 2008; 73(13):1405–1417. [PubMed: 19216708]

54. Lu M, Steitz TA. Structure of Escherichia coli ribosomal protein L25 complexed with a 5S rRNA fragment at 1.8-A resolution. Proc Natl Acad Sci U S A. 2000; 97(5):2023–2028. [PubMed: 10696113]

55. Fedorov R, Meshcheryakov V, Gongadze G, Fomenkova N, Nevskaya N, Selmer M, Laurberg M, Kristensen O, Al-Karadaghi S, Liljas A, Garber M, Nikonov S. Structure of ribosomal protein TL5 complexed with RNA provides new insights into the CTC family of stress proteins. Acta Crystallogr D Biol Crystallogr. 2001; 57(Pt 7):968–976. [PubMed: 11418764]

56. Selmer M, Dunham CM, Murphy FVt, Weixlbaumer A, Petry S, Kelley AC, Weir JR, Ramakrishnan V. Structure of the 70S ribosome complexed with mRNA and tRNA. Science. 2006; 313(5795):1935–1942. [PubMed: 16959973]

57. Corrigan RM, Grundling A. Cyclic di-AMP: another second messenger enters the fray. Nat Rev Microbiol. 2013; 11(8):513–524. [PubMed: 23812326]

58. Mehne FM, Gunka K, Eilers H, Herzberg C, Kaever V, Stulke J. Cyclic di-AMP homeostasis in bacillus subtilis: both lack and high level accumulation of the nucleotide are detrimental for cell growth. J Biol Chem. 2013; 288(3):2004–2017. [PubMed: 23192352]

59. Barb AW, Cort JR, Seetharaman J, Lew S, Lee HW, Acton T, Xiao R, Kennedy MA, Tong L, Montelione GT, Prestegard JH. Structures of domains I and IV from YbbR are representative of a widely distributed protein family. Protein science : a publication of the Protein Society. 2011; 20(2):396–405. [PubMed: 21154411]

60. Zhang G, Darst SA. Structure of the Escherichia coli RNA polymerase alpha subunit amino-terminal domain. Science. 1998; 281(5374):262–266. [PubMed: 9657722]

61. Westover KD, Bushnell DA, Kornberg RD. Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. Cell. 2004; 119(4):481–489. [PubMed: 15537538]

62. Boch J, Bonas U. Xanthomonas AvrBs3 family-type III effectors: discovery and function. Annu Rev Phytopathol. 2010; 48:419–436. [PubMed: 19400638]

63. Deng D, Yan C, Pan X, Mahfouz M, Wang J, Zhu JK, Shi Y, Yan N. Structural basis for sequence-specific recognition of DNA by TAL effectors. Science. 2012; 335(6069):720–723. [PubMed: 22223738]

64. Mak AN, Bradley P, Cernadas RA, Bogdanove AJ, Stoddard BL. The crystal structure of TAL effector PthXo1 bound to its DNA target. Science. 2012; 335(6069):716–719. [PubMed: 22223736]

65. Mahfouz MM, Li L, Shamimuzzaman M, Wibowo A, Fang X, Zhu JK. De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. Proc Natl Acad Sci U S A. 2011; 108(6):2623–2628. [PubMed: 21262818]

66. Garnham CP, Campbell RL, Davies PL. Anchored clathrate waters bind antifreeze proteins to ice. Proc Natl Acad Sci U S A. 2011; 108(18):7363–7367. [PubMed: 21482800]

67. Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. Nucleic acids research. 2010; 38:W545–W549. (Web Server issue). [PubMed: 20457744]

68. Li H, Marsolo K, Parthasarathy S, Polshakov D. A new approach to protein structure mining and alignment. 2004:1–10.

69. Montagne N, Desdevises Y, Soyez D, Toullec JY. Molecular evolution of the crustacean hyperglycemic hormone family in ecdysozoans. BMC Evol Biol. 2010; 10:62. [PubMed: 20184761]

70. Webster SG, Keller R, Dircksen H. The CHH-superfamily of multifunctional peptide hormones controlling crustacean metabolism, osmoregulation, moulting, and reproduction. Gen Comp Endocrinol. 2012; 175(2):217–233. [PubMed: 22146796]

71. Katayama H, Nagata K, Ohira T, Yumoto F, Tanokura M, Nagasawa H. The solution structure of molt-inhibiting hormone from the Kuruma prawn Marsupenaeus japonicus. J Biol Chem. 2003; 278(11):9620–9623. [PubMed: 12519766]

72. McCowan C, Garb JE. Recruitment and diversification of an ecdysozoan family of neuropeptide hormones for black widow spider venom expression. Gene. 2014; 536(2):366–375. [PubMed: 24316130]

73. Yang S, Liu Z, Xiao Y, Li Y, Rong M, Liang S, Zhang Z, Yu H, King GF, Lai R. Chemical punch packed in venoms makes centipedes excellent predators. Mol Cell Proteomics. 2012; 11(9):640–650. [PubMed: 22595790]

74. Neumann S, Fuchs A, Mulkidjanian A, Frishman D. Current status of membrane protein structure classification. Proteins. 2010; 78(7):1760–1773. [PubMed: 20186977]

75. Wendt KU, Schulz GE. Isoprenoid biosynthesis: manifold chemistry catalyzed by similar enzymes. Structure. 1998; 6(2):127–133. [PubMed: 9519404]

76. Gao Y, Honzatko RB, Peters RJ. Terpenoid synthase structures: a so far incomplete view of complex catalysis. Nat Prod Rep. 2012; 29(10):1153–1175. [PubMed: 22907771]

