Published in final edited form as:

Proteins. 2020 November ; 88(11): 1513–1527. doi:10.1002/prot.25970.

β-Strand-mediated interactions of protein domains

Archana S. Bhat¹, Lisa N. Kinch², Nick V. Grishin^{1,2}

¹Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050

²Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050

Abstract

Protein domains exist by themselves or in combination with other domains to form complex multi-domain proteins. Defining domain boundaries in proteins is essential for understanding their evolution and function but is not trivial. More specifically, partitioning domains that interact by forming a single β -sheet is known to be particularly troublesome for automatic structure-based domain decomposition pipelines. Here, we study edge-to-edge β -strand interactions between domains in a protein chain, to help define the boundaries for some more difficult cases where a single β -sheet spanning over two domains gives an appearance of one. We give a number of examples where β -strands belonging to a single β -sheet do not belong to a single domain and highlight the difficulties of automatic domain parsers on these examples. This work can be used as a baseline for defining domain boundaries in homologous proteins or proteins with similar domain interactions in the future.

Keywords

Protein domains; Domain interactions; Domain boundaries; β -sheets and strands; Structural folds; Homology

Introduction:

The introduction of high-throughput, whole-genome sequencing has led to revolutionizing the study of protein evolution, structure, and function ¹. In order to exploit this whole-genome sequencing data, comparing new genomes to known examples of well annotated genes is essential. Identification of sequence homology to functional domains has long been used for purposes of annotation ^{2,3}. Because such domains frequently represent the functional units of proteins, their definition becomes an important task. Although there is no universally accepted definition for a protein domain, it can be considered as a distinct functional and/or structural mobile evolutionary unit ^{4,5}. Based on this definition and evidence from previous studies ^{6,7}, domains can exist in different combinations forming complex multi-domain proteins. One of the major challenges for partitioning these complex proteins into simpler components is determining the domain boundaries ⁸.

Corresponding authors: Archana S. Bhat – archana.bhat@UTSouthwestern.edu, Nick V. Grishin – grishin@chop.swmed.edu.

The two main and often conflicting criteria for defining domain boundaries are sequence continuity and compactness of structure. Maximizing sequence continuity by partitioning domains into the least number of fragments tends to reflect biologically meaningful boundaries. Because domains are mobile evolutionary units that are found in multiple gene contexts, their sequence continuity in an alignment with homologous proteins often suggests boundaries⁹. On the other hand, domain boundary definition must also consider compactness of structure, although this concept often leads to over fragmenting⁸. Structurally compact, or globular domains that fold independently, usually have a hydrophobic interior and hydrophilic exterior, suggesting that the part of the protein which belongs to the same domain has more intra-domain contacts than inter-domain contacts⁹. Such definition of domains as independently folding units with compact globular structures have been suggested before^{10,11}, with automatic methods for domain decomposition such as DomainParser¹², PUU¹³ and Protein Domain Parser (PDP)¹⁴, utilizing this concept with varying degrees of success⁸.

Given the difficulty of balancing sequence and structure in domain parsing, manual evaluation of domain partitioning remains necessary. Structure topology classifications such as CATH¹⁵, SCOP¹⁶ and ECOD¹⁷ provide a solid basis for such evaluation by establishing the functional and evolutionary relationships of existing protein structures. Specifically, ECOD (Evolutionary Classification Of Protein Domains) uses protein sequence and structure to classify domains into a hierarchy of architecture (A), possible homology (X), homology (H), topology (T), and family (F) groups. The ECOD definition of domain tends to represent a happy medium with respect to the mainly structural definition provided by CATH (tends to have smaller fragments) and the functional definition provided by SCOP (tends to have larger fragments)¹⁸. When accounting for topology, splitting β -class domains is particularly problematic for automatic domain parsing given the complex interfaces that can occur between two all- β domains or between all- β and α/β domains⁸. As such, understanding the roles β -strands play in protein domain interactions should help establish better rules for defining their bounds.

β -sheet interactions can mediate protein dimerization, protein-protein interaction, or even protein aggregation observed in human disease^{19,20}. While there are published studies involving protein-protein interactions and secondary structure elements (SSEs) involved in those interactions^{19–23}, lesser work has been done on studying the nature of SSEs in domain interactions^{24,25}. For example, the study by Guharoy et al. compares SSE contribution to protein homo and heterodimer interactions, finding a preference for regular SSEs such as strands and helices over non-regular loops in homodimer interfaces²¹. While examples of protein-protein interactions that form extended β -sheets have been identified^{20,21}, examples of inter-domain β -strand interactions contributing to the tertiary structure of multidomain proteins are not discussed. This presents the opportunity to study β -strand interactions between domains within the same protein subunit, with the intention to help define some difficult domain boundaries that can be missed by automated programs.

Identification and partition of individual domains sharing a β -sheet can be particularly tricky due to the increased contacts between the residues within β -strands forming the β -sheet, giving the appearance of a single domain. In this article, we show that, although rare, all

strands of a continuous β -sheet do not necessarily belong to the same domain. Individual protein domains that interact by forming a continuous β -sheet occur in superfamilies with common folds as well as in less common/fast-evolving proteins. Furthermore, one possible mechanism of continuous β -sheet formation by individual domains could involve replacement of a β -strand from a known fold topology with a β -strand from the interacting domain, to create a stable protein structure. Understanding such structural attributes of β -strands involved in domain interactions can help improve establishment of domain boundaries for homologous protein structures that are solved in the future.

Methods:

A bioinformatics pipeline was developed to obtain all β -strand mediated edge-to-edge domain interactions. Firstly, ECOD¹⁷ was used as a reference for domain classifications and boundaries. ECOD defines domains based on evolutionary relationships and classifies them by architecture, X, H, T and F groups, as described in the introduction. The ECOD database is updated weekly and distributable files are provided. The distributable file with domain definition is called “ecod.latest.domains.txt” that can be obtained from ECOD download page (ECOD develop239, updated on June 21, 2019). Using the domain, chain and sequence range information in this file, a list of all protein subunits with multiple domains was compiled. This list consisted of the PDB ID of the protein subunits with more than one domain, along with the statistic of number of domains that were a part of that protein subunit. Following this, with Biopython version 1.7.2²⁶, PDB (Protein Data Bank)²⁷ files corresponding to these chains were retrieved, and the specific multi-domain subunits were extracted. Each chain was stored as an individual PDB format file.

Each PDB format file was then run with DSSP²⁸ (updated CMBI version by ElmK / April 1, 2000) to calculate secondary structure. The DSSP output was parsed for secondary structure information and presence of β -bridges, which are adjacent β -strands. For the purpose of this project, more than three consecutive residues assigned “E” by DSSP (i.e., residues characterized as β -strands based on H-bonding pattern), were considered as β -strands. Further, if a β -strand with more than three residues was found to interact with another β -strand, parallel or anti-parallel, it was recorded and kept track of. Using the β -strand/bridges information from DSSP and domain boundaries as defined by ECOD, all β -strand interactions that occurred between domains within the same chain were identified. The hierarchical ECOD classification for these domain pairs was obtained by mapping to the domain definition file mentioned above. Since a β -strand interaction between duplicate domains are easier to define boundaries with, only domain pairs with different topologies (ECOD T-groups) were analyzed. Lastly, these domain pairs were made non-redundant by grouping based on ECOD T-group classification. All T-group pairs were manually studied to identify the type of interaction between the two domains.

Homologs for the interacting domains that exist by themselves or in association with other domains were found using structure-based trees in ECOD T-groups²⁹. The structure-based tree lists all families within the T-group and clusters them based on structure similarity, using which, homologs with differing domain architectures were found. For cases where homologs were not found within the same T-group, DALI structure alignments³⁰ were used

to find possible homologs from PDB, and ECOD classification was used to verify domain combinations. Domain architecture and domain-interactions from the homologous proteins were used as a resource to justify the domain boundaries in the proteins sharing a β -sheet. In line with the goal of this study, all instances where a β -sheet is divided among multiple interacting domains were recorded. For these examples, partitioning into domains using methods that assess the globular structure might not work well due to the increased contacts within the β -strands that form the β -sheet. To test this hypothesis, we compared our evolutionary-based domain definitions to those partitions provided by PDP (2004 version). The same PDB format file that was used as input for getting secondary structure annotation from DSSP, was used as input for PDP as well. The domain partition results from PDP was recorded and compared to ECOD domain definitions and boundaries.

