Structural relationships among proteins with different global topologies and their implications for function annotation strategies

Donald Petrey, Markus Fischer, and Barry Honig¹

Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032

Contributed by Barry Honig, July 31, 2009 (sent for review March 22, 2009)

It has become increasingly apparent that geometric relationships often exist between regions of two proteins that have quite different global topologies or folds. In this article, we examine whether such relationships can be used to infer a functional connection between the two proteins in question. We find, by considering a number of examples involving metal and cation binding, sugar binding, and aromatic group binding, that geometrically similar protein fragments can share related functions, even if they have been classified as belonging to different folds and topologies. Thus, the use of classifications inevitably limits the number of functional inferences that can be obtained from the comparative analysis of protein structures. In contrast, the development of interactive computational tools that recognize the “continuous” nature of protein structure/function space, by increasing the number of potentially meaningful relationships that are considered, may offer a dramatic enhancement in the ability to extract information from protein structure databases. We introduce the MarkUs server, that embodies this strategy and that is designed for a user interested in developing and validating specific functional hypotheses.

protein fold space | protein function annotation |
protein structure alignment | protein structure similarity

The identification of structural and functional relationships between proteins based on similarities in their amino acid sequence is an essential component of modern biology. It has been recognized for some time that two proteins can be similar to one another in structure despite a lack of any detectable sequence similarity and that this information can be used to assign function. There has been considerable discussion over the past several years as to how structural similarities can most usefully be described (1–8). Widely used databases such as SCOP (9) and CATH (10) describe relationships between proteins using a hierarchy of classifications that reflect similarities in the spatial organization of secondary structure elements (SSEs). For example, proteins with the same overall SSE composition are described as belonging to the same “class,” and proteins with similar spatial arrangements of SSEs are described as belonging to the same “fold” or “topology.” Classification implies discreteness in the organization of structure space in that a protein that is assigned to one class or fold will not belong to another.

An alternative view suggests that protein structure space should be viewed as “continuous” rather than discrete (1, 2, 6, 8). Indeed, it has become apparent that structural relationships between protein domains exist at various scales; from small sets of SSEs (4), to larger fragments (1, 2), even when the proteins have been assigned to different folds and structural classes (3). Such structural and/or functional relationships between fragments of two different proteins have been extensively discussed (5, 7, 11, 12) and pose serious challenges to hierarchical classification schemes (13). For example, the fact that homologous proteins can belong to different folds raises the question of whether fold should be placed above or below homology in the

hierarchy. A number of solutions have been suggested. These include modifications and additions to SCOP (12), new classification levels in CATH (14) and the identification of fragments that connect different folds (5). We have suggested an alternative approach to the problem that involves abandoning structure-based classification and instead relying on structural alignments alone to identify geometric relationships, which are then used as a basis for function annotation (6).

It is possible that, in some instances, geometric similarity simply reflects structural and energetic constraints associated with the packing of SSEs (see e.g. ref. 15). However, there has been increased recognition that there are evolutionary processes that allow proteins to change their global topologies while still maintaining a functional relationship (5, 11, 16–18). A relevant example is provided by the family of phage Cro transcription factors; one member, P22 Cro, is an all- α protein whereas another member, λ Cro, is a mixed α/β protein. Although described as belonging to different folds (19–21), there are local structural similarities between these proteins that reveal important functional information: Both proteins contain helix–turn–helix DNA binding motifs that superimpose to 1.6 Å rmsd.

Although the existence of structural fragments with a common function is implicit in recent discussions of the evolution of protein folds (16, 18, 21), they have not, to our knowledge, been used as part of a general strategy for function annotation. Here, we suggest how this goal can be realized. We first demonstrate that the presence of structurally similar fragments, even in the absence of global sequence or structural similarity, often implies the existence of a functional relationship between two proteins. We illustrate our approach by choosing a number of query proteins, identifying structural neighbors defined by containing at least three aligned SSEs with the query (see *Materials and Methods*) and then selecting a subset of such neighbors that share a common function. Our results suggest that there are a significant number of functional relationships between proteins that have been classified differently. Obscured by classification schemes, these relationships have important implications for the practical goal of extracting function from structure.

