On the Universe of Protein Folds

Rachel Kolodny,[1] Leonid Pereyaslavets,[2]
Abraham O. Samson,[3] and Michael Levitt[2]

[1]Department of Computer Science, University of Haifa, Haifa 31905, Israel;
email: trachel@cs.haifa.ac.il

[2]Department of Structural Biology, Stanford University, Stanford, California 94305;
email: leonidp@stanford.edu, michael.levitt@stanford.edu

[3]Faculty of Medicine, Bar-Ilan University, Safed 13300, Israel; email: avraham.samson@biu.ac.il

Annu. Rev. Biophys. 2013. 42:559–82

First published online as a Review in Advance on March 20, 2013

The *Annual Review of Biophysics* is online at biophys.annualreviews.org

This article's doi:
10.1146/annurev-biophys-083012-130432

Keywords

protein fold, fold evolution, fold classification, fold use

Abstract

In the fifty years since the first atomic structure of a protein was revealed, tens of thousands of additional structures have been solved. Like all objects in biology, proteins structures show common patterns that seem to define family relationships. Classification of proteins structures, which started in the 1970s with about a dozen structures, has continued with increasing enthusiasm, leading to two main fold classifications, SCOP and CATH, as well as many additional databases. Classification is complicated by deciding what constitutes a domain, the fundamental unit of structure. Also difficult is deciding when two given structures are similar. Like all of biology, fold classification is beset by exceptions to all rules. Thus, the perspectives of protein fold space that the fold classifications offer differ from each other. In spite of these ambiguities, fold classifications are useful for prediction of structure and function. Studying the characteristics of fold space can shed light on protein evolution and the physical laws that govern protein behavior.

Contents

INTRODUCTION

Properties of Native Proteins

A native protein functions in a living cell and is characterized by three properties: (*a*) its amino acid sequence, which defines the atom types and how they are connected by chemical bonds; (*b*) its structure, which defines where every atom is positioned in three-dimensional space; and (*c*) its function or phenotype in the context of a living cell and indeed the entire organism.

A simple example helps illustrate the relationship between protein sequence, structure, and function. Myoglobin from sperm whale is a chain of 153 amino acids that folds into a three-dimensional structure consisting mainly of α-helices that bind a heme group. The heme group in turn binds oxygen and stores it in the whale's muscle, enabling the whale to dive deeply for an extended period of time and so survive. Hemoglobin in human blood cells is a related protein that consists of two different ($\alpha_2\beta_2$) polypeptide chains with a somewhat similar sequence and a three-dimensional structure almost identical to that of myoglobin. From a functional perspective, hemoglobin and myoglobin are also similar in that they bind and release oxygen, with minor differences such as the affinity for oxygen and cell type location.

Fold: characteristic of protein domains whereby they have the same major secondary structures arranged similarly in three dimensions and with similar order or path along the polypeptide chain

Underlying Assumptions About Native Proteins

Two important underlying assumptions regarding native proteins are (*a*) that a protein sequence adopts only one native structure and (*b*) that similar sequences fold into similar structures (5). Even though there are exceptions to both assumptions (61, 69, 79), they still hold in most cases.

Another assumption is that the important unit of structure is a structural domain. Structural domains can be defined in different ways, but there is widespread agreement that they are a unit formed from a single stretch of amino acid sequence and that it interacts weakly with adjacent domains. Domains are found to be from 50 to 300 amino acids long; if they are too short, they will not be stable in isolation, and if they are too long, their folding will be too slow (18, 34). A protein consists of several domains and there are many examples of different ways of combining the same domains in different protein chains (139). Function can be associated with one or more domains, and even with many chains, as seen for large protein machines made of dozens of different chains. Indeed, by relying on reuse of optimized domains, nature can explore far more efficiently the functions in the huge space of possibly longer chains (18, 70). Exactly how domains should be defined is a point of debate (see below), but once these units are defined and classified, we can identify representatives termed folds.

The Universe of Protein Folds

The three sets of properties of protein molecules mentioned above are related to one another. The amino acid sequence (or polypeptide chain) folds and adopts a particular three-dimensional shape, and the enzymatic function, solubility, and other properties depend on this three-dimensional structure. They are, however, different sets in that they describe different objects: strings of letters in sequence space, lists of atomic Cartesian coordinates in structure space, and lists of properties in function space. The universe of protein folds is a complicated object that is related to these three different sets or spaces.

Protein sequence space.

Protein sequence space is simplest, in that it is easily enumerated and there are only 20 different naturally occurring amino acids. In this space, a polypeptide chain of length 100 would be a string of 100 letters and a point in a 100-dimensional space where each axis has the 20 amino acids arranged along it (the order is arbitrary but an order that has chemically similar amino acids close together may be better than, say, alphabetical order). There are 20^{100} different amino acid sequences of this length, a number much larger than the number of electrons in all the galaxies of the universe. In this space, similar sequences are sets of points clustered close together.

RMSD: root mean square deviation

Family: the lowest SCOP level; groups proteins on the basis of their sequence similarity (at least 30% identity)

Protein fitness: how well a protein is suited to all aspects of its biological role

Superfamily: the SCOP level below fold that groups protein families that have low sequence identities but whose structures and functional features suggest a common evolutionary origin is probable

Protein structure space. Protein structure space includes the atomic coordinates of all the atoms in our hypothetical chain of 100 amino acids. As a typical amino acid has about 15 atoms, protein structure space would be approximately 4,500-dimensional (100 × 15 × 3) and each x-, y-, and z-axis would be able to take any value from say −200 to 200 angstrom units. Any one protein structure would be a point in this space, a protein vibrating would be described as a small cloud of points, and an unfolding protein would explore much of the space. Native proteins are a tiny fraction of the points in this space. In general, similar proteins will be close together in structure space.

Protein function space. Protein function space is least well-defined in that it depends on the physiology of the cell and indeed the entire organism. It is not a regular space that can be easily defined by axes, but one expects proteins with similar functions to be close together. For this reason, we focus here on protein sequence and protein structure spaces.

Visualizing Spaces

The high-dimensional sequence and structure spaces mentioned above cannot be visualized, but we can calculate the distances between any two sequences or any two structures. In both spaces there are many measures of distance (or alternatively, similarity) and we favor use of the simplest. For sequences, we line up the two strings and count the number of identical amino acids. For structures, we superimpose the two structures and calculate the root mean square deviation (RMSD) between corresponding C_α atoms.

If we look at the fitness (i.e., viability in the natural environment) of all sequences, this can be represented by a surface (**Figure 1**). This surface has many holes in it, meaning that many sequences cannot be accommodated in any stable protein structure and so have no measurable fitness. The regions of greatest fitness (i.e., the deepest wells in the surface) correspond to protein structures that are of most value to the living organism. Each of these regions corresponds to a sequence family, a sequence superfamily, and a fold. Large regions of sequence space map onto small local regions in structure space. These regions may be surrounded by sequences that cannot form any stable, unique protein structure. The sequences associated with the set of highly similar structures (say, with an RMSD less than 2 Å) are generally related to one another by just a single mutation, allowing evolution to easily explore all sequence variants and scientists to painstakingly classify protein folds.