Results:

Proteins frequently consist of more than one domain. Although such domains can be defined structurally as independent folding units with compact globular structure^{10,11}, the evolutionary pressure to maintain primary sequence continuity often contradicts simple structural domain partitions⁹. The goal of this study is to analyze a particularly difficult case of domain partitioning where multidomain proteins form inter-domain β -strand interactions. We utilize evolutionary-based domain definitions from ECOD with secondary structure predictions by DSSP to identify protein domains that interact by forming a continuous β -sheet and compare these domains to those produced by a structural domain parser.

Dataset, filtering and statistics:

Domain boundaries as defined by ECOD were used as an initial dataset to create a list of all multi-domain structures in the Protein Data Bank (PDB). This dataset included 152,729 PDB chains, and 372,185 ECOD domains (Fig 1a). A total of 133,920 chains, and 328,152 ECOD domains were extracted for further study (Fig 1a), in the form of PDB format structure files. The remaining 18,809 chains were excluded from the analysis due to very large PDB file size (>62 chains and/or 99999 ATOM lines), or discrepancies between the ECOD and PDB databases (ECOD keeps obsolete PDBs). Every multi-domain structure was assigned secondary structure annotations using DSSP to identify the β -strands that interact with another β -strand to form a β -bridge. Using the secondary structure and β -bridge annotation from DSSP, those β -strands that form inter-domain β -bridges were identified using ECOD domain definitions.

Once we define interacting β -strands (length >3 residues) between different ECOD domains within the same protein subunit, the 133,920 chains become 17,858 domain pairs from 13,833 chains (Fig 1b). The discrepancy between number of chains and domains arises due to multi-domain proteins having more than two domains interacting through β -strands. Further, interactions between identical, homologous domains were filtered as β -sheets formed from duplicated domains are easier to define. For example, the protein structure of Methylmalonyl-CoA epimerase (MMCE)³¹ includes duplicate domains interacting within the protein chain to form a β -sheet (Fig 2a and 2b). This protein structure is composed of four identical chains, with two $\beta\alpha\beta\beta$ domains (Fig 2b), that are a result of evolutionary

events like gene duplication, fusion and domain swapping³¹. The N-terminal β -strands of the two domains interact and extend the sheet into both the domains. The repeating $\beta\alpha\beta\beta$ domains that form inter-domain β -sheets are relatively easy to define given their identical topologies. Removing duplication examples like these reduced the dataset to 6325 domain pairs/5105 chains (Fig 1b).

To reduce manual inspection of inter-domain β -sheets in the dataset, ECOD classification was used to reduce redundancy. All domain-pairs with the same ECOD topologies (T-groups) were grouped together. A total of 340 unique ECOD “topology-pairs” were found after grouping (Supplementary Table I). One random representative from each group was manually analyzed for the type of β -strand interaction between domains. Manual inspection of the 340 representatives revealed that 86/340 constituted domain interactions that are mediated through β -strands (Supplementary Table I). The remaining ECOD classification pairs mainly resulted from poorly defined domain boundaries in ECOD (Supplementary Table I), arising from automated domain assignments that are heavily based on sequence continuity rather than globular structure of the protein. The 86 ECOD classification pairs with genuine inter-domain interactions occurring via a β -strand were further analyzed, to manually determine the type of β -strand interaction (Supplementary Table I).

More often, among the 86 representatives, the edge-to-edge interaction of β -strands between domains occur due to N/C-terminal extensions or insertions to evolutionarily defined domains, in the form of a β -hairpin or β -strand (Fig 2b, Supplementary Table I). These constituted of 70/86 cases. As seen in Fig 2c, the structure of N-acylamino acid racemase consists of a TIM-barrel domain with a C-terminal β -strand insertion. This inserted β -strand interacts with a β -sheet in the $\alpha+\beta$ domain and extends the overall β -sheet by one β -strand. Since such cases involve only a β -hairpin/strand in one or both domains, which does not form a β -sheet with more than two strands, these cases were recorded and not considered for further analysis. Notably, due to the sequence continuity between the TIM-barrel domain and C-terminal strand extension, ECOD automatic domain assignment works well to partition the domains correctly. However, without sequence continuity, there is an increased chance of domain boundary errors in automatic pipelines since the domain boundary can easily get extended by one β -strand/hairpin that does not involve many residues. This type of profile corruption resulting from iterative multiple sequence alignments and profile searches have been reported before^{32–34}, where domain boundaries are extended into non-homologous sequence regions.

For example, in Penicillin-binding protein 2X (PBP2X), a Profilin-like domain is inserted with a small α -helical domain (Fig 2d). As the automatic pipeline in ECOD is reliant on sequence continuity, this protein subunit is incorrectly partitioned (Fig 2e). Although the five stranded antiparallel sheet belongs to the Profilin-like domain, the discontinuous sequence causes the automatic ECOD pipeline to incorrectly define the protein subunit as two domains, divided between the N and C-terminal halves of the protein (Fig 2e). Our pipeline to detect inter-domain β -strand interactions helped recognize and fix such domain boundary issues in ECOD. Excluding the cases with ECOD domain boundary errors arising from the automatic pipeline, and domain interactions between a β -hairpin/strand in one or both domains, we found 16 representatives where two domains share a β -sheet, with more than

two β -strands in each interacting domain (Table I). Selective cases that embody a broad representation of evolutionary relationships are discussed in the following sections. While the examples where the interacting domains belonging to large super families suggests that these are genuine domains that occur as independently functioning units, some of the other examples discussed also aim to highlight structural attributes that could aid in the formation of stable interactions between the domains that share a β -sheet.

Evaluating domain boundary definition by PDP:

For the proteins that share a β -sheet between domains, defining domain boundaries using methods that utilize the globular structure might not work well due to the increased contacts within the β -strands that form the β -sheet. To test this hypothesis, the PDP tool was used on all 16 representative protein subunits that were found to involve a domain interaction mediated by a single continuous β -sheet, as annotated by the evolutionary-based domain definition by ECOD. PDP uses the globular structure of proteins to partition it into functional domains. Additionally, PDP is also used within the automatic pipeline in ECOD¹⁷, to eliminate small gaps between partitioned domains. Using PDP in this study, among the 16 representative protein subunits, 8 were correctly partitioned into domains (Supplementary Table 2). In the correctly defined cases, the presence of some evident structural features could be considered as the basis for domain partition by tools like PDP that use protein globularity and structure to define domains.

For instance, in Yerb protein (Table I, Supplementary Table 2, Fig 3a), β -strands that are involved in the interaction between the domains (orange and dark blue β -strands in Fig 3a) appear to diverge from each other within the β -sheet, forming a “V” shape between the integrating strands. While only four residues are annotated as a part of the inter-domain β -strand interaction by DSSP, six residues interact between β -strands within the same domain (purple and blue β -strands in Fig 3a). Similarly, in the CRP transcription regulator protein (Table I, Supplementary Table 2, Fig 3b), although the β -strands of the HTH and jelly-roll domains interact, the β -strands in the jelly-roll domain are much longer, and are pulled away from the HTH domain to complete the core of the jelly-roll domain. In this instance as well, four residues are involved in the inter-domain β -strand interaction, which is lesser than six residues that interact between the β -strands in jelly roll domain (orange and salmon colored β -strands in Fig 3b). In such instances, the β -strands appear to be pulled towards the domain they are a part of, which could help define the domain boundaries in these proteins.

While the presence of structural features such as these help in defining domain boundaries for some of these proteins, there are instances where similar structural features mislead the domain parser towards incorrect definition of domain boundaries. For instance, in prephenate dehydratase (Table I, Supplementary Table 2, Fig 3c and 3d), β 1 and β 4 (dark blue β -strand in Fig 3c) of the C-terminal ACT-like domain are extended and pulled towards the β -strands in the N-terminal Rossmann-like domain (orange β -strand in Fig 3c). Likely due to the direction of these strands, PDP incorrectly classifies most of the residues in β 1 and β 4 of the ACT-like domain, as a part of the Rossmann-like domain (Fig 3d). The ACT-like domain that is known to fold with the $\beta\alpha\beta\beta\alpha\beta$ topology³⁵, is incorrectly defined by PDP without β 1 and β 4, that are a part of the ACT-like domain fold definition. Overall

however, PDP correctly divides the protein into two domains (Fig 3d), although the boundaries are not precisely defined within $\beta 1$ and $\beta 4$.

