Characteristics of Protein Fold Space Exhibits Close Dependence on Domain Usage

Michael T. Zimmermann[1], Fadi Towfic[2], Robert L. Jernigan[3], and Andrzej Kloczkowski[4,5(✉)]

[1] Clinical and Translational Science Institute, Medical College of Wisconsin, Milwaukee, WI 53223, USA
[2] Celgene Corporation, Summit, NJ 07901, USA
[3] Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, USA
[4] Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, Columbus, OH 43205, USA
Andrzej.Kloczkowski@nationwidechildrens.org
[5] Department of Pediatrics, The Ohio State University, Columbus, OH 43205, USA

Abstract. With the growth of the PDB and simultaneous slowing of the discovery of new protein folds, we may be able to answer the question of how discrete protein fold space is. Studies by Skolnick et al. (PNAS, 106, 15690, 2009) have concluded that it is in fact continuous. In the present work we extend our initial observation (PNAS, 106(51) E137, 2009) that this conclusion depends upon the resolution with which structures are considered, making the determination of what resolution is most useful of importance. We utilize graph theoretical approaches to investigate the connectedness of the protein structure universe, showing that the modularity of protein domain architecture is of fundamental importance for future improvements in structure matching, impacting our understanding of protein domain evolution and modification. We show that state-of-the-art structure superimposition algorithms are unable to distinguish between conformational and topological variation. This work is not only important for our understanding of the discreteness of protein fold space, but informs the more critical question of what precisely should be spatially aligned in structure superimposition. The metric-dependence is also investigated leading to the conclusion that fold usage in homology reduced datasets is very similar to usage across all of PDB and should not be ignored in large scale studies of protein structure similarity.

Keywords: Protein structure · Protein structural alignment · Protein fold space · Graph theory · Protein structure universe · Protein families · Protein folds

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11465, pp. 356–369, 2019.
https://doi.org/10.1007/978-3-030-17938-0_32

1 Introduction

The three dimensional structures of proteins are often grouped into hierarchical classifications in order to facilitate our understanding of their relationships with each other. Using this concept, one can envision a "fold space" for protein structures where a fold is defined as a specific spacial arrangement of secondary structures. These folds of single domain proteins have been classified by a number of structural ontologies including CATH [1] and SCOP [2] that group most known protein structures based upon combinations of sequence homology, structural topology, and function. Pfam [3] is another well used resource, but focuses more on functional classification, rather than structural (though the two are often related). Presently, structural classifications such as CATH and SCOP still rely heavily upon expert manual curation. The prevailing view concerning protein fold space is that it is comprised of a finite number of discrete folds that are described by these structural ontologies. Recent updates have yielded increased coverage of the diverse types of folds that proteins can assume [4], with a noticeable saturation being reached. The results of such efforts, largely driven by structural genomics initiatives [6], may imply that we are reaching full enumeration of the single domain folds [5]. As such, one of the interesting and important implications that arise from these works is that fold space is discrete and not continuous.

Work by Skolnick *et al.* [7] challenges this view concerning the discreteness of protein structure space using a graph theory approach to analyzing the topological relatedness of protein structures. By considering a large representative set of structures, and a graph based on pair-wise structural relatedness judged by TM-score [8] of 5906 protein chains with low homology from the PDB [9], it was shown that the average shortest path in this network is seven. In the graph, a node represents each PDB file and edges are placed between them whenever pair-wise TM-score is greater than 0.4. We believe that this may not necessarily be informative about protein structure space [10], but instead is likely a general network property since the same result can be obtained in a simpler way using the approximation of Watts and Strogatz [11] for random small world graphs. Multiple questions still arise such as the metric-dependence of this conclusion, if state-of-the-art structure matching algorithms can distinguish topological diversity from conformational, and the overall role of domain architecture. In this work we seek a more detailed understanding of the properties of fold space graphs and their implications for our perceptive on protein structure relatedness. Our main contributions are as follows (1) Graphs generated based on various TM score cutoffs show a high degree of modularity, however (2) we show that the TM algorithm is not well suited for distinguishing topology from conformation based on our comparative analysis (utilizing TM align) of reverse transcriptase (RT) structures gathered from Pfam to manually curated categories in CATH. Thus (3) we explored structure space using a domain-based comparison utilizing CATH and SCOP categories. Our comparison showed that there exists one dominant, modular cluster with some discontinuities in structure space outside of the larger cluster. Thus, we conclude that the continuity (or discreteness) of protein structure fold space depends highly on the resolution one is willing to impose for distinguishing folds.