CLASSIFICATIONS OF FOLDS

Early Work on Classification

Early on, scholars concluded that one can catalog and classify all natural protein folds (71). This idea was supported by initial data: The structures solved in the early days of structure determination (1970s and 1980s) included many examples of the same folds, such as lysozyme-like folds, NAD(P)-binding Rossmann folds, globin-like folds, trypsin-like serine proteases, and immunoglobulin-like

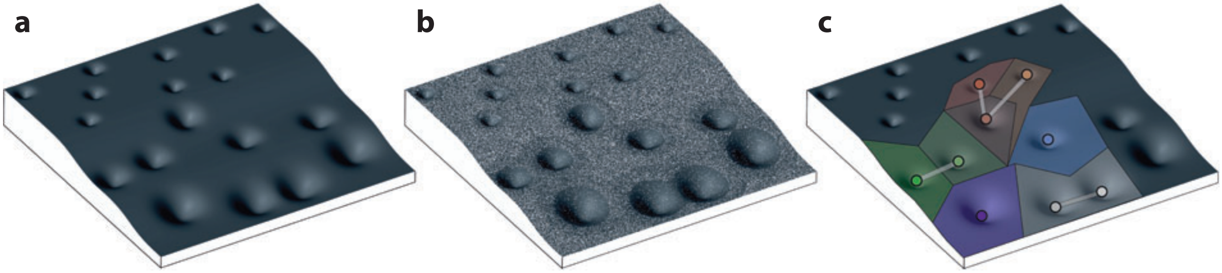


Figure 1

A schematic representation of sequence space, structure space, and function space. (*a*) The smooth dimpled surface plots the functional fitness for all possible protein sequences. Deeper minima are more fit; broader minima have more sequences associated with the structure that is most fit. (*b*) Many sequences cannot be accommodated in any stable protein structure and they are shown as white dots, which represent holes in the sequence surface. (*c*) Structures are shown as colored balls at some of the regions of sequence space that are most fit for the function at hand. Each structure is associated with a region of sequence space around it so that sequences in this region are more likely to adopt the particular most-fit structure. Some of the structures are similar to one another, and these are drawn in the same color and connected by a bar. In the parlance of SCOP (structural classification of proteins), these related structures could constitute a superfamily or fold depending on the level of sequence and structural similarity. The surface in panel *c* has the same holes as those shown in panel *b*, but the holes are omitted for greater clarity. It is not known whether the sequence regions associated with different structures are adjacent or separated by a white unoccupied area. More specifically, are there bridges between different folds that involve a single amino acid change?

β-sandwich folds. In theory, there is a general consensus on the level of similarity required for two proteins to share the same fold: The proteins must share (*a*) the same secondary structures with similar three-dimensional arrangement (denoted architecture) and (*b*) the same path through the structure taken by the polypeptide chain (denoted topology). Thus, in the 1990s, two teams headed by Murzin and Orengo, respectively, embarked on the heroic effort of building the SCOP (structural classification of proteins) (81) and CATH (class, architecture, topology, homology) (84) catalogs of all protein folds. For historical accuracy, we note that around that time FSSP (families of structurally similar proteins) (53), a database of protein structural similarities found automatically, was also created. These classifications provide an ordered view of structure space, with the goal of facilitating a better understanding of its characteristics and evolution.

Scenarios for clustering structures. The underlying scheme for clustering protein structure space is generally agreed upon (see **Figure 2**). There are two main scenarios for constructing a classification: (*a*) incremental classification, i.e., a new protein chain is added incrementally to domains already clustered into folds (an existing classification); and (*b*) full classification, i.e., clustering all protein domains into folds simultaneously. For an incremental classification, a newly solved protein chain is first partitioned into domains. Then, these domain structures are compared to all the existing folds in the classification. If structural similarity is high enough, the domain is added to one such fold, and if not, the domain is listed as a new category, i.e., a fold not observed previously. SCOP, CATH, and other classifications evolve this way (40, 116). For a full classification, all PDB chains are first partitioned into domains, then the similarity of their structures is quantified, and finally an automatic clustering method is used to cluster them based on these similarities (or distances). For example, Pascual-García et al. (89) and Daniels et al. (26) classify folds in this fashion.

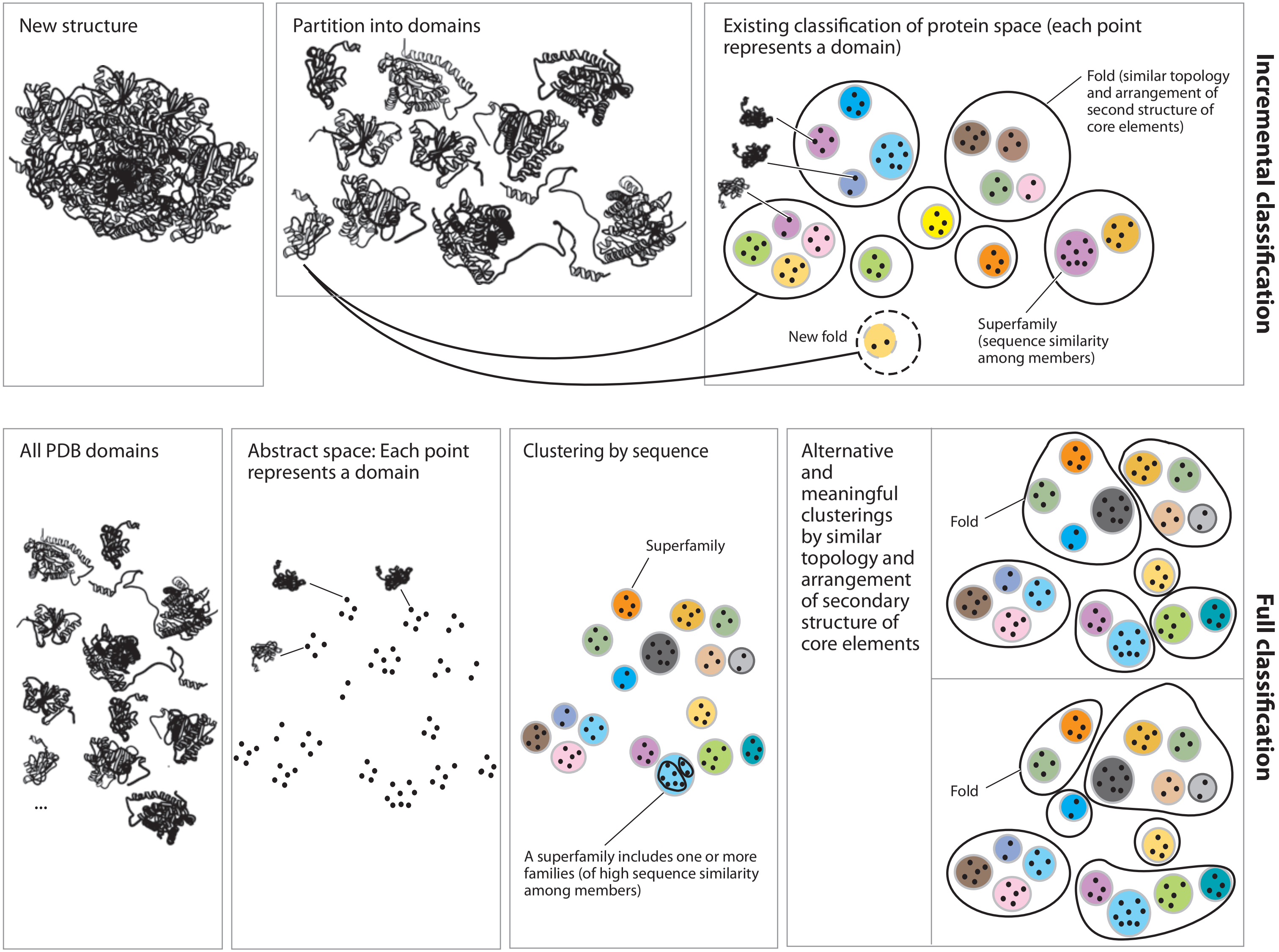
Classifications differ from one another. Classifications vary in their construction, and consequently offer different views of fold space. When dealing with the actual data, the strategy for