On the other hand, when the β -strands within the β -sheet are straight and continuous without evident structural features to distinguish domains, PDP domain boundaries are incorrect and the β -sheet that is shared between two evolutionary domains is annotated to be a part of a single domain (Table I, Supplementary Table 2, Fig 3e, 3f, 3g and 3h). As seen in the structure of Exoribonuclease 2 (XRN2), the N-terminal haloalkanoic acid dehalogenase (HAD) domain and a C-terminal $\alpha + \beta$ domain form a continuous β -sheet (orange and dark blue β -strands in Fig 3e), with a helical domain inserted between the N and C-terminal domains (gray domain in Fig 3e). However, PDP not only fails to partition the N and C-terminal domains (Fig 3f), but also considers most of the helical domain to be a part of one single domain that includes terminal domains (Fig 3f). Lastly, in some instances, the whole protein subunit is annotated as a single domain (Fig 3g and 3h). While ECOD divides the DUF2233 family protein into four functional domains (Fig 3g), PDP considers the whole protein as a single domain (Fig 3h). These incorrect domain definitions highlight the need to study these proteins at a deeper level in order to achieve the right domain annotations and boundaries.

Domains with common structural folds sharing a β -sheet:

Domains that function as independent folding units can exhibit mobility by occurring in diverse architectures. Those that exist as common structural folds with many evolutionary relationships (here on referred to as common structural folds, Table I), with multiple examples of differing architectures, provide the clearest examples of independent domain interaction by a shared β -sheet. Glycinamide ribonucleotide synthetase (PurD) is an enzyme involved in the purine biosynthetic pathway and is responsible for the formation of glycinamide ribonucleotide (GAR) using phosphoribosylamine, glycine and ATP³⁶. The two subunits of this protein contain 3 domains, an N-terminal Rossmann-like PreATP-grasp domain, a middle ATP-grasp domain that binds the nucleotide ligand, and a C-terminal $\alpha + \beta$ hammerhead domain (Fig 4a). The ATP-grasp and C-terminal hammerhead domains interact to form a continuous β -sheet (Fig 4a), both of which are common structural folds occurring within protein structures of various different architectures^{37–39}.

The ATP-grasp domain represents a combination of two $\alpha + \beta$ subdomains that “grasps” the ATP molecule in between (Fig 4a). While these subdomains appear as independent structural units forming two β -sheets, they function together harboring three conserved motifs corresponding to the metal-binding site and phosphate-binding loop^{37,40}. While the ATP-grasp domain and hammerhead domain combination are observed in other proteins from the carboxylate-amine/thiol ligase superfamily, like pyruvate carboxylase, glycinamide ribonuclease synthetase, carbamoyl-phosphate synthase, etc., the hammerhead domain is not present in most proteins from the D-Ala-D-X ligase family. For example, in D-Alanine-D-Lactate ligase⁴¹ (Fig 4b), the ATP-grasp domain that is in the same family as PurD, is preceded by a Pre-ATP grasp domain but lacks the hammerhead domain. The β -strand of the ATP-grasp domain in PurD that interacts with the hammerhead domain is comparatively longer than the corresponding β -strand in the homolog and is shifted towards the C-terminal

helix of the ATP-grasp domain (Fig 4a and 4b). This noticeable difference in structure of the PurD interacting β -strand could be a result of evolution where changes in conformation occur for favorable interactions with another domain⁴². However, taken together, these observations suggest that the ATP-grasp domain exists as a mobile evolutionary unit with conserved structure and function, with or without the hammerhead domain.

Domains with the β -hammerhead motif adopt a unique conformation that resembles a hammer, formed from two antiparallel strands in elongated loops. This motif was first noticed in acetyl-coenzyme A carboxylase, biotin carboxyl carrier protein (BCCP) subunit where its sequence and structure conservation were reported⁴³. Various proteins adopting different structural folds with the motif have been observed, including the $\alpha + \beta$ hammerhead³⁹, as seen in PurD. The presence of this domain in multiple proteins with different structural folds and architectures, indicates that domains with the hammerhead motif are also independently evolving protein domains. The hammerhead motif in PurD is sometimes deteriorated, as seen in N5-carboxyaminoimidazole ribonucleotide synthetase (PurK; PDB 1B6S), where a similar interaction occurs between the ATP-grasp and hammerhead domains. Despite this deterioration, the PurD and PurK hammerhead domains are homologous to that from 60S ribosomal protein (Fig 4c), as classified by ECOD, which includes the characteristic motif (Fig 4c, purple). In the 60S ribosomal protein, however, the $\alpha + \beta$ hammerhead domain occurs alone without the ATP-grasp domain, and is known to be a conserved fold within the ribosomal L16p/L10e family⁴⁴. Thus, PurD and related structures combine independent domains through the formation of a single β -sheet.

Another example of β -sheet domain interaction between domains from common structural folds is noticed in InternalinJ (InIJ). Internalins are exclusive *Listeria* species proteins that contain a leucine-rich repeat (LRR) domain with repetitions of 20–22 amino acids forming a solenoid shaped structure⁴⁵. InIJ in *Listeria monocytogenes* is a virulence-associated surface-protein with the hydrophobic cores of the N-terminal and C-terminal LRR repeats being shielded from the solvent by an α -helical EF-hand domain and a C-terminal Immunoglobulin-like (Ig) domain, respectively⁴⁶. The LRR domain and Ig-like domain together form a continuous β -sheet (Fig 5a). A similar domain interaction has been observed and analyzed by Schubert et. al for other internalin structures (InIB and InIA)⁴⁷, where they suggest that extended β -sheet formed by the interaction of the LRR and Ig domain provides a concave interaction surface for binding to the host cells during infection.

The large LRR domain superfamily consists of thousands of proteins with a wide array of functions involving protein-protein and protein-ligand interactions⁴⁸. One such protein is Phytosulfokine (PSK), which plays an important role in plant growth and development. It contains multiple α/β horseshoe fold LRR domains that are homologous to the LRR domain in InIJ (Fig 5b). The corresponding LRR domains belong to the same ECOD T-group (Table I), but in PSK, the LRR domain exists without the capping Ig domain. The LRR domain is also found with other domains in protein families like malectins (carbohydrate binding protein of endoplasmic reticulum), pleckstrin homology (PH-domain) and kinases, highlighting its existence as an independent functional unit. The Ig domain in InIJ also represents an independently folding unit, as it is homologous to Ig domains that exist alone and in combination with other domains, such as the Ig domain in Gingipain R (RGPB).

RGPB is a cysteine proteinase that acts as a virulence factor from *Porphyromonas gingivalis*⁴⁹. The C-terminal Ig domain of RGPB (Fig 5c) is homologous to InIJ as the two Ig domains are classified within the same ECOD T-group (Table I). However, RGPB contains 2 other Rossmann-like domains, an N-terminal alpha/beta domain followed by a Caspase-like catalytic domain, without the presence of an LRR domain as in InIJ. Interestingly, both InIJ and RGPB are virulence-associated proteins with an Ig domain. As Ig domains have previously been reported to play a role in cell-adhesion during infection^{47,50}, the Ig domains in these virulence-associated proteins may be seen combining with other domains for performing similar functions.

Moreover, focusing on the structural features of the β -strand inter-domain interaction, the last strand in the LRR domain appears to be shared with the first strand of the Ig domain (Fig 5a). The structure core common to all related Ig domains adopts a Greek-key sandwich having seven strands in two sheets, with some additional β -strands seen in many cases⁵¹. The Ig domain in InIJ shares its N-terminal core β -strand with the C-terminal β -strand of the LRR domain, making delineation of the domain bounds difficult. This sharing is more clearly visible in the tube representation of the two C-terminal repeats (Fig 5a, circled) where the first half of the β -strand includes the characteristic bulge of the LRR repeat, and the second half of the β -strand extends into the Ig-domain. Compared to the corresponding strand from the homologous Ig domain in RGPB (small orange strand in Fig 5c), the strand is longer, likely to allow for the interaction of the two domains in InIJ and for completing the repeat region of the LRR domain. Also, this N-terminal strand in InIJ is laterally displaced adjacent to the last β -strand in the Ig domain as opposed to its typical interacting with the opposite sheet in RGPB and other Ig domains. Such positioning of the β -strand is observed in LRR interacting InIB and InIA as well⁴⁷.

Absence of a β -bulge mediates interactions:

As mentioned in the introduction, protein aggregation can be a result of β -sheet formation. The CRP family transcription regulator proteins (CRP) provides a good example for avoiding such β -sheet formations through a β -bulge. CRP family proteins are comprised of two tandem domains, one of which has domain interactions that form a continuous β -sheet (Fig 6a). Both the N-terminal cNMP binding jelly-roll domain and the winged helix-turn-helix (HTH) domains represent common structural folds among protein structures. The cNMP binding domain belongs to a class of domains that bind to cyclic nucleotides like cAMP or cGMP, while the HTH domain mostly serves as a DNA binding domain. Proteins containing this domain combination serve as a versatile group of CRP/FNR regulatory transcription factors⁵². The jelly-roll and HTH motif domains also exist in other proteins, in combination with other domains, justifying their decomposition into domains.

