Published in final edited form as:

Proteins. 2016 December ; 84(12): 1859–1874. doi:10.1002/prot.25169.

Development of a motif-based topology-independent structure comparison method to identify evolutionarily related folds

Joseph M. Dybas^{a,b} and Andras Fiser^{a,b,*}

^aDepartment of Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue Bronx, NY 10461, USA

^bDepartment of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue Bronx, NY 10461, USA

Abstract

Structure conservation, functional similarities and homologous relationships that exist across diverse protein topologies suggest that some regions of the protein fold universe are continuous. However, the current structure classification systems are based on hierarchical organizations, which cannot accommodate structural relationships that span fold definitions. Here we describe a novel, supersecondary-structure motif-based, topology-independent structure comparison method (SmotifCOMP) that is able to quantitatively identify structural relationships between disparate topologies. The basis of SmotifCOMP is a systematically defined supersecondary-structure motif library whose representative geometries are shown to be saturated in the Protein Data Bank and exhibit a unique distribution within the known folds. SmotifCOMP offers a robust and quantitative technique to compare domains that adopt different topologies since the method does not rely on a global superposition. SmotifCOMP is used to perform an exhaustive comparison of the known folds and the identified relationships are used to produce a non-hierarchical representation of the fold space that reflects the notion of a continuous and connected fold universe. The current work offers insight into previously hypothesized evolutionary relationships between disparate folds and provides a resource for exploring novel ones.

Keywords

fold evolution; structure classification; Smotif; supersecondary structure motif; structure comparison

Introduction

Approximately 120,000 structures have been experimentally solved and deposited in the Protein Data Bank¹, and the majority of these structures can be classified into ~1,200–1,400 known folds^{2,3}. These fold definitions can be used to structurally characterize about 40% of the ~60 million known sequences, on the residue level⁴. The categorization and classification of the vast number of sequences and structures is facilitated by the observation

*Corresponding author: andras.fiser@einstein.yu.edu, phone: 718-678-1068, fax: 718-678-1019.

that related sequences adopt similar structures⁵. This approach implies an evolutionary view of monotonously diverging sequences within a discrete set of folds, where homology does not exist across fold definitions. This gave rise to fold classification systems that are organized along a strict hierarchy^{6,7}. However, several recent studies argue that at least some regions of the fold space are so dense that there is a continuum of structural relationships between folds^{8–14}. Homologous structures can adopt different folds^{15–19} and conserved functional site motifs can impart functional similarities that predict evolutionary relationships between folds where no significant global sequence or structure similarity exists^{20,21}. Evolutionary mechanisms that can cause fold change among homologues include circular permutations, insertions/deletions/substitution of secondary structures or rearrangements of β -strands to change β -sheet topologies¹⁶, gene duplication and subsequent divergence^{15,17}, differences in oligomeric states¹⁵ or conservation of “ancient peptides”¹⁸.

Structures that are related in evolution but have different folds often exhibit some degree of local sequence and structure conservation. One such example is the heterogeneous nuclear ribonucleoprotein K fold and the ribosomal protein S3 fold, which share a conserved KH domain, an RNA-binding helix-helix motif with a distinct sequence and structure signature²². These two KH domain folds adopt a similar architecture but different topologies, and are classified differently in both SCOP⁶ and CATH⁷. Another example is the swapped-hairpin fold and double-psi β -barrel fold, which are both built around a GD box, a $\beta\alpha\beta$ motif with a distinct structure and a conserved Gly-Asp sequence signature^{23–25}. These two GD box folds are not related by a direct evolutionary pathway but likely evolved from an ancestor fold containing the $\beta\alpha\beta$ motif²³. Homologous proteins may undergo sequence divergence that can induce changes in the structures that are drastic enough to produce different fold definitions. For example, the P22 Cro like fold (all- α) and the λ Cro like fold ($\alpha+\beta$) have up to 40% sequence identity and each fold contains a conserved helix-turn-helix DNA-binding motif in their N-terminal region. Yet, the C-terminal regions exhibit stark structural differences, where the P22 Cro like fold is composed of α helices while the λ Cro like fold β strands^{26–28}. Duplication and subsequent divergence of super-secondary structure elements can build folds that evolve from a common origin but diverge to adopt different global topologies. For example, repeated β -hairpin structures combine to build various outer membrane β -barrel domains²⁹ and amplification and differentiation of β -meander motifs act to build β -propellers³⁰ and related folds³¹. Additionally, the $(\alpha\beta)_4$ half-barrel is the common origin of the very diverse $(\beta\alpha)_8$ -barrel fold^{32–34} and is shown to link the $(\beta\alpha)_8$ -barrel and the flavodoxin-like superfolds in evolution^{35,36}. In-depth sequence and structure analysis showed that a minimal $(\alpha\beta)_2$ element is conserved in the $(\beta\alpha)_8$ -barrel and flavodoxin-like folds, as well as an intermediate structure, and this motif is the likely ligand binding site in many of the protein families. The $(\beta\alpha)_8$ -barrel and flavodoxin-like folds may have evolved divergently from a larger domain with a functional constraint on the $(\alpha\beta)_2$ element or from repetition and embellishment of the $(\alpha\beta)_2$ element³⁵.

Structure classification groups protein domains according to their structural and evolutionary relationships and can identify distant homologues that have diverged to such an extent that sequence-based homology detection is no longer possible³⁷. Thus, structure-based classifications are critical importance in light of structural genomics efforts since the most

interesting targets for experimental structure solutions are proteins with no known homologous sequences⁴. Furthermore, structure classification can guide the development of evolutionary based modeling approaches^{38–40}, aid in methods for targeted protein design^{41,42} and facilitate functional annotation⁴³.

SCOP⁶ and CATH⁷, the two most widely used protein structure classification systems, are both based on a hierarchical organization in which the top levels are structure-based groupings and the lower levels group domains with suspected or known evolutionary relationships. The SCOP “Fold” and “Superfamily” levels are analogous to the CATH “Topology” and “Homologous Superfamily” levels, respectively, and these are the most critical when evaluating evolutionary relationships between structures. The Fold/Topology is defined by the arrangement and connectivity of the secondary structure elements within a structure. The structural similarity found within the Fold/Topology groups includes examples of divergent evolution of homologous structures but also may be a product of physical and chemical properties that dictate the specific packing of the structures and, thus, may include convergently related domains⁶. The Superfamily/Homologous Superfamily is one level below the Fold/Topology in the hierarchy and groups domains where there are suspected evolutionary relationships as defined by significant sequence identity and/or functional similarity. The Superfamily/Homologous Superfamily level is considered the most important and challenging aspect of structure classification⁴⁴.

Structure classification is an outstanding problem that is continuously being addressed within the field of structural biology. The Evolutionary Classification Of protein Domains (ECOD) database⁴⁵ is a more recently developed hierarchical classification, which groups homologous structures without the requirement of fold similarity that is inherent in the SCOP and CATH classifications. In some cases, the homology-based groups merge SCOP Folds or CATH Topologies, which demonstrates that evolutionary relationships may be obscured when defined only in the context of fold similarity. SCOP2⁴⁶ is a prototype that is being developed as a successor to SCOP. The new database abandons the hierarchical classification of the original SCOP in favor of a graph-based organization where the nodes represent groups of structures that share specified relationships. The current beta release covers only a small subset of the PDB.

The two major components of structure classification are the determination of evolutionary relationships between structures and the grouping of structures based on the identified relationships. The current structure classification methods employ manual^{6,46} or semi-automatic^{7,45} approaches to identify the structural and evolutionary relationships that guide the classification of domains. Manual curation is advantageous in evaluating domains that are difficult to assign and can potentially scrutinize non-trivial fold relationships but it is not feasible to keep pace with the continually increasing number of new structures⁴. Also, manual curation is an inherently subjective approach and, as a consequence of this, there are a substantial number of differences in fold definitions and domain classifications between SCOP and CATH^{47–49}. On the other hand, automated approaches necessitate the use of structure comparison methods^{50–53}, which are appropriate for measuring similarities between structures with similar global topologies for the purpose of fold identification but are usually not designed to detect local relationships between domains that have dissimilar

folds. There is an inherent difficulty in comparing and classifying folds based only on a structure comparison metric and in the absence of an evolutionary context^{14,54–56}.

In addition to the challenge of addressing the comparison and grouping of structures within a structure classification, there is a conceptual shortcoming of doing classification within the framework of a hierarchical system^{6,7} because it can not accommodate structures that are related in evolution but have different folds. An important consequence of classification within a hierarchy is that an assignment of a domain into a branch of the hierarchy precludes the existence of structural or evolutionary relationships with domains in other branches. Moreover, differences between branches of the hierarchy are not quantified and are implicitly assumed to be equal. Hierarchical classifications impart a view of a discrete fold space, which appeared to be a logical and natural choice when structural coverage of the fold universe was limited but is not suitable to accommodate all of the complex relationships that can exist between protein folds. In order to automatically and quantitatively capture relationships among protein folds, a new approach to evaluating structure similarity and classifying the identified relationships is required.