CATH: class, architecture, topology, homology

Class: the highest level in the hierarchical protein classifications used in both SCOP and CATH; it characterizes the proportion and general arrangement of secondary structures at the coarsest level

SCOP: structural classification of proteins

Topology: level of CATH classification that corresponds to fold in SCOP; refers to the connections between chain segments and their order along the polypeptide chain


Incremental classification
New structure
Partition into domains
Existing classification of protein space (each point represents a domain)
Fold (similar topology and arrangement of second structure of core elements)
New fold
Superfamily (sequence similarity among members)
Full classification
All PDB domains
...
Abstract space: Each point represents a domain
Clustering by sequence
Superfamily
A superfamily includes one or more families (of high sequence similarity among members)
Alternative and meaningful clusterings by similar topology and arrangement of secondary structure of core elements
Fold
Fold

classifying protein folds depends on specific decisions, algorithms, and parameters. These decisions, algorithms, and parameters vary among different programs and scholars, and thus, even though all programs start from the same PDB data, with almost the same goal in mind (i.e., based on scenario *a* or *b*), the resulting classifications differ dramatically. The most cited classifications are SCOP and CATH, but there are others, e.g., DDD (DALI domain dictionary) (54), PDUG (protein domain universe graph) (29), and COPS (classification of protein structures) (122). To construct a classification, first, the domains of similar sequences are grouped together (i.e., family/superfamily in SCOP, and homology in CATH). For these domains, there is strong evidence for an evolutionary relationship and their grouping is clear-cut, with only a few exceptions (116). Next, the domains are grouped by structural similarities (i.e., class, architecture, and topology in CATH, and class and fold in SCOP). Whether constructed automatically or not, the grouping of domains depends on specific parameters and cutoff values. SCOP was the first classification and it was initially curated manually by Murzin (56), based on visual inspection of the structures, whereas CATH was constructed using automatic computer programs, with manual intervention only for resolving ambiguities (84). As the number of new experimental structures increased (currently thousands of chains are added annually to the PDB), it became more complicated to curate these data manually and now both SCOP and CATH (as well as all other classifications) rely on fully automatic or semiautomatic classification procedures.

Protein domain: the protein structural unit that has structural, biological, and evolutionary significance

PREREQUISITES FOR CLASSIFICATION

There are three essential prerequisites for the classification of folds: (*a*) the object of comparison, generally taken as the rather poorly defined protein domains mentioned above; (*b*) the measure of similarity used on structures and sequences; and (*c*) the way similar objects are grouped together in the classification. Unfortunately, these three prerequisites have not been agreed upon.

Domain Assignment Is Problematic

The first requirement for fold classification is partitioning of proteins into domains, a task that is neither easy nor trivial.

Dividing proteins into domains. It is widely agreed that identifying the domains is a necessary step for classifying multidomain proteins, because the domains are the evolutionary building blocks (63). The proportion of such multidomain proteins in the PDB is large (50%) and increasing (70, 98). One complication of defining domains as units is that approximately 20% of these domains are discontinuous along the chain (98). Determining domain boundaries is not easy, and there is neither a trivial automatic process nor a consensus on how best to do this (51, 63, 131). The extent

Figure 2

Scenarios for constructing an incremental classification and a full classification of all protein structures. The object clustered is a protein domain. In this schematic, each domain is represented by a point, and the structural distance between any two domains is described by their distance in the two dimensions of the schematic. In incremental classification, given a newly solved structure, the new protein structure is partitioned into domains, and each new domain is compared to domains of the existing classification to identify the most similar folds. If no such fold exists (dependent on the parameters of the classification), then a new fold is added to the classification and the particular domain is added to it. In full classification, the domains are first clustered by their sequence similarity, forming the protein families and superfamilies. Typically, the structures of the domains in a family are similar (i.e., close in space). Then, superfamilies that have similar structures (i.e., close in space) are clustered into folds. There are many meaningful ways to cluster similar sets. This is true even in the very simple setting of points in two dimensions, and we show two such meaningful clusterings.

of the complication can be appreciated both from the many solutions offered (63 and references therein) and by the fact that in the CASP structure prediction competition, partitioning the target proteins into domains is a responsibility of the judges (22, 127).

CASP: critical assessment of structure prediction

Structural alignment: the computational procedure that compares two protein structures to quantify their similarity

Methods for domain assignment. Methods for automatic domain assignment rely either on the comparison of the target protein to already identified domains or on the identification of geometric or physicochemical properties of the structures (2, 33, 54, 63, 98, 140, 143). Because different methods identify different domains, scholars take one of two paths. In the first, they resort to assigning the domains manually, as is the case for SCOP. Manual assignments are considered more reliable (51). In the second, scholars trust domain boundaries that have been identified by several different methods (because different approaches reached the same conclusion), as is the case for CATH. In CATH, domain assignment is done by a consensus procedure using three algorithms for domain recognition: If all algorithms concur, the common solution delineates the domains of that protein; if not, the assignment is done manually.

Comparing Structures Is Difficult

The second requirement for fold classification is comparison of protein sequence and structure. In contrast to the relative ease with which we compare two protein sequences, comparing two structures is much more challenging.

It is easy to compare sequences. Sequences can be compared by counting how many amino acids need to be changed to transform one sequence into the other. If the sequences are the same length, then this is the length of the sequence minus the number of identical amino acids. If the sequences are not the same length, then the sequences are aligned; this can be done easily using a dynamic programming algorithm, which runs in time proportional to the lengths of the sequences squared (82, 117). Other parameters, such as the penalty of inserting a gap into either sequence, remain to be specified, but a solid history of comparing sequences of different lengths has led to trusted and generally accepted procedures.

