Published in final edited form as:
Proteins. 2011 ; 79(Suppl 10): 21–36. doi:10.1002/prot.23190.

CASP9 Target Classification

Lisa N. Kinch[1,*], **Shuoyong Shi**[2], **Hua Cheng**[2], **Qian Cong**[2], **Jimin Pei**[1], **Valerio Mariani**[3], **Torsten Schwede**[3], and **Nick V. Grishin**[1,2]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center at Dallas, Texas 75390-9050, USA [2]Department of Biochemistry, University of Texas Southwestern Medical Center at Dallas, Texas 75390-9050, USA [3]Swiss Institute of Bioinformatics Biozentrum, University of Basel Klingelbergstrasse 50/70 CH-4056 Basel / Switzerland

Abstract

The Critical Assessment of Protein Structure Prediction round 9 (CASP9) aimed to evaluate predictions for 129 experimentally determined protein structures. To assess tertiary structure predictions, these target structures were divided into domain-based evaluation units that were then classified into two assessment categories: template based modeling (TBM) and template free modeling (FM). CASP9 targets were split into domains of structurally compact evolutionary modules. For the targets with more than one defined domain, the decision to split structures into domains for evaluation was based on server performance. Target domains were categorized based on their evolutionary relatedness to existing templates as well as their difficulty levels indicated by server performance. Those target domains with sequence-related templates and high server prediction performance were classified as TMB, while those targets without identifiable templates and low server performance were classified as FM. However, using these generalizations for classification resulted in a blurred boundary between CASP9 assessment categories. Thus, the FM category included those domains without sequence detectable templates (25 target domains) as well as some domains with difficult to detect templates whose predictions were as poor as those without templates (5 target domains). Several interesting examples are discussed, including targets with sequence related templates that exhibit unusual structural differences, targets with homologous or analogous structure templates that are not detectable by sequence, and targets with new folds.

Keywords

Protein Structure; CASP9; Classification; Fold space; sequence homologs; structure analogs; free modeling; template based modeling; structure prediction

Introduction

This report summarizes the Critical Assessment of Protein Structure Prediction (CASP9) target proteins, which included 129 experimental structures (designated T0515-T0643). The experimental structures were submitted from several structural genomics centers (117 targets) and other research groups (12 targets). Several of these targets were excluded from the assessment for a number of reasons: T0519 was cancelled by the organizers; T0549 was considered of insufficient quality for tertiary prediction assessment; experimental structures

[*]corresponding author: Lisa Kinch, UT Southwestern, Biochemistry, 5323 Harry Hines Blvd, Dallas, TX, USA 75390, 214 645-5951 phone, 214 645-5948 fax lkinch@chop.swmed.edu.

for T0546, T0554, T0556, T0577, T0583, T0587, T0595, T0631, T0633, and T0642 were not available in time for assessment; T0533, T0536, T0600, T0612, T0637 became available before the target expired for prediction and were defined as "server only"; and T0535 was determined for a different sequence than the one submitted. Table 1 outlines the sequence and structural information available for CASP9 proteins in the context of known folds and evolutionary relationships. This information provided the basis for a domain-based classification of the target structures into two assessment categories for tertiary structure prediction: template-based modeling (TBM) and template free modeling (FM). Examples of non-trivial links between CASP9 target domains and existing structures that support our classifications are provided. The details of this analysis, including descriptions of each target, can be found at http://prodata.swmed.edu/CASP9. The goal of this article is three-fold: 1) to discuss the strategies for defining evolutionary domains and CASP evaluation units; 2) to describe a general framework for attributing evaluation units to CASP categories (FM and TBM); and 3) to propose an evolutionary classification of CASP9 structures and highlight several challenging examples of target categorization. All three goals, however, are interrelated and should be considered together.

Domain Boundary Definition and Splitting

Domains represent the basic units of folding and evolution and are usually defined as globular units in protein structures. Criteria used for domain definition often differ, leading to inconsistencies among various strategies for domain parsing. We applied to CASP9 targets our conceptual view on protein domains[1], which correspond to structurally compact evolutionary modules. We took into account the following criteria for domain parsing: 1) similarity to other protein sequences and structures, 2) self-similarity or internal duplications, 3) structural compactness (globularity) and presence of a hydrophobic core, and 4) sequence continuity. To define precise boundaries between domains, we inspected side-chain orientations and their interactions with residues that belong to the domains bordering the boundary.

For some difficult cases, certain regions of the chain protruded away from the remainder of the structure. These regions were often involved in domain swaps, which are defined as either an exchange of domains or secondary structural elements between protein chains or between domains. Several CASP9 target structures exhibited domain swaps (T0530, T0600, T0637, T0521, T0579, T0602, and T0628). Although such swaps are well-documented phenomena in protein structure[2], they remain virtually impossible to predict without having precedence in the existing pool of templates and therefore require special considerations for assessment. For example, the swapped chains in the PAS domains[3] of T0600 are novel for this fold. Many of the swaps do not heavily influence scores of predictions, as they are formed from single secondary structure elements or the closest detectable templates for these targets are similarly swapped. One of the swaps was not apparent upon inspection of the target (T0602), which forms a dimer of helices lined up in a plane. However, comparison to the closest sequence-related structure (3a7m, identified with HHsearch at 100% probability) revealed a swap in the template relative to the target. The C-terminal helices of T0602 must swap to form the core of the template 3-helix bundle. In contrast to these examples of domain swaps, some protruding regions of target chains interacted with other chains by crystal packing. For example, target T0516 includes a C-terminal extension (6 residues) that lacks secondary structure and interacts with the neighboring chain. Such extended regions were removed from target structures. Currently, neither most predictors nor standard model evaluation processes pay significant attention to domain swaps. However, due to the commonality of such swaps and the difficulty in dealing with them, they deserve more attention than they are usually given.

CASP targets are traditionally evaluated as domains to allow for more discrete classification and to minimize scoring penalties arising from differences in the relative packing of individual units. For example, proper classification of target T0604 requires defining appropriate domain boundaries and assigning the resulting domains to different categories: T0604_1 and T0604_3 belong to FM, while T0604_2 belongs to TBM. Additional CASP9 targets required splitting based on category classification (T0529, T0608, and T0629). Another motivation for splitting targets into domains is when differences in mutual domain orientations exist between an experimental structure and a model or template. For such cases a single superposition does not adequately represent the similarities or differences between folds, and splitting targets into domains provides material for a more telling analysis of predictions. On the other hand, if all domains in a multidomain protein are from the same category and their mutual orientation does not differ much between the target and high-scoring predictions, performing domain based evaluation on such a target might not be necessary. Such multidomain proteins are better be treated as single evaluation units to promote development of methods that find correct domain assembly.

To decide whether or not splitting domains is required for evaluating multi-domain targets, we compared Global Distance Test (GDT_TS[4], referred to as GDT throughout text) scores computed on whole chains with those computed on parsed domains using first server models. We used only server models for this analysis, as we believe this sample of predictions to be the most objective and consistent, having about the same number of models per target. A weighted sum was calculated for individual domains as follows: GDT scores for each domain were multiplied by the domain length, summed, and the sum was divided by the sum of domain lengths. Correlation plots between this weighted sum (Y axis) and the whole chain GDT score (X axis) for all server predictions were examined. For example, the correlation plot for target T0521 (Figure 1A) suggests that the individual domains were predicted reasonably well (GDT scores approached 80); while their packing was not (GDT scores approached 40). The duplicated EF-hand domains in Target T0521 form an intertwined dimer through an N-terminal swap (Figure 1B, EF hand domains in blue and red, swap in slate). A close structure template exists with a similar EF hand duplication (Figure 1C, 2aao); however, the swapped dimers of the template are not intertwined. While domain-based evaluation appears to be beneficial for target T0521, it may not be necessary for all multi-domain targets. For example, the weighted sum and whole chain evaluation for T0515 is similar for all servers (Figure 1D) and does not reveal any interesting features of predictions. Indeed, target T0515 (Figure 1E) exhibits a close structure template (Figure 1F, 1f3t) with a similar domain orientation. Comparison of domain-based server predictions with whole chain server predictions reveals that 27 targets require domain-based evaluation (T0521, T0528, T0529, T0533, T0534, T0542, T0543, T0544, T0547, T0548, T0550, T0553, T0555, T0571, T0575, T0579, T0582, T0586, T0589, T0596, T0600, T0604, T0608, T0611, T0628, T0629, and T0632). The cutoff to decide which domain splits to use was found from a plot of the RMS of the difference between domain GDT and whole chain GDT versus the slope of the best-fit line, both computed on top 10 server predictions (results not shown here but available on the web-site http://prodata.swmed.edu/CASP9/evaluation/Domains.htm). This procedure resulted in 147 "domains" gathered from 116 targets. For some targets our suggested evaluation units differ from those defined by the prediction center. Comments for each of the discrepancies can be found in the footnotes for Table 1. For example, we suggest splitting out the N-terminal helical extension for Target T0632 since it forms a dimerization unit of intertwined helices, and we define similar duplicated domains for the homologous targets T0544, T0533, and T0555.

Structures that contained discontinuous domain boundaries with respect to their primary sequence structure pose a challenging problem for structure modeling. Such arrangements

may result from a domain insertion into the middle of an existing fold or from a swap of secondary structural elements between domains. Target T0534 provides an excellent example of a CASP9 target with discontinuous domain boundaries. This target includes a four helix up-and-down bundle inserted into another helical bundle arranged in a bromodomain-like fold. An additional insertion of three short helical segments in the middle of the last bromodomain-like helix further complicates the fold. Although usable templates exist for modeling each of the domains present in T0534 (GDT around 50 to closest templates, see Evaluation paper), sequence-based methods failed to identify the correct templates, most likely due to the complicated domain organization.

CASP Category Definition: TBM or FM

The main goal of CASP9 target classification was to evaluate the difficulty of targets and their relatedness to existing fold space so that predictions could be assessed according to categories: TBM and FM. The Free Modeling category was first introduced in CASP7 (2006) as a replacement for the historic '*Ab-intio* (New Fold)' category. The reasons for this replacement seem to be at least three-fold. First, prediction methods have evolved substantially, and distinguishing between *ab-initio* and hybrid approaches that partly derive strength from weakly similar templates is virtually impossible to accomplish. Second, with increasing coverage of the fold space, the number of targets with what was classically defined as "new folds" is typically very small, on the order of just a couple of structures per CASP experiment. Moreover, useful templates covering significant portions of structures classified as new folds typically exist[5], while the challenge remains in finding such templates by sequence. Third, an element of subjectivity subsists in defining what constitutes a new fold, and ensuring consistency of such definitions between CASP experiments is difficult when assessors change and express differing opinions on this question. Despite the label change from 'New Fold' to 'Free modeling', certain difficulties remained in its definition, including subjectivity. An ideal, although possibly not entirely practical, solution to defining categories more objectively is to base the classification on certain numeric criteria. In CASP9 we attempted a move towards such an objective definition.