77. Tarshis LC, Yan M, Poulter CD, Sacchettini JC. Crystal structure of recombinant farnesyl diphosphate synthase at 2.6-A resolution. Biochemistry. 1994; 33(36):10871–10877. [PubMed: 8086404]

78. Lesburg CA, Zhai G, Cane DE, Christianson DW. Crystal structure of pentalenene synthase: mechanistic insights on terpenoid cyclization reactions in biology. Science. 1997; 277(5333): 1820–1824. [PubMed: 9295272]

79. Liang PH. Reaction kinetics, catalytic mechanisms, conformational changes, and inhibitor design for prenyltransferases. Biochemistry. 2009; 48(28):6562–6570. [PubMed: 19537817]

80. Hosfield DJ, Zhang Y, Dougan DR, Broun A, Tari LW, Swanson RV, Finn J. Structural basis for bisphosphonate-mediated inhibition of isoprenoid biosynthesis. J Biol Chem. 2004; 279(10):8526–8529. [PubMed: 14672944]

81. Liang PH, Ko TP, Wang AH. Structure, mechanism and function of prenyltransferases. Eur J Biochem. 2002; 269(14):3339–3354. [PubMed: 12135472]

82. Cheng W, Li W. Structural insights into ubiquinone biosynthesis in membranes. Science. 2014; 343(6173):878–881. [PubMed: 24558159]

83. Huang H, Levin EJ, Liu S, Bai Y, Lockless SW, Zhou M. Structure of a membrane-embedded prenyltransferase homologous to UBIAD1. PLoS Biol. 2014; 12(7):e1001911. [PubMed: 25051182]

84. Fan J, Jiang D, Zhao Y, Liu J, Zhang XC. Crystal structure of lipid phosphatase Escherichia coli phosphatidylglycerophosphate phosphatase B. Proc Natl Acad Sci U S A. 2014; 111(21):7636–7640. [PubMed: 24821770]

85. Neuwald AF. An unexpected structural relationship between integral membrane phosphatases and soluble haloperoxidases. Protein science : a publication of the Protein Society. 1997; 6(8):1764–1767. [PubMed: 9260289]

86. Bischofberger M, Gonzalez MR, van der Goot FG. Membrane injury by pore-forming proteins. Current opinion in cell biology. 2009; 21(4):589–595. [PubMed: 19442503]

87. Iacovache I, Bischofberger M, van der Goot FG. Structure and assembly of pore-forming proteins. Current opinion in structural biology. 2010; 20(2):241–246. [PubMed: 20172710]

88. Parker MW, Feil SC. Pore-forming protein toxins: from structure to function. Progress in biophysics and molecular biology. 2005; 88(1):91–142. [PubMed: 15561302]

89. Anderluh G, Lakey JH. Disparate proteins use similar architectures to damage membranes. Trends in biochemical sciences. 2008; 33(10):482–490. [PubMed: 18778941]

90. Kinch LN, Grishin NV. Evolution of protein structures and functions. Current opinion in structural biology. 2002; 12(3):400–408. [PubMed: 12127461]

91. Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S, Yeats C, Dessailly BH, Rentzsch R, Orengo CA. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. Nucleic acids research. 2014; 42:D240–D245. (Database issue). [PubMed: 24270792]

92. Perry K, Mondragon A. Structure of a complex between E. coli DNA topoisomerase I and single-stranded DNA. Structure. 2003; 11(11):1349–1358. [PubMed: 14604525]

93. Dong KC, Berger JM. Structural basis for gate-DNA recognition and bending by type IIA topoisomerases. Nature. 2007; 450(7173):1201–1205. [PubMed: 18097402]

94. Nichols MD, DeAngelis K, Keck JL, Berger JM. Structure and function of an archaeal topoisomerase VI subunit with homology to the meiotic recombination factor Spo11. EMBO J. 1999; 18(21):6177–6188. [PubMed: 10545127]

95. Keck JL, Roche DD, Lynch AS, Berger JM. Structure of the RNA polymerase domain of E. coli primase. Science. 2000; 287(5462):2482–2486. [PubMed: 10741967]

96. Lee BI, Kim KH, Park SJ, Eom SH, Song HK, Suh SW. Ring-shaped architecture of RecR: implications for its role in homologous recombinational DNA repair. EMBO J. 2004; 23(10): 2029–2038. [PubMed: 15116069]

97. Janowski R, Kozak M, Jankowska E, Grzonka Z, Jaskolski M. Two polymorphs of a covalent complex between papain and a diazomethylketone inhibitor. J Pept Res. 2004; 64(4):141–150. [PubMed: 15357669]

98. Papadopoulos JS, Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics. 2007; 23(9):1073–1079. [PubMed: 17332019]

99. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004; 14(6):1188–1190. [PubMed: 15173120]