It is much harder to compare structures. A method for quantifying the similarity or distance between two structures is needed. Unfortunately, there is no such agreed-upon method or measure in the field. Rather, many methods compare protein structures, some are used in the classification schemes and some are developed independently of classification. The task of identifying and quantifying the similarity of domains is termed structural alignment. Structural alignment does not compare whole domains but rather equally sized substructures contained in them. The similarity of two substructures is measured with scores that balance the geometric distance between corresponding atoms (e.g., RMSD, the alignment length, and occasionally other parameters such as the number of gaps, and secondary structure agreement) (52, 64, 137). Furthermore, given a score, finding the optimal superposition and substructures quickly and accurately is a nontrivial technical challenge. Kolodny & Linial (65) proved that an alignment with an optimal score can always be found but their (polynomial) algorithm is slow and runs in time proportional to the sequence lengths to the eighth power. Many programs, including STRUCTAL, SSAP, CE, DALI, MAMOTH, Matt, and SSM, use their own heuristic solutions to obtain much faster structural alignments. The different programs identify different common substructures. Because the programs deduce the similarity of a domain pair from the similarity of the substructures, different programs reach different conclusions regarding the similarity of the pair of domains. Consequently, they identify different structurally similar pairs of domains. Thus, Kolodny et al. (64)

suggested using the combined results of multiple methods. For reviews of structural alignment methods see References 62, 97, and 112.

Clustering Is Tricky

The third requirement for fold classification is clustering of similar protein structures. Even if we decide on a measure of similarity/distance between protein structures, we still need methods to convert these pairwise relationships into a clustering of structure space.

Clustering is an art. Clustering domains, like the clustering of any dataset, is more of an art than a science. To cluster, one must define the distance or similarity measure between objects, in this case protein domains (or protein sequence superfamilies). Then, given these distances, the goal is to cluster the data so that the similarity within a cluster is greater than that between clusters. Unfortunately, in the case of protein structure, the measure of similarity is not unanimously agreed upon. Also, there is no consensus on a reasonable ratio between the inter- and intracluster distances or similarities; this is important because different ratios result in different numbers of clusters (see **Figure 2**). Nevertheless, once one assumes a distance measure and a suitable ratio, automatic clustering can be done, and there are different methods to do this.

Manual clustering. This is how Murzin and colleagues constructed SCOP: They inspected the structures one by one and determined to which fold each domain belongs. This is how the classifiers of SCOP interpret the term fold. In particular, they consider only core elements and decide which of the residues are in the core [up to 50% of the residues can be left out (56)]. A domain is deemed sufficiently similar to a fold if the core looks sufficiently similar to the cores of the domain elements already in the fold. The advantage of manual classification is that the immense expertise of the classifier is summarized in the database and made available to the biological community. The disadvantages are that it is difficult to classify large datasets, and that manual classification relies almost entirely on the knowledge accumulated in the mind of the individual who is the classifier. Further, one could argue that because this is done by a human, there is a limit to the number of folds that the classifier can remember/inspect, and that this is the effective limit of the number of folds in such a classification.

PROTEIN CLASSIFICATIONS ARE INCONSISTENT

Domain Boundaries Are Inconsistent

The domain boundaries are defined differently by SCOP, CATH, DDD, and automatic methods for domain assignment (25, 27, 49, 51, 105). For example, CATH tends to break protein chains into smaller domains than SCOP does (25, 27, 49), and a single domain in SCOP can be mapped to as many as six domains in CATH. Moreover, 28% of SCOP domains are mapped to more than one CATH domain, whereas only 14% of CATH domains are mapped to more than one SCOP domain (25). Overall, only 70–80% of the domains classified in SCOP and CATH have similar domain boundaries (80% overlap) (25, 105). Domains assigned by automatic methods differ from the domains classified by SCOP and CATH even more, and over 10-20% of the automatic-method domains are under- or overcut compared with the domains that the classifications agree upon (51). To deal with these discrepancies, several studies suggested using a consensus set (27, 105). These consensus datasets have the advantage that their domains are undisputed and thus useful for training and parameterizing new automatic methods for domain assignment. The disadvantage is that the ambiguous, and hence interesting, evolutionary relationships are missing.

Classification Hierarchies Are Inconsistent

Even when considering only the domains whose boundaries are similarly defined in SCOP and CATH, the grouping of domains at the fold level in SCOP and at the topology level in CATH differs. It is unclear what is the best way to compare two classifications that have a different number of clusters of different sizes. Several studies compared the number of times that two domains are clustered together in CATH (i.e., have the same class, architecture, and topology, or CAT, classification) yet clustered differently in SCOP (i.e., are not in the same fold), or vice versa (25, 89). The disagreement is significant: There are 3.9 times more pairs classified in the same fold and different superfamily by CATH than by SCOP. More than 94% of the domain pairs defined by SCOP in the same fold are also co-classified by CATH, but these commonly joined pairs represent only one-third of the pairs with the same CAT classification in CATH (25, 89). These calculations are heavily influenced by the fact that CATH has several very large clusters at the topology level (because all pairs within these clusters contribute to the count, their overall contribution is significant). Thus, many errors can be attributed to relatively few superfolds such as the Rossmann fold or the immunoglobulin fold.

Structure Similarity Measures Are Inconsistent

Automatic classifications rely on different structural alignment programs for identifying the structural similarity of domains and, as such, reach different results. In CATH, folds are clustered at the topology level; that is, domains of the same fold have the same C, A, and T levels. To determine whether two domains should have the same T classification, CATH relies on the structural alignment program CATHEDRAL (98) [which evolved from SSAP (Sequential Structure Alignment Program) (85)] and checks that the SSAP score is above a threshold value and that a significant portion of the domains are aligned with each other. In DDD (28), the similarity is detected by the structural alignment program DALI (52) and quantified via its Z-score. In the classification by Daniels et al. (26), the structural alignment program Matt is used (76), and in PDUG, the classification uses DALI Z-score (108). COPS (122) uses a different measure described by Sippl (114).

The Meaning of a Fold Is Inconsistent

In the automatic classifications, the definition of "fold" depends implicitly on the selection of the structural alignment program and on the particular threshold values used. Different structural alignment methods optimize different scores, with different weights of the geometric parameters. The methods also involve design decisions that have an impact on what is considered similar. For example, many methods use the algorithmic technique of dynamic programming to compare the two chains and identify the aligned substructures (121, 142). Such methods can only match residues in the same order along their polypeptide chains; in particular, these methods cannot detect circularly permuted similarities and thus such cases (72) are assigned to different folds (notice, however, that this is in agreement with the common definition of a fold). Finally, the sensitivity of structural alignment methods varies (64), and this also affects what it means to have the same fold.

Clustering Is Inconsistent

An insightful analysis by Pascual-García et al. (89) shows that a significant source of disagreement between SCOP and CATH is the procedure used when clustering the hierarchies. They quantify the agreement between different classifications, SCOP, CATH, and classifications calculated by automatic hierarchical clustering, and find that at the fold level, the disagreement between SCOP and CATH is greater than the disagreement with the results of their clustering procedure. The authors further show that the single-linkage clustering agrees more with CATH, and that the average-linkage clustering agrees more with SCOP, compared with the relative agreement between SCOP and CATH. Sam et al. (102) show that the grouping in SCOP is most consistent with automatic average-linkage clustering or with Ward's method clustering; these methods cluster so that each cluster (namely, fold) is cohesive as a whole. This makes sense, as SCOP uses a procedure that is effectively an average-linkage algorithm, whereas CATH uses something more like single linkage (no penalty for joining structurally distinct domains). Their conclusion is to consider consensus sets, or all pairs that are classified similarly by both SCOP and CATH (and perhaps DDD). Here, too, it is clear that these are the less disputed cases. Again, focusing on consensus sets may lead scholars to overlook interesting cases that are not errors, but ambiguities that shed light on evolutionary relationships.