Results

Generic Cation-Binding Fragment. Sporulation initiation phosphotransferase (Spo0F) (Fig. 1A), which binds a magnesium ion, is classified as a flavodoxin-like protein in SCOP1.71 and a 3-layered alpha/beta sandwich in CATH3.1.0. There are 87 proteins with this fold in SCOP1.71 and 1,336 proteins with this topology in CATH3.1.0, based on a non-redundant dataset at the 60%

Author contributions: D.P. and B.H. designed research; D.P. performed research; M.F. contributed new reagents/analytic tools; D.P. analyzed data; and D.P. and B.H. wrote the paper.

The authors declare no conflict of interest.

¹To whom correspondence should be addressed. E-mail: bh6@columbia.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907971106/DCSupplemental.

Fig. 1. Alignment of Spo0F with structurally similar proteins. (A) Backbone of Spo0F (PDB entry 1F51, chain E). (B) Backbone of AlaD (1EB3). (C) Backbone of the iron ABC transporter (1Y4T, chain A). The colored regions indicate the structurally similar subset of SSEs shared by 1F51, 1EB3 or 1Y4T. (D) Structural alignment of 1F51E (red), 1EB3 (blue) and 1Y4TA (green). Metals associated with each protein shown as spheres. Only regions that structurally align to 1F51 from either 1EB3 or 1Y4T are shown (also see Fig. 2D). (E) Structure-based sequence alignment of residues 1204–1212 in 1F51 to the structurally equivalent regions of 1EB3 and 1Y4T. Residues in color correspond to metal chelating residues using the coloring in D. Structure-based sequence alignments and rigid-body transformations that relate the proteins discussed in all figures are provided in [SI Appendix](#).

sequence identity level in both cases. In contrast, using our structural alignment program, Ska (1, 22), we find that Spo0F has 1866 neighbors belonging to 97 different SCOP folds, 5 different CATH architectures and 37 different CATH topologies (representative alignments to each fold or topology are provided in the [SI Appendix](#)). Many of these structural neighbors have metal binding sites that occupy the same position in space as the magnesium site in Spo0F, even if the proteins belong to very different folds.

Fig. 1 illustrates the relationship between Spo0F and two of its structural neighbors: 5-aminolaevulinic acid dehydratase (AlaD) from *S. cerevisiae*, which is a zinc binding protein (Fig. 1B), and the iron ABC transporter from *C. jejuni* (Fig. 1C). Seven of the 10 SSE's in Spo0F have equivalent SSE's in AlaD, superposing with an rmsd of 3.2 Å and 6 SSE's in Spo0F have equivalent SSE's in the iron ABC transporter superposing to 4.0 Å rmsd (Fig. 1D). Moreover, the three metals occupy approximately the same position in space within the aligned fragment. A structure-based sequence alignment reveals that some of the liganding residues also align (Fig. 1E) even though the metals are different and hence the identities of their ligands are different. In all, we identify 160 proteins with metal chelating residues that align to those in Spo0F, representing 23 different SCOP folds, 3 different CATH architectures and 11 different CATH topologies.

The structural fragment shown in Fig. 1 can clearly be used to house binding sites for very different metals, but its role appears to be even more general. Fig. 2A shows a structural superposition of Spo0F, a UDP-glucosyl transferase, spermidine synthase, and acetylcholinesterase. In each of these structures the residue occupying structurally equivalent position to the liganding residue D1254 of Spo0F (Fig. 2B and C) interacts with a positively charged amino group either belonging to another amino acid in the protein or to a bound ligand. These include a histidine from UDP-glucosyl transferase, a spermine and acetylcholine. In each case, the positively charged amino group is highlighted in Fig. 2B in sphere representation. The residues structurally equivalent to D1254 are shown in the alignment in Fig. 2C. These include: D121 of 2ACW, which forms an ion pair with His-22 [this pair is thought to act as a base in the catalytic reaction (23)]; D196 of spermine synthase, which interacts with atom N10 of spermine; and E199 in acetylcholinesterase, which is known to play a

role in stabilizing the positively charged acetylcholine substrate (24). As shown in Fig. 2D the structural fragments identified here have six SSE's in common with each of the cation binding fragments described in Fig. 1, suggesting the existence of a conserved motif that stabilizes positive charges.