What constitutes the FM category? Even the answer to this question varies between researchers. However, two points regarding FM target qualities remain generally constant. First, FM structures demonstrate a lack of sequence detectable templates. Second, FM targets exhibit difficulties for structure prediction. Thus, 'Free modeling' represents a state of predictors being free to do whatever they can to model structures. In practice, structure modeling does not work reliably in the absence of templates and poor prediction quality for a domain remains a good indicator of an FM target. In fact, the most objective and simple measure to define FM would arguably be a cutoff on prediction quality. For example, all targets with a few best predictions having GDT scores below a certain cutoff, let's say 35, could encompass the FM category. While being very practical, this definition tends to exclude the possibility of measured progress, as it does not allow assignment of quality predictions to the FM category. Clearly, the ability to detect a template by sequence, which represents the classic but not always fully objective definition, should be taken into account in assigning CASP targets to categories. However, an equally clear concept for consideration is that the presence of a template should not preclude the target from being considered in FM category. Some templates, while easily found by structure similarity search, are virtually impossible to detect by sequence, even with the most advanced methods like HHpred. Targets with such sequence undetectable templates, especially when the best models do not look to be template-based (and due to methods imperfection, a special "look" does exist for some template-free models), should be considered under FM.

Sequence-based methods are becoming more advanced, and they frequently find homologous templates (sometimes with marginal statistical significance, but predictors still use them) that are quite different structurally from the target. These differences are a consequence of significant evolutionary changes in structures that are accompanied by marginal conservation of certain, mostly functional, sequence motifs. The conserved motifs, especially when being enhanced by rich sequence profiles and largely correct secondary structure predictions, allow detection of such structurally divergent templates. Despite the presence and possible detectability of structurally distant templates, they offer very poor starting models and result in predictions of quality comparable to that of template-free modeling. In fact, a possibility exists for models produced through *ab-initio* protocols to be more accurate than models based on misleading templates. One further important consideration for category distinction is the requirement for different types of evaluation methodology. For low-scoring models typically found in the FM category (GDT below 30), automatic evaluation methods do not adequately rank models, and manual inspection is necessary to highlight problems with predictions and suggest paths to improvement. Current and largely automatic methods of TBM evaluation, combined with the large numbers of TBM targets, do not allow to special consideration of targets with very poor templates. Thus, consideration of such targets under FM category, where manual inspection is helpful to rectify possible problems with automatic evaluation, is provides an advantage for predictions whose good structural qualities may otherwise be overlooked.

Using these considerations, we defined CASP9 FM category to encompass target domains: 1) without templates detectable by advanced sequence methods; 2) with weakly detectable homologous templates that are distant enough for the prediction quality to be as low as for targets without templates. The majority of the FM targets (25) were without detectable templates, with only 5 targets being poorly modeled domains with detectable, but inadequate for modeling templates. The following semi-automatic procedure was used to formally define the targets. First, two members of the assessment team were performing template-based predictions during the CASP season. Their assignment of templates was not biased by the target structure, as some folds look so simple and trivial (e.g., immunoglobulin-like domains or OB folds) when they are known (‘postdiction’), but are sometimes extremely difficult to predict from sequence through template identification. These two team members provided an unbiased verdict at the time of prediction whether or not templates could be found by sequence, and those target domains lacking reliable homologous templates were definitively assigned to the FM category.

Second, when target structures became available, we compared them to all models with the purpose of defining prediction difficulty for each target. Similar to our automated procedure for domain splitting, we used server models for this analysis. Only first server models were used, as they were chosen by predictors to represent what they thought to be the best predictions. Similar to our analysis of CASP8 targets[6], we found that the average GDT score for models above a calculated random score effectively describes general prediction difficulty. The random score calculation was the same as in the CASP8 analysis[6] and was based on model-to-structure scores that were computed on sequence permutations of the target structure. For all targets except T0629_2, this random score was significantly lower than the best scoring model. The T0629_2 random score was better than any prediction simply because the structure was a non-globular extended hairpin, as the calculated random score assumes the shape of the model to be similar to the shape of the structure. The Gaussian kernel density estimation on average GDT scores of first server models above random resulted in a multimodal density (Figure 2A), with major modes being around GDT scores of 60–70. The lower non-trivial mode resulting at medium bandwidths (black and dark read curves) corresponded to difficult targets, mostly without detectable templates. This mode was relatively well separated from the rest of the distribution, with a natural cutoff

around GDT 32 (marked by the green line). Interestingly, a pronounced mode in the middle (GDT around 45) indicates a cluster of targets with intermediate difficulty, most of which possessed easily detectable templates. Nevertheless, for future CASP experiments it might be useful to carefully inspect these medium quality predictions.

Average values might not fully reveal the details of target difficulty. It is also interesting to see how many servers obtain predictions of reasonable quality for each target. To find the cutoff of what might be considered "reasonable" prediction quality, a histogram of GDT scores for all above random first models was constructed for all targets. Due to the large number of scores (10571 models) considered for this evaluation, there is no need for kernel density estimation and a simple histogram reveals the trends (Figure 3A). Interestingly, the distribution is at least bi-modal, with a prominent shallow minimum in the middle (GDT score around 36). The histogram suggested a cutoff for reasonable quality, and the number of server predictions with scores above this cutoff (GDT>36) was computed for each target. However, many hard targets lacked a single prediction above GDT 36. To discriminate the difficulty of such targets, we introduced an additional term to the calculation, averaging the number of predictions with GDT score above random to the number of predictions with GDT above 36. This score was computed for all targets, and Gaussian kernel density estimation resulted in a multimodal density similar to the one for GDT scores (Figure 2B). Similarly, the cutoff value of 31 separated the lower mode, which corresponded to targets with very few decent predictions.

A 2D plot of the data is shown in Figure 3B. Positions of FM targets are shown as target numbers, while TBM targets are plotted as black points. Cutoff values (light gray lines) identify boundaries from the gaussian kernel density estimates in Figure 2. A majority of FM targets without detectable templates cluster in the lower left quadrant. While this lower left quadrant contains most of the difficult targets with low average GDT scores and small number of decent models, it also contains targets with weakly detectable homologous templates (T0550_1, T0571_1, T0537, T0604_3 and T0621; shown in bold) that were equally hard for predictors and were included in the FM category. Interestingly, quite a number of targets without templates fell in the upper right quadrant. All of these examples corresponded to a set of homologous helical domains, which were modeled adequately by *ab-initio* approaches, and frequently through detection of weakly significant, possibly non-homologous helical templates (e.g. spectrin repeats and helix-turn helix domains). The lower right quadrant includes 3 FM targets. Models for these targets were quite reasonable, but very few servers submitted high quality models. Most of these models were probably obtained by *ab initio* assembly. Two domains (T0547_3 and T0547_4) were all helical (2 helical and 3 helical structures), which are probably the easiest to model *de novo*. The third domain (T0604_1) has a ferredoxin-like fold, which is one of the most common folds among proteins. However, sequence searches do not find templates for this domain and the best models were also obtained by *ab-initio* assembly. Interestingly, the two remaining (TBM) domains in this quadrant are from T0543, domains 1 and 2, which are consecutive Somatomedin B - like disulfide-rich domains with a very close (GDT ~85) template. However, very few servers found this template, and models were essentially random without its use. The template search was likely hindered by other three large domains in this protein and small size of the first two domains.

The cutoff definitions suggested by the data are not absolute. While the bulk of the category assignments remained robust to variations of the method and cutoffs (e.g. whether we use first or best models, server or all predictions, cutoff for "good" models in the range from 30 to 40 GDT, and not only 36, etc), several targets fluctuate in assignment. Borderline TBM structures near the intersection of the two cutoff lines (e.g. 3 black points in the upper left quadrant) might be considered within the FM category. These targets were excluded from

the FM evaluation because of their somewhat higher sequence-based template detection scores (HHpred probability scores above 60) and their somewhat better quality models than those of the four targets shown in blue in the lower left quadrant (Figure 3B). A histogram of combined score distributions for FM (Figure 3C, black bars) and TBM (Figure 3C, gray bars) targets further highlights the blurred boundary between CASP categories. The scores on the X axis represent the first principle component built on the following variables: number of groups scoring above random, number of groups scoring above a difficulty cutoff GDT 36, average of GDT scores above random, and GDT score between target and closest template. FM targets overlap TBM targets in the middle, creating a potential area interesting for further analysis. However, for the purpose of CASP9, our methodology ultimately yielded 30 FM and 117 TBM domains as detailed in Table 1.

Evolution-Based Domain Classification

To classify target domains, we used a combination of sequence/profile and structure database searching approaches to find the closest neighbors (templates) to target domains. We based the results on a classification scheme similar to that defined by the Structural Classification of Proteins (SCOP) database[7]. This procedure allowed us to hypothesize about the evolutionary relationships between CASP9 targets and existing protein structures. The results shown in Table 1 not only indicate FM/TBM classification but also assign CASP domains to existing SCOP folds and superfamilies and indicate whether the target is hypothesized as homologous to other proteins with known structures. Those target domains having templates that are reliably and easily identified by sequence-based methods (i.e. PSI-BLAST[8], HHPRED[9], or PROCAIN[10]) were assigned to the template-based modeling assessment category (TBM). For all of these cases, structural similarity to identified folds in the Protein Data Bank[11] was confirmed with inspection of Dali structure superpositions[12]. Several cases of unusual structural differences revealed in this inspection between the targets and templates with detectable sequence similarities are discussed below.

While this target difficulty-based assignment of domains to the TBM and FM categories comes out naturally from the data and allows for consistent evaluation of how methods deal with different target types, domains belonging to the FM category may not necessarily lack reasonable structure templates. To evaluate the relationships of FM domains to existing fold space, we used Dali[12] and VAST[13] to search the PDB for protein structures with similar folds and evaluated the potential evolutionary relationships of identified folds using data provided by the HorA server[14], which uses a Support Vector Machine to discriminate homology (structure similarity due to a common ancestor) from analogy (convergent structure similarity) for given structure pairs. Finally, we employed a secondary structure-based vector search program developed in our lab (ProSMoS [15,16]) to identify more distant protein structures in the PDB that displayed similar topologies to the target folds. We combined these automated search programs with manual inspection and a general knowledge of protein folds to produce the final classification. For cases with identified structural similarities (143 domains), analogy between the target and template was assumed unless there was enough compelling evidence to hypothesize descent from a common ancestor (see examples below). For those cases without clear similarities to known structures, a classification of new fold was assigned (4 domains).

Unusual Structure Differences in Sequence Related Targets

Perhaps one of the most interesting classification targets in CASP9 is T0604, the VP0956 protein from *Vibrio parahaemolyticus*. This target contains three structural domains (Figure 4A): an N-terminal domain with a unique ferredoxin-like fold (FM) is followed by a Rossmann-like FAD/NAD(P)-binding domain (TBM) with an inserted α+β sandwich (FM).