Modularity, graph partitioning efficiency, and community detection are three terms that refer to roughly the same concept; determining if there exist regions of a network with high connectivity of nodes within individual clusters, but relatively low connectivity between different clusters. For our application to fold space, this translates into groups of structures that are close structural matches to each other within a group, but not to members of other groups. Therefore, for community structure to be prevalent there must be groups of structures that are closely related, but few structures that are simultaneously similar to members of a different group. High modularity combined with a relatively large number of clusters would point to a discrete fold space. Low modularity or a high modularity with very few clusters would point toward a continuous view. Various metrics to evaluate the community structure in graphs have been developed including the modularity score of Newman and Girvan [12] that we apply here. The logic behind community structure and graph clustering to explain the small average shortest path is the following: Consider a cluster *A* that is well connected. That is, for every node n_i in *A*, any other node n_j in *A* is reachable, on average, via a greater number of shorter paths compared to another node, n_x that is a member of a different cluster B. This means that any neighbor of any node in *A* is quickly reachable from any node in the cluster. Strong community architecture does exist in fold space graphs and further analysis is performed by employing the Markov Cluster (MCL) Algorithm [13, 14]. If a large number of well-connected clusters exist in the graph and relatively few edges connect them, then either the dataset is not a complete representation of fold space or the space is not continuous.

Many methods to determine the relatedness of proteins and protein structures have developed. These are dominated by sequence algorithms because sequence data is abundant and sequence-based algorithms are computationally efficient and fairly intuitive. One such scheme is VAST, Vector Alignment Search Tool [15], which incorporates statistical significance thresholds and estimation of interactions chosen by chance. The widespread use of PSI-BLAST [16] and similar string algorithms in structure classifications like CATH and SCOP are further examples. Matches based on sequence homology represent a conservative subset of similar proteins due to the fact that the inverse folding problem, determining how many sequences can assume a given 3D shape (fold), is unsolved in general. Many cases exist where sequences with little to no homology assume nearly identical folds; i.e. Ubiquitin (1UBI) and SUMO (1WM2) have 15% sequence identity, but fold to practically similar structures differing only by 1.5 Å C^{α} RMSD. Many structure alignment procedures exist that are widely used in structural biology. In this work, we will primarily use TM-align [8], which has been shown to give excellent alignments and used for template detection in I-TASSER [17], currently ranked among the best performing 3D structure prediction servers.

Another fundamental question that needs to be addressed is exactly what should be compared? Proteins with different numbers of amino acids are, mathematically, objects with different conformational dimensions; therefore, we commonly simplify the problem to finding the best superimposition of two structures. Interestingly, there may be patterns in other mathematical spaces that simplify the analysis of structures, such as the relation between spectral dimension (related to energy transfer efficiency) and fractal dimension (related to packing density) in protein structures [18, 19]. However, the details of the structure can influence the energy transfer (allostery) pathways within

the structure [20] or their interactions [21]. A much coarser view could consider proteins as approximate globules – amorphous 3D blobs whose surfaces are semi-molten [22] and have mostly polar character but with some internal nonpolar groups exposed to water. We have just described two very different views on protein structures where the details can highly bias an analysis toward a specific conclusion. In the first case, it is intuitive that relatively few structures will be similar whereas in the second case, many proteins could resemble each other's shape. Thus, the resolution with which we consider structures will affect our conclusion about the discreteness of fold space. In this work, we investigate the structure superimposition using TM-align and domain similarity. Our application of TM-align procedure is similar to Skolnick *et al.* [7], except that we consider various thresholds instead of a single, fixed TM score cutoff. We also collect a representative from each known fold type and apply the same graph analysis to this smaller dataset. These representatives have been deemed by expert manual curation to symbolize distinct fold types and thus represent a best case scenario for concluding that fold space is discrete. The results of this analysis are utilized to interpret data obtained by using the complete protein dataset. For domain similarity, we analyze fold space independent of any structure-based comparison by connecting nodes if the proteins they represent share a common CATH or SCOP annotation. Annotations were taken from CATH at the Topology level and from SCOP at the Fold level. Such an analysis provides an impartial baseline for how any structure similarity metric that seeks to approximate CATH or SCOP-level fold similarity will perform on the dataset.

Defining the entities that are compared: complete PDB files, PDB chains (individual polypeptides), or single domains is an important problem. Much effort has been applied to developing methods for computational domain prediction. Early contributions such as FSSP using Dali [23, 24] have been very influential, while newer algorithms such as DomNet [25] show increased refinement and agreement with manual curation. However, in this study we will focus on the manual curation levels of CATH and SCOP. If whole PDB files or chains are used, there will be cases where the peptide chain folds to two or more domains. These structures can act as cluster-linkers in the fold space graph since one domain may have a significant score with structures in one cluster, while the other domain will have high structural relation to a different cluster. Alternatively, a single domain could require the interaction of more than one polypeptide. Such proteins complicate the relationship between sequence-homology reduced datasets and fold usage. Considering the size of a protein may also be important since a small protein is more likely to possess a topology that is some subset of a larger protein.