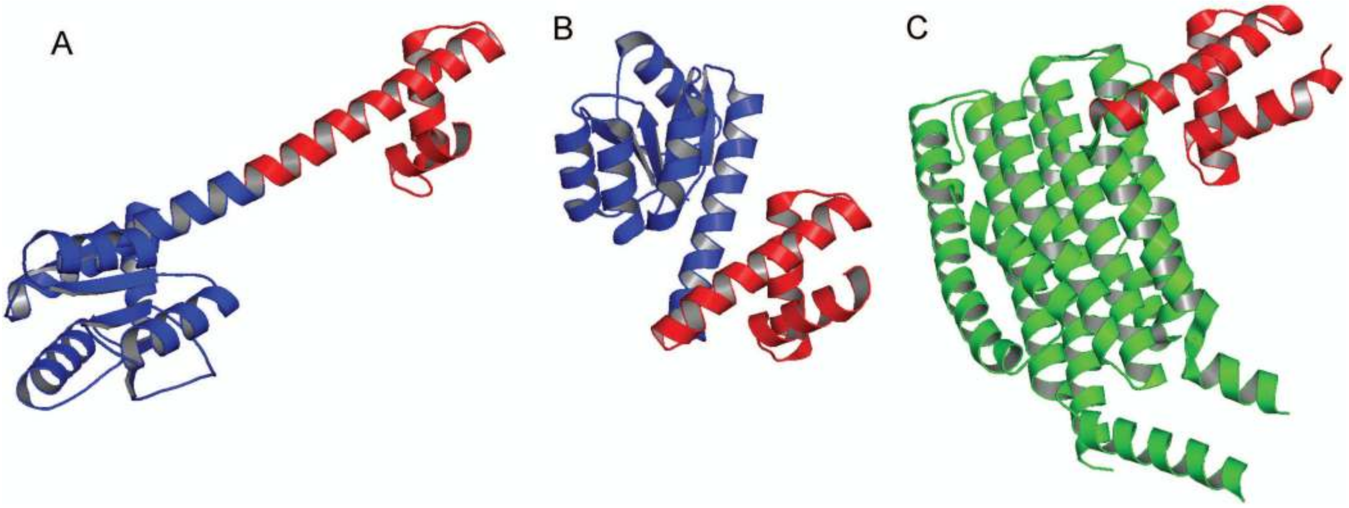

Figure 1.
ANTAR domains in different proteins. (A) RNA-binding positive regulator AmiR (1qo0, chain D). (B) Putative transcriptional antiterminator Rv1626 (1sd5, chain A). (C) NasR transcription antiterminator (4akk, chain A). C-terminal ANTAR domains are colored red, N-terminal Rossmann-like domains in AmiR and Rv1626 are colored blue, and N-terminal helical bundle domain in NasR is colored green. All structural diagrams in this manuscript are prepared with PyMOL (The PyMOL Molecular Graphics System, Schrödinger, LLC.).

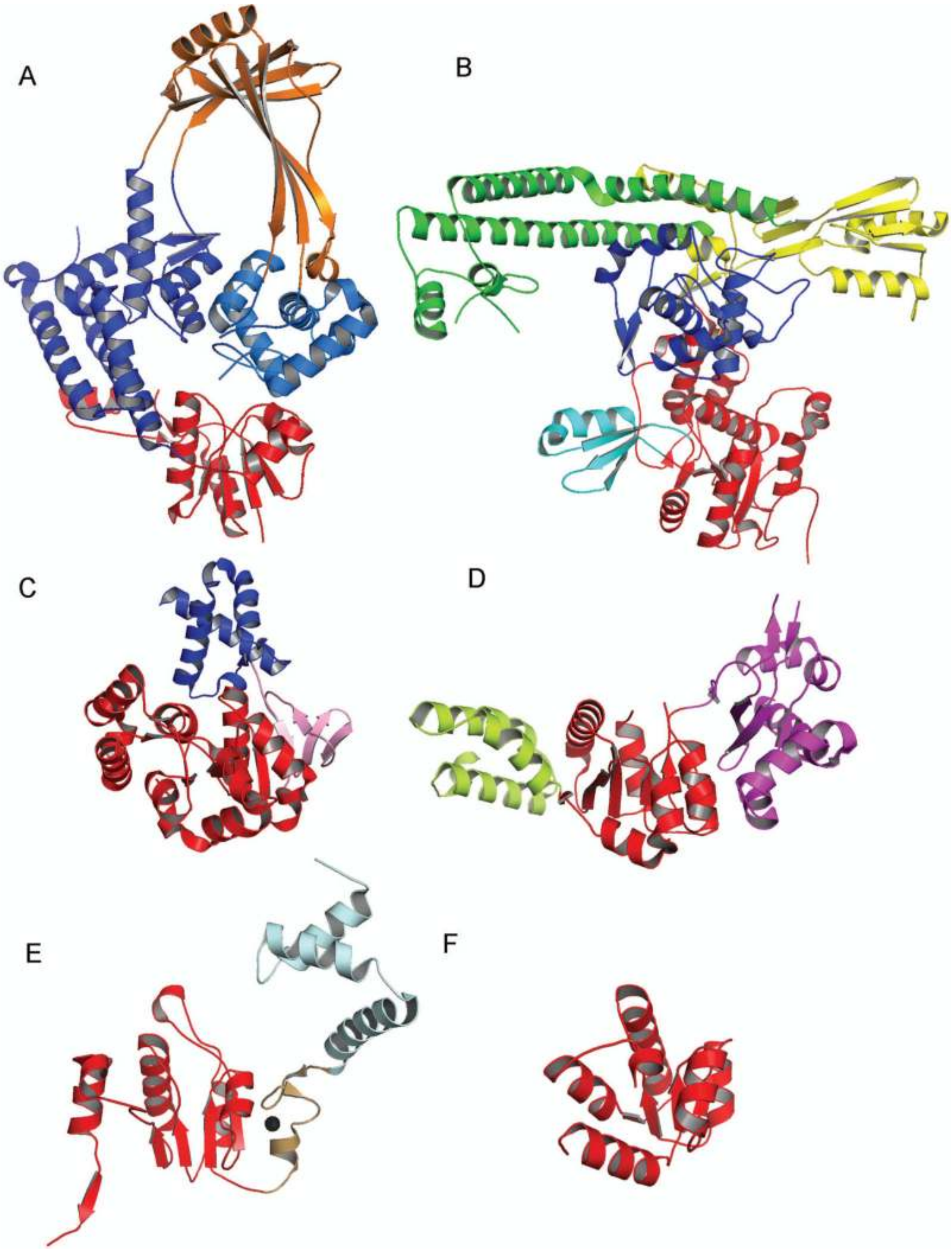

Figure 2.