Although the N-terminal T0604_1 domain has no detectable sequence relationships to known folds, it adopts a ferredoxin-like fold topology (Figure 4B). Among the numerous different ferredoxin-like folds, which currently include 59 superfamilies, no good template dictates the correct secondary structure interactions and orientations unique to T0604_1 (58.5 GDT to closest template 2w7a, Figure 4C). Accordingly, the best predictions for this domain were produced by free modeling. The second domain of this target represents an FAD/NAD(P)-binding domain that is easily detected by sequence (classified as TBM). The closest FAD/NAD(P)-binding domain sequence template (Figure 4D: 2i0z) retains 26.7% identity to the T0604_2 FAD/NAD(P)-binding domain (Figure 4E), while an alternative template FAD/NAD(P)-binding domain (Figure 4F: 1kdg) is more distantly related (16.2% sequence identity) and includes a number of decorations to the core fold. The closest FAD/NAD(P)-binding domain template (2i0z) has an insertion classified as HI0933 insert domain-like that forms a six-stranded barrel with an inserted four-helical bundle (Figure 4G). Alternatively, the T0604_3 domain insertion forms an α+β sandwich (Figure 4H) that retains a core antiparallel sheet (order 23415) flanked by a helix of the FAD-linked reductase C-terminal domain fold topology (Figure 4I) found as an insert to the more distantly related FAD/NAD(P)-binding protein 1kdg. When compared to the template FAD-linked reductase C-terminal domain, T0604_3 includes a number of decorations to the core that make up almost half of the target domain. These decorations include a twisted β-hairpin insertion in the strand 3–4 loop and a mixed α/β insertion following strand 4 that extends the core sheet by two strands.

Several targets (T0544, T0553, T0554(cancelled) and T0555) are homologous to each other, belonging to the Pfam family Phycobilisome linker polypeptide (PF00427). Their structures show a duplication of two related helical domains (Figure 4J, blue and red). Each domain consists of a three helical bundle (helices 1–3 in Figure 4K) followed by a short C-terminal α-helix (helix 4) that mainly interacts with the three-helical bundle in the other domain. Helix 4 can thus be viewed as swapped between the two domains. The interactions between the two domains are also reinforced by the N-terminal loop in the first domain and the C-terminal loop in the second domain. Weak sequence similarities between the two duplicated domains were recognized by HHpred searches (reasonable probability scores above 80). HHpred searches also gave marginal hits to EF-hands (unreliable probability scores less than 60). EF-hands are a group of helix-loop-helix domains that show high sequence and structural variability. Two EF-hand domains often pack together to form a four helical bundle (Figure 4L) and the loops in between their respective core helices can accommodate calcium-binding sites. The three helical bundle structure of each domain of T0553 can be structurally aligned to three helices in two packed EF-hands (for example, the best Dali Z-score between the first domain of T0553 to an EF-hand (pdb id: 2obh) is 4.0). EF-hands can thus be used as templates for T0553. However, the drawbacks of using EF-hands structures as templates are 1) low structural similarities to templates and difficulty in determining which EF-hand structures serve as best templates, 2) incorrect domain orientations in templates and difficulty in modeling correct domain interactions based on EF-hand structures. In fact, the best models of T0553 are produced by free modeling, and the reliability of sequence-based identification of EF hands as templates was marginal (HHpred probability score around 50 or less).

Examples of Homologous Domains not Detectable by Sequence

In classifying individual CASP9 target domains, we sought to establish potential evolutionary relationships to existing folds wherever possible. First, we defined as a homolog any target whose sequence detected its corresponding template sequence using any available method. We attempted to remain unbiased by classifying template sequences during CASP9 *prior* to the release of the target structure for our analysis. For targets lacking

detectable sequence similarity, we considered the various alignments and scores produced by HorA[14], combined with additional structural and functional considerations as evidence for homology. Examples of such additional considerations included similarities in the organization of domain structure, the sharing of unusual structural features, the sharing of local structural motifs, and the placement of active sites. CASP9 target domains with difficult to identify homology include T0531, T0537, T0540, T0544_1, T0544_2, T0553_1, T0553_2, T0555_1, T0555_2, T0550_1, T0571_1, T0561, T0604_3, T0616, T0621, and T0624.

For some of the cases with suggested homology, relatively small structural units link the target domains to existing fold space. For example, the target T0531 3-strand β-meander (Figure 5A) resembles the midkine fold (Figure 5B). While the target includes an additional inserted helix, the two folds retain a similar curvature of the sheet and two of three disulfide bond pairs (Figure 5A/B alignment, magenta residues). While the main difference between T0531 and its related midkine is limited to a single loop, some of the remaining cases of potential homology embody more significant changes in the core secondary structure elements and elaborations to the core fold. The structure of Target T0561 (Figure 5C) includes a central 3-helical bundle with a near perpendicular arrangement of the helices, reminiscent of a DNA-binding 3-helical bundle superfold dictated by the helix-turn-helix motif (HTH). HorA identifies the C-terminal domain of replication initiation factor DnaA (1l8q) with a combined score suggestive of homology. Although the core 3-helical bundle of DnaA differs from that of the target (Figure 5D, HTH helices are all longer), the core includes conserved functional residues (Figure 5C/D alignment, blue residues) that traditionally mediate interactions with DNA; and both are elaborated with N-terminal and C-terminal helices arranged on a similar face of the HTH and in a similar orientation to each other (colored slate and salmon, respectively). The core 3-helix bundle of an additional HTH-containing template more closely resembles the target core in terms of helix size (MogR repressor, not shown), but has different helical elaborations. Finally, the function of the target in binding J-containing DNA[17] supports this homologous relationship.

Mapping Unclassified TBM Targets to Fold Space: Potential New Folds

TBM domains include several examples of folds with sequence-related templates that are not yet classified in SCOP and offer interesting structure classification challenges. Target T0603_1 and its closest templates (e.g. 3god domain 1) cannot be attributed to any existing SCOP fold and should be considered “new” for the purpose of structure classification. Thus, CASP9 “new folds” are not limited to the FM category. We explored the relationship of these targets to existing fold space using data generated by HorA[14], VAST[13], and ProSMoS[16]. Target T0557 (Figure 6A) represents the N-terminal domain of a putative ATP-dependent DNA helicase RecG-related protein from *Nitrosomonas europa*. The T0557 sequence identifies the N-terminal α+β sandwich domain of a divergent AAA (PF04326) as a closely-related template (Figure 6B: 3lmm). The target and the template define a three-layer αβα fold that includes an N-terminal helix in different orientations covering the central mixed sheet (order 1243) and a β-hairpin like insertion following the second β-strand. Ignoring the N-terminal helix, the AAA N-terminal domains follow the same core topology as IF3-like folds (Figure 6C). The HorA server identified a number of different IF-3 like folds as potential target domain homologs, with the top scores assigned to AlbA-1-like superfamily members that function to bind DNA. The fusion of the target domain to a putative ATP-dependent DNA helicase would be consistent with a similar functional role of the target domain in DNA binding, further supporting a homologous relationship to IF-3-like fold assignment for this unclassified target domain.

Target T0540 (Figure 6D) belongs to a family defined in PFAM as Fas apoptotic inhibitory molecule (FAIM1). The FAIM1 family includes a single structure representative (Figure 6E: 2kd2) described as a novel 7-stranded β-sandwich [18]. Target T0540 forms a β-sandwich of a similar topology to the template. However, the presence of an edge strand that was previously defined as a crossover loop in the template yields an 8-stranded β-sandwich. Although the FAIM1 structures form two somewhat flattened sheets, their β-meander topology is reminiscent of the 8-stranded barrel displayed by streptavadin-like folds (Figure 6F). Indeed, HorA identified this link as a potential homolog (as a top hit), followed by two 8-stranded β-sandwiches: the HSP70 fold, which adopts a different topology (includes a partial meander with the sheet completed by a hairpin insertion), and the NusG fold, which adopts a meander that is circularly permuted with respect to FAIM1. When compared to the streptavadin-like fold, the positions of the crossover strand and the C-terminal loop found in the FAIM1 template could be interpreted as beginning to form a barrel. Likewise, the corresponding crossover strand in streptavadin-like folds can often be broken, yielding a somewhat flattened barrel (Figure 6F, 1ei5). Despite these tendencies towards similarity, the difference in the structural features of sheets (FAIM1) versus barrels (streptavadin-like) suggests a distinction between the two folds, and the relationship of the target sandwich to existing folds more resembles the analogous NusG circular permutation.

The two-domain TBM target T0603 belongs to the CRISPR associated protein Cas1 family, which includes close structural templates identified by sequence (3god, 3lfx, and 2yzs). Although the domains do not require splitting for assessment, we consider each separately for classification. The Cas1 domains were described as novel [19], having a unique N-terminal β-strand domain (Figure 7A) followed by a C-terminal α-helical domain (Figure 7C). Despite the predominance of β-strands in the N-terminal domain, HorA SVM scores consistently identified α/β class folds as potential homologs. For example the α/β RNase H-like motif fold (Figure 7B) can superimpose with the template domain, with an RNase H helix being partially replaced by a β-meander of the Cas1 T0603_1 template. Although the β-meander can frequently replace a helix in homologous structures [20], the remaining topological differences (for example, a hairpin insertion in RNaseH and a C-terminal helix replacing a strand) support an assignment of new fold to the target domain T0603_1. The C-terminal domain of the same target assembles into 10 α-helices (4 split + 2 helices) with four conserved residues (E141, H208, D218, and D221) that contribute to a DNA-specific endonuclease activity (Figure 7C). HorA identified a number of fertilization protein folds as potential homologs of T0603_2. These hits map to the C-terminal helices of Cas1, which form an up-and-down bundle of three split helices (Figure 7C, rainbow). The same helices include three of the active site residues (H208, D218, and D221) and part of a positively charged surface patch (K211 and K224) thought to bind DNA substrate [21]. Interestingly, a VAST search with the Cas1 α-helical domain (3nkeB) identified among the top hits an O-phosphoseryl-tRNA kinase C-terminal domain (CTD). The CTD mediates binding to its tRNA substrate (Figure 7D: 2adb) in a similar position as proposed for Cas1 DNA binding. In the resulting structure superposition, the positively charged side chain of a presumed Cas1 DNA binding residue (K224) is positioned near some positively charged tRNA binding residues of the kinase domain (R195 and R219). Despite a similar structural positioning of these residues, they are not contributed from the same position in the 3-helix bundle, and the two structures display enough divergence to suggest an analogous relationship.