2 Results

For any approach that relies on graph theory, understanding the structure of the graphs used is necessary. Figure 1 shows us that, for any TM-score threshold, there exist a relatively small number of nodes possessing a high degree of connectivity. We have investigated these hub nodes and draw two conclusions. Some of them are hubs because they are among the smallest proteins in the dataset with approximately 50 residues. It is more likely for a small protein to be a topologically similar subset of a

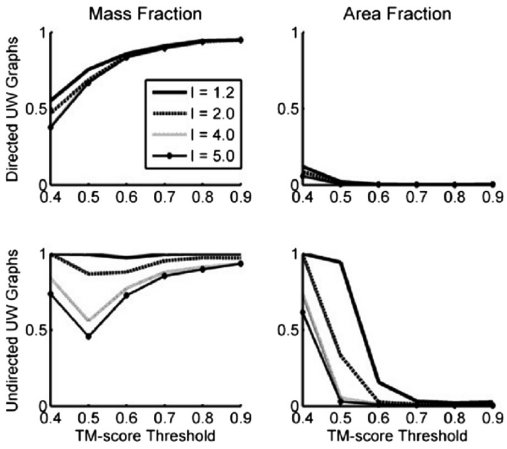

Fig. 1. Metrics evaluating MCL clustering on TM-score graphs. UW stands for uniformly weighted meaning that if an edge exists in the graph, we assign it a weight of one. Area Fraction is defined by Eq. 2 and relates to cluster size. Mass Fraction is the fraction of total edge weight that is captured within clusters and is formally defined in Eq. 3. Including the edge weights does not impact these metrics (see Tables 1 and 2).

larger protein than for two proteins of equal size to match. Others have high connectivity because they have multiple domains. Each domain can individually have a significant alignment with other structures, which inflates the connectivity relative to single domain chains.

In our previous work [10], the relationship between average shortest path computed using the WattsStrogatz approximation [11] and the TM-score threshold for retaining edges in the graph was investigated. We found that the average shortest path is less than seven for cutoffs below 0.75. Stricter cutoffs result in large areas of the graph becoming disconnected. With increasing TM-score, the edge set gets sparser, approaching a cardinality of zero. This is shown in Fig. 2 where the number of nodes with no edges increases as the TM-score threshold increases.

Since TM-scores are numerical, defined on the interval from zero to one, and are not symmetric, pairwise scores can easily be interpreted as a directed graph where we use TM-scores as edge weights. In the MCL algorithm edge weight is the probability of a random walk traversing along a given edge. We construct unweighted graphs by assigning all edges a weight of one and undirected graphs by linking nodes (with or without edge weights) based on the larger of their two TM-scores. From Supplemental Tables 1 and 2, it is evident that the edge treatment makes a minimal impact upon MCL clustering.

Since the sequence-structure relationship is not fully understood, sequence-homology reduced datasets are not necessarily the same as topology reduced datasets. The effect on graph behavior of a topologically reduced dataset is of interest for comparison to the homology reduced dataset. Parameter choices that yield expected results in the topologically reduced dataset will help us to better interpret the meaning of clusters for the homology reduced set. For this reason, we also compare distinct