The jelly-roll domain is described as a double-stranded β -helix formed by two four-stranded antiparallel β -sheets, which together pack a hydrophobic interface⁵³, and their presence has been noted in a wide range of proteins⁵⁴. Furthermore, the CRP jelly-roll domain is very similar to the nucleotide binding jelly-roll domain in the potassium channel of *Mesorhizobium loti* (Fig 6b). The corresponding jelly-roll domains belong to the same T-group in ECOD (Table I) and the β -strands that form the core of the jelly-roll domain

superimpose well (Z-score 12.8). The presence of homologous jelly-roll domains in the potassium channel of *Mesorhizobium loti*, without an interacting HTH domain, implies that these jelly-roll domains exist by themselves with independent nucleotide binding functions. The HTH motif is found in three super-kingdoms of life, to mainly function as the DNA-binding domain via the third α -helix⁵⁵. Two, winged-HTH domains exist in plasmid protein TubR as an intertwined dimer⁵⁶, and the HTH domains in TubR (Fig 6c) and CRP are homologs as they are classified under the same ECOD T-group (Table I). Structurally, the TubR HTH domain consists of some additional insertions in the form of α -helices, but the HTH motif region is very similar to the corresponding HTH domain in CRP.

While all the CRP/FNR transcription factors include the same tandem domain combination, the domains do not necessarily interact in the same orientations, further hinting at these domains being independent mobile units. The HTH domain of more typical CRP/FNR family structures interacts with the cNMP extended β -hairpin using the C-terminus of the first HTH motif helix. Such an interaction results in an approximate 180° rotation of the HTH domain with respect to the cNMP domain (Fig 6d). The unusual orientation forming the continuous β -sheet with the HTH domain (Fig 6a) has the β -hairpin extending slightly away from the core, compared to the orientation lacking the inter-domain β -sheet formation. Interestingly, while the typical orientation of the HTH domain interaction with the jelly-roll domain lacks the inter-domain β -sheet formation, the corresponding β -strand in the jelly-roll domain that makes the interaction with HTH domain in the typical orientation, consists of a β -bulge (Fig 6d and 6f). On the other hand, the unusual orientation that is forming the β -sheet (Fig 6a and 6e) between the two domains lacks the β -bulge. Such β -bulges are known to be present in natural proteins to avoid edge-to-edge β -sheet aggregation and amyloid formation⁵⁷, which in this case too, may be present to avoid an edge-to-edge interaction with another domain. The unusual orientation forming a β -sheet with the HTH domain, which lacks the β -bulge, may not be a favorable orientation for the protein.

β -sheet domain interactions with replaced β -strands from less common folds:

A second category of shared β -sheet interactions involves clear mobile domains from large superfamilies interacting with less common domains that appear as independent folding units, but do exist alone or in alternate topologies exemplified in existing structures. Interestingly, finding possible homologs for the interacting domains from less common folds suggested that the interaction could replace or compensate for a β -strand included in the definition of a common fold. As an example, the *Caenorhabditis elegans* Exoribonuclease 2 (XRN2) functions in RNA metabolism⁵⁸. The XRN2 subunit in the crystallized protein structure contains 3 domains, an N-terminal haloalkanoic acid dehalogenase (HAD) domain, a middle helical domain, and a C-terminal $\alpha + \beta$ domain. A continuous β -sheet forms between the N-terminal HAD and the less common C-terminal domain fold (Fig 7a).

The core of XRN2 is formed by the N-terminal HAD-like domain and the middle helical domain, which together encloses the nucleotide binding site at the center, as seen in homologous protein XRN1-substrate complex from *Drosophila melanogaster* (PDB 2Y35)⁵⁹. While the XRN2 HAD domain is shown to possess 5'→3' exonuclease activity⁵⁸, proteins in the HAD superfamily display varied cellular activities but are broadly considered

as catalysts for phosphate ester hydrolysis in metabolic pathways⁶⁰. The presence of the HAD domain across all superkingdoms of life, along with its variability in structural folds and conformations⁶⁰, implies the prominence and functional importance of this domain in proteins. The HAD domain exhibits promiscuity, as it also exists together with other domains such as the SAM-like domain, as in the case of protein T4 RNase H (Fig 7b). Overall, the core HAD domain structure is described as a five-stranded parallel sheet with repeating β - α units, which adopt the Rossmannoid topology⁶¹. The HAD domains in XRN2 and the homolog in T4 RNase H adopt similar topologies and are classified under the same T-group in ECOD (Table I). Although the HAD domain in XRN2 has additional secondary structure that decorate the core fold, like the N-terminal extension with a β -strand and a much longer α -helix following P3, the rest of the structure closely resembles the HAD domain in T4 RNase H. Also, in this case of a shared β -sheet, the β -strand making the interaction with the C-terminal domain is longer compared to corresponding strand in the homolog, suggesting fold adaptability in the HAD domain to make this interaction possible.

The C-terminal $\alpha + \beta$ subunit of XRN2 can be considered as a domain, due to its globular nature, although it does not have homologs in the PDB. The first best structural hit with DALI structure alignment against this C-terminal domain was a deteriorated fragment from a homologous exoribonuclease (PDB 3FQD, Z-score 3.9), followed by a PAS domain hit (PDB 4L9E, Z-score 3.1). In general, the PAS domain is a mobile unit present by itself or in other enzymes like diguanylate cyclase⁶², histidine kinase⁶³, etc., that usually function as a molecular velcro for other molecules to hold on to⁶⁴. The PAS domain fold exists as a five-stranded, β -pleated and α -helical structure⁶⁵. The DALI hit PAS domain was that of the transcriptional regulator PpsR, which contains all the β -strands describing the core elements of the fold (Fig 7c). This PpsR PAS domain can be considered as a possible homolog of the C-terminal $\alpha + \beta$ domain in XRN2, which is missing one of the strands defining the typical PAS fold. It is possible that the interaction between the HAD and $\alpha + \beta$ domain in this case is driven by the HAD strand replacing the missing strand in XRN2 (Fig 7c). Such an interaction would lead to a domain that has become immobile in evolutionary terms, as the interacting domain completes the fold.

Another example of a domain interaction potentially replacing a core β -strand of a fold exists in the first structural representative from the DUF2233 family (Fig 8a). The crystal structure of protein BACOVA_00430⁶⁶ from the human gut bacterium *Bacteroides ovatus* includes domains from the DUF2233 family present in bacterial and viral proteins, as well as in a homologous family represented by mammalian transmembrane glycoprotein N-Acetylglucosamine-1-phosphodiester α -N-acetylglucosaminidase (NAGPA)⁶⁶. The NAGPA enzyme converts N-acetylglucosamine-P-mannose diester to mannose-6-P monoester, and disruption of the NAGPA gene has been associated with excessive secretion of acid hydrolases by cells⁶⁷. The crystal structure of BACOVA_00430 revealed 4 domains. The first domain (not included in DUF2233) adopts an $\alpha + \beta$ fold without clear homologs, followed by three $\alpha + \beta$ two-layer DUF2233 domains (possible gene duplications), where the C-terminal region of the protein is inserted into the second domain⁶⁶. The domain architecture for the NAGPA family suggests that DUF2233 is mobile, as it exists with Copper amine oxidase N-terminal domain, Calcium-binding EGF-like domain, etc., in other proteins⁶⁸.

A triangular core is formed by the β -sheets of the three DUF2233 domains with α -helices of these domains covering the β -sheets from the outer side (Fig 8a). The third DUF2233 domain, along with the N-terminal domain, forms a continuous β -sheet comprising seven β -strands (Fig 8a). The N-terminal domain structure is similar to proteins from the cystatin/monellin superfamily that contain domains with the cystatin-like fold⁶⁶, which adopts an α -helix packed against a curved antiparallel β -sheet with five strands⁶⁹. DALI structure alignment using the N-terminal domain identified proteins with the cystatin-like fold, such as YmpB (PDB 2GU3; DALI Z-score 5.2) and RAS GTPase-activating protein (PDB 5FW5; DALI Z-score 4.8). Another protein with a cystatin-like fold domain is the human latexin protein (Fig 8b), which serves as an endogenous carboxypeptidase inhibitor⁷⁰. This human latexin protein could also be considered as a possible homolog to the N-terminal cystatin-like domain in BACOVA_00430 (DALI Z-score 4.1)⁶⁶, and the interaction between the potential N-terminal cystatin-like domain and the DUF2233 domain may be replacing the additional strand found in human latexin (Fig 8b). The DUF2233 domain could be considered as a possible homolog to proteins of the NosL/MerB-like superfamily, which are classified under the same X-group in ECOD. For example, in the apo-NosL protein from *Achromobacter cycloclastes* (Fig 8c), where the NosL domain exists as a duplication, without an N-terminal cystatin-like domain. The NosL/MerB superfamily contains domains with $\beta\beta\alpha\beta$ motifs and includes the NosL (copper metallochaperone) and MerB (organomercurial lyase) proteins⁷¹. The DUF2233 domain in BACOVA_00430 also contains a similar $\beta\beta\alpha\beta$, suggesting its similarity to NosL/MerB proteins and its mobile nature.