Conserved Sugar-Binding Fragment. Ligand binding sites also appear to be conserved across folds. Fig. 3 shows the structures of three carbohydrate binding proteins: the sialic acid-binding VP8 domain of the capsid protein from the CRW-8 strain of rotavirus (“jelly roll” fold), mannose-binding garlic lectin (“β-prism II”), and protein RSC2107, a methyl-fucose binding protein from *Ralstonia solanacearum* (“β-propeller”). Despite their different classifications, all three proteins share a conserved substructure (see Fig. 3D) consisting of a three-stranded and a four-stranded β-sheet packed together. The C_α rmsd for the superposition of any two of these fragments is <4.5 Å and, as is evident from Fig. 3A–C, each is associated with carbohydrate binding. The binding sites appear at different locations on the surfaces of the fragment, either packing against the faces of one of the two β-sheets (sites 1 and 3) or in between the two sheets (site 2). The capsid protein binds sialic acids at sites 1 and 2; garlic lectin binds mannose at sites 1 and 3 (as well as a third site not contained in the conserved substructure); and RSC2107 binds fucose at sites 2 and 3.

The results summarized in Fig. 3 suggest that there is a considerable amount of information available in existing databases that could be exploited to infer the location of binding sites. We have developed an approach to identify the location of binding sites based on those observed in structural neighbors. The underlying idea is that the same geometric transformation that aligns one protein, A, that contains a ligand, with another, B, whose ligand binding sites are unknown, will place the ligand of A in the coordinate system of B, suggesting a possible binding site on B (6, 25, 26). Fig. 3E provides an example of this procedure applied to the structural neighbors of the VP8 domain. The ligands from the structural neighbors are colored according to the fold classification of the proteins to which they bind. Three clusters of ligands are identified, corresponding to sites 1, 2 and 3. The VP8 domain only binds sialic acids (shown in blue) in sites 1 and 2 but the location of each of these sites

Fig. 2. A generic cation binding fragment. (A) Multiple structure alignment with Ska of SpoOF (red) acetylcholinesterase (blue, PDB entry 2ACE), spermidine synthase (yellow, 3B7P), and a UDP-glycosyl transferase from *M. truncatula* (green, 2ACW). Only the structurally equivalent residues as determined by the structure alignment are shown in worm representation. (B) Cationic moieties from structural neighbors of SpoOF are shown in sphere representation, colored as in A. These correspond to acetylcholine from 2ACE, a spermine from 3B7P, and a histidine side chain from 2ACW. The strand containing the conserved acidic residue is shown in wire representation and the residue itself is shown at the top of this strand in stick representation. The magnesium contained in 1F51 is shown as a sphere and the positively charged amino group from ligands and H22 from 2ACW are also shown as spheres. (C) Structure based sequence alignment of the strand shown in B with the conserved acidic residue shown in red. (D) The set of SSEs common to all of the proteins shown in this figure and to the metal binding proteins shown in Fig. 1.

can clearly be identified both from proteins in the same fold (ligands in magenta) and different folds (ligands colors in green and red). It is clear from the figure that proteins in the β -prism and β -propeller folds can be used to predict the location of carbohydrate binding sites in proteins in the jelly roll fold, and vice versa.

In addition to the structural relationships presented here, there is evidence for a more general functional relationship between “jelly-roll,” “ β -propeller,” and “ β -prism” proteins, that goes beyond just carbohydrate binding. The jelly-roll proteins facilitate viral entry into bacterial cells whereas the β -propeller proteins facilitate bacterial entry into eukaryotic cells. Although the specific function of the β -prism proteins has not been identified, it is known that they are involved in apoptosis (27). An intriguing possibility, suggested by the common sugar binding function of the three groups of proteins, is that β -prism proteins play a role in autophagy, a mechanism of apoptosis that involves fusion of the phagosome with the lysosome. Autophagy is mediated by a similar mechanism to that used in sugar-mediated viral and bacterial entry; binding to sugar modified proteins on a target membrane.