The current work describes a supersecondary-structure motif (Smotif) based, topology-independent structure comparison method (SmotifCOMP) that was developed to address the systematic identification of non-trivial evolutionary relationships between protein domains, including between disparate topologies, in an automated and quantitative manner. The Smotif-based comparison is rooted in the observation that there is a finite set of Smotif building blocks that is sufficient to build all possible known and novel folds⁵⁷. SmotifCOMP finds conserved Smotifs within compared domains by implementing an Smotif-based local alignment. The Smotif similarity is converted to a statistical Z-score, in order to quantitatively define significant fold relationships as opposed to trivial fragments of structure similarity. Relationships that are identified within an exhaustive comparison of SCOP Superfamilies are represented in a non-hierarchical, network-based system that reflects the notion of a continuous fold space more accurately than hierarchical classification schemes.

When inferring evolutionary relationships based on structural conservation alone, it is difficult to discriminate instances of homology versus analogy. A detectable sequence similarity is generally indicative of a homologous relationship between structures but in some cases proteins may have diverged to such an extent that a sequence signal cannot be detected^{37,58}. Quantitatively identifying structural relationships between domains is a necessary step for structure classification and offers the utility of informing a functional or evolutionary analysis, even if homology or analogy (convergent evolutionary event⁵⁹) is not immediately obvious. However, conservation of local structure motifs is most often the hallmark of predicted evolutionary relationships between disparate topologies^{20,22,23,28,31,35,60}, therefore, it is expected that SmotifCOMP detects structural similarities that are typically based on evolutionary relationships. SmotifCOMP and the network-based representation of the fold universe offer further insight into many previously reported examples of evolutionary relationships between folds and provide a systematic resource to explore new ones.

Methods

Smotif definition and representation

An Smotif is defined through the spatial orientation of two secondary structures that are consecutive in primary sequence, connected by a loop. Any protein structure can be characterized by the sequence of overlapping Smotifs that is generated from the specific secondary structure geometry of the respective structure (Fig. S1). The length requirement for a secondary structure to define an Smotif is 5 residues for a helix or 2 residues for a strand. There is no limit on the length of the intervening loop and a valid Smotif may include unresolved loop residues but a single Smotif cannot span multiple chains within a structure. Smotifs are classified by their secondary structure type; helix-loop-helix (HH), helix-loop-strand (HE), strand-loop-helix (EH) and strand-loop-strand (EE).

An Smotif is represented by two vectors corresponding to the two secondary structures that comprise the Smotif and are calculated as the principal eigenvectors of the moments of inertia of the backbone atoms of the respective secondary structures. The complete secondary structure is not considered in cases where there is a severe bend in the helix or strand. To transform the secondary structure to a vector representation, an initial vector is calculated using the first nine residues from the loop for a helix or three residues from the loop for a strand (the entire secondary structure is considered if it is shorter than the respective thresholds). Next, a subsequent vector is calculated in the same manner as the original vector but by extending the length from the loop by one residue. If the angle between the two vectors is less than five degrees the longer vector is kept. The process iterates over each residue of the secondary structure until all residues are considered or until a subsequent vector deviates by more than five degrees. Visual analysis suggested that the five-degree threshold was generous enough to allow for the natural curvature of strands and helices but was sufficient to terminate the vector elongation step in the case of sever bends or kinks in secondary structures. The length of the final vector is transformed to equal the length of the secondary structure.

Secondary structure assignment for Smotif characterization

DSSPcont⁶¹ is used to assign the secondary structure type for each residue of the structure. DSSPcont mimics the inherent dynamics of a protein by assigning a confidence level for each of eight secondary structure types (G, H, I, T, E, B, S, L) using a weighted average of DSSP⁶² results for a range of hydrogen bond distance thresholds. Residue-level DSSPcont confidence measurements (a percentage between 0 and 100) are transformed into secondary structure assignments using two criteria. First, the DSSPcont confidence level must be greater than 10 for either a helix (G, H or I) or strand (B, E) residue assignment. In addition, the potential helix or strand residue must be adjacent to a residue of similar secondary structure assignment with a confidence level greater than 50. If the residue does not meet the above two requirements it is assigned as a loop residue. The residue level secondary structure assignments are then filtered in a three-state classification: helix (alpha (H), 3_{10} (G) and pi (I)), strand (beta-bridge (B) and extended beta strand (E)) and loop (turn (T), bend (S), other/loop (L) and unassigned residues). Any stretches of consecutive helix or strand

residues that do not meet the length requirements for the Smotif definition (5 or 2 residues, respectively) are reassigned as loops.

Comparisons of Smotifs by RMSD measurement

Structural similarity is measured only between Smotifs with the same secondary structure types. Each vector of the compared Smotifs is transformed to a uniform length equal to the average distance between the midpoints of the two vectors of each Smotif. The transformation is performed so that neither the lengths of the secondary structures nor the distance between them overwhelms the Smotif comparison. The pair of vector systems are optimally superposed using a quaternion-based method to calculate the largest eigenvalue of the characteristic polynomial of the key matrix⁶³. The RMSD is measured for the four corresponding vector end-points in the vector representations of the compared Smotifs.

Smotif library

A library of Smotifs was constructed by decomposing each structure in the January 2012 PDB¹ into its constituent Smotifs (the PDB was filtered to include only X-ray structures with resolution $\leq 3 \text{ \AA}$ and R-factor ≤ 0.3 , structures were not filtered based on sequence redundancy). There are 2,052,870 Smotifs from 130,555 protein chains (56,158 structures) contained in the library. The data characterizing each Smotif (type, start and end residue number, secondary structure lengths, vector representations, etc.) is stored in a MySQL relational database.

The Smotifs are clustered by secondary structure type (HH, HE, EH or EE) and, within each secondary structure type, by geometry. Due to the size of the library, which would necessitate $\sim 4 \times 10^{12}$ pairwise comparisons to generate a full distance matrix, a complete hierarchical clustering cannot be performed. Instead, a representative set of Smotifs is collected from the library and clustered and the remaining Smotifs are added to the cluster with the most similar representative Smotif. Specifically, a representative set of 11,068 representative Smotifs (2,775 HH Smotifs, 2,167 HE, 2,063 EH, 4,063 EE) is established by transforming one randomly selected domain from each SCOP Fold group into its constituent Smotifs. An all-by-all RMSD measurement is made on these representative Smotifs, within each supersecondary structure type. The representative set is clustered, at $2.5 \text{ \AA}$ RMSD threshold, using the “neighbor” algorithm implemented in the PHYLIP software package⁶⁴. The $2.5 \text{ \AA}$ threshold was determined by performing a manual analysis of corresponding Smotifs in pairs of domains in ten different Superfamilies and examining the distribution of corresponding Smotif pair RMSD measurements. Next, an RMSD comparison is calculated for each of the remaining 2,041,802 Smotifs in the library compared to each Smotif in the representative set. The remaining Smotifs are added to the cluster of the most similar representative Smotif. The representative set of Smotifs cluster into 1,780 representative geometries. The initial 1,780 representative clusters classify 2,047,902 Smotifs (99.76% of the library) without the need to establish a new cluster. The remaining 4,968 Smotifs, which are different from all representative Smotifs by $> 2.5 \text{ \AA}$, are re-clustered at $2.5 \text{ \AA}$ threshold to form 515 additional clusters. A total of 2,295 Smotif cluster geometries are present in the library (822 HH clusters, 461 HE, 401 EH, 611 EE).

Smotif-based topology-independent structure comparison (SmotifCOMP) method

The SmotifCOMP method consists of the following main steps: all-by-all comparison of Smotifs between two structures; alignment of Smotifs; alignment optimization; scoring of Smotif matches (Fig. 1).

All-by-all individual Smotif similarity scores—The two structures being compared are decomposed into their Smotif components and an all-by-all Smotif RMSD measurement is performed across the structures. Each raw RMSD value is converted to a percentile based on a comparison to a background distribution of RMSD values for Smotif comparisons of the same secondary structure type as the original comparison. The background distribution is generated from Smotif pair comparisons derived from all-by-all Smotif-based comparisons of 500 randomly selected SCOP Superfamily domains (a total of 8,401,452 Smotif pair comparisons). Smotif similarity scores are produced by weighting the RMSD percentile values according to the relative lengths of the corresponding secondary structures of the compared Smotifs. The secondary structure length ratios are calculated for the corresponding Smotif N-terminal secondary structures and for the corresponding C-terminal secondary structures. The smaller of the two secondary structure length ratios (larger difference in corresponding secondary structure lengths) is used as the input to a sigmoid function

$$\frac{1}{1+e^{-7(\text{SecStrLengthRatio}-0.5)}}$$

to generate a weight value in the range of zero to one. The weight is multiplied by the RMSD percentile to generate the Smotif similarity score. Large differences in corresponding secondary structure lengths cause the RMSD percentile values to be down-weighted, which generate lower Smotif similarity scores, compared to Smotifs with similar corresponding secondary structure lengths.