Toprim domains in various proteins. Each panel shows one SCOP entry that contains a Toprim domain. Multi-domain entries are split into individual domains in ECOD as shown by different colors. In all diagrams, the Toprim domain is colored in red. (A) DNA topoisomerase I (PDB 1MW9, chain X[92]). SCOP entry d1mw9x_, in 'Multi-domain proteins' class and 'Prokaryotic typ I DNA topoisomerase' fold. (B) DNA topoisomerase 2 (PDB 2RGR, chain A[93]). SCOP entry d2rgra1, in 'Multi-domain proteins' class and 'Type II DNA topoisomerase' fold. (C) DNA topoisomerase VI A subunit (PDB 1D3Y, chain A[94]).

SCOP entry d1d3ya_, in ‘Multi-domain proteins’ class and ‘DNA topoisomerase IV, alpha subunit’ fold. (D) DnaG catalytic core (PDB 1DD9, chain A[95]). SCOP entry d1dd9a_, in ‘Multi-domain proteins’ class and ‘DNA primase core’ fold. (E) Recombinational repair protein RecR (PDB 1VDD, chain A[96]). SCOP entry d1vdda_, in ‘Multi-domain proteins’ class and ‘Recombination protein RecR’ fold. (F) Putative protein aq_2086 (PDB 1T6T, chain 1, unpublished). SCOP entry d1t6t1_, in ‘Alpha and beta proteins (a/b)’ class and ‘Toprim domain’ fold.

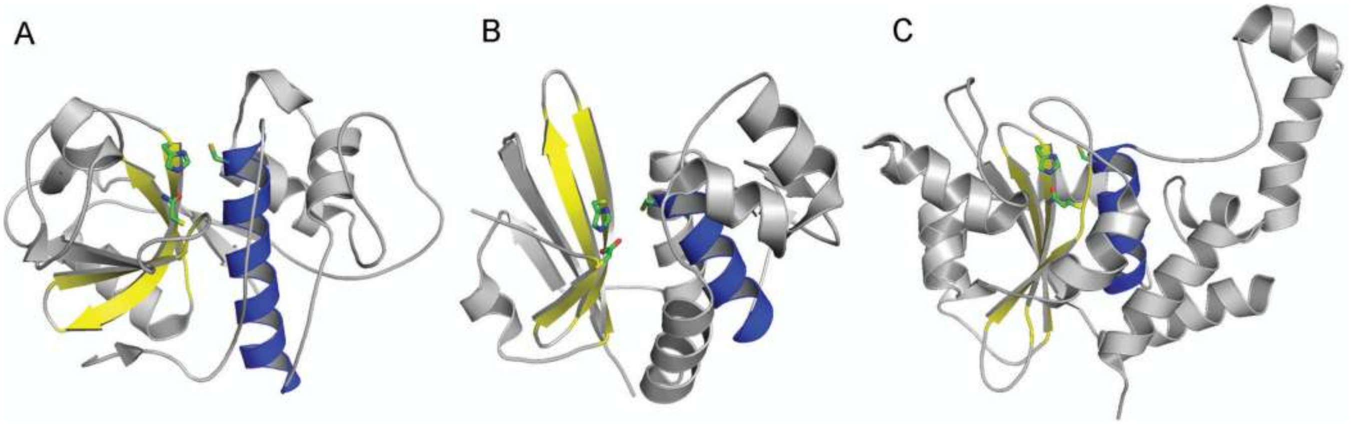

Figure 3.

Structural comparison of papain, Tae4, and Cif. (A): Papain (PDB 1KHQ, chain A)[97]. (B): Tae4 (PDB 4HFL, chain A)[31]. (C): Cif (PDB 3GQJ, chain A)[35]. In all three structures, the structural core of cysteine proteinase superfamily is colored, whereas the rest of the protein is in gray. The catalytic triad is identified according to the above references and shown in sticks. The alpha-helix bearing the Cys in the catalytic triad is colored in blue, and the beta-sheet bearing the His and the polar residue in the catalytic triad is colored in yellow.