Hierarchical clustering: an automatic procedure for clustering protein domains that uses a similarity measure between pairs of domains to place similar domains in the same cluster

Architecture: an intermediate level in CATH between class and topology; groups protein domains with similar arrangement of secondary structures in three-dimensional space, but not necessarily the same topology

OTHER ISSUES

Estimates of the Number of Folds in Nature Vary Widely

Even when estimating a single parameter, such as the number of folds in nature, the results range from 1,000 to 10,000, depending on the classification used. Initially, Chothia (17) estimated 1,000 folds; a later and more detailed analysis of statistical sampling using SCOP resulted in an estimate of 4,000 folds (44). Then, using CATH, Orengo et al. estimated an even larger number of 8,000 folds (83). Most recently, relying on SCOP and the change in the number of observed folds over time, Coulson & Moult (24) estimated over 10,000 folds. As indicated by Grant et al. (45), there is an inherent difficulty in estimating this number because of the vast amount of genomic data that we have not seen yet. Further, as Sippl (115) pointed out, this number is sensitive to the parameters of the classification, i.e., how widely or narrowly fold is defined. Different definitions result in dramatically different estimates.

Nonmetric Distance Measure

Another fundamental issue with fold classification is the use of a distance measure that is based on the similarity of substructures rather than on complete domains. Such distance measures are nonmetric, implying that they do not follow our intuitive notion of a distance and, as noted by several authors, are inappropriate for clustering (99, 101). Sippl (114) also suggested a measure of protein structural distance that is metric. In CATH, SSAP is used for clustering, so that the equivalence associates 70% of the residues of the smaller domain (84). In SCOP, the similarity of the architecture and topology is assessed over the cores of the proteins, and different instances of the same fold may have so-called peripheral elements of secondary structure and turn regions that differ in size and conformation, and may consist of as much as half of the structure (56). The problem with a nonmetric classification is twofold. First, transitive inference of similarity fails. Imagine domains A and B are similar and domains B and C are similar; then domains A and C should also be similar. However, when the similarity is defined only on the basis of substructures, domains A and C can have nothing in common (see **Figure 3**). Pascual-García et al. (89) show that the number of transitivity violations in the context of clusters is significant.

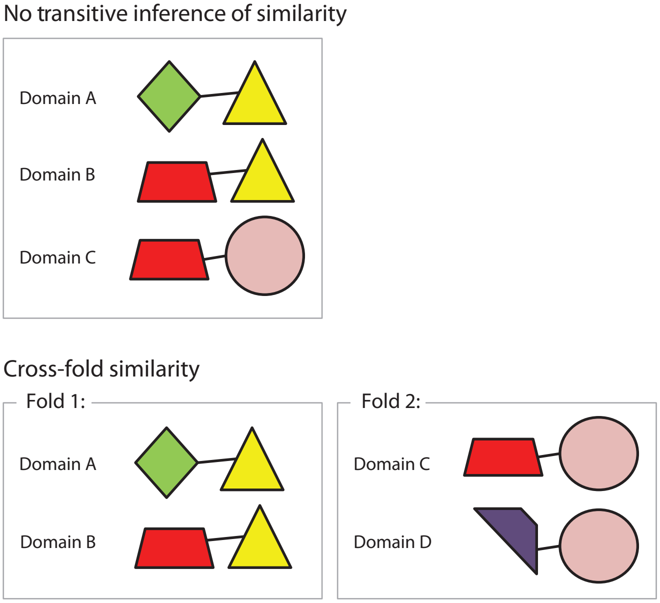


Figure 3

Measuring structure similarity over substructures implies that similarity cannot be inferred by transitivity: Domain A is similar to domain B (they share the yellow triangle) and domain B is similar to domain C (they share the red trapezoid), and yet domains A and C have nothing in common. A demonstration of cross-fold similarity when Fold 1 has domains A and B and fold 2 has domains C and D (they share the pink circle): Domains B and C are classified differently yet are similar.

Second, cross-fold similarities are abundant. Imagine domains A and B have one fold and domains C and D have another fold; then one would expect domains A and C, which have different folds, not to share significant substructures. However, cross-fold similarities in SCOP and CATH are abundant and demonstrate an inherent ambiguity in the data. The existence of domains that are classified differently yet have significant geometrically similar substructures was already mentioned in the original CATH paper (84, 89), and cases in CATH were later surveyed more systematically (50, 60, 64). Similar evidence is available for SCOP (37, 89, 100, 110, 137). Cross-fold similarities remain even after a classifier resolves (possibly in an arbitrary manner) ambiguities in the data. Indeed, clustering the domains into folds in the presence of such ambiguities is a very complicated task, and the result is sensitive to the particular clustering algorithm used. It may be that the dataset of all domains does not easily lend itself to clustering, at least when relying on similarity measures derived from structural alignments. The high frequency of cross-fold similarities interferes with a hierarchical classification and has been described as the continuous nature of fold space (66, 84).

Continuous nature of fold space: describes the existence of numerous similarities in fold space that are not implied by the particular hierarchical classification

Cross-Fold Similarities

One should keep in mind that partitioning the protein structure universe into discrete folds does not exclude the similarities between protein domains that occur in different folds (**Figure 3**). Namely, the debate on whether to view structure space as discrete or continuous is, in a way, the debate on whether these other, additional cross-fold structural similarities should be ignored in that they are masked by the fold classifications. In some cases it is beneficial to focus only on some of the similarities and count on a classification, whereas in other cases it is not. As these noncompliant similarities are not easy to detect and can be of evolutionary or functional significance, several groups have collected them in publicly available databases: FSSP (53), Fragnostic (38), and SISYPHUS (5). Sadowski & Taylor (100) suggest characterizing structures with (one or more) labels, rather than a hierarchical classification (similar to function classification using GO), to account for these cross-fold similarities.

Another complication in the construction of fold classifications arises from the assumption that proteins of similar sequences always have similar structures. The assumption is employed when proteins with similar sequences are first grouped into family and superfamily levels in SCOP and the homology level in CATH. After this is done, proteins with different sequences but with similar structures are grouped into the fold level. There are two reasons why the same sequence has different structures: conformational changes needed for function (e.g., induced fit) and conformational changes caused by a changing environment (e.g., pH change).

Conformational Changes Due to Function

There are many examples of conformational changes that are related to the mechanics of the functioning protein. The three-dimensional structure of a protein is, of course, not static; the structure can change to accommodate the function of the protein. There are numerous examples of large conformational changes of proteins upon binding to ligands, DNA, or metals. Other well-known examples are the conformational changes of myosin and of membrane channel proteins. These are, in a sense, mechanistic conformational changes involving proteins that have more than one stable conformation, e.g., depending on their environment or their binding partner. In many cases, these conformational changes involve relative movement of rigid domains and so do not undermine the idea that a protein domain with a particular sequence has a particular structure. There are more surprising cases in which small domains adopt very different folds, such as hemagglutinin conformational changes with pH. Other examples are often associated with pathologies and include the prion protein involved in mad cow disease (103) as well as the amyloid peptide involved in Alzheimer's disease (107). In both cases, the alternative fold is stabilized by aggregation.