Identifying Potential Ligands in Proteins of Unknown Function. To further illustrate how structural relationships between protein fragments can be used to infer function, we have predicted the binding function of a structural genomics target that has not yet

Fig. 3. A carbohydrate binding fragment. (A–C) The structures of three carbohydrate binding proteins: the VP8 domain of the capsid protein from the CRV-8 strain of porcine rotavirus (PDB entry 2I2S, magenta and gray ribbon representation) (A), garlic lectin (green and gray, 1KJ1) (B), and protein RSC2107 from *Ralstonia solanacearum* (red and gray, 2BT9) (C). Colored regions are structurally conserved between the three proteins. Cocrystallized ligands are shown in stick representation. (D) The conserved substructure present in all three of the proteins shown in A. The structurally equivalent strands from each protein (i.e., each strand that aligns to a strand from 2I2S based on a structure-based sequence alignment) are colored identically. The largest rmsd between any neighbor and 2I2S was 4.4 Å. (E) The conserved substructure of 2I2S shown in magenta. Carbohydrate ligands from structural neighbors of 2I2S are shown in stick representation and colored according to the fold to which the protein belongs using the color code of 3A. Two sialic acids cocrystallized with 2I2S are shown as blue sticks. The ligands and PDB files from which they are derived are provided in *SI Appendix*.

been annotated. The structure of TM1055 from *Thermatoga maritima* was determined by the Northeast Structural Genomics Consortium (NESG) and there is currently no publication that describes this protein's structure or function. TM1055 has a deep cavity on its surface (Fig. 4A) that is recognized by our SCREEN program (28) as the most likely location on the protein surface to bind a ligand. Following the procedure described in *Materials and Methods*, we identified structural neighbors of TM1055 and rotated any ligands into the coordinate frame of TM1055, retaining those ligands that were close to the cavity. In all, we found 1793 structural neighbors belonging to 70 different SCOP folds, 3 different CATH architectures, 10 different CATH topologies and 48 different CATH homologous superfamilies. These proteins collectively bind nearly 500 distinct ligands as judged by their having unique identifiers in the Protein Data Bank (PDB) file.

Clustering these ligands based on a standard measure of small molecule similarity (see *Materials and Methods*) showed that the largest cluster (259 ligands) contained ligands that were enriched in rings and double bonds. Fig. 4B shows the structure of TM1055 with four of the ligands from this cluster, each of which is associated with a protein in a different SCOP fold and different CATH homologous superfamily. The ligands have been placed within the structure of TM1055 using the same

Fig. 4. Identifying a potential ligand that binds to a protein of unknown function. (A) Molecular surface of protein TM1055 highlighting a cleft identified by the program SCREEN (28) as the most likely ligand binding site on the protein surface, colored by solvent accessibility (42). (B) The structure of TM1055 (PDB entry 1RCU) shown as an orange worm. Four ligands from structural neighbors of TM1055 are shown as colored sticks, oriented in the coordinate system of TM1055 by transforming the coordinates of the ligands according to the transformation that relates the structural neighbor to TM1055. These are a tyrosine from a tyrosyl-tRNA synthetase (red, 1WQ4), an AMP from *M. methylotrophus* electron transfer flavoprotein (green, chain C), an S-adenosylhomocysteine from a methyltransferase (yellow, 9MHT), and a CoA from formyl-CoA transferase (blue, 1VGR). (C) The molecular surface of TM1055 with the aromatic moieties of the ligands from B as magenta sticks. Each aromatic moiety occupies an approximately equivalent position in the cavity identified in A. (D) The set of nine structurally equivalent SSE shared by TM1055 and all four structural neighbors.