Smotif alignment—Local dynamic programming alignment is used to locate the largest continuous stretch of Smotifs between structures that otherwise may exhibit different global topologies. The topology independence of the method is guaranteed at the level of the Smotif comparisons, while the local dynamic programming explores expanding all Smotif matches to the largest possible substructure of Smotifs. The Smotif similarity scores are used as the substitution values in evaluating the recurrence relations to generate the F-matrix. Gaps are defined in the F-matrix if the aligned Smotifs are of different secondary structure types. Optimal local alignments of Smotifs are generated using the Smith-Waterman algorithm⁶⁵, with an affine gap score based on the number of predicted embellishment secondary structures within an inserted Smotif. Since the Smotif similarity scores range between 0 and 1, a low similarity score (<0.1), indicating a suboptimal alignment, is transformed to a gap in the F-matrix in order to generate only high quality local alignments. The gap score and the local suboptimal alignment score cutoff were optimized to maximize the performance of the method in recapitulating the SCOP classification of a 500 Superfamily test set.

In a small number of cases (<10%) the dynamic programming creates an invalid alignment due to a misscorrespondence in overlapping secondary structures in the Smotif characterization of the structures, where a secondary structure in one structure to be aligned to two different secondary structures in the compared structure, resulting in an invalid Smotif alignment. These alignments are discarded and an exhaustive alignment search is performed. The exhaustive search quickly becomes intractable for large structure comparisons but this problem is alleviated by dynamically filtering invalid and sub-optimal alignments. In practice, a vast majority (>90%) of the alignments are successfully implemented by the dynamic programming algorithm without the need to perform an exhaustive search.

Alignment optimization by identifying and eliminating embellishment secondary structures—Structures with a similar classification may contain “considerable elaboration”⁴⁴ of their fold by the presence of extraneous secondary structures. These additional secondary structures do not contribute to the core of the fold but do have a significant effect on the Smotif characterization of the structure. In order to identify embellishment secondary structures in an automated manner, the algorithm optimizes the comparison by exhaustively testing all possible alignments with any combination of up to three secondary structures ignored in each structure. If the structure comparison alignment score is improved when a secondary structure is ignored, the ignored secondary structures are predicted to be embellishments and are eliminated from the Smotif characterization. Only helices less than or equal to ten residues and strands less than or equal to four residues are candidates to be ignored. Additionally, not more than 50% of the elements can be ignored in any given fold.

Structure comparison scoring—The raw local Smotif alignment scores are a function of the size of the compared structures. Therefore the raw alignment scores are converted to statistical Z-scores, based on the total number of Smotifs in the compared structures, in order to identify significant relationships. Background distributions for alignment scores were established by generating and aligning 500 pairs of random sequences of Smotifs, which represent hypothetical structures of specified sizes. To generate a hypothetical structure, individual Smotifs are randomly selected and appended with the requirement that the overlapping secondary structures must be the same type (helix or strand) and length. The observation that the current Smotif library is saturated and sufficient to build any existing or novel fold (shown here and in⁵⁷) ensures that the Z-score calculation will remain robust even if and when novel folds are discovered in the future, since novel folds are generated by a unique combination of known Smotifs and are not expected to introduce novel Smotif geometries.

SmotifCOMP Superfamily comparisons

Redundant proteins are removed within each SCOP Superfamily by performing CD-HIT⁶⁶ clustering at 40% sequence identity level. Three non-redundant domains are randomly selected as the representatives of their respective SCOP Superfamilies. If there are only two structures available they are both selected. Superfamilies with only one non-redundant structure are not analyzed. To compare two Superfamilies, an all-by-all SmotifCOMP

comparison of the representative domains is performed and the individual comparison Z-scores are averaged to get an overall Superfamily similarity score.

Comparing SmotifCOMP to SCOP Superfamily classifications

A 500 Superfamily test set was randomly selected from SCOP Classes *a*, *b*, *c*, *d*, *e*, *f*, and *g*. An all-by-all SmotifCOMP Superfamily comparison was performed and each intra-Superfamily comparison score was ranked within the distribution of the test set of inter-Superfamily scores. An “identical” classification was defined if the intra-Superfamily score was the top ranked within the distribution. A “similar” classification was defined if the intra-Superfamily score was ranked within the top 1% of the distribution.

Generating the non-hierarchical representation of fold space

An all-by-all SmotifCOMP Superfamily comparison was performed on all SCOP Superfamilies from Classes *a-g* with at least two non-redundant (<40% sequence identity) domains (860 Superfamilies). A network of Superfamily connections was generated to display the predicted evolutionary relationships within the Superfamily universe. The network nodes represent Superfamilies and edge connections are defined for inter-Superfamily Z-scores greater than 3.0. Superfamilies for which the SmotifCOMP intra-Superfamily Z-score was less than 3.0 were removed from the analysis due to being very diverse groups (86 Superfamilies removed). The network representation was generated using Cytoscape⁶⁷ with the “prefuse force directed layout” option used to represent the edge lengths based on the SmotifCOMP Z-score, where shorter edges (closer nodes) represent higher scoring Superfamily comparisons.

Results

The number of unique Smotif geometries is saturated

Smotifs are defined systematically through the spatial orientation of two regular secondary structures that are consecutive in primary sequence and connected by a loop. An exhaustive set of Smotifs, derived from approximately 130,000 PDB chains, was clustered by secondary structure type and RMSD measurements. Since the inception of the PDB, the number of representative Smotif clusters continually increased until approximately the year 2000, after which it remained constant despite the continuous discovery of new protein folds (Fig. 2). The saturation of the number of clusters suggests that the library of existing Smotifs is essentially complete and sufficient to recapitulate any existing or novel fold from a specific subset of Smotifs⁵⁷. The general observation that all the possible fold “building blocks” appear to be already available has been successfully utilized in developing a hybrid protein structure modeling method⁶⁸, fragment based loop modeling^{69,70} and protein design methods⁷¹.

Smotifs are not uniformly distributed within protein folds

An Smotif can be part of a single topology (unique) or can appear repeatedly in evolution in a variety of topologies (ubiquitous). To quantify this feature, we define the “Interfold Count” (IC) as the number of different topologies that an Smotif appears in (Fig. S2). To calculate the IC, one structure was randomly selected from each SCOP Fold and, for each Smotif

cluster, the number of unique folds that contain an Smotif from that cluster was counted. The IC was averaged over 100 random selections. The distribution of the IC of Smotifs follows a power law. Most of the Smotif geometries exist in one or a very small number of folds while a smaller but non-negligible number of Smotifs occur repeatedly in as many as ~80 different folds.

Comparing SmotifCOMP with existing structure comparison methods

One of the main motivations behind the development of SmotifCOMP is to systematically identify novel evolutionary relationships among different topologies. However, it is also expected that SmotifCOMP be able to distinguish structural similarities between domains with global structure similarity. In order to evaluate how well SmotifCOMP identifies similarity within similar fold structures, the performance of SmotifCOMP was compared to other structure comparison methods. For this analysis 500 intra-Superfamily domain pairs (<40% sequence identity) were randomly selected and compared using the SmotifCOMP Z-score, CE Z-score⁵⁰, DALI Z-score⁵¹ or LGA_S score⁵². The resulting similarity scores (Fig. S3) suggest that SmotifCOMP is able to replicate established methods when comparing domains with similar Superfamily classifications, even with low sequence similarity. The SmotifCOMP Z-scores for the Superfamily comparisons are well correlated with CE Z-scores ($\rho=0.71$), DALI Z-scores ($\rho=0.68$) and LGA_S scores ($\rho=0.51$). These correlations are comparable to the correlations among the three reference methods (DALI-CE $\rho=0.90$, DALI-LGA_S $\rho=0.52$, CE-LGA_S $\rho=0.49$).