Domain interactions between fast evolving folds:

Fast evolving folds are the most difficult for defining domains given their propensity to change⁷². Colicins are toxic plasmid-encoded proteins produced by *E. coli*, regulated by SOS response to kill competing *E. coli* strains during unfavorable growth conditions^{73,74}. The general mechanism followed by colicins to kill host cells can be through pore formation (pore-forming toxins) or by digesting host nucleic acids (nuclease colicins)⁷⁵. More specifically, the structure of the colicins typically contain an N-terminal translocation domain, central receptor binding domains and a C-terminal activity domain. Firstly, the receptor binding domain binds to the target cell via membrane receptors, following which, the intra periplasmic proteins of the Tol/Ton families bind to the translocation domain to create a translocon for transport of the pore-forming domain into the inner membrane of the host cell⁷⁵.

The protein structure for Colicin S4⁷³ contains such a domain architecture, with a partial N-terminal translocation domain, followed by two receptor binding domains (as a result of duplication) and a C-terminal pore-forming domain. The receptor binding domains have a hydrophobic core with a small β -sheet covering the longest helix, and the β -sheet in the second receptor binding domain interacts with β -strands in the N-terminal translocation domain, forming a continuous sheet (Fig 9a). The N-terminal translocation domain is typically unresolved in crystal structures and high flexibility of this domain has been seen through NMR spectroscopy⁷⁶. Here too, the first ~60 residues of colicin S4 structure is unresolved, but, it has been observed that the N-terminal region of Colicin S4 and Colicin K

have 94% sequence identity⁷⁷. While this similarity suggests that the N-terminal region of colicin S4 is conserved within colicin family, colicin proteins are also known to exhibit domain modularity by recombining with other domains for evolving new toxic functions in competitive environments⁷⁸. Structure search with the available N-terminal translocation domain did not yield any significant hits, possibly due to the availability of only a small portion of the structure.

Among the domains in colicins, some of the highest variability is observed in the receptor binding domains, in order to bind various receptors with high affinity⁷⁴. The presence of two receptor-binding domains that arose from a recent duplication event is unique for the colicin family proteins^{73,77}, and, as colicin S4 is the only known colicin that binds to OmpW, these receptor-binding domains do not display sequence similarity with other colicin receptor binding domains⁷³. Based on structure however, these receptor binding domains can be considered possible homologs to ribonuclease PH domains (RNase PH). In ECOD, RNase PH domains and Colicin S4 receptor binding domains are present in the same X-group. Furthermore, a DALI structure search using the receptor binding domain also resulted in RNase PH domain hits with Z-score = 4.7, in proteins such as RNA exosome complex (PDB 6H25) and protein with unknown function, YbaB (PDB 1PUG). The RNase PH domain is described as $\beta\alpha\beta\alpha$ fold⁷⁹, and the three stranded β -sheet in YbaB RNase PH domain, along with the long helix superimpose well with the receptor binding domain of Colicin S4 (Fig 9b), apart from some terminal α -helical regions that differ between the two domains (Fig 9b). Lastly, two similar RNase PH domains are present with C-terminal RNA-binding S1 and KH domains in polynucleotide phosphorylase (PNPase)⁸⁰. Interestingly, in both PNPase and Colicin S4, the domains that are postulated to be possible homologs occurs as duplicates within the protein^{73,79}. In PNPase, the duplication of the RNase PH domain helps with the trimer assembly of the protein⁸¹, suggesting that the duplication of the receptor binding domain in Colicin S4 could also help with overall stability of the protein. Taken together, the N-terminal translocation domain and receptor binding domains are both mobile units found in combination with different domains.

Conclusions:

Protein domain boundaries are defined taking into account both their globular structure and sequence continuity. As these domains usually represent functional and evolutionarily mobile units, their definition becomes important for annotating genomes and assigning function. The examples discussed in this article represent some compelling cases where a β -sheet, which is usually used to define a single domain core, is instead shared between two domains within the same protein structure. When these examples include domains with common structural folds, their definition is comparatively easier, as seen in ATP-grasp and hammerhead domain interaction in PurD, and LRR and Ig domain interaction in InlJ. However, in InlJ, identifying the precise domain boundary with respect to the C-terminal β -strand of LRR domain was harder because the β -strand was being shared by the LRR and Ig domains. The first half of the C-terminal β -strand in LRR domain was assigned to the LRR domain, due to the presence of the characteristic bulge resembling that in the other repeats of the LRR domain. Another example of domain interaction between common structural folds is observed in CRP family transcription regulator protein, which also serves as an interesting

case where the presence of a β -bulge avoids the formation of a β -sheet to prevent protein aggregation.

While domain interactions between common structural folds indicates that such situations where a β -sheet is shared between domains does occur in nature, the existence of homologs with variable domain architectures also helps justify and define domain boundaries in less common folds. In XRN2, the C-terminal domain was recognized as a possible homolog to a PAS domain and partitioned from the preceding helical domain that is unique to exoribonucleases. This relationship highlights the importance of identifying potential homologs for the purpose of defining domain boundaries. Although the proposed PAS-like domain in XRN2 is missing a β -strand that is characteristic to the PAS domain fold, its interaction with the HAD domain could be replacing the additional strand in XRN2. A similar interaction is also noticed between DUF2233 and a potential cystatin-like domain, where the interaction suffices the additional β -strand that is a part of the cystatin-like fold. Such interactions, while adding to the stability of the overall protein structure, tend to disallow further domain mobility given the absence of core evolutionary β -strands. Interestingly, we noticed examples of domain deterioration in the PurK hammerhead and the XRN2 α + β subunit that might arise from pressures to evolve function in the context of structure stability arising from forming of a β -sheet. Such dead ends in terms of domain recombining, could be the reason for the existence of relatively few inter-domain β -sheet examples in the protein structure universe.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement:

This work was supported by the National Institutes of Health [GM127390 to NVG] and the Welch Foundation [I-1505 to NVG].

References:

1. Thornton JW, DeSalle R. Gene family evolution and homology: genomics meets phylogenetics. *Annu Rev Genomics Hum Genet.* 2000;1:41–73. [PubMed: 11701624]
2. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* 2019;47(D1):D427–D432. [PubMed: 30357350]
3. Marchler-Bauer A, Bo Y, Han L, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 2017;45(D1):D200–D203. [PubMed: 27899674]
4. Janin J, Chothia C. Domains in proteins: definitions, location, and structural principles. *Methods Enzymol.* 1985;115:420–430. [PubMed: 4079796]
5. Majumdar I, Kinch LN, Grishin NV. A database of domain definitions for proteins with complex interdomain geometry. *PLoS One.* 2009;4(4):e5084.
6. Marsden RL, Lee D, Maibaum M, Yeats C, Orengo CA. Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res.* 2006;34(3):1066–1080. [PubMed: 16481312]
7. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol.* 2004;14(2):208–216. [PubMed: 15093836]