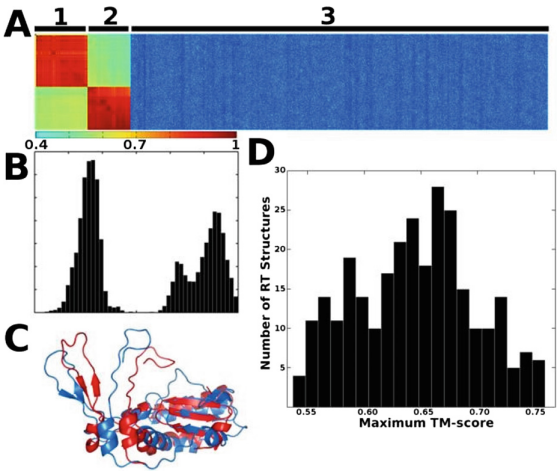

Fig. 2. (A) Heat map of TM-scores between 283 reverse transcriptase (RT) structures and 1233 topology representatives from CATH v3.3. Rows are arranged in the same order as the columns of Sects. 1 and 2. Sections 1 and 2 correspond to two orientations of the RT "fingers," while Sect. 3 is the TM score between the topology representatives and the RT structures. (B) Histogram of the TM-scores within the set of 283 reverse transcriptase structures. The set can easily be split into structures that are related to each other at a TM-score of greater or less than 0.7. No pairwise scores are below 0.4. Low scores correspond to the two finger domains being in different positions, while the higher scores correspond to the two fingers in roughly the same orientation. The high scoring population can be somewhat thought of as two groups; one where the fingers are in a more closed conformation, and one where they are both extended. (C) We show a representative of the lower TM-score population; 1RW3 aligned to 1JLA with a TM-score of 0.56. The view shown highlights the different finger positions that are characteristic of the lower scoring group. (D) Histogram of maximum TM-score between each reverse transcriptase domain and the topology representatives from CATH (max for each row of Sect. 3). Each reverse transcriptase domain has a TM-score between itself and a topology representative of at least 0.53, but none are higher than 0.76. There are 277 topologies matched to the 283 reverse transcriptase structures. Thus, large TM-scores, while relatively sparse, are not because of any single (or even a small set) of reverse transcriptase like topologic representatives.

topologies to each other by gathering 1233 CATH version 3.3 topology representatives; a collection of manually curated topology representatives that span all of PDB, performing the same procedure. Interestingly, this dataset of distinct topology representatives exhibits a high modularity, indicative of community structure (Supplemental Table 2). We calculate a modularity score defined in [12] by comparing the number of edges within clusters to the number of edges that are linking clusters to each other. At low TM thresholds (0.4), the graph exhibits high connectivity (57550 edges) and the majority of the nodes included in the largest cluster.

The extent of community structure is less than for the PDB300 dataset (as judged by F_{Mass} and F_{Area} – see Methods), but remains high. The MCL inflation parameter determines granularity of the clustering with a low inflation yielding few large clusters and high inflation producing many small clusters. Even for the high inflation value of 5,

the largest cluster still contains 855 structures, whereas a low value of 1.2 retains 1217. At a TM score cutoff of 0.6 we find that MCL consistently distinguishes many of the topologies from each other (only 334 edges between the 1233 nodes). Thus, these graphs may either be modular because they are significantly related (pointing to structure space being continuous) or because they are mutually distantly related (pointing to a discrete fold space).

Conformational variability is also an important consideration for comparing topologies. Are our structure comparison metrics able to distinguish between conformational variation and topological? To address this question, we will compare 283 reverse transcriptase (RT) structures gathered from Pfam [3] family PF00078 to each other and to the CATH topology representatives to investigate the ability for structure comparison metrics to distinguish between conformational (within the RT family) and topological (between RT and fold representatives) differences. RT structure is often described by analogy to a human hand where the active site is in the center of the palm and the fingers and thumb "grip" the substrate. The Pfam family set used corresponds to two fingers and the palm, thus containing sequence (average sequence identity of 67%) and conformational variants. A TM-score above 0.4 is regarded as a significant topological relationship. We find that all members of the reverse transcriptase family have TM-scores above this threshold, but there is significant diversity of scores within the family (Fig. 2). Roughly half of the pairwise comparisons are between 0.4 and 0.7 corresponding to different finger conformations (generalized from visualizing 100 randomly chosen pairs from this group). Higher scoring pairs are characterized by the structures having the same general finger conformation. The subgroup at about 0.82 has a higher representation of a more extended finger conformation (again from visualization of 100 randomly chosen pairs; data not shown). A representative pair is shown in Fig. 2. Further, each reverse transcriptase domain has a TM-score between itself and a topology representative of at least 0.53, but none are higher than 0.76. Therefore, all RT structures have a significant structure alignment to a topology representative. One might expect that because all RT structures share a common fold, one topology would be the best match to most of the RT structures. However, matching each of the 283 RTs to its highest scoring topology yields 277 different topologies. Thus, large TM-scores, while relatively sparse, are not because of any single (or even a small set) of RT- like topologic representatives. Further, TM (and likely any rigid superimposition algorithm) is, in general, unable to distinguish between conformational and topological variation. Methods like Fr-TM-align [26] or FATCAT [27] that are capable of accounting for flexibility of the biomolecule may perform better in this specific test, but fast and accurate methods for incorporating flexibility in structure matching are still being improved. Current structure comparison algorithms have difficulty in distinguishing between conformational and topological differences.

Metrics similar to Silhouettes [28] have also been generated (not shown). These are basically average path length from a node to any other node within a given cluster compared to the average path length from a node to every node that is not in that cluster. Evidence of the high number of connections within each cluster exists in that the average out-of-cluster path is only slightly longer than the average within-cluster path.