FM Targets with New Folds

For many targets in the FM category, usable templates exist among known structures. The major obstacle for improving prediction for these difficult targets is detection of the templates by sequence. However, targets exist for which even having a 3D structure at hand

leaves significant challenges for template identification. To make sure we did not miss any potential template folds of target domains and to help make the distinction between structure analogs and new folds, we employed a secondary structure vector search program ProSMoS[16]. This program finds topological and architectural similarities including circular permutations, but is not sensitive to structural details such as packing, length of secondary structure elements or large insertions. To perform this vector search, secondary structural elements belonging to each FM target domain with a potential new fold (T0529_1, T0581, T0608_1, T0618, and T0624) were defined, and interaction matrices between these elements were used to search for exact matches in a database of similar matrices defined for existing PDB structures. We omitted target T0629_2 from the search, due to its unusual elongated structure and oligomerization state for which we could define no vectors. To our knowledge, the only other fibril like fold with an elongated arrangement of β-strands is the collagen fibre. However, the arrangement of collagen fibre strands into a trimer of single elongated strands differs from that of the hairpin trimer organization of T0629_2, which forms a unit of six elongated strands and is not related.

Some peripheral secondary structures were omitted from the vector searches to maximize the number of hits found. For example, we limited our search for T0529_1 to include an interaction matrix defined for a group of helices (Figure 8A, rainbow) that form local contacts and also contribute to the function of the protein (binding mRNA cap), ignoring a number of secondary structures with mainly long range contacts (Figure 8A, slate and salmon) that are unlikely to form a structure core. One of the hits identified by this vector search was also found by a VAST search, identifying the C-terminus of the gamma subunit of dissimilatory sulfite reductase I (DsrC: 3or1_C). DsrC has a 3-stranded β-meander packed against an array of five helices. The VAST hit covers the entire DsrC helical array (Figure 8B), although the orientations of the helices are not all identical. Despite the presence of this small analogous structure core, the many secondary structure decorations present in T0529_1 define the majority of the structure and are required to complete the functional site. Given these considerations T0529_1 is better assigned as a new fold. After complete analysis of all CASP9 target domains, we designated 4 targets as new folds (T0529_1, T0581, T0603_1, and T0629_2). T0581 forms a novel four-stranded α+β sandwich. T0603 is a TBM target with a close template not yet classified in SCOP (discussed above). T0629_2 forms a long extended tail through trimeric antiparallel β-strands that organize around seven iron atoms. The iron-coordinating histidine residues are also present in the C-terminal sequence of a domain homologous to the N-terminal domain of this target (T0629_1), the receptor-binding domain of short tail fibre protein gp12. Many predictions extended the alignments of the N-terminal domain to include sequence regions with these histidines, although the C-terminal structure of the tail fibre protein gp12 is not similar. However, this region might be homologous to the non-globular iron-binding domain of T0629.

Conclusion

The experimental structures forming the basis of CASP9 tertiary structure prediction assessment have been evolutionary classified by defining sequence and structure relationships to existing folds. This classification aided in assigning target domains into two assessment categories: TBM and FM. The TBM category included target domains with templates that could be detected by sequence (117 target domains), while the FM category included target domains without sequence-detectable templates (25 target domains). Due to a blurred boundry between the two categories, some target domains were included in the FM category (5 target domains) that had weakly sequence-detectable templates but displayed poor server performance similar to those FM target domains without templates. In addition to providing a basis for tertiary structure prediction assessment, the classification of CASP9

targets provided a number of interesting examples of evolutionary relationships among protein structures.

Acknowledgments

This work was supported in part by the National Institutes of Health (GM094575 to NVG) and the Welch Foundation (I-1505 to NVG).

REFERENCES

1. Majumdar I, Kinch LN, Grishin NV. A database of domain definitions for proteins with complex interdomain geometry. PLoS One. 2009; 4(4):e5084. [PubMed: 19352501]
2. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci. 2002; 11(6): 1285–1299. [PubMed: 12021428]
3. Moglich A, Ayers RA, Moffat K. Structure and signaling mechanism of Per-ARNT-Sim domains. Structure. 2009; 17(10):1282–1294. [PubMed: 19836329]
4. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003; 31(13):3370–3374. [PubMed: 12824330]
5. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci U S A. 2006; 103(8):2605–2610. [PubMed: 16478803]
6. Shi S, Pei J, Sadreyev RI, Kinch LN, Majumdar I, Tong J, Cheng H, Kim BH, Grishin NV. Analysis of CASP8 targets, predictions and assessment methods. Database (Oxford). 2009 2009: bap003.
7. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247(4):536–540. [PubMed: 7723011]
8. 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 Res. 1997; 25(17):3389–3402. [PubMed: 9254694]
9. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005; 33(Web Server issue):W244–W248. [PubMed: 15980461]
10. Wang Y, Sadreyev RI, Grishin NV. PROCAIN: protein profile comparison with assisting information. Nucleic Acids Res. 2009; 37(11):3522–3530. [PubMed: 19357092]
11. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003; 10(12):980. [PubMed: 14634627]
12. Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics. 2008; 24(23):2780–2781. [PubMed: 18818215]
13. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol. 1996; 6(3):377–385. [PubMed: 8804824]
14. Kim BH, Cheng H, Grishin NV. HorA web server to infer homology between proteins using sequence and structural similarity. Nucleic Acids Res. 2009; 37(Web Server issue):W532–W538. [PubMed: 19417074]
15. Shi S, Chitturi B, Grishin NV. ProSMoS server: a pattern-based search using interaction matrix representation of protein structures. Nucleic Acids Res. 2009; 37(Web Server issue):W526–W531. [PubMed: 19420061]
16. Shi S, Zhong Y, Majumdar I, Sri Krishna S, Grishin NV. Searching for three-dimensional secondary structural patterns in proteins with ProSMoS. Bioinformatics. 2007; 23(11):1331–1338. [PubMed: 17384423]
17. Heidebrecht T, Christodoulou E, Chalmers MJ, Jan S, Ter Riet B, Grover RK, Joosten RP, Littler D, van Luenen H, Griffin PR, Wentworth P Jr, Borst P, Perrakis A. The structural basis for recognition of base J containing DNA by a novel DNA binding domain in JBP1. Nucleic Acids Res.

18. Hemond M, Rothstein TL, Wagner G. Fas apoptosis inhibitory molecule contains a novel beta-sandwich in contact with a partially ordered domain. J Mol Biol. 2009; 386(4):1024–1037. [PubMed: 19168072]

19. Wiedenheft B, Zhou K, Jinek M, Coyle SM, Ma W, Doudna JA. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure. 2009; 17(6):904–912. [PubMed: 19523907]

20. Grishin NV. Fold change in evolution of protein structures. J Struct Biol. 2001; 134(2–3):167–185. [PubMed: 11551177]

21. Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, Gagarinova A, Pogoutse O, Brown G, Binkowski A, Phanse S, Joachimiak A, Koonin EV, Savchenko A, Emili A, Greenblatt J, Edwards AM, Yakunin AF. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol. 79(2):484–502. [PubMed: 21219465]

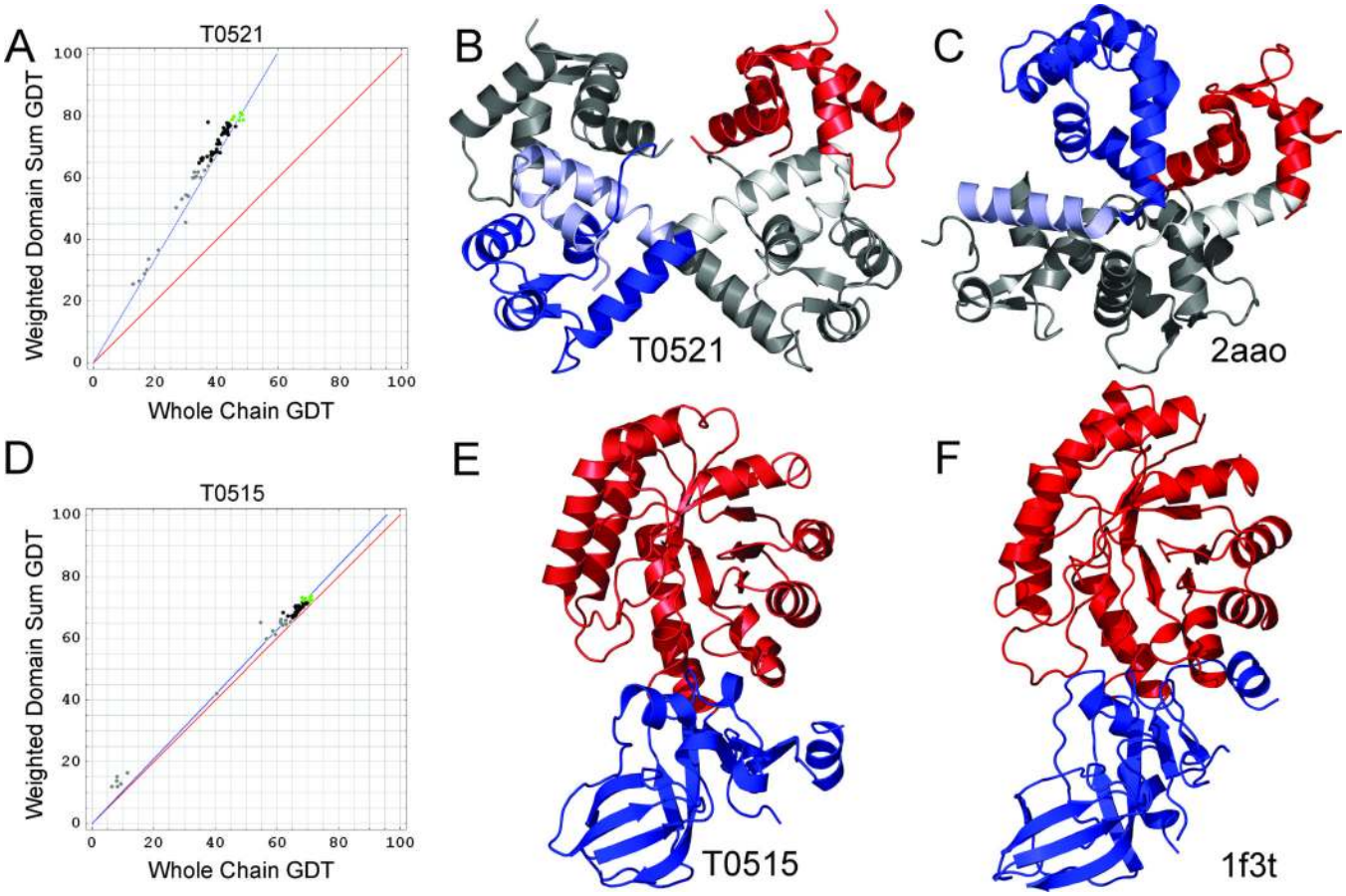

Figure 1. CASP9 domains

CASP9 target structures with defined domains (colored red and blue) are split according to graphs of whole chain (x axis) vs. weighted sum of domain (y axis) GDT scores. **(A)** Slope of GDT scores for target T0521 suggests splitting. **(B)** Target T0521 forms a swapped (slate and white helix) and intertwined dimer (gray second chain) of duplicated EF Hand domains. **(C)** Template for T0521 (2aao) forms a non-intertwined swapped dimer of duplicated EF Hand domains colored as above. **(D)** Graph Slope of GDT scores for target T0515 suggests no split. **(E)** The domains of target T0515 are arranged in a similar orientation as **(F)** the template (1f3t).