8. Holland TA, Veretnik S, Shindyalov IN, Bourne PE. Partitioning protein structures into domains: why is it so difficult? *J Mol Biol.* 2006;361(3):562–590. [PubMed: 16863650]
9. Dawson N, Sillitoe I, Marsden RL, Orengo CA. The Classification of Protein Domains. *Methods Mol Biol.* 2017;1525:137–164. [PubMed: 27896721]
10. Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl AcadSci U S A.* 1973;70(3):697–701.
11. Galzitskaya OV, Melnik BS. Prediction of protein domain boundaries from sequence alone. *Protein Sci.* 2003;12(4):696–701. [PubMed: 12649427]
12. Xu Y, Xu D, Gabow HN. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics.* 2000;16(12):1091–1104. [PubMed: 11159328]
13. Holm L, Sander C. Parser for protein folding units. *Proteins.* 1994;19(3):256–268. [PubMed: 7937738]
14. Alexandrov N, Shindyalov I. PDP: protein domain parser. *Bioinformatics.* 2003;19(3):429–430. [PubMed: 12584135]
15. Sillitoe I, Dawson N, Lewis TE, et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.* 2019;47(D1):D280–D284. [PubMed: 30398663]
16. 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]
17. Cheng H, Schaeffer RD, Liao Y, et al. ECOD: an evolutionary classification of protein domains. *PLoS Comput Biol.* 2014;10(12):e1003926.
18. Cheng H, Liao Y, Schaeffer RD, Grishin NV. Manual classification strategies in the ECOD database. *Proteins.* 2015;83(7):1238–1251. [PubMed: 25917548]
19. Cheng PN, Pham JD, Nowick JS. The supramolecular chemistry of beta-sheets. *J Am Chem Soc.* 2013;135(15):5477–5492. [PubMed: 23548073]
20. Dou Y, Baisnee PF, Pollastri G, Pecout Y, Nowick J, Baldi P. ICBS: a database of interactions between protein chains mediated by beta-sheet formation. *Bioinformatics.* 2004;20(16):2767–2777. [PubMed: 15166020]
21. Guharoy M, Chakrabarti P. Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions. *Bioinformatics.* 2007;23(15):1909–1918. [PubMed: 17510165]
22. Remaut H, Waksman G. Protein-protein interaction through beta-strand addition. *Trends Biochem Sci.* 2006;31(8):436–444. [PubMed: 16828554]
23. Watkins AM, Arora PS. Anatomy of beta-strands at protein-protein interfaces. *ACS Chem Biol.* 2014;9(8):1747–1754. [PubMed: 24870802]
24. Argos P. An investigation of protein subunit and domain interfaces. *Protein Eng.* 1988;2(2):101–113. [PubMed: 3244692]
25. Raghavachari B, Tasneem A, Przytycka TM, Jothi R. DOMINE: a database of protein domain interactions. *Nucleic Acids Res.* 2008;36(Database issue):D656–661. [PubMed: 17913741]
26. Cock PJ, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 2009;25(11):1422–1423. [PubMed: 19304878]
27. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28(1):235–242. [PubMed: 10592235]
28. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 1983;22(12):2577–2637. [PubMed: 6667333]
29. Liao Y, Schaeffer RD, Pei J, Grishin NV. A sequence family database built on ECOD structural domains. *Bioinformatics.* 2018;34(17):2997–3003. [PubMed: 29659718]
30. Holm L. Benchmarking Fold Detection by DaliLite v.5. *Bioinformatics.* 2019.
31. McCarthy AA, Baker HM, Shewry SC, Patchett ML, Baker EN. Crystal structure of methylmalonyl-coenzyme A epimerase from *P. shermanii*: a novel enzymatic function on an ancient metal binding scaffold. *Structure.* 2001;9(7):637–646. [PubMed: 11470438]

32. Gonzalez MW, Pearson WR. Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.* 2010;38(7):2177–2189. [PubMed: 20064877]
33. Li W, McWilliam H, Goujon M, Cowley A, Lopez R, Pearson WR. PSI-Search: iterative HOE-reduced profile SSEARCH searching. *Bioinformatics.* 2012;28(12):1650–1651. [PubMed: 22539666]
34. Kim BH, Cong Q, Grishin NV. HangOut: generating clean PSI-BLAST profiles for domains with long insertions. *Bioinformatics.* 2010;26(12):1564–1565. [PubMed: 20413635]
35. Chipman DM, Shaanan B. The ACT domain family. *Current Opinion in Structural Biology.* 2001;11(6):694–700. [PubMed: 11751050]
36. Sampei G-i, Baba S, Kanagawa M, et al. Crystal structures of glycinamide ribonucleotide synthetase, PurD, from thermophilic eubacteria. *The Journal of Biochemistry.* 2010;148(4):429–438. [PubMed: 20716513]
37. Galperin MY, Koonin EV. A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity. *Protein Sci.* 1997;6(12):2639–2643. [PubMed: 9416615]
38. Balaji S, Aravind L. The RAGNYA fold: a novel fold with multiple topological variants found in functionally diverse nucleic acid, nucleotide and peptide-binding proteins. *Nucleic Acids Res.* 2007;35(17):5658–5671. [PubMed: 17715145]
39. Schaeffer RD, Kinch LN, Liao Y, Grishin NV. Classification of proteins with shared motifs and internal repeats in the ECOD database. *Protein Sci.* 2016;25(7):1188–1203. [PubMed: 26833690]
40. Grishin NV. Phosphatidylinositol phosphate kinase: a link between protein kinase and glutathione synthase folds. *J Mol Biol.* 1999;291(2):239–247. [PubMed: 10438618]
41. Kuzin AP, Sun T, Jorczak-Baillass J, Healy VL, Walsh CT, Knox JR. Enzymes of vancomycin resistance: the structure of D-alanine-D-lactate ligase of naturally resistant *Leuconostoc mesenteroides*. *Structure.* 2000;8(5):463–470. [PubMed: 10801495]
42. Grishin NV. Fold change in evolution of protein structures. *J Struct Biol.* 2001;134(2–3):167–185. [PubMed: 11551177]
43. Athappilly FK, Hendrickson WA. Structure of the biotinyl domain of acetyl-coenzyme A carboxylase determined by MAD phasing. *Structure.* 1995;3(12):1407–1419. [PubMed: 8747466]
44. Nishimura M, Kaminishi T, Takemoto C, et al. Crystal structure of human ribosomal protein L10 core domain reveals eukaryote-specific motifs in addition to the conserved fold. *J Mol Biol.* 2008;377(2):421–430. [PubMed: 18258260]
45. Dortet L, Veiga-Chacon E, Cossart P. *Listeria Monocytogenes*. In: Schaechter M, ed. *Encyclopedia of Microbiology* (Third Edition). Oxford: Academic Press; 2009:182–198.
46. Bublitz M, Holland C, Sabet C, et al. Crystal Structure and Standardized Geometric Analysis of InlJ, a Listerial Virulence Factor and Leucine-Rich Repeat Protein with a Novel Cysteine Ladder. *Journal of Molecular Biology.* 2008;378(1):87–96. [PubMed: 18343406]
47. Schubert W-D, Gobel G, Diepholz M, et al. Internalins from the human pathogen *Listeria monocytogenes* combine three distinct folds into a contiguous internalin domain. Edited by T. Richmond. *Journal of Molecular Biology.* 2001;312(4):783–794. [PubMed: 11575932]
48. Norio Matsushima PE, Masakatsu Kamiya, Mitsuru Osaki and Robert H. Kretsinger. Leucine-Rich Repeats (LRRs): Structure, Function, Evolution and Interaction with Ligands. *Drug Design Reviews - Online* (Discontinued). 2005; 2: 305..
49. Eichinger A, Beisel HG, Jacob U, et al. Crystal structure of gingipain R: an Arg-specific bacterial cysteine proteinase with a caspase-like fold. *EMBO J.* 1999;18(20):5453–5462. [PubMed: 10523290]
50. Krishnan V, Gaspar AH, Ye N, Mandlik A, Ton-That H, Narayana SV. An IgG-like domain in the minor pilin GBS52 of *Streptococcus agalactiae* mediates lung epithelial cell adhesion. *Structure.* 2007;15(8):893–903. [PubMed: 17697995]
51. Bork P, Holm L, Sander C. The Immunoglobulin Fold: Structural Classification, Sequence Patterns and Common Core. *Journal of Molecular Biology.* 1994;242(4):309–320. [PubMed: 7932691]
52. Korner H, Sofia HJ, Zumft WG. Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol Rev.* 2003;27(5):559–592. [PubMed: 14638413]