procedure described for the sugar binding proteins. Although no ligand fits into the cleft of TM1055 in its entirety, each contains an aromatic moiety that occupies a position that could be fully accommodated by the cleft with minimal conformational change (Fig. 4C) suggesting that TM1055 binds a ligand with an aromatic moiety. Fig. 4D shows the fragment that is common to TM1055 and the four structural neighbors (Fig. 4B). Despite the difference in overall topologies these proteins share a fragment that, in all cases, produces a cleft that can accommodate aromatic moieties. Strong support for the prediction that TM1055 binds an aromatic group is provided by the recently determined structure of the molybdenum cofactor (MOCO) binding protein (MBP) from *C. reinhardtii* (PDB entry 2IZ6), which was solved without a bound ligand. This protein is quite similar to TM1055 and the two proteins share 26% sequence identity. MBP binds MOCO (which contains an aromatic moiety) in a cleft that is structurally equivalent to the prominent cleft on the surface of TM1055; and a residue known to be critical for MOCO binding (M50) is conserved in TM1055 (M48).

Using Relationships Between Fragments to Extract Functional Information. The results presented in the previous sections indicate that structural alignments between proteins that have been classified differently can be used to identify structurally similar fragments that in many cases share a common function. Although the use of only three query proteins may suggest that our results may represent special cases, we emphasize that each choice was essentially random. Spo0F was chosen as an example of a broadly represented set of α/β proteins. VP8 was chosen so as to determine if a conclusion based on its crystal structure, that it contains two sialic binding sites, could have been independently deduced. Finally, TM1055 is a structural genomics target with no known function and it was of interest as a test of how functional hypotheses could be derived from remote structural homologs. Despite a limited number of examples, our results consistently indicate that there are a large number of structurally and potentially functionally related fragments common to proteins classified differently, which can be used to extract functional information from structure. A difficulty with such an approach, however, is the large number of false positives that will inevitably be associated with a less stringent definition of what constitutes “significant” structural similarity. For example, although many different folds contain a metal binding site that is structurally equivalent to the one in Spo0F, the absolute number of proteins containing such a site is small compared with the

total number of structural neighbors identified. How can this difficulty be overcome?

In general, starting with a large set of structural neighbors of a query protein identified independent of classification, a combination of other computational tools can then be used to filter and analyze this set. For example, in the Spo0F analysis, aligned metal liganding residues can be identified from a list of structural neighbors using UniProt sequence “features” (functional annotations that are associated with specific residues in a sequence). Such features can also be used to define a set of functions that occur at a particular location in a structure. In another type of application, patterns in the location of ligand binding sites can be found by restricting the original set of structural neighbors to proteins with a particular GO annotation. For example, all of the proteins shown in Fig. 3 correspond to structural neighbors of VP8 that have the GO annotation, “sugar binding.” The insights that were obtained for these examples suggest the value of a flexible strategy for function annotation that can, under user control, be adapted to the needs of a particular problem.

We have developed a function annotation server, MarkUs (29), that is designed to facilitate a dynamic, interactive strategy that, as suggested by the analyses described above, has the potential to discover functional relationships. MarkUs allows a researcher interested in exploring a particular hypothesis, or developing a new one, to ask specific functional questions (e.g., “What are the possible functions of a region of a protein believed to be functionally important?” or “Where are the likely binding sites on a protein believed to bind carbohydrates?”). Such questions can be addressed through comparative analysis and filtering based on conservation patterns, biophysical properties and existing functional annotation databases. We illustrate this process in Fig. 5, which provides a detailed description of the MarkUs functionality in the context of the analysis of the sugar-binding properties of the viral VP8 domain discussed in Fig. 3.

Discussion

Are the structural and functional similarities described above the result of convergent or divergent evolution? The existence of common structural fragments in proteins with very different global topologies is consistent with recently discussed hypotheses (21, 30, 31) that a relatively small number of ancestral structural modules, which were associated with particular functions (e.g., metal binding, sugar binding, aromatic binding) may each have diverged into a large set of structurally related fragments with