Comparing SmotifCOMP with existing hierarchical structural classifications

In addition to quantitatively evaluating relationships between different topologies, it is expected that SmotifCOMP can recapitulate known, well-defined, domain relationships. To examine this, the method was compared to the SCOP (version 1.75) Superfamily classification. An all-by-all “Superfamily comparison” was performed, using SmotifCOMP, for 500 Superfamilies randomly selected from SCOP Classes *a-g*. The SmotifCOMP intra-Superfamily Z-scores were ranked within the distribution of inter-Superfamily Z-scores (Fig. 3). SmotifCOMP classifies 75.30% of the Superfamilies identically to the SCOP classification (the intra-Superfamily Z-score ranked highest out of the 500 candidates). Furthermore, 88.26% of the cases ranked in the top 1% of the test set (rank 1–5) and 97.57% of the cases ranked within the top 10% (rank 1–50) of the test set. The average rank for the Superfamilies that are not identically classified compared to SCOP is 18.11 (top 3.6%). Among these cases, a different Superfamily from the same Fold or Class is identified as the most similar in 18.03% and 64.75% of the times, respectively. The ability of the SmotifCOMP method to recapitulate the SCOP Superfamily classifications does not depend on the SCOP Class of the respective Superfamily since the percent of cases identically identified is consistent within each SCOP Class (Table S1). Further, including the Smotif IC in the SmotifCOMP scoring function did not substantially alter the results with respect to the comparison with SCOP (data not shown), which suggests that the Smotif structural similarity was the driving force to predict a relationship. It is likely that the IC does not appear to contribute significantly to the groupings due to the fact that the Smotif similarity and the IC signals are largely overlapping since the preponderance of Smotif geometries have low IC values (Fig S2).

Generally, the most trivial identification of Superfamily relationships in the test set occur for cases in which the Superfamilies are unique within their respective fold and, thus, the background distribution is composed of domains that are of different Superfamily and different Fold classifications. However, when there are multiple distinct Superfamilies within the same Fold classification there are expected to be domains for which a high degree of structural similarity exists (different Superfamilies within the same Fold) and SmotifCOMP must discriminate between these cases based on the subtle evolutionary differences within the respective Superfamilies. For the cases of multiple Superfamilies within the same Fold (187 Superfamilies representing 50 Folds), SmotifCOMP classified 72.19% the Superfamilies identically to SCOP (compared to 75.30% for the complete test set), which suggests that SmotifCOMP can discriminate between Superfamilies within the same Fold classification. The ability to distinguish between Superfamily and Fold groups indicates that SmotifCOMP can discriminate homologous relationships from analogous relationships in a manner similar to the manually curated SCOP assessments. A more robust analysis would require a test on large sets of known homologous and analogous structure groups but no such sets exist to our knowledge, aside from the implicit relationships in the Fold (combination of homologous and analogous relationships) and Superfamily (only homologous relationships) groupings in SCOP.

The cases for which SmotifCOMP identifies an identical or similar relationship to that of SCOP are likely the most straightforward and non-ambiguous domain relationships within the fold universe. Superfamilies for which there is a disagreement between the identified SmotifCOMP relationship and the SCOP classification are the most interesting cases, which likely encompass instances of debatable classifications or situations where the prevalence of structural or evolutionary relationships that span traditional fold definitions (“dense” areas of the fold universe) preclude a straightforward grouping into a single classification, as occurs by default in hierarchical schemes. These cases likely represent structural or evolutionary connections between Superfamilies that would suggest that these groups could be merged into a single group or, at least, that a hierarchical classification separating the groups is not adequate for these cases.

Notably, a substantial number of the cases for which the SmotifCOMP results and the SCOP classification differ are also inconsistent in SCOP and CATH. The discrepancy between the two established classification systems^{47–49} indicates an inherent difficulty in generating an unambiguous assignment for some structural relationships. We analyzed the degree of inconsistent domain definitions and inconsistent classifications within the Superfamilies of the current set in order to explore the level of ambiguity and its effect on the SmotifCOMP identification results. We define an inconsistent domain definition if equivalent SCOP and CATH domains differ by more than 20 residues in the start or end position or in the total coverage of the PDB chain, or if there is no common mapping between the SCOP and CATH domains. We define an inconsistent classification if there are multiple CATH Homologous Superfamily classifications for group of equivalent domains within a single SCOP Superfamily (i.e. CATH splits the group of SCOP Superfamily domains into multiple CATH Homologous Superfamilies). For the SCOP Superfamilies that are properly identified by SmotifCOMP (75.30% of the test set), 39.25% have both consistent domain definitions and consistent classifications in SCOP and CATH. In contrast, for the SCOP Superfamilies

that were not classified similarly by SmotifCOMP, only 20.49% have both consistent domain definitions and consistent classifications in SCOP and CATH. The fact that the SCOP Superfamilies that are identified differently by the SmotifCOMP method also have more inconsistencies between SCOP and CATH, suggests that these are ambiguous, divergent groups that are difficult to classify and/or they represent rather continuous transitions between different topologies.

Non-hierarchical representation of SmotifCOMP-based Superfamily relationships

SmotifCOMP was used to perform an all-by-all Superfamily comparison of all SCOP Superfamilies from Classes *a-g* with at least two non-redundant (<40% sequence identity) domains (a total of 860 Superfamilies). The distinct separation of Z-score distributions between the intra-Superfamily (peak of distribution approximately Z-score=6) and inter-Superfamily (peak of distribution approximately Z-score=1) comparisons indicates that the method properly discriminates between related and unrelated groups (Fig. 4). The SmotifCOMP intra-Superfamily Z-scores are greater than 3.0 for all but the most ambiguously ranked cases in the comparison with the SCOP classification (Fig. 3), which suggests that this threshold is suitable to identify possible evolutionary relationships between Superfamilies.

The SmotifCOMP-based Superfamily comparisons are used to generate a non-hierarchical, network-based representation of SCOP Superfamily relationships (Fig. 5). The nodes of the network represent the SCOP Superfamilies and the edges connect Superfamilies that have a structural relationship identified by SmotifCOMP (SmotifCOMP Z-score ≥ 3.0). The lengths of edges are inversely related to the SmotifCOMP Z-score of the Superfamily comparison (closer nodes represent more highly related Superfamilies). The data used to generate the network, including Superfamily comparison scores and representative domains, can be accessed from http://www.fiserlab.org/our_programs.htm.

The SCOP Superfamilies in the network are generally clustered according to the four major SCOP Classes (*a*, *b*, *c*, *d*). The other SCOP Classes included in the network (*e*, *f* and *g*) are less well represented in the dataset and more dispersed in the network. However, Class *d* ($\alpha + \beta$) is intermediate to Class *b* (all- β) and Class *c* (α/β), while Class *a* (all- α) is well separated from the rest.

Singleton Superfamilies in the SmotifCOMP classification

At the SmotifCOMP Z-score ≥ 3.0 threshold to predict relationships between Superfamilies, there are only 36 Superfamilies that are “singletons”, which have no connection to any other nodes in the network, a surprisingly small fraction of the fold space. There are some singleton Superfamilies that adopt unusual folds and represent possibly interesting evolutionary cases. For example the LDH C-terminal domain-like Superfamily (d.162.1) and the Methionine synthase activation domain-like Superfamily (d.173.1) are each classified as an “unusual fold” in the SCOP description of the topology. Additionally, the Peptide deformylase Superfamily (d.167.1) has a unique configuration of a beta-sheet wrapping around a single helix. Furthermore, the 14kDa protein of cytochrome bc1 complex (f.27.1) and the Non-heme 11kDa protein of cytochrome bc1 complex (f.28.1) Superfamilies each

have an uncommon arrangement of helices in which multiple helices are arranged in what is nearly a single layer. However, the remaining singleton Superfamilies have structural characteristics that preclude SmotifCOMP from finding a significant comparison due to the design of the method and may not necessarily represent interesting biological cases. For example, there are 6 Superfamilies that consist of extremely small structures, such as a single helix-helix motif, which will not produce a significant SmotifCOMP Z-score (a.137.3, a.2.1, a.2.5, a.38.1, f.17.2, g.38.1). On the other hand, there are 8 Superfamilies that are very large and/or have extremely long secondary structures, which makes it very difficult to find a comparison with a significant SmotifCOMP Z-score (a.127.1, a.250.1, b.121.2, d.258.1, d.83.1 e.1.1, e.10.1, e.18.1). In six cases, irregular secondary structures, such as unusually curved strands, likely prevent high Smotif RMSD similarity scores, thus limiting the resulting Smotif alignment and associated SmotifCOMP Z-score (b.106.1, b.110.1, c.91.1, d.230.1, d.61.1, e.5.1). In seven cases, a large structure appears to be composed of two or more distinct domains, any of which may be related to another Superfamily if considered in isolation, but the total size of the structure prevents a significant SmotifCOMP Z-score (interestingly none of these domains are classified as Class *e* “multi-domain proteins” in SCOP: a.264.1, d.133.1, d.142.1, d.142.2, d.143.1, d.265.1, d.283.1). There are 4 Superfamilies that are very diverse, encompassing a group of very divergent domains, which likely prevents the SmotifCOMP “Superfamily comparison” from being significant (b.131.1, c.43.1, d.284.1, d.43.1).