A critical point of the above analysis hinges on the efficacy of the TM score algorithm in quantifying the fold space of proteins. Thus, it is reasonable to ask: are these investigations into protein fold space dependent upon the metric used? We have already shown that the state-of-the-art structure comparison method has difficulty in distinguishing topological and conformational differences, but can we explore fold space independent from structure superimposition? One way is to make a graph where each protein chain is represented by a node and nodes are connected by edges if the two proteins share a common fold. Common folds are determined by a shared CATH topology or SCOP fold using CATH version 3.4 and SCOP version 1.75. In the PDB300 dataset 90% of the protein chains are annotated by at least one of these ontologies, while all of the PISCES proteins are annotated (see Methods for dataset details). Unannotated nodes are neglected in the following analysis. Using the same graph analysis procedure, we find that this domain based graph also has a very high degree of connectivity and modularity. See Table 1 for details. We again find that there exists one dominant cluster. It has been shown that MCL usually generates a dominant cluster and for some applications modifications that generate a more even granularity are preferred [29]. However, using these approaches would be equivalent to assuming fold space is discrete. Another explanation for the dominant cluster is the imbalance in topology usage. Table 2 summarizes the usage of the ten most used topologies across all of CATH, PDB300, and the PISCES dataset. Seven of the ten most used annotations across all of CATH are also in the top ten most used topologies in the datasets used here. Further, if we sort the topology classes by their use across all of CATH, and compare with the topology use in each of our datasets, PDB300 and PISCES have a correlation coefficient with CATH of 0.94 and 0.93, respectively. Thus, the relative distribution of domain types is similar in these datasets compared to the whole PDB. We conclude that the reason for the observed shortest paths in TM-score based graphs is the modularity of proteins and the bias in topology usage. Protein structures exhibit variation upon themes – stable domains develop and are embellished upon for further modification of function.

Viksna and Gilbert [30] proposed a new method of assessment of domain evolution by measuring the rate of certain kinds of structural changes that can lead to novel fold development. Birzele *et al.* [31] find fascinating evidence that alternative splicing plays a role in protein structure evolution by developing transitional structures between fold types. Fong and colleagues [32] emphasize the modularity and importance of domain fusion events in the evolution of protein domains. Meier *et al.* [33] suggest a link between conformational flexibility and domain evolution where the native state ensemble can partially occupy at least two intermediate fold types and the relative population of each may be influenced by single amino acid mutations. The results presented here combined with these studies point to the importance of considering protein folds more rigorously in structure matching. It is not only important for our understanding of the discreteness of protein fold space, but informs the more critical question of what precisely should be spatially aligned in structure superimposition.

3 Discussion

We have shown that graphs generated either from TM scores or domain annotation show a high degree of community structure (modularity). TM-scores alone are not able to fully distinguish manually curated topology representatives from each other at the same threshold levels that have been used to analyze fold space across the PDB. This is partly due to the effect of conformation on TM-score. It is shown here that conformational variability within a set of reverse transcriptase structures can lead to very different conclusions about which CATH topology is a closest representative. It is important to realize that since we do not fully understand the relationship between protein sequence and structure, homology reduced datasets may not be topology reduced. This has been shown by analyzing graphs generated by connecting nodes if they share any common CATH topology or SCOP fold and showing that they have similar modularity and graph structure compared to graphs based on structure superimposition (see Fig. 3). It is possible that improving coarse-grained representations like TOPS strings [34, 35] will be useful in the future for handling the multi-resolution complexities of structure comparison.

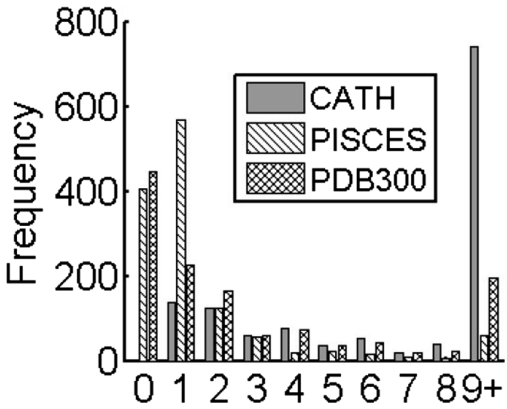

Fig. 3. Frequency of CATH topology usage in three datasets. Using CATH version 3.4, we consider the 1282 topology IDs, counting the number of times a structure in each dataset is annotated with each topology ID. For both of the datasets used in this study, we find that many topologies are not used at all (abscissa value of 0), and that relatively few topologies have a high rate of use. In the CATH database, most topologies have a high rate of use. Interestingly, the PISCES dataset is more topologically diverse than the PDB300 dataset.