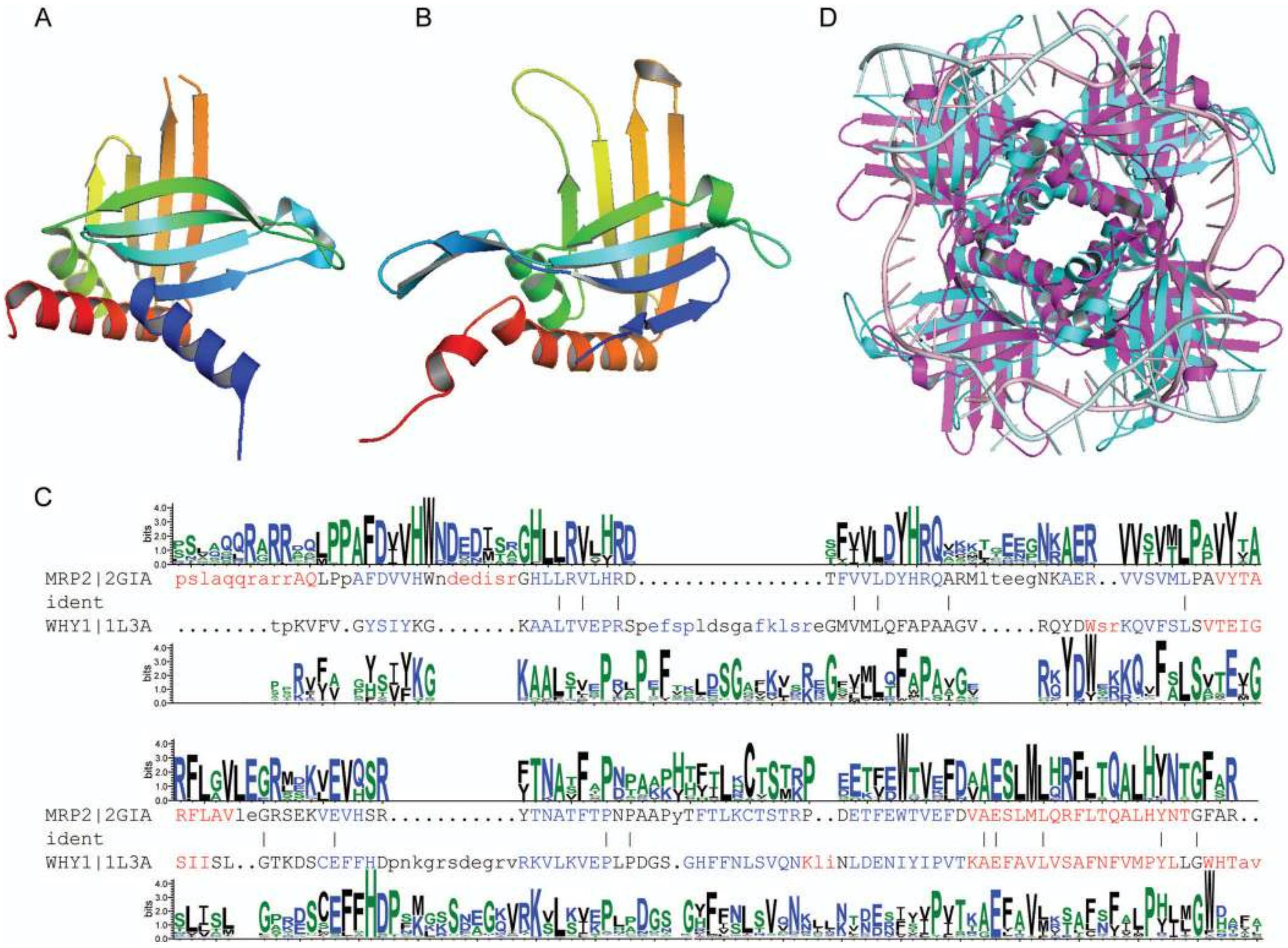
A
B
D
C
bits
MRP2|2GIA psIaqqrarrAQLPpAFDVVHWndedisrGHLLRVLHRD................TFVVLDYHRQARMlteegNKAER..VVSVMLPAVYTA
ident
WHY1|1L3A ....tpKVFV.GYSIYKG.......KAALTVEPRSpefspldsgafklsreGMVMLQFAPAAGV.....RQYDWsrKQVFSLSVTEIG
MRP2|2GIA RFLAVleGRSEKVEVHSR...........YTNATFTPNPAAPyTFTLKCTSTRP..DETFEWTVEFDVAESLMLQRFLTQALHYNTGFAR..
ident
WHY1|1L3A SIISL..GTKDSCEFFHDpnkgrsdegrvRKVLKVEPLPDGS.GHFFNLSVQNKliNLDENIYIPVTKAEFAVLVSAFNFVMPYLLGWHTav

Figure 4.

Structure and sequence comparisons of MRP and Whirly. (A) Structure of MRP2 (PDB 2GIA, chain A). (B) Structure of WHY1 (PDB 1L3A, chain A). Both structures are shown in cartoon and colored in a rainbow. (C) Dali structure alignment of MRP2 and WHY1. Residues are colored red for α-helices and blue for β-strands. Sequence profiles are represented by sequence logos generated from multiple sequence alignment of BLAST hits[98] by WebLogo[99]. (D) Superposition of MRP1/MRP2 (PDB 2GJE) and WHY2 (PDB 3N1K) tetramers with bound nucleic acids. MRP1 and MRP2 are colored cyan with RNA in pale cyan. WHY2 is colored magenta with DNA in light pink. Crystallography symmetry was applied to generate the biological units.