53. Chelvanayagam G, Heringa J, Argos P. Anatomy and evolution of proteins displaying the viral capsid jellyroll topology. *Journal of Molecular Biology*. 1992;228(1):220–242. [PubMed: 1447783]
54. Hutchinson EG, Thornton JM. The Greek key motif: extraction, classification and analysis. *Protein Eng*. 1993;6(3):233–245. [PubMed: 8506258]
55. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. The many faces of the helix-turn-helix domain: Transcription regulation and beyond*. *FEMS Microbiology Reviews*. 2005;29(2):231–262. [PubMed: 15808743]
56. Ni L, Xu W, Kumaraswami M, Schumacher MA. Plasmid protein TubR uses a distinct mode of HTH-DNA binding and recruits the prokaryotic tubulin homolog TubZ to effect DNA partition. *Proc Natl Acad Sci U S A*. 2010;107(26):11763–11768. [PubMed: 20534443]
57. Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A*. 2002;99(5):2754–2759. [PubMed: 11880627]
58. Richter H, Katic I, Gut H, Grosshans H. Structural basis and function of XRN2 binding by XTB domains. *Nat Struct Mol Biol*. 2016;23(2):164–171. [PubMed: 26779609]
59. Jinek M, Coyle SM, Doudna JA. Coupled 5' nucleotide recognition and processivity in Xrn1-mediated mRNA decay. *Mol Cell*. 2011;41(5):600–608. [PubMed: 21362555]
60. Allen KN, Dunaway-Mariano D. Markers of fitness in a successful enzyme superfamily. *Curr Opin Struct Biol*. 2009;19(6):658–665. [PubMed: 19889535]
61. Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L. Evolutionary Genomics of the HAD Superfamily: Understanding the Structural Adaptations and Catalytic Diversity in a Superfamily of Phosphoesterases and Allied Enzymes. *Journal of Molecular Biology*. 2006;361(5):1003–1034. [PubMed: 16889794]
62. Gomelsky M, Klug G. BLUF: a novel FAD-binding domain involved in sensory transduction in microorganisms. *Trends Biochem Sci*. 2002;27(10):497–500. [PubMed: 12368079]
63. An SQ, Allan JH, McCarthy Y, Febrer M, Dow JM, Ryan RP. The PAS domain-containing histidine kinase RpfS is a second sensor for the diffusible signal factor of *Xanthomonas campestris*. *Mol Microbiol*. 2014;92(3):586–597. [PubMed: 24617591]
64. Hennig S, Strauss HM, Vanselow K, et al. Structural and Functional Analyses of PAS Domain Interactions of the Clock Proteins *Drosophila* PERIOD and Mouse PERIOD2. *PLOS Biology*. 2009;7(4):e1000094.
65. Hefti MH, Frangoijs K-J, de Vries SC, Dixon R, Vervoort J. The PAS fold. *European Journal of Biochemistry*. 2004;271(6):1198–1208. [PubMed: 15009198]
66. Das D, Lee WS, Grant JC, et al. Structure and function of the DUF2233 domain in bacteria and in the human mannose 6-phosphate uncovering enzyme. *J Biol Chem*. 2013;288(23):16789–16799. [PubMed: 23572527]
67. Boonen M, Vogel P, Platt KA, Dahms N, Kornfeld S. Mice lacking mannose 6-phosphate uncovering enzyme activity have a milder phenotype than mice deficient for N-acetylglucosamine-1-phosphotransferase activity. *Mol Biol Cell*. 2009;20(20):4381–4389. [PubMed: 19710420]
68. Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. *Genome Res*. 2002;12(10):1619–1623. [PubMed: 12368255]
69. Janowski R, Kozak M, Jankowska E, et al. Human cystatin C, an amyloidogenic protein, dimerizes through three-dimensional domain swapping. *Nature Structural Biology*. 2001;8(4):316–320. [PubMed: 11276250]
70. Pallares I, Bonet R, Garcia-Castellanos R, et al. Structure of human carboxypeptidase A4 with its endogenous protein inhibitor, latexin. *Proc Natl Acad Sci U S A*. 2005;102(11):3978–3983. [PubMed: 15738388]
71. Kaur G, Subramanian S. Repurposing TRASH: Emergence of the enzyme organomercurial lyase from a non-catalytic zinc finger scaffold. *Journal of Structural Biology*. 2014;188(1):16–21. [PubMed: 25220669]
72. Medvedev KE, Kinch LN, Grishin NV. Functional and evolutionary analysis of viral proteins containing a Rossmann-like fold. 2018;27(8):1450–1463.

73. Arnold T, Zeth K, Linke D. Structure and function of colicin S4, a colicin with a duplicated receptor-binding domain. *J Biol Chem*. 2009;284(10):6403–6413. [PubMed: 19056731]
74. Cascales E, Buchanan SK, Duche D, et al. Colicin biology. *Microbiol Mol Biol Rev*. 2007;71(1):158–229. [PubMed: 17347522]
75. Ridleya H, Johnson CL, Lakey JH. Interfacial interactions of pore-forming colicins. In: *Proteins Membrane Binding and Pore Formation*. Springer; 2010:81–90.
76. Deprez C, Blanchard L, Guerlesquin F, et al. Macromolecular import into *Escherichia coli*: the TolA C-terminal domain changes conformation when interacting with the colicin A toxin. *Biochemistry*. 2002;41(8):2589–2598. [PubMed: 11851406]
77. Pils H, Smajs D, Braun V. Characterization of colicin S4 and its receptor, OmpW, a minor protein of the *Escherichia coli* outer membrane. *J Bacteriol*. 1999;181(11):3578–3581. [PubMed: 10348872]
78. Riley MA, Wertz JE. Bacteriocins: evolution, ecology, and application. *Annu Rev Microbiol*. 2002;56:117–137. [PubMed: 12142491]
79. Aravind L, Koonin EV. [1] - A Natural Classification of Ribonucleases. In: Nicholson AW, ed. *Methods in Enzymology*. Vol 341. Academic Press; 2001:3–28. [PubMed: 11582786]
80. Shi Z, Yang WZ, Lin-Chao S, Chak KF, Yuan HS. Crystal structure of *Escherichia coli* PNPase: central channel residues are involved in processive RNA degradation. *RNA*. 2008;14(11):2361–2371. [PubMed: 18812438]
81. Symmons MF, Jones GH, Luisi BF. A Duplicated Fold Is the Structural Basis for Polynucleotide Phosphorylase Catalytic Activity, Processivity, and Regulation. *Structure*. 2000;8(11):1215–1226. [PubMed: 11080643]

Figure 1: Statistics on dataset depicting data filtering at each step in the pipeline

Plots of **(a)** Initial number of multi-domain PDB chains in ECOD with counts for domains that are a part of each of those chains, followed by the number of chains and domains that were analyzed. **(b)** Number of PDB chains and domain-pairs that are involved in edge-to-edge β -strand interactions with more than three residues, followed by reduced counts after excluding domain interactions between duplicate domains.

Figure 2: Examples of β -strand mediated domain interactions that were excluded from the dataset

(a) Crystal structure of Methylmalonyl-CoA epimerase (PDB 1JC4) colored by domains. The two duplicate PaPPP domains are colored in wheat and light blue. The residues in the β -strands interacting between the domains are highlighted in orange and dark blue in the respective domains. Polar interactions between the residues involved in the domain interactions are represented as black dashes. (b) Duplicate domains in the same orientation, colored as in (a). (c) Crystal structure of N-acylamino acid racemase (PDB 2ZC8) colored by domains. The TIM barrel domain colored light blue has a C-terminal β -strand extension (highlighted in dark blue), which interacts with a β -sheet in the wheat colored α + β domain. Color coding for residues and polar interactions are same as (a). The bar representation under the protein structure depicts sequence continuity within the protein, numbered by domains. (d) Penicillin-binding protein 2X (PDB 2ZC3), colored by domains based on correct domain definition that uses globular protein structure. Strands incorrectly assigned to be a part of different domains in (e) are highlighted in orange, with polar interactions between strands represented as black dashes. The bar representation under the protein shows sequence discontinuity within the domains, where the second domain is inserted into the first domain. (e) Same Penicillin-binding protein 2X from (d), colored based on incorrect domain boundaries as obtained from the automatic ECOD pipeline. Strands incorrectly assigned to be a part of different domains are highlighted in orange and dark blue in the respective incorrectly defined domains, with polar interactions between strands represented as black dashes. The bar representation under the protein structure depicts the incorrect domain boundary that is based on sequence continuity.

Figure 3: PDP and ECOD domain annotations for examples where domains interact to share a β -sheet

(a) Crystal structure of YerbB protein (PDB 2PSB) colored by domains. The β -strands involved in the interaction between the domains are highlighted in dark blue and orange in the respective domains. The purple strand is highlighted to show the increased number of residue interactions between intra-domain β -strands. (b) Crystal structure of CRP family transcription regulator protein (PDB 1ZYB) colored by domains. The β -strands involved in the interaction between the domains are highlighted in dark blue and orange in the respective

domains. The salmon colored strand is highlighted to show the increased number of residue interactions between intra-domain β -strands. **(c and d)** Crystal structure of Prephenate dehydratase (PDB 4LUB) colored by domains based on annotations from ECOD in (c) and PDP in (d). The β -strands of interest are highlighted based on domain boundary predictions by ECOD and PDP, with the polar interactions between the β -strands shown as black dashes. **(e and f)** Crystal structure of Exoribonuclease 2 (PDB 5FIR) colored by domains based on annotations from ECOD in (e) and PDP in (f). The β -strands of interest are highlighted based on domain boundary predictions by ECOD and PDP, with the polar interactions between the β -strands shown as black dashes. **(g and h)** Crystal structure of BACOVA_00430 (PDB 3OHG) colored by domains based on annotations from ECOD in (e) and PDP in (f). The β -strands of interest are highlighted based on domain boundary predictions by ECOD and PDP, with the polar interactions between the β -strands shown as black dashes.