Conformational Changes Not Due to Environment

Alexander et al. (1) engineered an important recent example that shows how a single amino acid substitution can change the fold of a protein. The reported change is dramatic. One structure is a three-helix bundle, and the other has a four-stranded β-sheet with a single α-helix; 85% of the residues change their secondary structure, with only eight residues in the central α-helix plus one or two turn residues retaining the same conformation in both structures (111).

Overall, many pairs of domains in the PDB have similar sequences and nonsimilar structures, as identified by Kosloff & Kolodny (69) and subsequently by Burra et al. (13). Murzin (79) discusses cases of conformational changes due to mutations. To accommodate such cases, Alva et al. (3) suggest adding the level metafold to the hierarchical classifications; metafold would be a level in the hierarchy above a fold that is the collection of all such related folds. The authors offer a clear illustration of this idea with the cradle-loop-barrel metafold.

USEFULNESS OF FOLD DEFINITION

Do Fold Classifications Help Solve Problems?

Given the multitude of problems associated with protein fold classification, the reader may well be surprised to learn that the hierarchical classification of folds is of great practical value. The first and obvious measure of the value of fold classification techniques is whether they improve problem solving. The following applications illustrate the usefulness and contributive aspect of fold classifications (i.e., SCOP and CATH) as an important tool for the scientific community.

Elucidate rules. Classifications help find fold principles and increase our understanding of the rules governing the high frequency of occurrence of favorable structural motifs such as the Greek key motif and the immunoglobulin superfold. These favorable motifs, which are suited to many amino acid sequences and therefore highly populate fold space, have helped describe the structural principles of these folds, giving statistically significant rules meaningful to protein experts (23). The structural principles underpinning much of the fold space can thus be described with respect to different fold categories. Significantly, fold classifications do not provide information on the kinetic pathways of folding, and proteins with identical folds could fold through different pathways (104).

Predict structure. Fold classification techniques are also useful for structure prediction and determination. Most important in this category is the contribution of fold classification in providing a database used by fold recognition techniques (92, 118). Such techniques have been used to successfully predict structures in more recent CASP competitions. In addition, these techniques may be used to derive amino acid similarity matrices and substitution tables for sequence comparison and fold recognition methodologies (32, 109). As solved protein structures span more and more of the fold universe, they may soon be expected to encompass all possible natural folds. Once this admirable goal is reached, all subsequent protein structures must, by definition, adopt one of the existing folds. This would greatly facilitate structure prediction and determination.

Predict function. Fold classification databases are widely used to predict the function of proteins. As noted by several researchers (4, 59, 104), the variation of local structure caused by small changes in sequence is what gives rise to independent homologous and analogous proteins. Such variability often leads to the domain combination, permutation, and decomposition found in multidomain proteins (6, 128, 135). As a matter of fact, fold classification databases enable us to predict function of proteins in 95% of folds that have only one associated superfamily (11, 20, 51). Folds within these superfamilies are usually functionally related (86). Indeed, this is often why they are assigned to the same superfamily in SCOP. In such cases, knowing the fold of a domain could tell us much about protein function (20, 86, 130). Function prediction is particularly useful if the amino acid sequences seem unrelated and only the protein folds remain conserved.

Find homologous structures. The fold classification databases facilitate our access to information on structural homology. This is easily seen from a review of the literature that reveals that the techniques have been used as the basis for comparative structural analysis in thousands of articles (48 and references therein). Remarkably, classifications have been helpful to the investigation of distantly related proteins with similar folds (46, 55, 75, 80, 126). Comparative structural analysis is particularly useful when used with sequence homology, i.e., when designing fold-specific hidden Markov models for comparing sequences to the fold families of structures (32). Thus, fold classification databases are detailed and comprehensive descriptors of structural homology.

Protein engineering. Classification techniques simplify protein engineering and design. For instance, if a stable fold is required, then it can be based on conserved sequence characteristics of protein families and superfamilies with stable folds. This has been a convenient approach particularly for designing enzymes that adopt a stable fold (7, 35). Thus, the classification techniques increase our understanding of the structural principles underlying folds and domains and assist our endeavors in protein engineering and design.

Other databases. Fold classifications are useful reference datasets for constructing other structural databases. It is perhaps ironic that existing databases (i.e., SCOP and CATH) are utilized to generate other databases useful for integrative structural data mining (8–10, 12, 21, 58, 94, 120, 125, 133) and helpful for studying quaternary protein-protein interactions (31). Such studies make up a large number of the citations for the SCOP and CATH classifications.

Evolution. Fold classification techniques are widely used to better understand the evolution of protein enzymatic functions (39, 41, 67, 78, 90), evolutionary changes of protein structures (14, 47, 73, 87), and hierarchical structural evolution (30, 88). This application is discussed in detail below.

The applications of classification techniques listed above represent a select few and many more are conceivable. From a user's perspective, this is only a partial list of subfields of protein structural bioinformatics where SCOP and CATH have been used extensively. Much gratitude is owed to the authors of fold classification techniques for providing such a rich resource that has great significance in structural biology.

EVOLUTION OF PROTEIN FOLDS

Studying Protein Evolution

Studies of protein evolution rely on the relationships between the sequences, structures, and functions of current-day proteins. Such relationships can be evaluated on the basis of perceived similarity. Strong sequence similarity is considered sufficient evidence of common ancestry; medium-sized domains are considered homologous if more than 25% of their residues are identical, although statistically more sound methods based on expectation of errors (E-values) are also used (91). When sequence identity is too weak to be detected, significant structural and functional similarity can also provide evidence of remote homology (78). This assumes that the structures and function diverge more slowly than sequence and hence provide evidence of the common ancestry after sequence similarity has disappeared. Understanding of fold evolution also comes from simple "toy models" of the theoretical protein universe (sequence and structure space) and their comparison to the observed or natural protein universe (29, 77, 95, 134). **Figure 4** illustrates some of the evolutionary processes involving protein folds.

Starting Point of Evolution

Scholars do not agree on what constitutes the starting point in the evolution of proteins. In the single-birth model (19), all present-day protein families evolved from the proteins that existed in LUCA, the last universal common ancestor. In the multiple-birth model (16), the ancestral proteins emerged at different times. Scholars have reconstructed the evolutionary trees of proteins using phylogenetic analysis (15, 16, 138). These trees were also used to quantify the age of different folds (16, 132), and α/β proteins emerged as the oldest proteins in nature. As Taylor (123) points out, α/β proteins have a clear N-terminal folding bias, which is to be expected for a nascent chain translated from ribosomes, and suggests that advanced cellular machinery existed when they evolved. Scholars have also estimated the size of the initial set of proteins from simulations (95)