Fig. 5. Representative web page of the MarkUs protein function annotation server highlighting a subset of MarkUs functionalities used in the analysis of ligand binding sites of the structural neighbors of the VP8 domain discussed in Fig. 3. The “annotation map” (A) allows the visualization and analysis of functional data from different sources. The gray lines represent the sequences of a query protein (first line) and its structural neighbors. The magenta rectangles indicate functional residues, in this case, ligand-contacting residues as determined from cocrystallized proteins/ligands available in the PDB. The shaded rectangle indicates a structurally conserved region containing residues that bind ligands in “site 1” of the VP8 domain (see Fig. 3). The figure clearly indicates the conservation, both in overall location and in certain cases individual ligand contacting residues between folds (the last two lines represent β -prism and β -propeller proteins). Hovering the mouse over different areas of the annotation map will display “tool tips” (B) that provide additional functional details. For example, hovering the mouse over the icons in the area to the left of each sequence (C) provides details about each individual protein, including protein name from UniProt, source organism, EC class, and full GO tree. Other types of information can be displayed as well. In this annotation map, hovering over the magenta rectangles would display the identity of the residue/ligand pair. By clicking on the GO annotation within these tool tips (B, underlined), the set of proteins displayed can be restricted to those that share that particular annotation (“sugar binding” in this figure). (D and E) The information displayed on the annotation map can be changed using the controls to the left. In this case, all ligand contacts (excluding solvent) are displayed but this can be restricted to ligands of a certain type based on the ChEBI ligand classification (D). The menu (E) at the top left can be used to display a wide array of other structural and functional properties including UniProt sequence features, sequence conservation, protein–protein interactions, SNPS, and secondary structure. (F) Colored boxes in this region indicate residues lining cavities identified by the program SCREEN colored by conservation (dark red, highly conserved; blue, conserved; and black, unconserved).

diverse but related sets of functionalities. This picture of the evolution of protein structure and function suggests that homologous proteins undergo structural changes that result in their potentially being classified differently even though they all retain a structurally conserved functional fragment (see also refs. 7, 16–18, and 21). Such fragments would then be represented in the current repertoire of proteins either as isolated domains with a single functionality or as components of more complex domains with multiple functions.

However, we do not exclude the possibility that some of the relationships we have identified are the result of convergent evolution. For example, in the case of the Spo0F, the cation binding sites are located in “crevices” that are formed naturally when loops connecting β -strands appear on opposite sides of a β -sheet (32). Thus, such sites could in principle have evolved independently. More generally, Russell et al. (25) identified “supersites” that are structurally equivalent ligand binding sites shared by a group of proteins, which in some cases were classified as belonging to a different fold. Although this was used as an argument that supersites arose from convergent evolution, the process of evolutionary drift (18) suggests that this need not be the case.

Independent of evolutionary origins, the fact that structurally similar fragments often share a common function even if they are incorporated in proteins with different global topologies indicates that valuable information will be lost if annotation is based on discrete classifications. Thus, searching for structural relationships independent of classification should significantly increase the number of functional inferences that can be derived. Of course, one could consider defining a new category of fold-independent structural modules with a common function but we believe that any structure-based classification scheme will

necessarily be restrictive. Rather, our analysis suggests the need for a new generation of computational and data management tools that allow a user to explore sequence, structural and functional databases in an interactive fashion and to develop and validate hypotheses without the limitations imposed by pre-defined decisions about which relationships are meaningful. We have outlined a general strategy, embodied in the MarkUs server, which is specifically designed with this goal in mind.

We have focused here on the use of structural similarity between local regions of the protein backbone. The many tools that have been developed recently to identify functional sites based, for example, on specific configurations of small sets of amino acids (8, 33–35) or similarities in local surface patches (36–38) suggest that even more remote structural similarities could be usefully exploited. A potentially powerful strategy would be one that combines searching broadly for possible structural relationships with filtering based on specific local features (e.g., ligand binding residues, active site and cavity shape comparisons). However, this kind of residue-specific information is not generally used as component of function annotation strategies. Although UniProt sequence features do provide residue-specific information, they lack the much richer and versatile descriptions of function provided by databases such as GO. An explicit association of such “controlled vocabularies” with particular structure/sequence features would overcome some of the inherent shortcomings of annotation transfer based solely on overall sequence/structural similarity (39) and allow the primarily manual analyses we describe here to be carried in a more automated way.