Classifying protein tertiary structures into a discrete set of domains is useful in that it helps our conceptual understanding of protein structures, aids in reducing the possible outcomes for sequence based folding procedures, adds to our understanding of the structure-function relationship, as well as many other applications. Whether protein fold space is continuous or discrete depends upon the resolution with which it is viewed. We believe the more fundamental observation is the usage of topology types.

4 Methods

Datasets

We use the PDB300 dataset from [7], which consists of 5906 protein chains of lengths between 40 and 300 residues sharing less than 35% sequence identity. Here, the word "chain" refers to a single polypeptide. It is worth noting that numerous individual protein chains in this dataset contain more than one domain (topology or fold) as defined by CATH or SCOP. Also, the PDB has been updated since this dataset was gathered; 3 structures have become obsolete and were not superseded by a new ID, while 41 have been updated to new IDs. For the purpose of comparison, we continue to use the original version of each PDB ID when employing TM-align.

Domain centric datasets are constructed in two ways. The first is to cut the CATH hierarchy at the topology level resulting in 1233 or 1282 representative domains for version 3.3 and 3.4 respectively. The second begins by using the PISCES server [36] to gather a representative set of chains that are of better than 2.5Å resolution, less than 20% mutual sequence identity, and a crystallographic R-factor of less than 0.25. This dataset contains 4750 PDB chains. We then use CATH to identify individual domains at the Topology level within this set.

The final dataset used is Pfam family PF00078, corresponding to reverse transcriptase (RT). At the time of data download, 283 members with full 3D coordinates were available in the PDB. These structures were downloaded and the subset of points agreeing with the Pfam family definition was retained.

Protein Structure Evaluation Metrics

In this paper, the TM-score defined in [8], is used to analyze the structural similarity of protein structures. This metric is interesting in that it is not symmetric; TM (A,B) does not necessarily equal TM (B,A) particularly when proteins A and B are of different lengths. The TM-score is defined as:

$$TM - score = \mathrm{Max}\left[\frac{1}{\mathrm{L_{Target}}}\sum_{i}^{L_{ali}}\frac{1}{1+\left(\frac{d_i}{d_0\left(L_{T\arg et}\right)}\right)^2}\right] \tag{1}$$

Where L_{ali} is the length of the alignment, L_{Target} is the sequence length of the target structure, d_i is the Euclidean distance between aligned points, and d_0 is a normalization factor based on L_{Target}.

Analysis of Structural Classification

A number of structural classification schemes exist including the CATH database (1), SCOP (2), and PFAM (3). Both CATH and SCOP are hierarchical in nature and utilize a combination of homology, topology, and biochemical function to organize protein structures. The first level of CATH and SCOP classification consists of 4 classes, binning structures into predominantly α-helix or β-sheet content, presence of both, or lack of secondary structure elements. The second level of SCOP, as well as the second and third levels of CATH, is based on overall secondary structure orientation. These

levels are manually curated and place proteins into general categories like beta-barrel and two-layer sandwich. The third level of SCOP takes into account the topology and function of a given protein to decide how related they are evolutionarily. All subsequent levels in both classifications are decided by sequence identity or, in some cases, other sequence based scoring schemes. Pfam families are generated by manual functional curation, multiple sequence alignments, and Hidden Markov Models and come in two varieties: Pfam-A for only manually curated entries and Pfam-B where automated methods are also used to extend the sequence space covered by classification. All three of these databases rely on sequence homology and biochemical functions to group proteins into fold types rather than directly comparing quantitatively the topology of the biomolecules.

Graph Construction

We define a graph based on TM-score as $G_t = \{E, V\}$ where $e_i \in E$ is an edge in G_t if it connects two vertices $a \in V$ and $b \in V$ and $TM(a,b) > t$. Each PDB chain in the dataset is represented by a single node. The TM-score threshold t is initially set at 0.4 as in [7], but values up to 0.9 are also considered to further analyze the graph structure. Graphs are either undirected or directed. To make the directed graphs we consider $t = max(t_1,t_2)$ where $TM(a,b) = t_1$ and $TM(b,a) = t_2$.

Cluster Generation and Comparison

To investigate the community structure of graphs we first employ the Markov Cluster Algorithm (MCL) [13, 14]. In this procedure, graphs are clustered based on random walks that simulate flow along the graph's edges. Nodes that are well connected will exhibit more flow between them than nodes with few connections; the probability of selecting an edge to walk along within the cluster is higher than choosing an edge that leads you out of the cluster (provided the probability of selecting any edge at random is uniform). As the algorithm progresses, nodes that share high amounts of flow (many common walks) are grouped together into clusters.