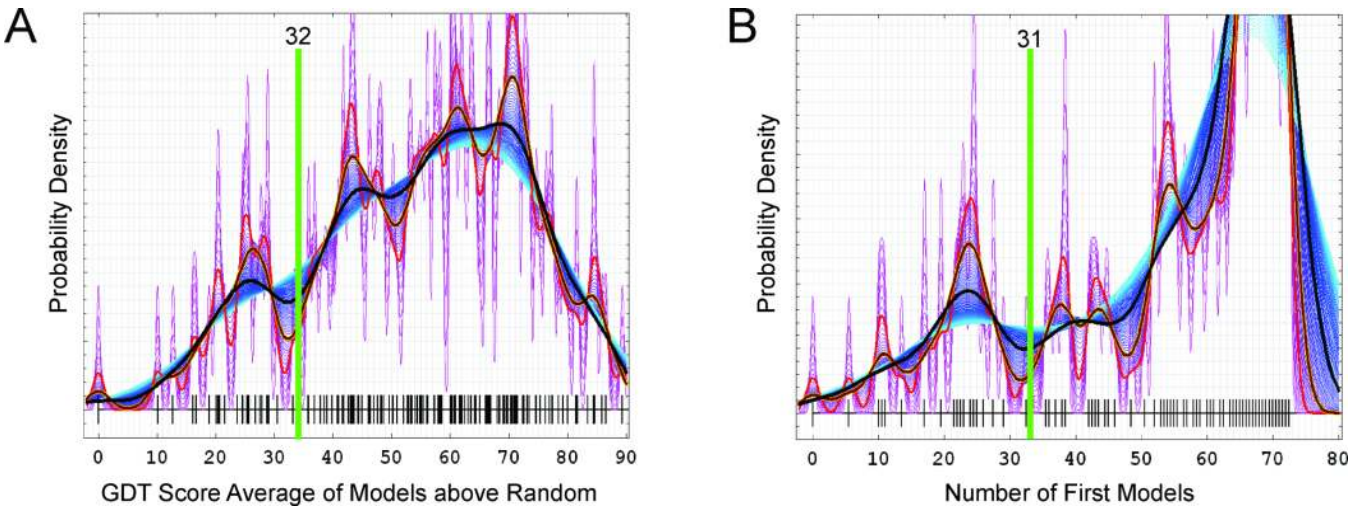

Figure 2. Gaussian kernel density cutoffs

Gaussian kernel density estimates for various bandwidths (small to large – magenta to blue, densities at representative intermediate bandwidths are shown as thicker red, brown and black curves) built on **(A)** first sever models for the GDT-TS scores above random and **(B)** "Number of first models", respectively. This "Number" was computed by averaging the number of first models above random with the number of first models above a difficulty cutoff of 36 and can be thought of as a number of reasonably good models for a given target. Long ticks on the X-axis mark the position of corresponding score for each target. Green vertical lines mark the data-suggested cutoff.

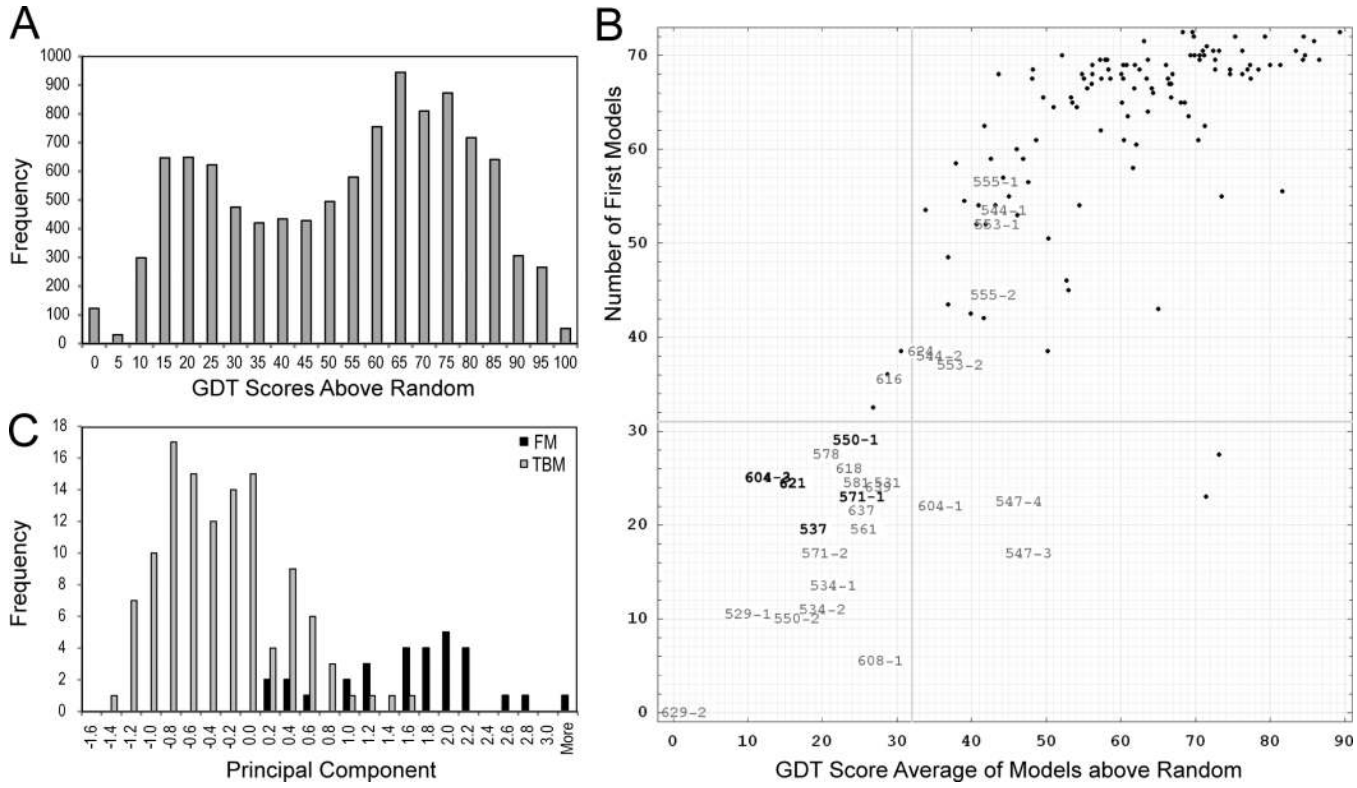

Figure 3. CASP9 target score distributions

(A) A histogram depicts T0571 GDT scores above random for all CASP9 first server models and suggests a difficulty cutoff around GDT score 36. **(B)** A scatter plot of "Number of first models" vs. average GDT scores depicts the distribution of CASP9 target domains. Positions of FM targets are shown by target number, bolded numbers are for targets with templates detectable by sequence. Positions of TBM targets (templates are readily detectable by sequence methods) are show as black dots. Gray lines correspond to cutoffs from gaussian kernel density estimates. **(C)** A histogram of first principal component scores, which combine four different individual scores that are calculated for each target domain (number of groups scoring above random, number of groups scoring above difficulty cutoff 36, average of GDT scores above random, and highest GDT score between target and closest template as found by LGA program[4]), shows incomplete separation for FM (black) and TBM (gray) targets.

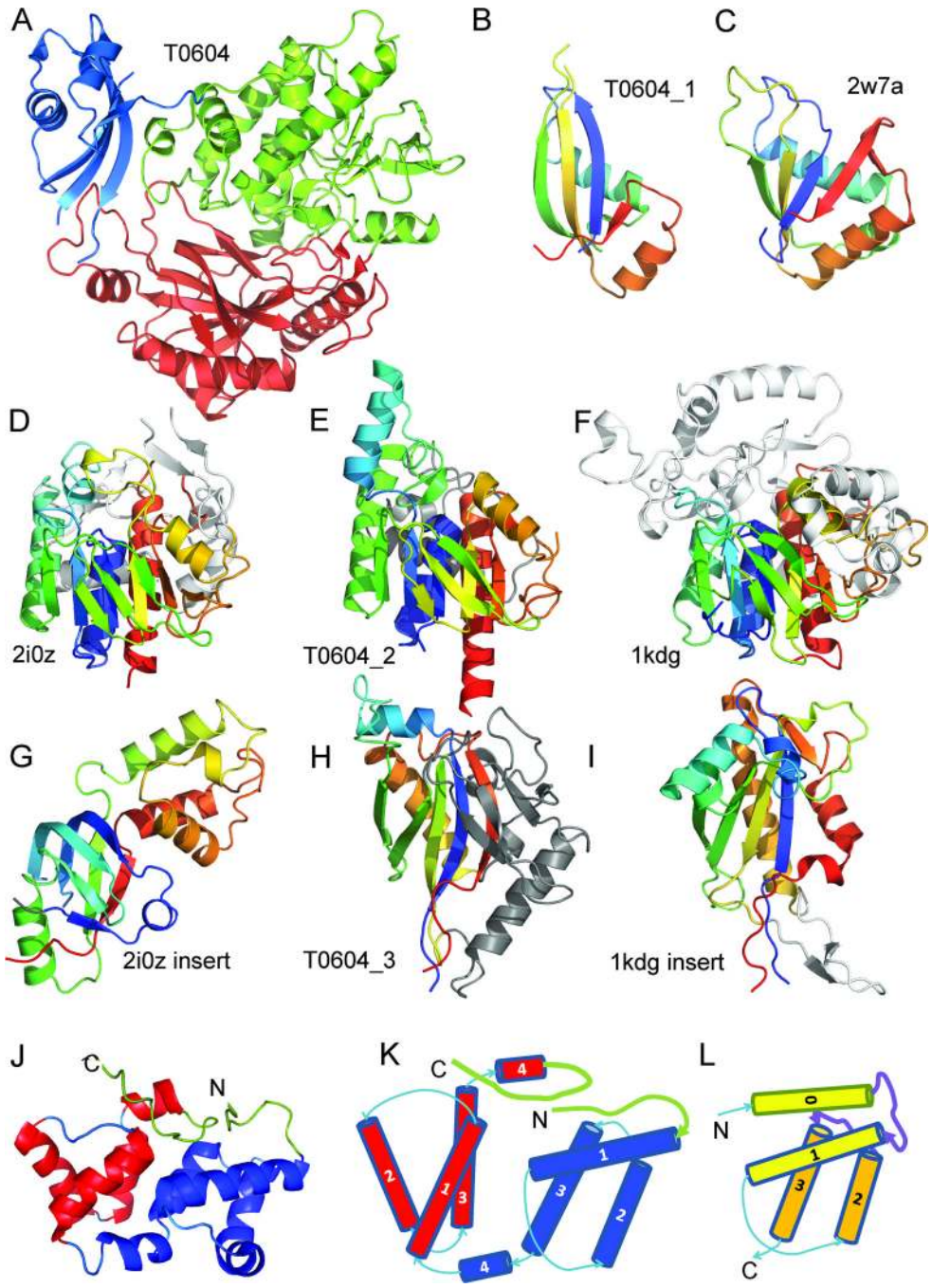

Figure 4. Unusual structure differences between targets and sequence-related templates