Finally, it is important to compare the approach proposed here with that embodied in the function annotation servers that

have been developed in recent years [e.g., ProFunc (40) and ProKnow (41)]. The design of these servers reflects, to some extent, the fact that structural genomics initiatives have generated a large number of protein structures for which there is little or no functional information. Function annotation servers apply a variety of advanced sequence and structure analysis tools to these proteins and provide a user with web pages that contain functional inferences. In contrast, the strategy we are proposing is intended primarily for a researcher with expertise in a particular protein or family of proteins who wishes to develop or validate specific functional hypotheses. This type of goal can best be addressed with interactive computational tools that do not rely on predefined classifications and which allow a researcher to make decisions about which relationships are likely to be meaningful in the context of a particular problem. We believe that this perspective will also enhance the value of structural genomics initiatives, by maximizing the number of relationships between proteins that can be discovered and by facilitating a more unrestricted navigation in sequence–structure–function space.

Materials and Methods

All proteins were structurally aligned with the program Ska (22), a version of the structure alignment program PrISM (1), which was modified to allow alignments to be considered significant even if only a fragment of one protein is aligned to the other. We require a minimum of three secondary structure

elements to define a fragment. A 60% non-redundant database of proteins from the PDB was searched for structural homologs of each query protein. Proteins with a protein structural distance (PSD) (1) <0.8 were kept for functional analysis. Alignments and transformations between Spo0F and representative structures classified as belonging to different folds/topologies are provided in the *SI Appendix*. Ligands from structural neighbors were placed into the coordinate system of each query protein by transforming the coordinates of the ligand using the transformation that structurally superimposes the query protein and structural neighbor. To identify proteins with similar metal binding sites to Spo0F, structural neighbors were examined to identify which ones had at least one metal chelating residue (taken to be Asp, Glu, His and Cys within 4 Å of Ca, Mg, Zn, K, Na, Mn, Cu, or Fe ion) that aligned to one of the metal chelating residues of Spo0F. The ligands in Fig. 2 were identified by manually examining a subset of the structural neighbors of Spo0F. Ligands shown in Fig. 3 were derived from structural neighbors of the VP8 domain, which had the GO annotation “sugar binding.” Ligands shown in Fig. 4 were identified by considering only those structural neighbors that had ligands with atoms that fell within 4 Å of any solvent accessible atom of the residues lining the cavity identified by SCREEN (28) after transformation into the coordinate system of TM1055. Clustering of the ligands for Fig. 4 was carried out using the Jarvis-Patrick algorithm with ligand similarity determined using a Tanimoto distance (see *SI Appendix* for details).

ACKNOWLEDGMENTS. We thank Fabian Dey for carrying out the clustering of ligands identified in the structural neighbors of TM1055. This work was supported by National Institutes of Health Grants GM030518, GM074958, and CA121852.