SmotifCOMP identifies related Superfamilies within SCOP Classes

The graph theoretical measure of “indegree” is calculated for each Superfamily in the network and averaged within each SCOP Class (Fig. S4). The “indegree” is a measure of the number of connections of a Superfamily to other Superfamilies in the network and, therefore, is used as a quantitative measure of the number of predicted evolutionary relationships. The most striking characteristic of the network of Superfamily relationships is the number of connections within the Superfamilies of Class *c* (α/β). The average “indegree” in Class *c* is significantly higher than that of Class *a* ($p=2.66\times 10^{-49}$), *b* ($p=7.32\times 10^{-22}$) and *d* ($p=9.19\times 10^{-35}$). Since the numbers of Superfamilies represented in the network are roughly equivalent for the four major Classes (Class *a* = 153 Superfamilies, Class *b* = 154, Class *c* = 173, Class *d* = 240), the higher “indegree” of Class *c*, compared to Class *a*, *b* and *d* is not due to an artifact of the number of available nodes in the network. There are two evolutionary scenarios that could explain the relatively higher number of evolutionary relationships for the Superfamilies of Class *c*. Class *c* may have gone through a recent expansion, leading to a preponderance of relationships among the folds. Alternatively, the higher number of relationships may suggest that the domains of Class *c* arose relatively earlier in evolution and, thus, have had more opportunity to diverge compared to the other major Classes. Previous work that estimated the relative ages of the known folds supports the second scenario, which suggests that Class *c* includes the most ancient domains^{72,73}.

Examples of SmotifCOMP-based Superfamily and Fold relationships

TIM-barrel and flavodoxin-like fold relationships—The SmotifCOMP-derived subnetwork containing the $(\beta\alpha)_8$ -barrel (c.1.-) and flavodoxin-like (c.23.-) Superfamilies (Fig. 6) shows that the Superfamilies are connected both within their respective Folds and

across the Folds and, therefore, SmotifCOMP predicts that these two SCOP Folds are related in evolution.

The SmotifCOMP analysis of the ($\beta\alpha$)₈-barrel fold (25 Superfamilies in the network) shows that all Superfamilies except for two (c.1.16 and c.1.20) have a significant (Z-score = 3) relationship to each other. The near-complete connectivity within the c.1.- Superfamily network supports a previously suggested hypothesis that these Superfamilies are related by divergence from a common origin^{32–34}.

SmotifCOMP analysis also shows nearly complete connectivity within the flavodoxin-like fold (10 Superfamilies in the network) as the method identifies significant relationships (Z-score = 3) between all but three pairs of Superfamilies (c.23.14 and c.23.12, c.23.14 and c.23.15, c.23.15 and c.23.5). The high degree of connectivity within the flavodoxin-like fold suggests that the Superfamilies of the flavodoxin-like fold originated from a common ancestor.

A previous study used profile-based sequence comparisons to analyze the relationships within the ($\beta\alpha$)₈-barrel fold and the flavodoxin-like fold³⁵. The sequence-based analysis found that the ($\beta\alpha$)₈-barrel fold shows nearly complete connectivity among its Superfamilies, which is similar to the SmotifCOMP results. However, in contrast to the SmotifCOMP results, the sequence profile-based analysis showed a more limited relationship among the Superfamilies within the flavodoxin-like fold (only 8 out of 15 analyzed Superfamilies had connections), which was attributed to the flavodoxin-like fold either arising several times independently or evolving from a common ancestor but diverging beyond the point of recognition of sequence analysis methods³⁵. The SmotifCOMP results suggest that the Superfamilies of the flavodoxin-like fold do indeed have a common origin.

The SmotifCOMP method also predicts an evolutionary relationship between the ($\beta\alpha$)₈-barrel fold and the flavodoxin-like fold since the method identifies significant relationships between all flavodoxin-like Superfamilies (10 total) to at least one ($\beta\alpha$)₈-barrel Superfamily (25 total), and vice-versa (Fig. 6). Similarly, the aforementioned sequence profile-based study³⁵ also predicts an evolutionary relationship between the ($\beta\alpha$)₈-barrel and flavodoxin-like folds by showing sequence similarities between the Superfamilies of the respective folds. However, the sequence-based study found that only two flavodoxin-like Superfamilies (of 15 analyzed) were connected to only 12 (of 33) ($\beta\alpha$)₈-barrel Superfamilies, in contrast to the nearly complete connectivity that is identified by SmotifCOMP.

A previous study showed that a conserved cation-binding motif defines a functional relationship between the sporulation initiation phosphotransferase protein (Spo0F), a flavodoxin-like fold that binds a magnesium ion, and five of its structural neighbors (as defined by an application of the Ska⁷⁴ structure alignment program), each with different folds²⁰. These structural neighbors adopt different topologies and different SCOP fold classifications but exhibit similar functions, related to stabilizing a positive charge: 5-aminolaevulinic acid dehydratase (AlaD) from *S. cerevisiae* (SCOP: d1eb3a_, c.1.10.3), iron ABC transporter from *C. jejuni* (SCOP: d1y4ta_, c.94.1.1), UDP-glucosyl transferase (SCOP: d2acwa1/d2acwb1, c.87.1.10), spermidine synthase (PDB: 3b7p, unclassified in

SCOP), acetylcholinesterase (SCOP: d2acea_, c.69.1.1). The functional relationships found in the aforementioned study are all identified by SmotifCOMP (with the exception of the undefined Superfamily of the spermidine synthase protein) and are included in the subnetwork containing the flavodoxin-like Superfamilies.

β -propeller fold relationships—The SmotifCOMP-derived subnetwork containing the β -propeller folds (b.66.-, b.67.-, b.68.-, b.69.-, b.70.-) (Fig. 7) shows that there is nearly complete connectivity between the β -propeller Superfamilies. This suggests that these folds are highly related despite being classified separately by both SCOP and CATH. The SmotifCOMP analysis suggests that the β -propeller folds are related in evolution and supports the hypothesis that these folds arose from a common ancestor by the amplification and divergence of an ancestral β -meander blade fragment³⁰.

SmotifCOMP identifies relationships between β -propeller Superfamilies and both the β -pinwheel and WW domain Superfamilies. As discussed in the Introduction, discriminating homology from analogy in the predicted evolutionary relationships is not always possible. However, the β -pinwheel is classified as a Superfamily of the 6-blade β -propeller fold in SCOP (GyrA/ParC C-terminal domain-like Superfamily). Additionally both the β -pinwheel and WW domains showed sequence similarity to β -propeller folds in a prior study investigating evolutionary relationships between β -propeller families and other all- β (but non- β -propeller) families³¹. Taken together, it is reasonable to infer a homologous evolutionary relationships between β -propellers and both β -pinwheel domains and WW domains, as is predicted by SmotifCOMP.

Also included in the β -propeller Superfamily subnetwork are Superfamilies from the SH3-like barrel (b.34.-), the β -trefoil fold (b.42.-) and the Streptavidin-like fold (b.61.-), which were identified by SmotifCOMP being related to the β -propeller folds. Similar to the β -propeller domains, these three folds are also based on β -meander structures and, thus, may have followed a similar path in evolution as hypothesized for the β -propeller folds.

UDP-binding fold relationships—The SmotifCOMP analysis predicts a statistically significant relationship between the MurCD N-terminal domain (c.5.1) Superfamily and the UDP-glucose/GDP-mannose dehydrogenase C-terminal domain (c.26.3) Superfamily (Fig. 8). The c.5.1 fold (c.5.-) is described (in the SCOP database) as binding a UDP group. Thus, a functional similarity between these two Superfamilies/Folds can be inferred, which further supports the evolutionary relationship that is predicted by the SmotifCOMP analysis. Interestingly, the domains analyzed from these two SCOP Superfamilies are merged into a single Homologous Superfamily, the NAD(P)-binding Rossmann-like Domain (3.40.50.720), in the CATH database (Table S2). A search of the SCOP Fold descriptions for “UDP” identifies 16 separate Superfamilies with an annotation involving UDP, 11 of which are included in the SmotifCOMP-derived subnetwork that includes the c.5.1–c.26.3 Superfamilies (c.2.1, c.26.3, c.4.1, c.5.1, c.59.1, c.68.1, c.72.2, c.87.1, c.98.1, d.159.1, d.68.2). This suggests that this region of the network contains other similar functional and evolutionary relationships, in addition to the c.5.1–c.26.3 relationship that is identified by the SmotifCOMP analysis.

Comparison with existing structure classification analyses

There have been a limited number of studies attempting to deliver updated representations of the relationships within the fold universe. Various strategies have been used to produce a more representative mapping of the fold universe including a classification that is based on an idealized version of the structure⁷⁵ or using structure similarity metrics to generate two-dimensional^{76,77} or three-dimensional^{72,78–80} domain mappings. Importantly, these representations of fold space have lead to interesting insights into evolutionary^{72,77} and functional relationships^{78,80} within the fold universe. In fact, the functional diversity within fold space⁸⁰ was analyzed using FragBag⁸¹, a fragment-based method of structure comparison in which a protein is represented by fragments derived from a library of 400 12mer fragments. The FragBag method has important differences compared to SmotifCOMP, including the fragment definition, the way structures are decomposed into fragments and the method to compare fragment representations of structures.