(A) Target T0604 is comprised of three domains: T0604_1(blue), T0604_2 (green), and T0604_3 (red). **(B)** T0604_1 forms a unique ferredoxin-like fold with a longer sheet than **(C)** the closest template 2w7a. **(D)** The FAD/NAD(P)-binding domain of 2i0z is the closest template to **(E)** the FAD/NAD(P)-binding domain of Target T0604_2. **(F)** A more distantly related alternate template for T0604_2 (1kdg) possesses more insertions (white). **(G)** A six-stranded barrel with an inserted four-helical bundle (HI0933 insert domain-like) is inserted in the closest FAD/NAD(P)-binding domain template 2i0z. **(H)** The target T0604_3 domain insert forms an α+β sandwich that resembles (**I**) the FAD-linked reductase C-terminal domain insertion of the more distantly related FAD/NAD(P)-binding domain (1kdg). **(J)**

Target T0553 α-Helices in the N- and C- terminal domains in are in blue (T0553_1) and red (T0553_2) colors respectively. **(K)** Four α-helices are labeled 1 to 4 for each of the duplicated domains. The N- and C-terminal loops are colored in green. **(L)** Packed EF hands consist of two α-helices each (labeled 0 and 1 for the first EF-hand shown in yellow, and 2 and 3 for the second EF-hand shown in orange) and a loop (colored purple) in between them.

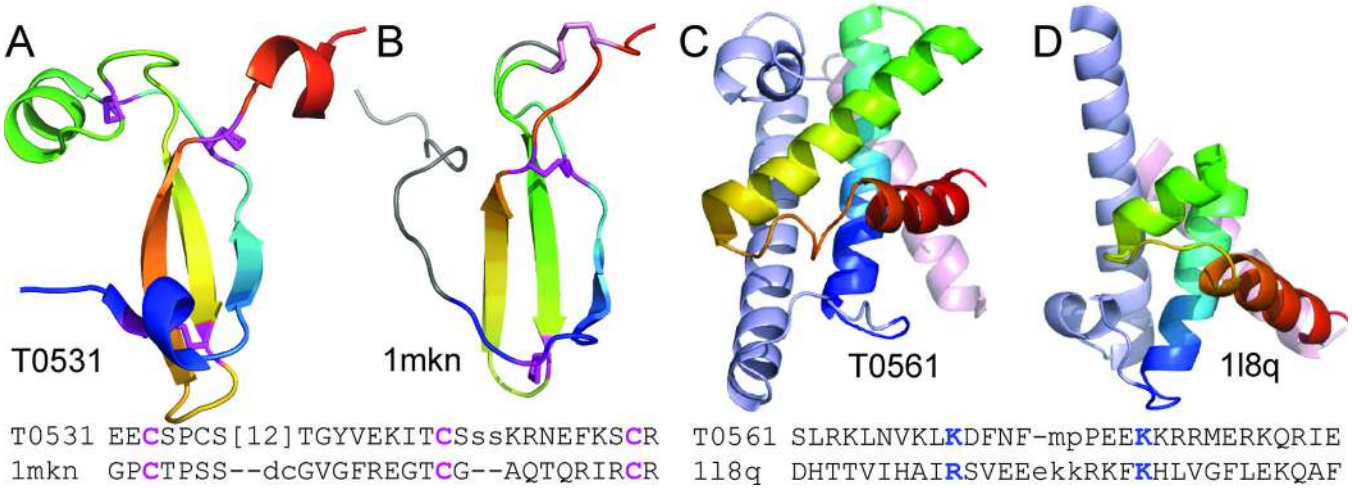

Figure 5. Domains with templates not detectable by sequence

(A) Target T0531 3-strand β-meander resembles the **(B)** midkine fold, with a structural alignment, preserving conserved disulfide pairs (magenta) that are important to the fold. **(C)** Target T0561 includes a central 3-helical bundle (rainbow) with N-terminal (slate) and C-terminal (salmon) elaborations that are similar to **(D)** the core helix-turn-helix motif (HTH) of the C-terminal domain of replication initiation factor DnaA, with the core HTH alignment including conserved functional residues.

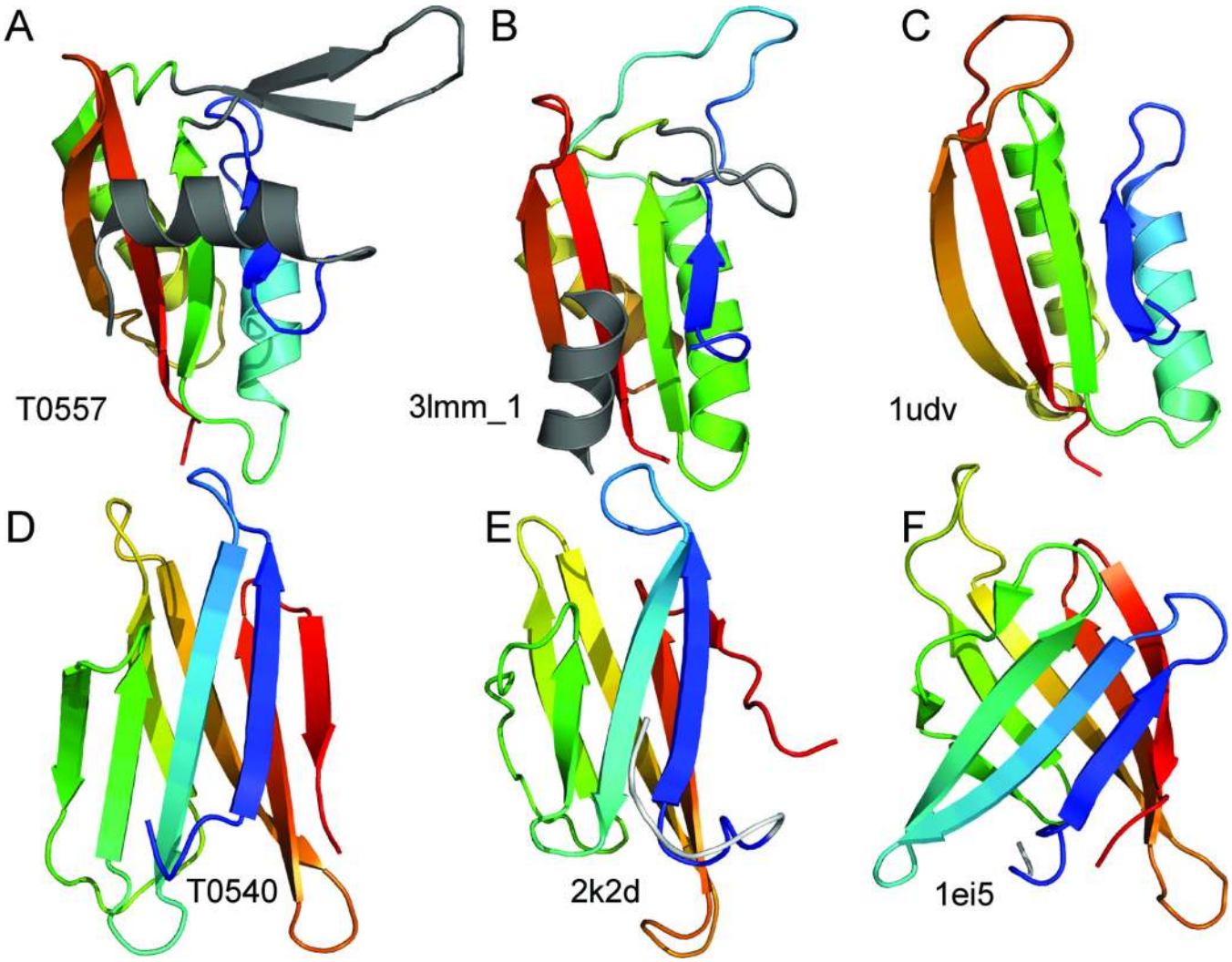

Figure 6. Mapping unclassified TBM targets to fold space

(A) Target T0557 α+β sandwich identifies **(B)** the N-terminal domain of a divergent AAA (3lmm_1) as a close homolog with similar N-terminal helical extensions and β-hairpin like insertions (gray) as compared to **(C)** the related IF3-like SCOP fold (1udv). **(D)** Target T0540 β-sandwich belongs to the FAIM1 family, which includes **(E)** a structure representative (2k2d) described as a 7-stranded β-sandwich. The common β-meander of the two flattened sandwiches resembles **(F)** the β-meander topology of streptavadin-like barrels (1ei5).