MCL has evaluation protocols to explore the relatedness of clustering with different parameters. In MCL each edge has a weight. Here we use uniformly weighted (UW) graphs or we use the TM-score as the edge weight. Defining cluster size as the number of nodes within a cluster, MCL computes the Area Fraction (F_{Area}) defined by Eq. 2. This metric gives an indication for the size of clusters as many small clusters or isolated nodes will result in a low F_{Area}. The Mass Fraction (F_{Mass}) is the sum of all edge weights within clusters and is shown in Eq. 3 where w_i is the edge weight of edge i such that edge i is in Cluster c.

$$F_{Area} = \frac{\sum clusterSize^2}{N(N-1)} \tag{2}$$

$$F_{Mass} = \sum_{c=1}^{|c|} \sum_{i=1}^{|E|} w_i \;\; s.t. w_i \in C_c \tag{3}$$

Having F_{Area} close to zero implies that the graph has been clustered into many small clusters, while a value of one implies that all nodes occupy one cluster.

Possessing F_{Mass} close to one indicates that clusters are tightly connected with relatively few edges connecting them. How the algorithm treats the length of a walk (number of edges traversed) is very important to the process and is controlled by a parameter called Inflation, I. Penalizing longer walks produces a large number of small clusters. Allowing longer walks generates fewer, but larger, clusters. It is informative to compare results across multiple inflation values to better understand the organization of the graph.

Table 1. Clustering metrics for graphs of common CATH or SCOP annotations

Dataset	Inflation	Mod_5	Mod_{10}	Mod_{all}	Eff	F_{Mass}	F_{Area}	#C	Max	Avg
PDB300	1.2	.98	.98	.99	.21	.98	.58	137	3617	34.8
	2.0	.89	.90	.92	.33	.91	.25	183	2155	26.1
	4.0	.75	.78	.82	.49	.81	.10	325	1429	14.7
	6.0	.51	.54	.59	.49	.71	.07	450	1179	10.6
	8.0	.37	.40	.45	.49	.66	.05	519	1006	9.2
	12.0	.26	.28	.33	.47	.62	.04	580	931	8.2
PISCES	1.2	.96	.97	.97	.34	.97	.27	138	864	15.7
	2.0	.89	.91	.94	.59	.93	.09	174	465	12.4
	4.0	.78	.82	.86	.66	.88	.06	236	357	9.2
	6.0	.61	.64	.68	.65	.82	.04	289	303	7.5
	8.0	.59	.62	.65	.64	.81	.04	302	296	7.2
	12.0	.55	.57	.61	.64	.80	.04	316	281	6.9

Table 2. Top 10 CATH topology usage

CATH				PDB300			PISCES		
ID	Usage	Architecture	Topology	ID	Usage	C_r	ID	Usage	C_r
3.40.50	19229	3-Layer(aba) sandwich	Rossmann fold	3.40.50	1339	1	3.40.50	379	1
2.60.40	13806	Sandwich	Immunoglobulin-like	2.60.40	678	2	2.60.40	132	2
3.20.20	6106	Alpha-beta barrel	TIM Barrel	1.10.10	384	11	3.20.20	118	3
3.30.70	4236	2-layer Sandwich	Alpha-Beta Plaits	3.30.70	334	4	3.30.70	101	4
2.40.10	3954	Beta barrel	Thrombin	2.60.120	312	6	2.60.120	96	6
2.60.120	3433	Sandwich	Jelly Rolls	2.40.50	268	7	1.10.10	81	11
2.40.50	2244	Beta barrel	OB fold	1.20.5	263	15	1.20.5	65	15
3.30.200	2012	2-layer sandwich	Phosphorylase Kinase	3.20.20	250	3	1.10.287	57	19
1.10.510	1992	Orthogonal bundle	Phosphotransferase	2.40.10	206	5	2.40.50	49	7
1.10.490	1983	Orthogonal bundle	Globin-like	2.30.30	205	16	1.20.120	44	29

Column titles are: Inflation for the MCL parameter that determines granularity of the clustering, Mod_5 - modularity using the 5 largest clusters, Eff - efficiency of the clustering, F_{Mass} and F_{Area} are given in Eqs. 2 and 3, #C - number of clusters, Max - number of nodes in the largest cluster, Avg - average number of nodes across all clusters.

Usage is the number of protein chains that are annotated with the given CATH topology ID. C_r is the rank of this topology ID in CATH.