Figure 4: Structure of PurD with homologous domains

(a) Crystal structure of PurD (PDB 2YW2) colored by domains, with ATP-grasp domain colored wheat and hammerhead domain in light blue. β -strands involved in the interactions are colored orange and dark blue in the corresponding domains. Polar interactions between the residues in the interacting β -strands are highlighted in black as dashes. Molecules binding the ATP-grasp domain are colored grey, with ATP displayed in sticks representation and PO_4^{+} as sphere. β -strands characteristic to the hammerhead motif are highlighted in purple. (b) ATP-grasp domain in D-Alanine-D-Lactate ligase (PDB 1EHI) colored in wheat. The corresponding β -strand involved in the interaction in (a) is highlighted in orange. Coloring scheme of binding molecules is borrowed from (a). (c) Hammerhead domain in 60S ribosomal protein (PDB 2PA2) colored light blue. β -strands characteristic to the hammerhead motif are highlighted in purple and β -strand involved in the interaction in (a) is highlighted in dark blue.

Figure 5: Structure of InlJ with homologous domains

(a) Crystal structure of InlJ (PDB 3BZ5) colored by domains, with LRR domain colored wheat and Ig domain in light blue. β -strands involved in the interactions are colored orange and dark blue in the corresponding domains. The orange strand is proposed to be shared between the LRR and Ig domain, with a section of the strand represented as tube to highlight the similarity between the repeat region in the strand with the rest of the domain (enclosed in black circle). Polar interactions between the residues in the interacting β -strands are highlighted in black as dashes. **(b)** LRR domain in PSK (PDB 4Z61) colored in wheat, with the repeat region represented as tube in the black circle. **(c)** Ig-domain in RGPB (PDB 1CVR) colored in light blue. β -strand that is covering the β -barrel and completing it is highlighted in orange and β -strand involved in the interaction in (a) is highlighted in dark blue. Polar interactions between the residues in the interacting β -strands are highlighted in dark gray as dashes.

Figure 6: Crystal structures of CRP transcription regulator proteins with homologs

(a) Crystal structure of CRP transcription regulator (PDB 1ZYB) colored by domains, with jelly-roll domain colored wheat and winged HTH domain in light blue. β -strands involved in the interactions are colored orange and dark blue in the corresponding domains. Polar interactions between the residues in the interacting β -strands are highlighted as black dashes and the corresponding residues are represented as sticks. (b) Cyclic-nucleotide binding jelly-roll domain of *Mesorhizobium loti* CNG potassium channel (PDB 1VP6) colored in wheat. Corresponding β -strand making the interaction in (a) is colored orange. CMP binding the domain is represented as sticks in gray. (c) TubR HTH domain (PDB 3M8E) colored in light blue. β -strand making the interaction in (a) is colored in dark blue with red segments representing some additional features absent in the HTH domain in (a). (d) CRP transcription regulator (PDB 2ZCW) colored by domains, with jelly-roll domain colored wheat and winged HTH domain in light blue. Used as a comparison to display the variation in orientation of the HTH domain. (e and f) Zoomed in view of the β -hairpin making the interaction in (a), versus the β -hairpin without the interaction in (d), which also consists of a β -bulge. The β -hairpins enclosed in black ellipse for clarity.

Figure 7: Structure of XRN2 with homologous domains

(a) Crystal structure of XRN2 (PDB 5FIR) colored by domains, with HAD domain colored wheat and postulated PAS domain in light blue. β -strands involved in the interactions are colored orange and dark blue in the corresponding domains. Polar interactions between the residues in the interacting β -strands are highlighted as black dashes. Sulphate ions binding the HAD domain are colored gray and represented as spheres. Sections colored in red are parts not present in the homologous domain in (b). **(b)** T4 RNase H protein (PDB 1TFR) colored by domains, with HAD domain in wheat and SAM-like domain in pale yellow. The corresponding β -strand involved in the domain interaction in (a) is highlighted in orange. Mg^{2+} ions binding the HAD domain are represented as spheres in gray. **(c)** PAS domain in PpsR (PDB 4L9E) colored in light blue. β -strand absent in postulated PAS domain in XRN2 is highlighted in purple, while the corresponding β -strand making the interaction in (a) is colored dark blue.

Figure 8: Structure of BACOVA_00430 with homologous domains

(a) Crystal structure of BACOVA_00430 (PDB 3OHG) colored by domains, with cystatin-like domain colored wheat and DUF2233 domain in light blue. β -strands involved in the interactions are colored orange and dark blue in the corresponding domains. Polar interactions between the residues in the interacting β -strands are highlighted as black dashes. (b) Human latexin protein (PDB 2BO9), with cystatin-like domain colored in wheat. Additional β -strand that is replacing the interaction with DUF2233 domain in (a) is colored in red and the corresponding β -strand making the interaction in (a) is colored orange. (c) NosL domain in apo-NosL protein from *Achromobacter cycloclastes* (PDB 2HPU) colored in light blue. β -strand making the interaction in (a) is colored in dark blue.

Figure 9: Structure of Colicin S4 with homologous domain

(a) Crystal structure of Colicin S4 (PDB 3FEW) colored by domains, with duplicate receptor binding domains colored light blue and gray, and the N-terminal translocation domain colored wheat. β -strands involved in the interactions are colored orange and dark blue in the corresponding domains. Polar interactions between the residues in the interacting β -strands are highlighted as black dashes. Secondary structure elements not superimposing with possible homolog in (b) is colored in red. **(b)** RNase PH domain of YbaB (PDB 1PUG) colored in light blue. Secondary structure elements not superimposing with possible homolog in (a) is colored in red. Corresponding β -strand making the interaction in (a) is highlighted in dark blue.

Table 1:Protein representatives with domain interactions mediated by a β -sheet

Representative protein name	Representative PDB ID	ECOD T-group names of domains sharing a β -sheet	
		Domain 1	Domain 2
Glycinamide ribonucleotide synthetase, PurD	2YW2_A	ATP-grasp ^{†‡}	CO dehydrogenase molybdo-protein N-domain-like ^{†‡}
Internalin-J	3BZ5_A	Leucine-rich repeats ^{†‡}	Immunoglobulin/Fibronectin type III/E set domains/PapD-like ^{†‡}
5'->3' Exoribonuclease 2 (XRN2)	5FIR_G	PN domain-like ^{†‡}	C-terminal segment in 5'->3' exoribonucleases
Acetone carboxylase beta subunit	5SVC_B	Ribonuclease H-like ^{†‡}	beta sandwich domain in acetophenone carboxylase (Apc)
Poly(ADP-ribose) glycohydrolase	5LHB_A	Macro domain-like ^{†‡}	Poly(ADP-ribose) glycohydrolase helical domain
Colicin S4	3FEW_X	Colicin S4 translocation domain	Colicin S4 receptor-binding domain
Uncharacterized protein from DUF2233 family	3OHG_A	DUF2233 N-terminal domain	DUF2233
Type II secretion system protein D	5MP2_A	NO domain in phage tail proteins and secretins [†]	Ring-building motif II in type III secretion system [†]
YerB protein	2PSB_A	N-terminal domain in YerB-like proteins	C-terminal domain in YerB-like proteins
Transcription regulator, CRP family	1ZYB_A	Double-stranded beta-helix ^{†‡}	winged HTH ^{†‡}
D-Amino acid oxidase	2E82_B	Nucleotide-binding domain ^{†‡}	FAD-linked reductases-C [†]
Urocanase	1X87_B	N-terminal a+b domain in urocanase (Pfam 01175)	C-terminal domain in urocanase
Methyltransferase	3TMA_A	Ferredoxin-like domain in ThiI [†]	THUMP domain-like [†]
GH86A beta-porphyranease	4AW7_A	Glycosyl hydrolase domain [†]	Galactose-binding domain-like ^{†‡}
Prephenate dehydratase	4LUB_A	Periplasmic binding protein-like II [†]	ACT-like [†]
Putative membrane-associated protein	3G3L_A	Prealbumin-like ^{†‡}	Aerolysin family of pore-forming toxins [†]

Footnotes:

[†] superfamily - multiple families (F-groups) in topologies (T-group)[‡] structural fold - multiple topologies (T-groups) within homologs (H-group)