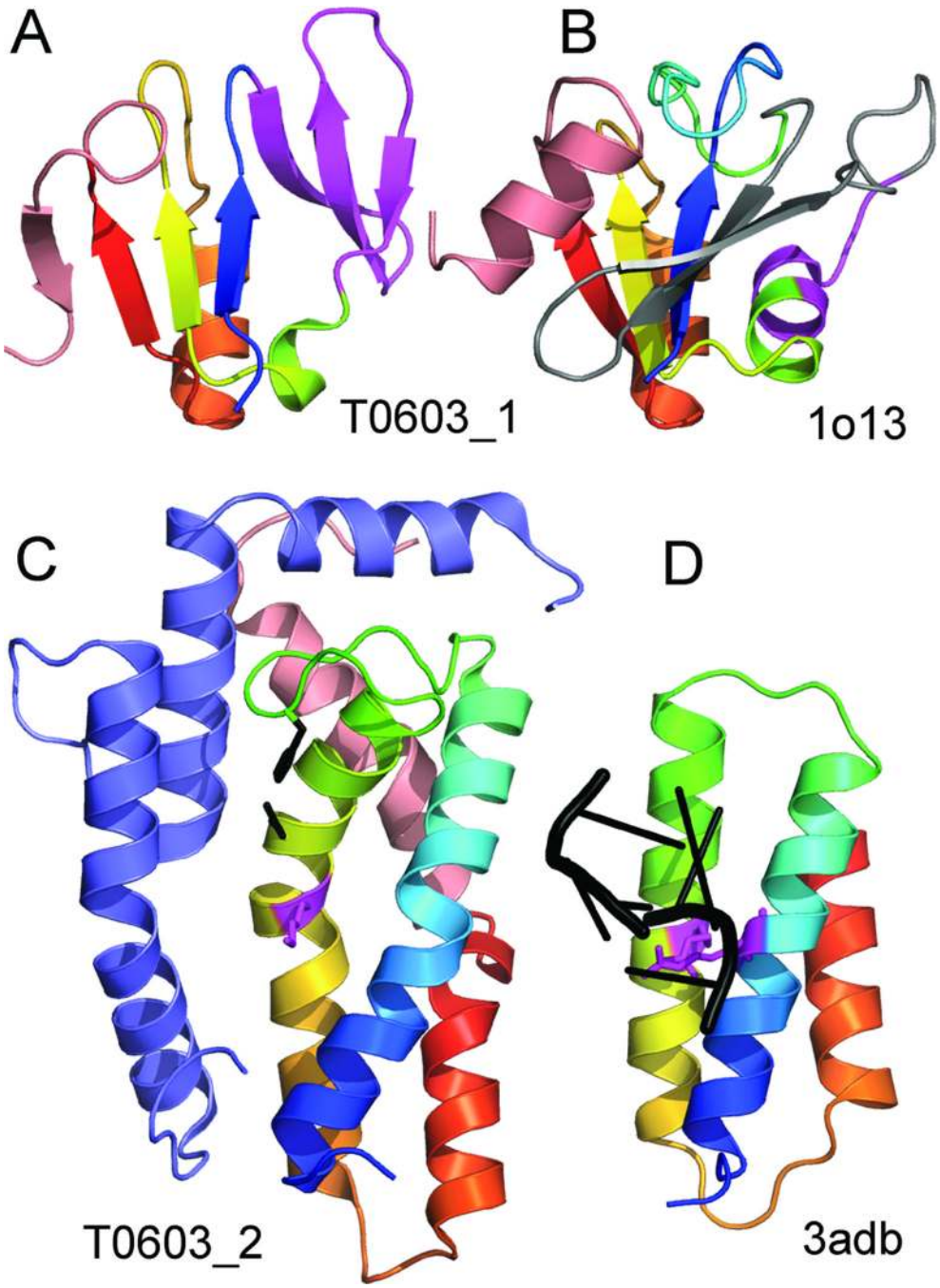

Figure 7. Structure analogs identified for T0603 TBM domains

(A) The N-terminal structural domain of T0603 has a small 3-stranded β-meander (magenta) that could replace **(B)** the helix (magenta) of α/β folds like the RNase H-like motif (1o13). **(C)** Three split helices (rainbow) of the T0603 C-terminal helical domain form an endonuclease active site (black) with a positively charges side chain (magenta) positioned similarly as **(D)** positively charged side chains (magenta) from the O-phosphoseryl-tRNA kinase C-terminal domain (3adb) nucleotide-binding site (black).

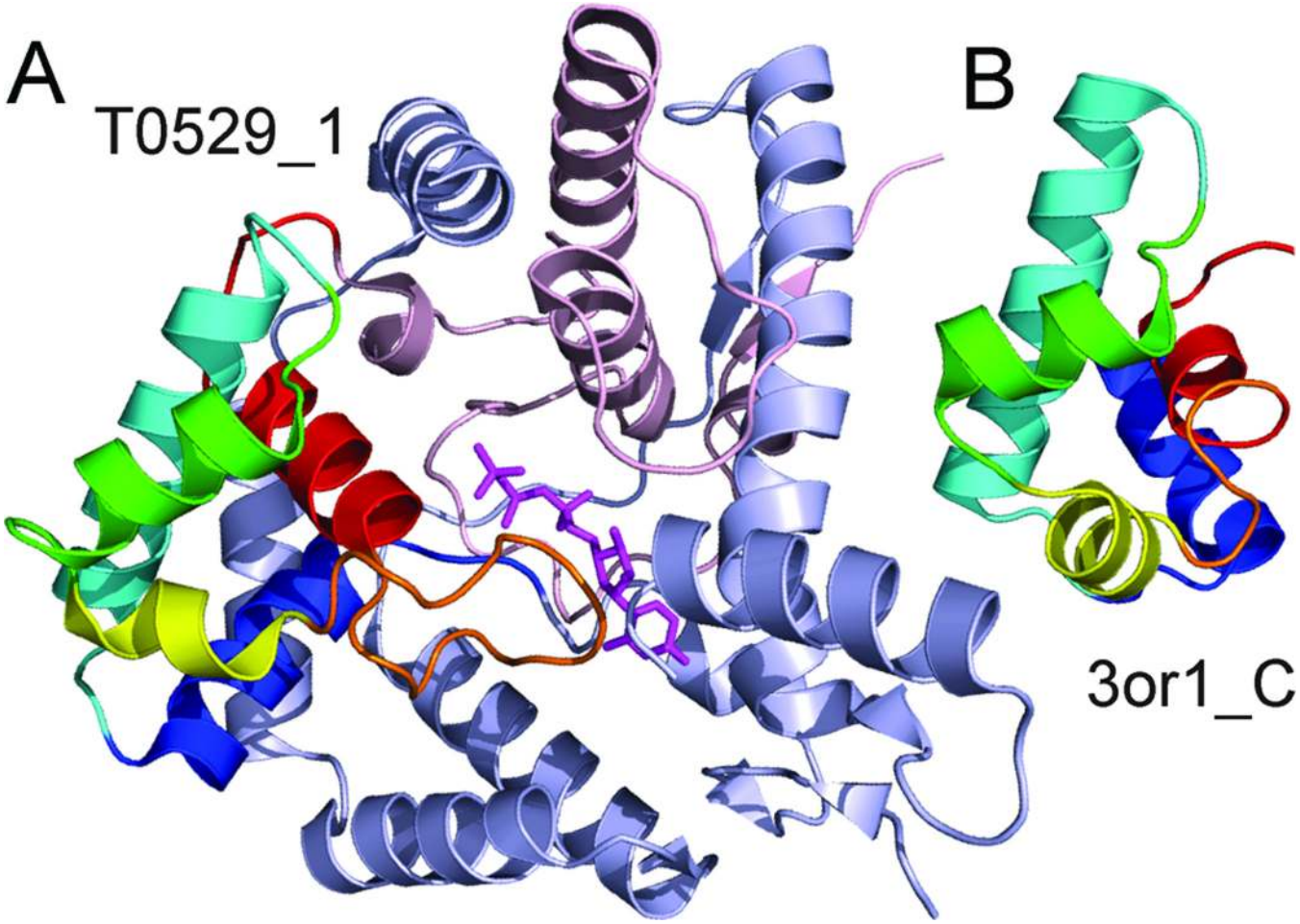

Figure 8. New fold includes small subdomain with local contacts

(A) The target T0529_1 includes a small set of helices (rainbow) that display local contacts and form part of the functional site (magenta). A large number of additional N-terminal (slate) and C-terminal (salmon) secondary structural elements decorate this core. **(B)** ProSMoS and HorA identify an array of helices the C-terminus of the gamma subunit of dissimilatory sulfite reductase I (3or1_C) that are arranged in a similar topology to the defined target sub domain.

Table 1

Overview of CASP9 Targets

Target	Class[a]	Fold[b]	Superfamily[c]	Evolution[d]	Comments[e]
T0515 T0547_1 T0547_2	TBM	Domain of alpha and beta subunits of F1 ATP synthase-like; TIM-barrel	Alanine racemase C-terminal domain -like; PLP-binding barrel	Homolog	unique T0547 domain orientation due to insertions
T0516	TBM	Heme oxygenase-like	Heme oxygenase-like	Homolog	C-terminal extension removed
T0517	TBM	DsrEFH-like	DsrEFH-like	Homolog	
T0518	TBM	Concanavalin A-like lectins/glucanases	Concanavalin A-like lectins/ glucanases	Homolog	
T0520	TBM	Ferredoxin-like	Adenylyl and guanylyl cyclase catalytic domain	Homolog	
T0521_1 T0521_2	TBM	EF Hand-like	EF-hand	Homolog	swap in T0521, altered domain orientation
T0522	TBM	HIT-like	HIT-like	Homolog	
T0523 T0527 T0536 T0600_1 T0600_2	TBM	Profilin-like	PYP-like sensor domain (PAS domain)	Homolog	swap in T0600 not present in existing templates and requires splitting[f]
T0524 T0526 T0609	TBM	Supersandwich	Ga lactose mutarotase-like	Homolog	
T0525	TBM	YktB/PF0168-like	YktB/PF0168-like	Homolog	
T0528_1 T0528_2	TBM	Periplasmic binding protein-like I	Periplasmic binding protein-like I	Homolog	
T0529_1	FM	New		New	Small subdomain resembles DsrC helical array, additional elements needed for function
T0529_2	TBM	Ribonuclease H-like motif	Ribonuclease H-like	Homolog	
T0530	TBM	OB-fold	BC4932-like	Homolog	Swapped N-terminal strand
T0531	FM	Midkine		Homolog	three-stranded

Target	Class[a]	Fold[b]	Superfamily[c]	Evolution[d]	Comments[e]
					meander core; disulfides (2) match
T0532	TBM	alpha-alpha superhelix	TPR-like	Homolog	
T0533_1[g] T0606	TBM	Periplasmic binding protein-like II	Periplasmic binding protein-like II	Homolog	
T0533_2	TBM	Ferredoxin-like	ACT-like	Homolog	
T0534_1	FM	bromodomain-like		Analog	3 helical insert into C-terminal helix; signal peptide
T0534_2	FM	four-helical up-and-down bundle		Analog	insertion into discontinuous domain 1
T0537[h]	FM	Single-stranded right-handed beta helix	Pectin lyase-like	Homolog	Repeated sequence signature
T0538	TBM	Histone-fold	Histone-fold	Homolog	Template not in SCOP;missing N-terminal helix; unswapped
T0539	TBM	RING/U-box	RING/U-box	Homolog	
T0540	TBM	NusG		Analog	circular permutation of NusG fold; template not in SCOP
T0541	TBM	Immunoglobulin-like beta-sandwich	E set domains	Homolog	
T0542_1	TBM	Carbon-nitrogen hydrolase	Carbon-nitrogen hydrolase	Homolog	
T0542_2[g]	TBM	Adenine nucleotide alpha hydrolase-like; HTH	Adenine nucleotide alpha hydrolase-like; HTH	Homolog	
T0543_ 1 T0543_2	TBM	Somatomedin B domain	Somatomedin B domain	Homolog	
T0543_3[g]	TBM	Alkaline phosphatase-like	Alkaline phosphatase-like	Homolog	
T0543_4	TBM	His-Me finger endonucleases	His-Me finger endonucleases	Homolog	