Acknowledgements. AK and RLJ acknowledge support from the National Science Foundation (DBI 1661391) and from National Institutes of Health (R01GM127701 and R01GM127701-01S1).

References

1. Cuff, A.L., et al.: Nucleic Acids Res. **37**, D310–D314 (2009)
2. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: J. Mol. Biol. **247**, 536–540 (1995)
3. Finn, R.D., et al.: Nucleic Acids Res **36**, D281–D288 (2008)
4. Cuff, A.L., et al.: Nucleic Acids Res. **39**, D420–D426 (2011)
5. Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E., Skolnick, J.: Proc. Natl. Acad. Sci. U.S.A. **103**, 2605–2610 (2006)
6. Grabowski, M., Joachimiak, A., Otwinowski, Z., Minor, W.: Curr. Opin. Struct. Biol. **17**, 347–353 (2007)
7. Skolnick, J., Arakaki, A.K., Lee, S.Y., Brylinski, M.: Proc. Natl. Acad. Sci. U.S.A. **106**, 15690–15695 (2009)
8. Zhang, Y., Skolnick, J.: Nucleic Acids Res. **33**, 2302–2309 (2005)
9. Berman, H.M., et al.: Nucleic Acids Res. **28**, 235–242 (2000)
10. Zimmermann, M., Towfic, F., Jernigan, R.L., Kloczkowski, A.: Proc. Natl. Acad. Sci. U. S. A **106**, E137 (2009)
11. Watts, D.J., Strogatz, S.H.: Nature **393**, 440–442 (1998)
12. Newman, M.E., Girvan, M.: Phys. Rev. E **69**, 026113 (2004)
13. Van Dongen, S.: Technical Report INS-R0010. National Research Institute for Mathematics and Computer Science in the Netherlands (2000)
14. Van Dongen, S.: Ph.D. Thesis, Univ Utrecht, The Netherlands (2000)
15. Gibrat, J.F., Madej, T., Bryant, S.H.: Curr. Opin. Struct. Biol. **6**, 377–385 (1996)
16. Altschul, S.F., et al.: Nucleic Acids Res. **25**, 3389–3402 (1997)
17. Zhang, Y.: BMC Bioinf. **9**, 40 (2008)
18. de Leeuw, M., Reuveni, S., Klafter, J., Granek, R.: PLoS One **4**, e7296 (2009)
19. Reuveni, S., Granek, R., Klafter, J.: Proc. Natl. Acad. Sci. U.S.A. **107**, 13696–13700 (2010)
20. Lee, J., et al.: Science **322**, 438–442 (2008)
21. Guntas, G., Purbeck, C., Kuhlman, B.: Proc. Natl. Acad. Sci. U.S.A. **107**, 19296–19301 (2010)
22. Zhou, Y., Vitkup, D., Karplus, M.: J. Mol. Biol. **285**, 1371–1375 (1999)
23. Holm, L., Sander, C.: Nucleic Acids Res. **25**, 231–234 (1997)
24. Holm, L., Sander, C.: Nucleic Acids Res. **26**, 316–319 (1998)
25. Yoo, P.D., Sikder, A.R., Taheri, J., Zhou, B.B., Zomaya, A.Y.: IEEE Trans. Nanobiosci. **7**, 172–181 (2008)
26. Pandit, S.B., Skolnick, J.: BMC Bioinf. **9**, 531 (2008)
27. Ye, Y., Godzik, A.: Bioinformatics **19**(Suppl 2), ii246–ii255 (2003)

28. Horimoto, K., Toh, H.: Bioinformatics **17**, 1143–1151 (2001)
29. Satuluri, V., Parthasarathy, S., Ucar, D.: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, BCB 2010, pp. 247–256 (2010). https://dl.acm.org/citation.cfm?doid=1854776.1854812
30. Viksna, J., Gilbert, D.: Bioinformatics **23**, 832–841 (2007)
31. Birzele, F., Csaba, G., Zimmer, R.: Nucleic Acids Res. **36**, 550–558 (2008)
32. Fong, J.H., Geer, L.Y., Panchenko, A.R., Bryant, S.H.: J. Mol. Biol. **366**, 307–315 (2007)
33. Meier, S., et al.: Curr. Biol. **17**, 173–178 (2007)
34. Gilbert, D., Westhead, D., Nagano, N., Thornton, J.: Bioinformatics **15**, 317–326 (1999)
35. Torrance, G.M., Gilbert, D.R., Michalopoulos, I., Westhead, D.W.: Bioinformatics **21**, 2537–2538 (2005)
36. Wang, G., Dunbrack Jr., R.L.: Bioinformatics **19**, 1589–1591 (2003)