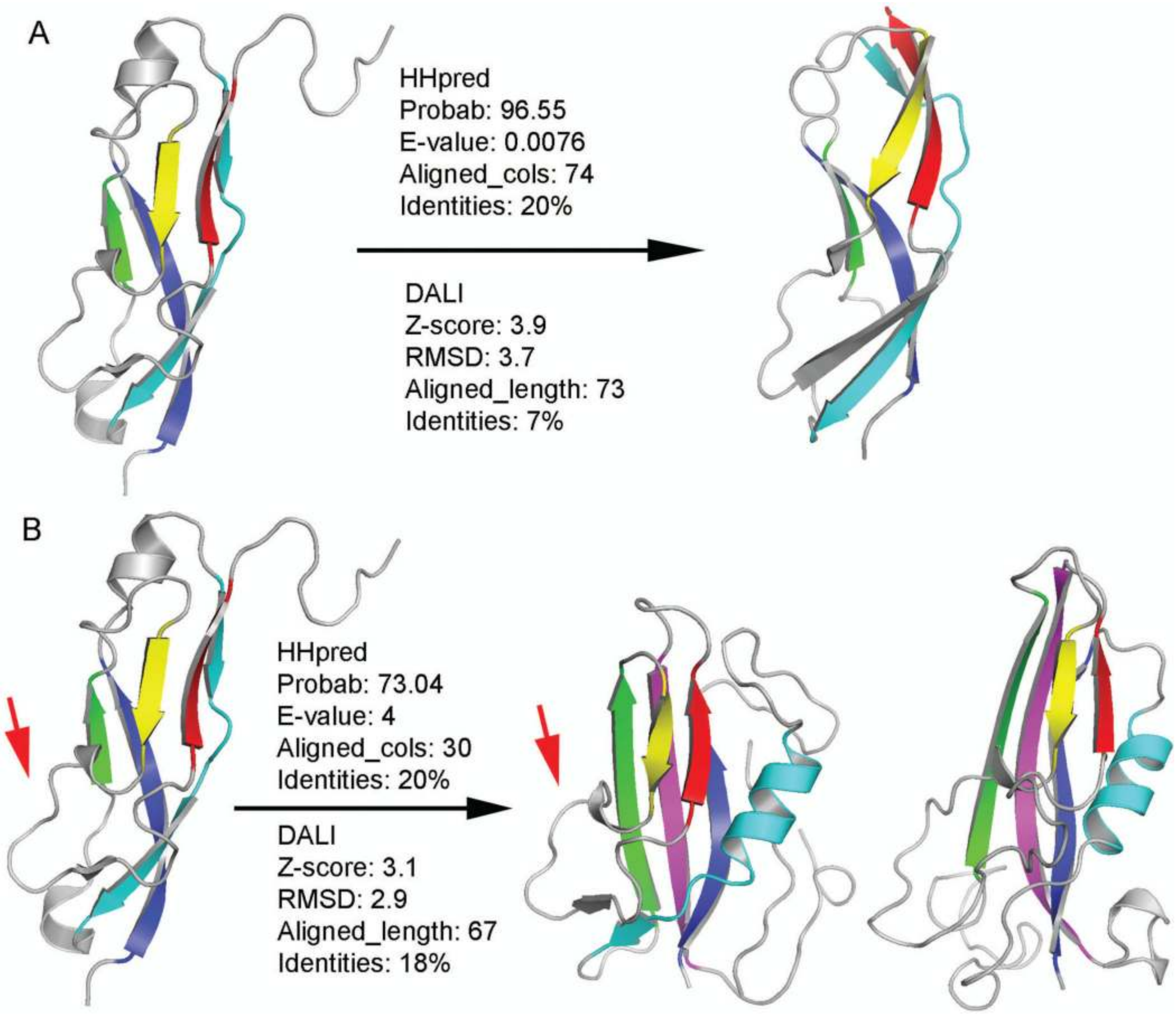

Figure 5.

Homology (A) and possible homology (B) for ribosomal protein TL5/L25 C-terminal domain. (A) Left: Ribosomal protein TL5/L25 C-terminal domain (PDB 1feu, chain A, residue 94–185). Right: YbbR domain (PDB 3lyw). (B) Left: Ribosomal protein TL5/L25 C-terminal domain from *T. thermophilus* (PDB 1feu, chain A, residue 94–185). Middle: RNA polymerase alpha subunit insert domain from *E. coli* (PDB 1bdf, chain C, residue 53–178). Right: RNA polymerase II subunit RPB3 insert domain from yeast (PDB 1twf, chain C, residue 42–172). Each structure in this figure is colored in a rainbow from N-terminus (blue) to C-terminus (red). Structurally corresponding SSEs in each domain are in the same color. The additional C-terminal strands in RNA polymerase insert domains are colored in magenta. Loops are colored gray. HHpred and DALI scores are shown with an arrowhead pointing from the query to the target.