Target	Class[a]	Fold[b]	Superfamily[c]	Evolution[d]	Comments[e]
T0544 _1[i] T0544_2[i] T0553_1 T0553_2 T0555_1[i] T0555_2[i]	FM	EF Hand-like		Homolog	Duplication of two EF Hand-like folds
T0545	TBM	Uracil-DNA glycosylase-like	Uracil-DNA glycosylase-like	Homolog	
T0547_3	FM	Spectrin repeat-like		Analog	3-Helical bundle insertion
T0547_4	FM	N/A		Analog	2-helixpair insertion
T0548_1	TBM	DNA breaking-rejoining enzymes	DNA breaking-rejoining enzymes	Homolog	Deteriorated; Template not in SCOP
T0548_2	TBM	DNA/RNA-binding 3-helical bundle	N-terminal Zn binding domain of HIV integrase	Homolog	Template not in SCOP
T0550_1 T0571_1	TBM/ FM	Immunoglobulin-like beta-sandwich	CalX-like	Homolog	
T0550_2 T0571_2	FM	Streptavidin-like		Analog	8-stranded beta meander barrel
T0551	TBM	ssDNA-binding transcriptional regulator domain	ssDNA-binding transcriptional regulator domain	Homolog	
T0552	TBM	Immunoglobulin-like beta-sandwich	Immunoglobulin	Homolog	
T0557	TBM	IF3-like		Homolog	Template not assigned in SCOP
T0558	TBM	6-bladed beta-propeller	YWTD domain	Homolog	
T0559 T0560 T0586_1 T0617	TBM	DNA/RNA-binding 3-helical bundle	“Winged helix” DNA-binding domain	Homolog	
T0561	FM	DNA/RNA-binding 3-helical bundle		Homolog	bundle elaborated with N and C-terminal helices, HTH functional residues
T0562	TBM	SufE/NifU	SufE/NifU	Homolog	

Target	Class[a]	Fold[b]	Superfamily[c]	Evolution[d]	Comments[e]
T0563 T0573	TBM	Double-stranded beta-helix	Clavaminate synthase-like	Homolog	
T0564	TBM	OB-fold	Nucleic acid-binding proteins	Homolog	Identified through related PF10694
T0565	TBM	SH3-like barrel; Cysteine proteinases	GW domain; Cysteine proteinases	Homolog	
T0566	TBM	Aha1/BPI domain-like	Activator of Hsp90 ATPase, Aha1	Homolog	
T0567	TBM	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases	Homolog	
T0568	TBM	Common fold of diphtheria toxin/transcription factors/cytochrome f	Alpha-macroglobulin receptor domain	Homolog	
T0569	TBM	beta-Grasp (ubiquitin-like)	Immunoglobulin-binding domains	Homolog	
T0570	TBM	TIM beta/alpha-barrel	PLC-like phosphodiesterases	Homolog	
T0572	TBM	Ribosomal protein L25-like	Ribosomal protein L25-like	Homolog	
T0574	TBM	Immunoglobulin-like beta-sandwich	Antigen MPT63/MPB63	Homolog	
T0575_1 T0596_1 T0611_1	TBM	DNA/RNA-binding 3-helical bundle	Homeodomain-like	Homolog	
T0575_2 T0596_2 T0611_2	TBM	Tetracyclin repressor-like, C-terminal domain	Tetracyclin repressor-like, C-terminal domain	Homolog	
T0576	TBM	Heme iron utilization protein-like	Heme iron utilization protein-like	Homolog	
T0578	FM	Restriction endonuclease-like		Analog	Similar topology; different active sites
T0579_1 T0579_2	TBM	SH3-like barrel	Tudor/PWWP/MBT	Homolog	swapped duplication

Target	Class[a]	Fold[b]	Superfamily[c]	Evolution[d]	Comments[e]
T0580	TBM	Phosphotyrosine protein phosphatases I-like	PTS system, Lactose/Cellobiose specific IIB subunit	Homolog	
T0581	FM	Template not in SCOP		New	Similar topology to Fatty acyl adenylate ligase C-terminal domain
T0582_1 T0582_2 T0625	TBM	Double-stranded beta-helix	RmlC-like cupins	Homolog	
T0584	TBM	Terpenoid synthases	Terpenoid synthases	Homolog	
T0585	TBM	Phosphorylase/hydrolase-like	Zn-dependent exopeptidases	Homolog	
T0586_2	TBM	N/A (extension)		Homolog	SCOP classified as extension of HTH
T0588	TBM	alpha/alpha toroid	Chondroitin AC/alginate lyase	Homolog	
T0589_1[g]	TBM	Class II aaRS and biotin synthetases	Class II aaRS and biotin synthetases	Homolog	
T0589_2	TBM	Anticodon-binding domain-like	Anticodon-binding domain of Class II aaRS	Homolog	
T0590	TBM	Immunoglobulin-like beta-sandwich	PKD domain	Homolog	
T0591	TBM	PLP-dependent transferases; DcoH-like	PLP-dependent transferases; PCD-like	Homolog	
T0592	TBM	Rhodanese/Cell cycle control phosphatase	Rhodanese/Cell cycle control phosphatase	Homolog	
T0593	TBM	Nucleotide-diphospho-sugar transferases	Nucleotide-diphospho-sugar transferases	Homolog	
T0594	TBM	TBP-like	Bet v1-like	Homolog	
T0597	TBM	Protein kinase-like (PK-like)	Protein kinase-like (PK-like)	Homolog	
T0598	TBM	Ligand-binding domain NO 4ignaling and Golgi transport	Ligand-binding domain NO 4ignaling and Golgi transport	Homolog	

Target	Class[a]	Fold[b]	Superfamily[c]	Evolution[d]	Comments[e]
T0599	TBM	ADC synthase	ADC synthase	Homolog	
T0601	TBM	Putative modulator of DNA gyrase, PmbA/TldD	Putative modulator of DNA gyrase, PmbA/TldD	Homolog	
T0602	TBM	STAT-like (from SCOP)	Flgn-like	Homolog	Template 3-helix bundle formed from target dimer C-terminal helix swap
T0603 T0603a[j]	TBM (TBM)	New Fold; Fertilization protein		New Analog	Two domains for classification; template not in SCOP
T0604_1	FM	Ferredoxin-like	RNA-binding domain, RBD	Homolog	Characteristic RBD helix packing
T0604_2	TBM	FAD/NAD(P)-binding domain	FAD/NAD(P)-binding domain	Homolog	
T0604_3	FM	FAD-linked reductases, C-terminal domain	FAD-linked reductases, C-terminal domain	Homolog	fused domains, common core with elaborations
T0605	TBM	N/A		N/A	one helix
T0607	TBM	Phosphorylase/hydrolase-like; Ferredoxin-like	Zn-dependent exopeptidases; RNA-binding domain, RBD	Homolog	
T0608_1	FM	Lysozyme-like		Analog	Missing active site β-meander
T0608_2	TBM	Barrel-sandwich hybrid	Duplicated hybrid motif	Homolog	
T0610	TBM	Restriction endonuclease-like	Restriction endonuclease-like	Homolog	
T0612	TBM	Immunoglobulin-like beta-sandwich	Transglutaminase, two C-terminal domains	Homolog	
T0613 T0626	TBM	Ferredoxin-like; Formyltransferase	ACT-like; Formyltransferase	Homolog	
T0614	TBM	PH domain-like	PH domain-like	Homolog	
T0615	TBM	HD-domain/PDEase-like	HD-domain/PDEase-like	Homolog	
T0616	TBM/ FM	HMG-box	HMG-box	Homolog	includes the EEEK motif

Proteins. Author manuscript; available in PMC 2012 October 14.

Target	Class[a]	Fold[b]	Superfamily[c]	Evolution[d]	Comments[e]
T0618	FM	All-alpha NTP pyrophosphatases		Analog	similar topology, unsimilar helix packing
T0619	TBM	SAM domain-like	lambda integrase-like, N-terminal domain	Homolog	
T0620	TBM	Dnase I-like	Dnase I-like	Homolog	
T0621	FM	Galactose-binding domain-like		Homolog	Structure core elaborated with helical inserts
T0622 T0640	TBM	NAD(P)-binding Rossmann-fold domains	NAD(P)-binding Rossmann-fold domains	Homolog	
T0623	TBM	DNA breaking-rejoining enzymes	DNA breaking-rejoining enzymes	Homolog	
T0624	FM	Domain of alpha and beta subunits of F1 ATP synthase-like		Homolog	Cradle-loop barrel unites homologous folds[23]
T0627	TBM	Heme oxygenase-like	Heme oxygenase-like	Homolog	
T0628_1 T0628_2	TBM	Ribonuclease H-like motif	Actin-like ATPase domain	Homolog	helix swapped duplicates
T0629_1	TBM	Receptor-binding domain of short tail fibre protein gp12	Receptor-binding domain of short tail fibre protein gp12	Homolog	
T0629_2	FM	New		New	Unique, elongated fold stabilized by dimerization, Fe binding
T0630	TBM	Cysteine proteinases	Cysteine proteinases	Homolog	active site cys
T0632_1	TBM	Thioesterase/thiol ester dehydrase-isomerase	Thioesterase/thiol ester dehydrase-isomerase	Homolog	An N-terminal helical extension allows dimerization[k]
T0634	TBM	Flavodoxin-like	CheY-like	Homolog	
T0635	TBM	HAD-like	HAD-like	Homolog	
T0636	TBM	PLP-dependent transferases	PLP-dependent transferases	Homolog	
T0637	FM	Fertilization protein		Analog	3 helix bundle,

Target	Class[a]	Fold[b]	Superfamily[c]	Evolution[d]	Comments[e]
					swapped C-terminal helix
T0638	TBM	Ribonuclease H-like motif		Analog	Template not defined in SCOP
T0639	FM	All-alpha NTP pyrophosphatases		Analog	Helix pair interacts to form four helix bundle, requires crossover, longer split helices
T0641	TBM	DAK1/DegV-like	DAK1/DegV-like	Homolog	
T0643	TBM	lambda repressor-like DNA-binding domains	lambda repressor-like DNA-binding domains	Homolog	

[a] See text for discussion of class.

[b] SCOP fold class; some folds are assigned N/A when the structures do not have enough information, secondary structure or interaction information to assign as a fold, for example the single helix in T0605.

[c] SCOP superfamily is assigned for all homologs.

[d] Evolutionary relationship to existing fold classifications: Homologs are linked by sequence, significant structure similarity, or similar functions/active site placement. Analogs display similar topologies and lack evidence for homology. New folds arrange secondary structures in a novel topology.

[e] Comments on unusual structural features that help evolution-based classification. Some targets have close sequence-based templates that are not yet classified in SCOP, we attempted to classify these domains in the spirit of the SCOP database.

[f] Swap evaluated as two domains.

[g] Evaluation unit comprised of united domains.

[h] Evaluation unit combined two domains due to small size of second structural domain and absence of correct predictions.

[i] Evaluated as a single unit by the prediction center, we split the domains to be consistent with the homolog.

[j] This domain is only defined for evolutionary-based classification and is considered together with the rest of the template as a single domain for evaluation.

[k] Potential for evaluating the N-terminal extension as a helical dimerization domain.