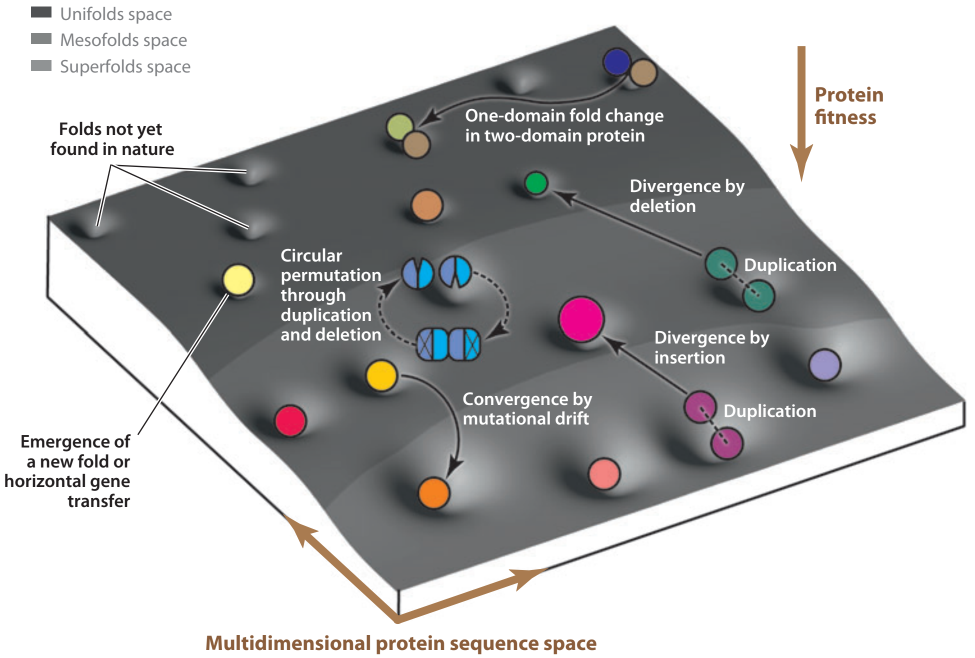

Unifolds space
Mesofolds space
Superfolds space
Folds not yet found in nature
One-domain fold change in two-domain protein
Protein fitness
Divergence by deletion
Circular permutation through duplication and deletion
Duplication
Divergence by insertion
Convergence by mutational drift
Duplication
Emergence of a new fold or horizontal gene transfer
Multidimensional protein sequence space

Figure 4

Schematic representations of different processes of protein fold evolution. The landscape surface represents a projection of the multidimensional protein sequence space. Funnels of different size correspond to varying degrees of fold fitness (deeper is more fit), and superfolds, mesofolds, and unifolds occupy regions of sequence space that are progressively less fit (higher above the base plane). Colored circles represent protein domains of different sizes. A common process in protein evolution that is responsible for new fold emergence is duplication followed by divergence. Rarer processes in protein fold evolution are convergence, circular permutation, and emergence of a new fold by occasional events. Certain folds that are physically possible may not exist because they have not yet been found in nature.

and phylogenetic analysis (96, 138), as well as the occurrence of supersecondary structures (73, 119) based on the evolutionary processes of fold replication summarized in **Figure 4**.

Duplication and Mutation

The most common process in protein evolution is duplication followed by divergence (18, 73). The advantage and beauty of this process are that it removes the functional pressure from the protein domain, as the original copy maintains the original function while the new divergent copy is free to explore alternative functions. Duplication is common in all types of species, and using the SUPERFAMILY database it was estimated that the proportion of duplicated domains in animal, fungi, and bacteria genomes is at least 93%, 85%, and 50%, respectively (18).

Insertion and Deletion

The sequence of a protein can diverge more by the insertion/deletion of larger segments. Importantly, the intermediate folds encountered during the evolutionary path cannot suffer a significant loss in stability. Therefore, different evolutionary processes differ in their effect on core residues and fold stability. Murzin (78) showed cases of limited change in different topological isomers in which only the relative position of the loops differs and which would not be expected to affect fold stability. On the contrary, changes to the core residues, such as β-strand insertion and deletion, β-hairpin flip and swap, accretion (piecemeal growth; 74), and helix-strand transitions (61, 93, 119, 124), are expected to affect fold stability more. They occur about an order of magnitude less frequently than mutations (47). Errors in the translation of the protein sequence from the DNA sequence can also facilitate the emergence of a new fold through a frameshift or a mutation of the stop codon (124).

Circular Permutations

Circular permutations are an alternative example of a change that has only a minimal impact on the structure and stability of a domain, as only the gap position between N and C termini, which are close in space, changes; 5% of all domains are estimated to be a result of such permutations (129).

Multiple Structures

Several studies have suggested that metamorphic proteins with multiple conformations, and possibly multiple functions, have an evolutionary advantage (42, 57, 106) and, in particular, that these metamorphic proteins facilitate the development of new folds (136). In addition, simulations confirmed that proteins that are bistable (i.e., that have multiple stable conformations) have an evolutionary advantage (113).

Convergent Evolution

Convergent evolution is the acquisition of the same biological characteristics in evolutionary unrelated lineages. Convergent evolution has been suggested for especially popular protein folds (42, 134). In this view, the converging superfolds are highly designable folds in that they constitute the stable three-dimensional structure of many different protein sequences. Such folds are characterized by sequences that diverge widely while maintaining similar structures. Nonetheless, several studies suggest that cases of convergent evolution are rare. The frequency of convergent evolution based on superfamily domains assignments was estimated to be a mere 0.4–4% (43), and in subsequent research based on PFAM domain assignments, was revised to be between 5.6% and 12.4% (36). Convergent evolution was also studied by analyzing simulations of two- and three-dimensional lattice protein models, where the computational model is sufficiently simple to allow in silico enumeration of all sequences (134, 141).

DISCUSSION

Protein fold space is shaped by physical restrictions and the course of evolution. Unfortunately, we understand the physical restrictions of protein chains only at a general level and we know even less about the course of evolution. We study the properties of current-day protein fold space, and its relationships to sequence and function, to better characterize these physical restrictions and to unravel the path of evolution. This also has important practical implications.

Classifications Offer the Scientific Community an Ordered Perspective of Fold Space

Identifying meaningful patterns in the large dataset of the PDB is a formidable challenge. In particular, it requires overcoming two nontrivial technical hurdles: identifying domains and comparing structures. The classifications are important because they have overcome these challenges and thus offer a shortcut to identifying and validating characteristics of structure space.

Restricted Repertoire of Folds

For example, several studies suggest that the repertoire of observed folds is fairly restricted. The estimated number of folds used by nature varies between 1,000 and 10,000, depending on how they are clustered and classified, but there is general agreement that the number is bounded (17, 24, 45). The relative frequency of sequences in a fold, or the number of superfamilies constituting a fold, is highly nonuniform. Coulson & Moult (24) characterized unifolds, mesofolds, and superfolds, which are folds with low (a single family of sequences), medium, and high numbers of sequences associated with them, respectively. Others (29, 68, 95) have characterized the number of sequences associated with a fold as a power-law distribution. This fundamental characterization would have been difficult to see without the ordered perspective of structure space that the classifications provide. We are left with a key question: Why is the repertoire of folds so limited?