1. Yang AS, Honig B (2000) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol* 301:665–678.
2. Shindyalov IN, Bourne PE (2000) An alternative view of protein fold space. *Prot: Struct Func Gen* 38:247–260.
3. Kihara D, Skolnick J (2003) The PDB is a covering set of small protein structures. *J Mol Biol* 334:793–802.
4. Szustakowski JD, Kasif S, Weng Z (2005) Less is more: Towards an optimal universal description of protein folds. *Bioinformatics* 21:i166–171.
5. Friedberg I, Godzik A (2005) Connecting the protein structure universe by using sparse recurring fragments. *Structure* 13:1213–1224.
6. Kolodny R, Petrey D, Honig B (2006) Protein structure comparison: Implications for the nature of ‘fold space’, and structure and function prediction. *Curr Opin Struct Biol* 16:393–398.
7. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA (2006) Structural diversity of domain superfamilies in the CATH database. *J Mol Biol* 360:725–741.
8. Xie L, Bourne PE (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci USA* 105:5441–5446.
9. Andreeva A, et al. (2004) SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucl Acids Res* 32:D226–229.
10. Pearl FMG, et al. (2003) The CATH database: An extended protein family resource for structural and functional genomics. *Nucl Acids Res* 31:452–455.
11. Grishin NV (2001) Fold change in evolution of protein structures. *J Struct Biol* 134:167–185.
12. Andreeva A, Plic A, Hubbard TJP, Murzin AG (2007) SISYPHUS—structural alignments for proteins with non-trivial relationships. *Nucl Acids Res* 35:D253–259.
13. Taylor WR (2007) Evolutionary transitions in protein fold space. *Curr Opin Struct Biol* 17:354–361.
14. Cuff AL, et al. (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res* 37:D310–314.
15. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc Natl Acad Sci USA* 103:2605–2610.
16. Alva V, Koretke KK, Coles M, Lupas AN (2008) Cradle-loop barrels and the concept of meta-folds in protein classification by natural descent. *Curr Opin Struct Biol* 18:358–365.
17. Andreeva A, Murzin AG (2006) Evolution of protein fold in the presence of functional constraints. *Curr Opin Struct Biol* 16:399–408.
18. Krishna SS, Grishin NV (2005) Structural drift: A possible path to protein fold change. *Bioinformatics* 21:1308–1310.
19. Newlove T, Konieczka JH, Cordes MH (2004) Secondary structure switching in Cro protein evolution. *Structure* 12:569–581.
20. Roessler CG, et al. (2008) Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proc Natl Acad Sci USA* 105:2343–2348.
21. Davidson AR (2008) A folding space odyssey. *Proc Natl Acad Sci USA* 105:2759–2760.
22. Petrey D, Honig B (2003) GRASP2: Visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol* 374:492–509.
23. Shao H, et al. (2005) Crystal structures of a multifunctional triterpene/flavonoid glycosyltransferase from *Medicago truncatula*. *Plant Cell* 17:3141–3154.
24. Wlodek ST, Antosiewicz J, Briggs JM (1997) On the mechanism of acetylcholinesterase action: The electrostatically induced acceleration of the catalytic acylation step. *J Am Chem Soc* 119:8159–8165.
25. Russell RB, Sasienski PD, Sternberg MJ (1998) Supersites within superfolds. Binding site similarity in the absence of homology. *J Mol Biol* 282:903–918.
26. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci USA* 105:129–134.
27. Karasaki Y, Tsukamoto S, Mizusaki K, Sugiura T, Gotoh S (2001) A garlic lectin exerted an antitumor activity and induced apoptosis in human tumor cells. *Food Res Intl* 34:7–13.
28. Nayal M, Honig B (2006) On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins* 63:892–906.
29. <http://luna.bioc.columbia.edu/honiglab/mark-us>
30. Dokholyan NV, Shakhnovich B, Shakhnovich EI (2002) Expanding protein universe and its origin from the biological Big Bang. *Proc Natl Acad Sci USA* 99:14132–14136.
31. Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol* 134:191–203.
32. Branden CI (1980) Relation between structure and function of alpha/beta proteins. *Q Rev Biophys* 13:317–338.
33. Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis. *Bioinformatics* 19:1644–1649.
34. Wang K, Samudrala R (2005) FSSA: A novel method for identifying functional signatures from structural alignments. *Bioinformatics* 21:2969–2977.
35. Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. *J Mol Biol* 285:1887–1897.
36. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* 21:2347–2355.
37. Ferre F, Aiusiello G, Zanzoni A, Helmer-Citterich M (2005) Functional annotation by identification of local surface similarities: A novel tool for structural genomics. *BMC Bioinformatics* 6:194.
38. Binkowski TA, Adamian L, Liang J (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol* 332:505–526.
39. Petrey D, Honig B (2009) Is protein classification necessary? Toward alternative approaches to function annotation. *Curr Opin Struct Biol* 19:363–368.
40. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: A server for predicting protein function from 3D structure. *Nucl Acids Res* 33:W89–93.
41. Pal D, Eisenberg D (2005) Inference of protein function from protein structure. *Structure* 13:121–130.
42. Nicholls A, Sharp KA, Honig B (1991) Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11:281–296.