Another study used analysis of transitivity violations within structure comparisons to determine an appropriate clustering level for confident classification of structures¹¹. It was found that approximately 85% of Superfamilies clustered in single clusters, which indicates that they can be uniquely and unambiguously identified and classified. This fraction is similar to the classification results using the SmotifCOMP method and, in general, the two studies show a good agreement in terms of specific Superfamilies as well. Within a subset of 375 Superfamilies that were part of both studies, 229 Superfamilies can be uniquely defined by both methods (ranked number 1 by SmotifCOMP or uniquely clustered by transitivity criteria). A further 48 Superfamilies within the common set are similarly predicted by each method to be related to other Superfamilies or otherwise ambiguously ranked. Therefore, in total, the two approaches agree for 229+48=277 out of 375 cases (74%) on the level of specific Superfamilies.

Another recent study generated a domain-based network defined by structural alignments with specific sequence identity and RMSD similarity requirements⁷⁶. It was shown that the network can shift from complete connectivity to nearly isolated domains by varying the alignment thresholds (structure similarity, sequence similarity and alignment length) from more inclusive to more stringent. For the thresholds that best reproduce the SCOP Fold classification, the network exhibits a continuous region, composed of mostly α/β domains, and a discrete region, composed of mostly all- α , all- β and $\alpha+\beta$ domains. There are two major differences between the previous study and the current study that could explain the increased connectivity within the Superfamily network that is predicted by SmotifCOMP. First, the SmotifCOMP method is a topology-independent comparison of Smotifs, which allows for a direct comparison of different global topologies in order to identify relationships between disparate folds. This is in contrast to the previous study that uses a structure alignment method that was not developed to quantitatively detect similarities across folds. Furthermore, the SmotifCOMP comparisons do not rely on a sequence similarity input, whereas the previous study used sequence requirements to identify related structures. Eliminating sequence identity requirements allows for identification of relationships where the Superfamilies may be homologous but have diverged to such an extent that a sequence similarity measurement is no longer meaningful. We note that the limitation of the structure-

based method is that it is often difficult to differentiate homology from convergent evolution based on structural similarity alone, without a significant sequence similarity.

The ECOD database⁴⁵ is a hierarchical classification scheme that groups structures by predicted homology without the requirement for global topology (fold) similarity. A systematic comparison of the SmotifCOMP results and the ECOD database is not appropriate because of the different nature of structure comparisons and classifications. Specifically, ECOD is a hierarchical classification that incorporates sequence similarity measures while SmotifCOMP is a non-hierarchical, network classification that is derived from strictly structure-based comparisons. However, a review of the ECOD classification of the three examples of SmotifCOMP-based fold relationships might be useful.

The first example illustrates that SmotifCOMP predicts an evolutionary relationship between the ($\beta\alpha$)₈-barrel (c.1.-) and flavodoxin-like (c.23.-) Superfamilies (Fig. 6). The ECOD domains equivalent to the representative SCOP domains selected for the SmotifCOMP analysis of these two folds (69 domains from 25 SCOP c.1.- Superfamilies, 27 domains from 10 SCOP c.23.- Superfamilies) are split into two distinct branches of the hierarchy based on the delineation within the SCOP classification. In the ECOD classification, the ($\beta\alpha$)₈-barrel domains are grouped within the “ α/β barrel” Architecture, the “TIM β/α barrel” X-group and the “TIM barrel” Topology group. These domains are not separated until the Family level, except for the domains e1vlpA2 and e1yirB2, which are classified in the “Nicotinate/Quinolinate PRTase C-terminal domain-like” Topology group. In the ECOD classification, the flavodoxin-like domains are grouped with the “ α/β three-layered sandwich” Architecture and “Flavodoxin-like” X-group classification. These domains are then classified into nine separate Topology groups. In summary, while ECOD does apparently group the Superfamilies within their respective folds it does not predict a relationship between the ($\beta\alpha$)₈-barrel and flavodoxin-like Folds, which is predicted by SmotifCOMP.

The second example describes that SmotifCOMP predicts an evolutionary relationship between the SCOP β -propeller folds (b.66.-, b.67.-, b.68.-, b.69.-, b.70.-) (Fig. 7). The ECOD classification also groups the β -propeller structures (36 representative domains from 14 SCOP β -propeller Superfamilies) as homologous. Specifically, these structures are all classified within the “beta duplicates or obligate multimers” Architecture, “beta-propeller-like” X-group and “beta-propeller” Homology group. The domains are then divided in different Topology groups based on their respective number of blades. Thus, SmotifCOMP and the ECOD classification similarly group all β -propeller folds, which are classified differently by SCOP, into a predicted evolutionary relationship.

In our third example, SmotifCOMP predicts an evolutionary relationship between 11 SCOP Folds annotated to have a UDB-binding function (c.2.1, c.26.3, c.4.1, c.5.1, c.59.1, c.68.1, c.72.2, c.87.1, c.98.1, d.159.1, d.68.2) (Fig. 8). The ECOD domains equivalent to the representative SCOP domains from these folds (31 domains from 11 SCOP Folds) are classified into two different ECOD Architectures. The SCOP Class *c* Folds are grouped into the “ α/β three-layered sandwich” Architecture while the SCOP Class *d* Folds are grouped into the “ $\alpha+\beta$ four layers” Architecture. Within these Architectures, the SCOP c.2.1, c.4.1 and c.5.1 Superfamily domains are grouped the same X-groups (“Rossmann-like”) and

Homology groups (“Rossmann-related”) and different Topology and Family groups. The SCOP c.59.1, c.68.1 and c.87.1 Superfamily domains are grouped within the same X-group (“Other Rossmann-like structures with the crossover”) but different Topology groups. The remaining domains are classified into different X-H-T-F groups with the same groupings delineated by the SCOP Fold classifications. While SmotifCOMP predicts an evolutionary relationship between all 11 Folds, ECOD predicts relationships between subsets of these Folds.

Discussion

The widely used structure classification schemes were established when there were relatively few experimentally solved protein structures. The sparse coverage of the structure space intuitively implied a discrete nature of the fold universe, which led naturally to the development of classifications with hierarchical organizations. However, hierarchical systems cannot accommodate the increasing number of recent observations where structural^{8–12}, functional^{20,21} and evolutionary relationships^{22,23,26,30,35} are known to span traditional fold definitions. A hierarchical system precludes the existence of relationships between different “fold” classifications and provides no quantitative metric to measure the differences between and within the branches of the hierarchy. Additionally, traditional structure comparison methods^{50–52}, which must constitute the basis of any automated classification method, are optimized for the purpose of fold identification, but are not necessarily suitable for quantitatively evaluating dissimilar global topologies in order to define evolutionary relationships across fold definitions.

The method introduced here, SmotifCOMP, is a topology-independent motif comparison method that provides a robust and quantitative way to compare structures and predict evolutionary relationships within the context of a continuous fold universe. SmotifCOMP has the advantage of providing measures of similarity even when structures contain substantial embellishments or have diverged into different global topologies.

The SmotifCOMP method was used to perform an exhaustive SCOP Superfamily comparison. The identified relationships are represented in a hierarchy-independent network and show that a strong connectivity exists between Superfamilies both within fold definitions and across different folds. We show that the SmotifCOMP method is able to recapitulate examples of known evolutionary and functional relationships between disparate folds, and it provides further insight into hypothesized relationships between Superfamilies and predicts novel evolutionary relationships between disparately classified domains.

In theory, it is not possible to differentiate between convergent and divergent evolutionary relationships (analogy vs homology) based on the SmotifCOMP analysis alone. The difficulty in discriminating between homology and analogy from a structure-based analysis has lead to the suggestion that interpreting the relationships within the structure universe (i.e. classifications) should be based on either structure similarities or evolutionary relationships, without attempting to combine the two components^{13,82}. However, the identification of structural similarities and the prediction of evolutionary relationships often must be inextricably linked due to the difficulty in elucidating sequence homology between

highly diverged protein families and the observation that structure is more conserved than sequence in evolution^{5,58}. In fact, it has been shown that the distribution of the number of domains and diversity of functions within SCOP Superfamily and Fold groups can be appropriately described by a divergent evolutionary model^{83,84}. Furthermore, it has been shown that sequence comparison by profile-HMM reveals homologous connections between domains that span both the SCOP Superfamily and Fold definitions⁸⁵. Therefore it is reasonable to assume that divergent evolution is a predominant force in shaping the structure space and, thus, the known homologous relationships that connect disparate Superfamilies and Folds^{22,23,26,30–32,35} are not likely to be isolated examples.