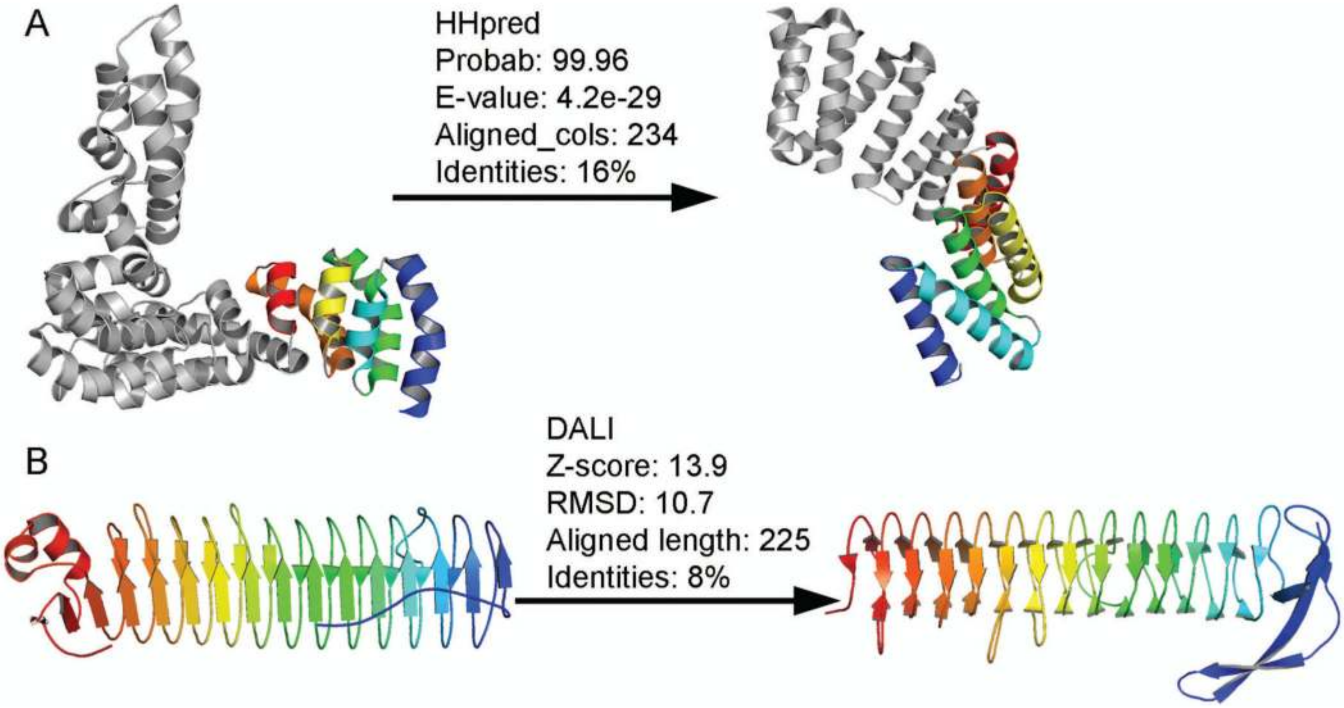

Figure 6.

High scores, opposite handedness. (A) Left: Transcription activator-like (TAL) repeats in dHax3 form a left-handed alpha-alpha superhelix (PDB 3v6p). Right: Tetratricopeptide repeats (TPR) in kinesin light chain 1 form a right-handed alpha-alpha superhelix (PDB 3nf1). In each structure, the N-terminal three alpha hairpins are colored in a rainbow from N-terminus (blue) to C-terminus (red). (B) Left: Antifreeze protein (AFP) is a right-handed beta-helix (PDB 3p4g). Right: Hypothetical protein YDCK is a left-handed beta-helix (PDB 2f9c). Each structure is shown in a rainbow from N-terminus (blue) to C-terminus (red). In both (A) and (B), comparison program and scores are shown above an arrowhead pointing from query to subject.

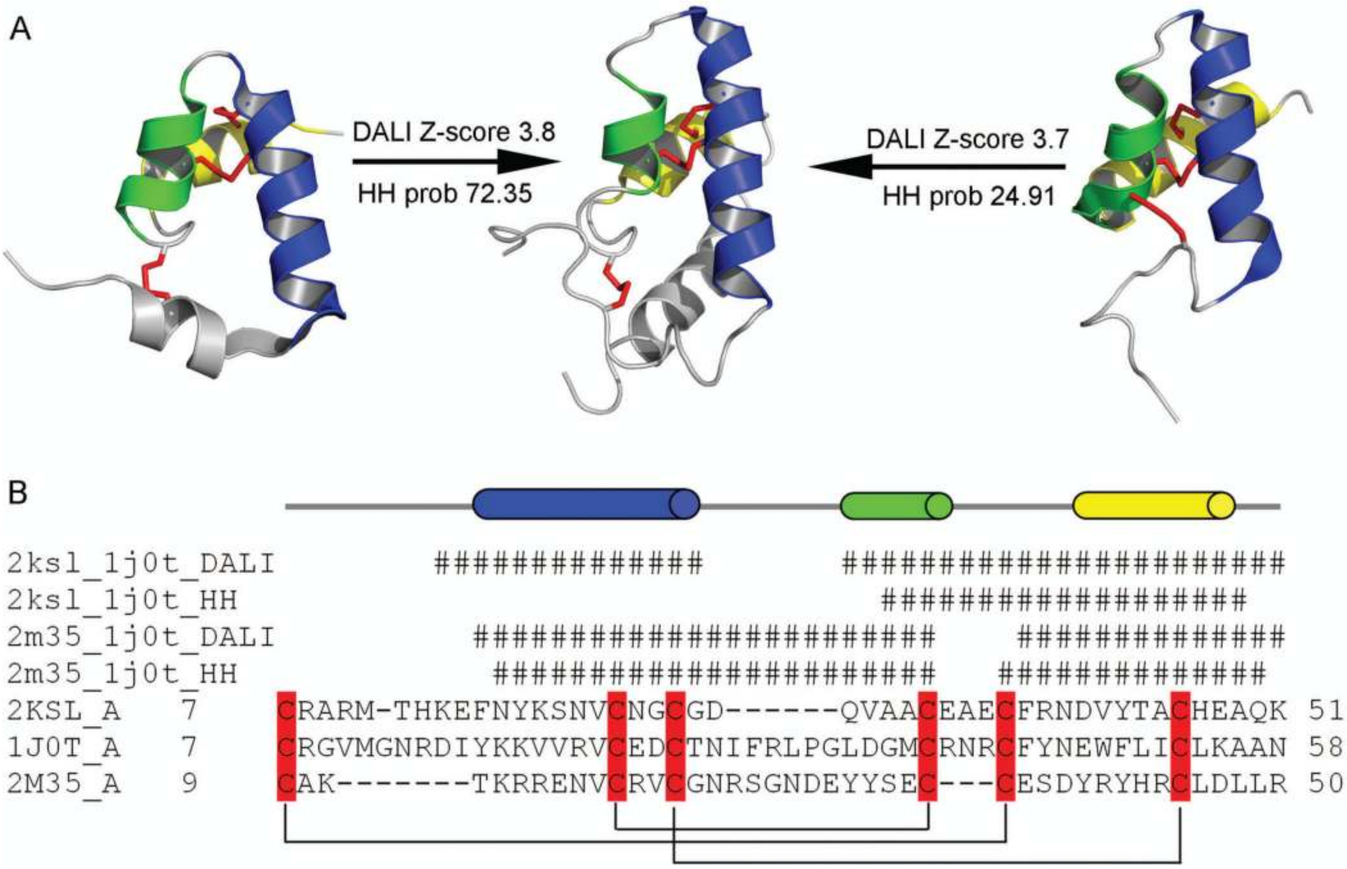
A
DALI Z-score 3.8
HH prob 72.35
DALI Z-score 3.7
HH prob 24.91
B
2ksl_1j0t_DALI ############## #######################
2ksl_1j0t_HH ###################
2m35_1j0t_DALI ######################## ##############
2m35_1j0t_HH ####################### ##############
2KSL_A 7 CRARM-THKEFNYKSNVCNGCGD------QVAACEAECFRNDVYTACHEAQK 51
1J0T_A 7 CRGVMGNRDIYKKVVRVCEDCTNIFRLPGLDGMCRNRCFYNEWFLICLKAAN 58
2M35_A 9 CAK-------TKRRENVCRVCGNRSGNDEYYSEC---CESDYRYHRCLDLLR 50