The primary sequence signal, especially for highly diverged cases, is often not adequate to identify homology between proteins. This necessitates a strategy in which a structure-based method is used to identify structure similarities between proteins – particularly between those that do not share a common global topology. The current study provides a new concept for structure classification that is guided by the notion that the fold universe can be continuous and evolutionary relationships often exist between proteins with different topologies. The SmotifCOMP method provides a novel approach to identify fold relationships in an automated, quantitative manner and can easily accommodate newly solved structures. The network-based representation of the Superfamily relationships is a resource to explore structural evolutionary events that shape the current universe of known topologies, to guide structure based functional assignment studies and to aid method development in protein modeling and design.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grant GM096041 and GM118709. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research*. 2000; 28(1):235–242. [PubMed: 10592235]
2. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*. 2008; 36(suppl 1):D419–D425. [PubMed: 18000004]
3. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Research*. 2013; 41:D490–D498. [PubMed: 23203873]
4. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111(10):3733–3738. [PubMed: 24567391]
5. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*. 1986; 5(4):823–826. [PubMed: 3709526]

6. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology*. 1995; 247:536–540. [PubMed: 7723011]
7. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH – a hierarchic classification of protein domain structures. *Structure*. 1997; 15(5):1093–1108.
8. Cuff A, Redfern OC, Greene L, Sillitoe I, Lewis T, Dibley M, Reid A, Pearl F, Dallman T, Todd A, Garratt R, Thornton J, Orengo C. The CATH Hierarchy Revisited – Structural Divergence in Domain Superfamilies and the Continuity of Fold Space. *Structure*. 2009; 17:1051–1062. [PubMed: 19679085]
9. Friedberg I, Godzik A. Connecting the Protein Structure Universe by Sparse Recurring Fragments. *Structure*. 2005; 13:1213–1224. [PubMed: 16084393]
10. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. Quantifying the Similarities within Fold Space. *Journal of Molecular Biology*. 2002; 323:909–926. [PubMed: 12417203]
11. Pascual-Garcia A, Abia D, Ortiz AR, Bastolla U. Cross-Over between Discrete and Continuous Protein Structure Space: Insights into Automatic Classification and Networks of Protein Structures. *PLoS Computational Biology*. 2009; 5(3)
12. Shindyalov IN, Bourne PE. An Alternative View of Protein Fold Space. *Proteins*. 2000; 38:247–260. [PubMed: 10713986]
13. Sadowski MI, Taylor WR. On the evolutionary origins of “Fold Space Continuity”: A study of topological convergence and divergence in mixed alpha-beta domains. *Journal of Structural Biology*. 2010; 172:244–252. [PubMed: 20691788]
14. Edwards H, Deane CM. Structural Bridges through Fold Space. *PLoS Comput Biol*. 2015; 11(9):e1004466. [PubMed: 26372166]
15. Andreeva A, Murzin AG. Evolution of protein fold in the presence of functional constraints. *Current Opinion in Structural Biology*. 2006; 16:399–408. [PubMed: 16650981]
16. Grishin NV. Fold Change in Evolution of Protein Structures. *Journal of Structural Biology*. 2001; 134:167–185. [PubMed: 11551177]
17. Kinch LN, Grishin NV. Evolution of protein structures and functions. *Current Opinion in Structural Biology*. 2002; 12:400–408. [PubMed: 12127461]
18. Lupas AN, Ponting CP, Russell RB. 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. *Journal of Structural Biology*. 2001; 134:191–203. [PubMed: 11551179]
19. Murzin AG. How far divergent evolution goes in proteins. *Current Opinion in Structural Biology*. 1998; 8:380–387. [PubMed: 9666335]
20. Petrey D, Fischer M, Honig B. Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proceedings of the National Academy of Sciences of the United States of America*. 2009; 106(41):17377–17382. [PubMed: 19805138]
21. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(14):5441–5446. [PubMed: 18385384]
22. Grishin NV. KH domain: one motif, two folds. *Nucleic Acids Research*. 2001; 29(3):638–643. [PubMed: 11160884]
23. Coles M, Hulko M, Djuranovic S, Truffault V, Koretke K, Martin J, Lupas AN. Common Evolutionary Origin of Swapped-Hairpin and Double-Psi β Barrels. *Structure*. 2006; 14:1489–1498. [PubMed: 17027498]
24. Coles M, Diercks T, Liermann J, Groger A, Rockel B, Baumeister W, Koretke KK, Lupas A, Peters J, Kessler H. The solution structure of VAT-N reveals a ‘missing link’ in the evolution of complex enzymes from a simple $\beta\alpha\beta\beta$ element. *Current Biology*. 1999; 9:1158–1168. [PubMed: 10531028]
25. Coles M, Djuranovic S, Koretke K, Truffault V, Martin J, Lupas AN. ArbB-like Transcription Factors Assume a Swapped Hairpin Fold that Is Evolutionarily Related to Double-Psi β Barrels. *Structure*. 2005; 13:919–928. [PubMed: 15939023]
26. Roessler CG, Hall BM, Anderson WJ, Ingram WM, Roberts SA, Montfort WR, Cordes MHJ. Transitive homology-guided structural studies lead to discovery of Cro proteins with 40%

- sequence identity but different folds. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(7):2343–2348. [PubMed: 18227506]
27. Dorn LOV, Newlove T, Chang S, Ingram WM, Cordes MHJ. Relationship between Sequence Determinants of Stability for Two Natural Homologous Proteins with Different Folds. Biochemistry. 2006; 45:10542–10553. [PubMed: 16939206]
28. Newlove T, Konieczka JH, Cordes MHJ. Secondary Structure Switching in Cro Protein Evolution. Structure. 2004; 12:569–581. [PubMed: 15062080]
29. Remmert M, Biegert A, Linke D, Lupas AN, Soding J. Evolution of Outer Membrane β -Barrels from an Ancestral $\beta\beta$ Hairpin. Molecular Biology and Evolution. 2010; 27(6):1348–1358. [PubMed: 20106904]
30. Chaudhuri I, Soding J, Lupas AN. Evolution of the β -propeller fold. Proteins. 2008; 71:795–803. [PubMed: 17979191]
31. Kopec KO, Lupas AN. β -Propeller Blades as Ancestral Peptides in Protein Evolution. PLoS One. 2013; 8(10)
32. Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M. Structural Evidence for Evolution of the β/α Barrel Scaffold by Gene Duplication and Fusion. Science. 2000; 289:1546–1550. [PubMed: 10968789]
33. Gerlt JA, Babbitt PC. Barrels in pieces? Nature Structural Biology. 2001; 8(1):5–7. [PubMed: 11135656]
34. Hocker B, Beismann-Driemeyer S, Hettwer S, Lustig A, Sterner R. Dissection of a $(\beta\alpha)_8$ -barrel enzyme into two folded halves. Nature Structural Biology. 2001; 8(1):32–36. [PubMed: 11135667]
35. Farias-Rico JA, Schmidt S, Hocker B. Evolutionary relationship of two ancient protein superfolds. Nature Chemical Biology. 2014; 19:710–715.
36. Hocker B, Schmidt S, Sterner R. A common evolutionary origin of two elementary enzyme folds. FEBS Letters. 2002; 510:133–135. [PubMed: 11801240]
37. Rost B. Twilight zone of protein sequence alignments. Protein Engineering. 1999; 12(2):85–94. [PubMed: 10195279]
38. Fiser A. Protein structure modeling in the proteomics era. Expert Review of Proteomics. 2004; 1(1):97–110. [PubMed: 15966803]
39. Fiser, A. Comparative protein structure modeling. In: Ridgen, DJ., editor. From protein structure to function with bioinformatics. Springer; 2008. p. 57–81.
40. Rykunov D, Fiser A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins. 2007; 67(3):559–568. [PubMed: 17335003]
41. Summa CM, Rosenblatt MM, Hong J-K, Lear JD, DeGrado WF. Computational de novo Design, and Characterization of an A_2B_2 Diiron Protein. Journal of Molecular Biology. 2002; 321(5):923–938. [PubMed: 12206771]
42. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D. Principles for designing ideal protein structures. Nature. 2012; 491(7423):222–227. [PubMed: 23135467]
43. Zhan C, Fedorov EV, Shi W, Ramagopal UA, Thirumuruhan R, Manjasetty BA, Almo SC, Fiser A, Chance MR, Fedorov AA. The ybeY protein from *Escherichia coli* is a metalloprotein. Acta Crystallographica Section F Structural Biology and Crystallization Communications. 2005; 61(Pt 11):959–963.
44. Brenner SE, Chothia C, Hubbard TJP, Murzin AG. Understanding protein structure: Using SCOP for fold interpretation. Methods in Enzymology. 1996; 266:635–643. [PubMed: 8743710]
45. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim B-H, Grishin NV. ECOD: An Evolutionary Classification of Protein Domains. PLoS Computational Biology. 2014; 10(12)
46. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Research. 2014; 42:D310–D314. [PubMed: 24293656]
47. Csaba G, Birzele F, Zimmer R. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Structural Biology. 2009; 9(23)

48. Day R, Beck DAC, Armen RS, Daggett V. A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Science*. 2003; 12:2150–2160. [PubMed: 14500873]
49. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*. 1999; 7:1099–1112. [PubMed: 10508779]
50. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*. 1998; 11(9):739–747. [PubMed: 9796821]
51. Holm L, Sander C. Protein Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology*. 1993; 233:123–138. [PubMed: 8377180]
52. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*. 2003; 31(13):3370–3374. [PubMed: 12824330]
53. Taylor WR, Orengo CA. Protein Structure Alignment. *Journal of Molecular Biology*. 1989; 208:1–22. [PubMed: 2769748]
54. Kolodny R, Petrey D, Honig B. Protein structure comparison: implications for the nature of ‘fold space’, and structure and function prediction. *Current Opinion in Structural Biology*. 2006; 16(3): 393–398. [PubMed: 16678402]
55. Taylor WR. Evolutionary transitions in protein fold space. *Current Opinion in Structural Biology*. 2007; 17:354–361. [PubMed: 17580115]
56. Valas RE, Yang S, Bourne PE. Nothing about protein structure classification makes sense except in the light of evolution. *Current Opinion in Structural Biology*. 2009; 19:329–334. [PubMed: 19394812]
57. Fernandez-Fuentes N, Dybas JM, Fiser A. Structural Characteristics of Novel Protein Folds. *PLoS Computational Biology*. 2010; 6(4)
58. Rost B. Protein structures sustain evolutionary drift. *Folding and Design*. 1997; 2(Supplement 1):S19–S24. [PubMed: 9218962]
59. Wu G, Fiser A, Kuile Bt, Sali A, Muller M. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proceedings of the National Academy of Sciences of the United States of America*. 1999; 96(11):6285–6290. [PubMed: 10339579]
60. Krishna SS, Grishin NV. Structural drift: a possible path to protein fold change. *Bioinformatics*. 2005; 21(8):1308–1310. [PubMed: 15604105]
61. Carter P, Andersen CAF, Rost B. DSSPcont: continuous secondary structure assignments for proteins. *Nucleic Acids Research*. 2003; 31(13):3293–3295. [PubMed: 12824310]
62. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983; 22:2577–2637. [PubMed: 6667333]
63. Theobald DL. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallographica*. 2005; A61(4):478–480.
64. Felsenstein, J. Distributed by the author Department of Genome Sciences. University of Washington; Seattle: 2005. PHYLIP (Phylogeny Inference Package) version 3.6.
65. Smith TF, Waterman MS. Identification of common molecular subsequences. *Journal of Molecular Biology*. 1981; 147(1):195–197. [PubMed: 7265238]
66. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006; 22(13):1658–1659. [PubMed: 16731699]
67. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003; 13(11):2498–2504. [PubMed: 14597658]
68. Menon V, Vallat BK, Dybas JM, Fiser A. Modeling Proteins Using a Super-Secondary Structure Library and NMR Chemical Shift Information. *Structure*. 2013; 21(6):891–899. [PubMed: 23685209]
69. Fernandez-Fuentes N, Oliva B, Fiser A. A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Research*. 2006; 34(7):2085–2097. [PubMed: 16617149]
70. Fernandez-Fuentes N, Zhai J, Fiser A. ArchPRED: a template based loop structure prediction server. *Nucleic Acids Research*. 2006; 34(suppl 2):W173–W176. [PubMed: 16844985]

71. Bonet J, Segura J, Planas-Iglesias J, Oliva B, Fernandez-Fuentes N. Frag'r'Us: knowledge-based sampling of protein backbone conformations for *de novo* structure-based protein design. *Bioinformatics*. 2014; 30(13):1935–1936. [PubMed: 24603983]
72. Choi I-G, Kim S-H. Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103(38):14056–14061. [PubMed: 16959887]
73. Winstanley HF, Abeln S, Deane CM. How old is your fold? *Bioinformatics*. 2005; 21(Suppl. 1):i449–i458. [PubMed: 15961490]
74. Petrey D, Honig B. GRASP2: Visualization, Surface Properties, and Electrostatics of Macromolecular Structures and Sequences. *Methods in Enzymology*. 2003; 374:492–509. [PubMed: 14696386]
75. Taylor WR. A 'periodic table' for protein structures. *Nature*. 2002; 416:657–667. [PubMed: 11948354]
76. Nepomnyachiy S, Ben-Tal N, Kolodny R. Global view of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111(32):11691–11696. [PubMed: 25071170]
77. Holm L, Sander C. Mapping the Protein Universe. *Science*. 1996; 273:595–602. [PubMed: 8662544]
78. Hou J, Jun S-R, Zhang C, Kim S-H. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(10):3651–3656. [PubMed: 15705717]
79. Hou J, Sims GE, Zhang C, Kim S-H. A global representation of the protein fold space. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100(5):2386–2390. [PubMed: 12606708]
80. Osadchy M, Kolodny R. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108(30):12301–12306. [PubMed: 21737750]
81. Budowski-Tal I, Nov Y, Kolodny R. FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107(8):3481–3486. [PubMed: 20133727]
82. Sadreyev RI, Kim B-H, Grishin NV. Discrete-continuous duality of protein structure space. *Current Opinion in Structural Biology*. 2009; 19:321–328. [PubMed: 19482467]
83. Koonin EV, Wolf YI, Kerev GP. The structure of the protein universe and genome evolution. *Nature*. 2002; 420:218–223. [PubMed: 12432406]
84. Goldstein RA. The structure of protein evolution and the evolution of protein structure. *Current Opinion in Structural Biology*. 2008; 18:170–177. [PubMed: 18328690]
85. Alva V, Remmert M, Biegert A, Lupas AN, Soding J. A galaxy of folds. *Protein Science*. 2010; 19:124–130. [PubMed: 19937658]

Fig. 1. SmotifCOMP Method Flowchart
Flowchart describing the steps in the SmotifCOMP method.

Fig. 2. Smotif Geometry Saturation

The cumulative number of Smotif clusters (left y-axis) in the PDB since its inception (red: HH clusters, blue: HE clusters, cyan: EH clusters, purple: EE clusters). The number of SCOP Fold classifications (right y-axis) for the year of the database release (grey diamonds).

Fig. 3. SmotifCOMP Superfamily Identification

The intra-Superfamily average Z-score for each Superfamily in the 500 Superfamily test set (blue diamonds, left y-axis) along with the corresponding rank in the test set distribution (red diamonds, right y-axis). A rank of 1 indicates that the Superfamily was similarly identified by SmotifCOMP and the SCOP classification.

Fig. 4. SmotifCOMP Superfamily Comparison Z-score Distribution

Distribution of Z-scores produced from an all-by-all comparison of 860 Superfamily groups. The normalized distribution is shown for intra-Superfamily individual domain comparison Z-score (grey solid line), intra-Superfamily average Z-score (blue solid line), inter-Superfamily individual domain comparison Z-score (dashed grey line) and inter-Superfamily average Z-score (dashed blue line).

Fig. 5. Non-hierarchical Representation of Fold Space

Network of SmotifCOMP-derived Superfamily relationships. The nodes represent Superfamily groups colored by Class definition (red: Class *a*, yellow: Class *b*, blue: Class *c*, lavender: Class *d*, cyan: Class *e*, grey: Class *f*, purple: Class *g*). The edges represent relationships between the Superfamilies as defined by SmotifCOMP.

Fig. 6. Subnetwork of $(\beta\alpha)_8$ -barrel and Flavodoxin-like Superfamilies
 SmotifCOMP-derived Superfamily relationships for the $(\beta\alpha)_8$ -barrel Superfamilies (red border) and flavodoxin-like Superfamilies (blue border).

Fig. 7. Subnetwork of β -propeller Superfamilies

SmotifCOMP-derived Superfamily relationships for the 4-blade (b.66.-) (red border), 5-blade (b.67.-) (lavender border), 6-blade (b.68.-) (purple border), 7-blade (b.69.-) (blue border) and 8-blade (b.70.-) (cyan border) β -propeller Folds.

Fig. 8. Subnetwork of UDP-related Superfamilies

Subnetwork of SmotifCOMP-derived Superfamily relationships for the UDP function-related Superfamilies. The MurCD N-terminal domain Superfamily (red border) and the UDP-glucose/GDP-mannose dehydrogenase C-terminal domain (blue border) are shown along with other UDP function-related Superfamilies that are identified by SmotifCOMP (yellow border).