Figure 7.

Structure and sequence comparisons of CHH-like superfamily members. (A) Structures of spider toxin TaITX1 (left, PDB 2KSL), prawn MIH (middle, PDB 1J0T), and centipede neurotoxin k-Ssm1a (right, PDB 2M35). The three structures are first superimposed and then separated for clarity. The three structurally equivalent helices are colored in blue, green, and yellow, respectively, and other parts of the structures are colored in gray. Disulfide bonds are shown in sticks and colored in red. DALI z score and HHsearch probability between two proteins are shown with an arrow pointing from the query to the hit. (B) Manually made structure-based multiple sequence alignment of the three proteins in (A). Cysteines forming disulfide bonds are highlighted in red and connected by lines. PDB ID, chain ID, and starting and ending residue numbers are shown for each sequence. The approximate positions of the three structurally equivalent helices depicted in (A) are indicated by cylinders of the same color. Positions that are also aligned by HHsearch or DALI are marked by a '#' symbol with the query, the hit, and the aligner indicated at the beginning of each line.

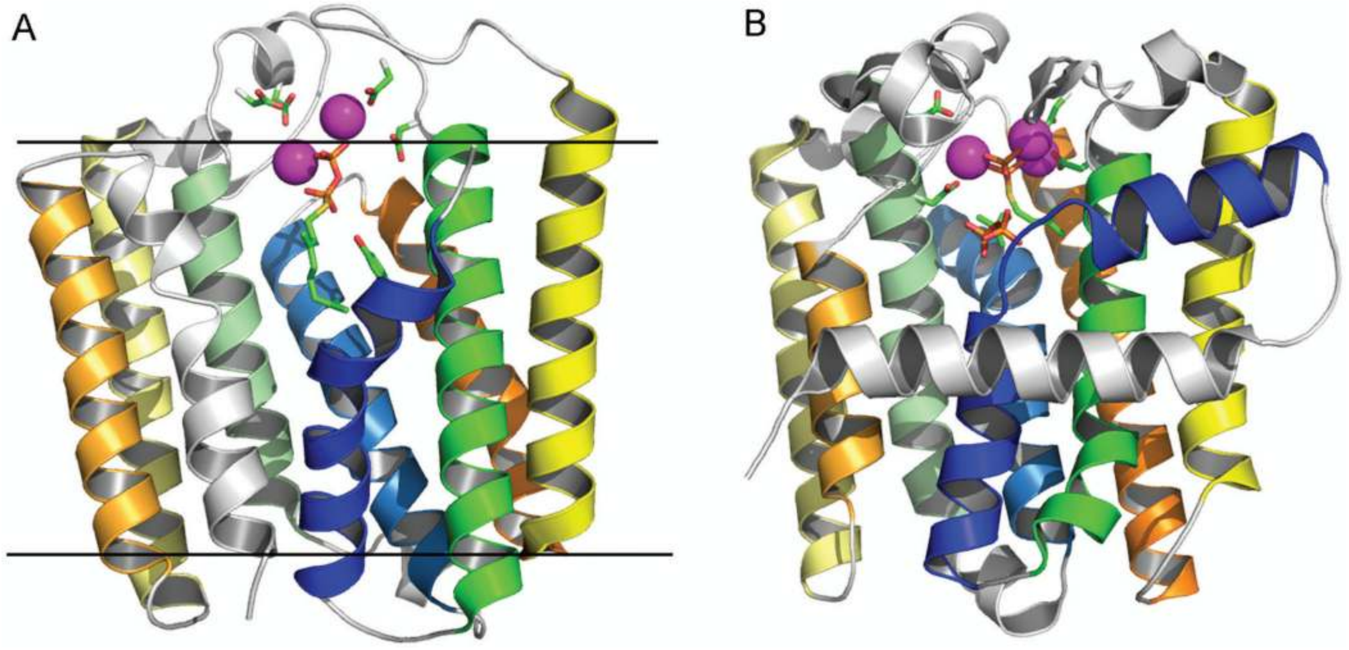

Figure 8.

Homology between soluble and intramembrane prenyltransferases (A) UbiA (PDB 4od5, chain A). (B) Farnesyl diphosphate synthase (FPS) (PDB 1rqi, chain A). Structurally correspondent helices in UbiA and FPS are in the same color, and extra helices are in gray. Both UbiA and FPS folds probably result from a duplication of a 4-helix bundle; thus the four helices in the first potential duplicate are colored in blue, green, yellow, and orange, respectively (from N-terminus to C-terminus), and the four helices in the second potential duplicate are colored in similar but pale shades. The aspartate side chains in the Asp-rich motifs are shown in sticks. Magnesium ions are depicted as magenta balls. Bound substrates are shown in sticks. Approximate positions of the membrane boundaries are indicated by two lines.