Classifications Might Mask Similarities in Fold Space

To provide an ordered and useful hierarchical perspective of structure space, the designers of a classification should resolve ambiguities in the data, and as a side effect, they might mask alternative and acceptable solutions. Indeed, there are multiple valid and meaningful characterizations of fold space, including SCOP, CATH, and possibly other classifications. Importantly, the fact that the definition of folds is not unique and objective does not diminish its usefulness, especially because many observations are revealed, e.g., both SCOP and CATH reveal the highly nonuniform nature of structure space. Nonetheless, when relying on a classification, it is important to keep in mind these hidden alternatives and, in particular, the existence of alternative domain definitions and cross-fold similarities (i.e., structurally similar proteins that have different folds).

The arbitrary nature of classification was also noted by Darwin in *On the Origin of Species*: "Finally, with respect to the comparative value of the various groups of species, such as orders, suborders, families, subfamilies, and genera, they seem to be, at least at present, almost arbitrary. . . Instances could be given among plants and insects, of a group of forms, first ranked by practiced naturalists as only a genus, and then raised to the rank of a subfamily or family; and this has been done, not because further research has detected important structural differences, at first overlooked, but because numerous allied species, with slightly different grades of difference, have been subsequently discovered."

SUMMARY

In this review, we have attempted to portray the concept of a fold in the universe of proteins. We began with broad definitions of sequence, structure, fold, and function space. We reviewed how the classification of folds began and what parameters were used. Then we reviewed the meaning of a fold biologically, physically, and evolutionarily and discussed the meaningfulness of each. We find that a protein fold, even if inconsistently and arbitrarily defined, is very useful to the scientific community.

SUMMARY POINTS

1. Protein folds are related.
2. Protein folds are arbitrarily and inconsistently defined.
3. Classifying protein folds is complicated.
4. Protein folds are useful for predicting structure and function.
5. Protein folds can help infer evolutionary relationships.
6. Protein folds are subject to natural laws governing their stability, unique conformations, and sequence repertoire.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by NIH award GM063817 to M.L., by Marie Curie IRG grant 224774 to R.K., and by Marie Curie CIG grant 322113 to A.O.S. Michael Levitt is the Robert W. and Vivian K. Cahill Professor of Cancer Research. The authors thank Dr. Sergio Moreno-Hernandez for feedback on the manuscript.

LITERATURE CITED

1. **Alexander PA, He Y, Chen Y, Orban J, Bryan PN. 2009. A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci. USA* 106:21149–54**
2. Alexandrov N, Shindyalov I. 2003. PDP: protein domain parser. *Bioinformatics* 19:429–30
3. Alva V, Koretke KK, Coles M, Lupas AN. 2008. Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. *Curr. Opin. Struct. Biol.* 18:358–65
4. Andreeva A, Murzin AG. 2010. Structural classification of proteins and structural genomics: new insights into protein folding and evolution. *Acta Crystallogr. F* 66:1190–97
5. **Andreeva A, Prlić A, Hubbard TJP, Murzin AG. 2007. SISYPHUS—structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res.* 35:D253–59**
6. Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* 310:311–25
7. Aroul-Selvam R, Hubbard T, Sasidharan R. 2004. Domain insertions in protein structures. *J. Mol. Biol.* 338:633–41
8. Aung Z, Tan KL. 2006. MatAlign: precise protein structure comparison by matrix alignment. *J. Bioinforma. Comput. Biol.* 4:1197–216
9. Bertone P, Gerstein M. 2001. Integrative data mining: the new direction in bioinformatics. *IEEE Eng. Med. Biol. Mag.* 20:33–40
10. Bhaduri A, Pugalenthi G, Sowdhamini R. 2004. PASS2: an automated database of protein alignments organised as structural superfamilies. *BMC Bioinforma.* 5:35
11. Brenner SE, Chothia C, Hubbard TJ. 1997. Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* 7:369–76
12. Bukhman YV, Skolnick J. 2001. BioMolQuest: integrated database-based retrieval of protein structural and functional information. *Bioinformatics* 17:468–78

1. Shows how a pair of engineered proteins with a single-residue difference can have dramatically different structures.

5. Presents cases that, although interesting, are actually masked by the decisions made in curating SCOP.

13. Burra PV, Zhang Y, Godzik A, Stec B. 2009. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc. Natl. Acad. Sci. USA* 106:10505–10
14. Caetano-Anollés G, Caetano-Anollés D. 2005. Universal sharing patterns in proteomes and evolution of protein fold architecture and life. *J. Mol. Evol.* 60:484–98
15. Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. 2009. The origin, evolution and structure of the protein world. *Biochem. J.* 417:621–37
16. Choi IG, Kim SH. 2006. Evolution of protein structural classes and protein sequence families. *Proc. Natl. Acad. Sci. USA* 103:14056–61
17. Chothia C. 1992. One thousand families for the molecular biologist. *Nature* 357:543–44
18. Chothia C, Gough J. 2009. Genomic and structural aspects of protein evolution. *Biochem. J.* 419:15–28
19. Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. *Science* 300:1701–3
20. Chothia C, Hubbard T, Brenner S, Barns H, Murzin A. 1997. Protein folds in the all-beta and all-alpha classes. *Annu. Rev. Biophys. Biomol. Struct.* 26:597–627
21. Chou KC, Maggiora GM. 1998. Domain structural class prediction. *Protein Eng.* 11:523–38
22. Clarke ND, Ezkurdia I, Kopp J, Read RJ, Schwede T, Tress M. 2007. Domain definition and target classification for CASP7. *Proteins* 69(Suppl. 8):10–18
23. Cootes AP, Muggleton SH, Sternberg MJ. 2003. The automatic discovery of structural principles describing protein fold space. *J. Mol. Biol.* 330:839–50
24. **Coulson AF, Moult J. 2002. A unifold, mesofold, and superfold model of protein fold use. *Proteins* 46:61–71**

24. Reviews folds grouped by the size of their sequence population and gives several possible explanations of these nonuniform distributions.

25. Csaba G, Birzele F, Zimmer R. 2009. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Struct. Biol.* 9:23
26. Daniels N, Kumar A, Cowen L, Menke M. 2010. Touring protein space with Matt. In *Bioinformatics Research and Applications*, ed. M Borodovsky, J Gogarten, T Przytycka, S Rajasekaran, pp. 18–28. Berlin/Heidelberg: Springer
27. Day R, Beck DAC, Armen RS, Daggett V. 2003. A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci.* 12:2150–60
28. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. 2001. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* 29:55–57
29. Dokholyan NV, Shakhnovich B, Shakhnovich EI. 2002. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl. Acad. Sci. USA* 99:14132–36
30. Dokholyan NV, Shakhnovich EI. 2001. Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* 312:289–307
31. Douguet D, Chen HC, Tovchigrechko A, Vakser IA. 2006. DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics* 22:2612–18
32. Dunbrack RL Jr. 2006. Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.* 16:374–84
33. Emmert-Streib F, Mushegian A. 2007. A topological algorithm for identification of structural domains of proteins. *BMC Bioinforma.* 8:237
34. Finkelstein AV, Badretdinov AY. 1997. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold. Des.* 2:115–21
35. Floudas CA, Fung HK, McAllister SR, Monnigmann M, Rajgaria R. 2006. Advances in protein structure prediction and de novo protein design: a review. *Chem. Eng. Sci.* 61:966–88
36. Forslund K, Henricson A, Hollich V, Sonnhammer EL. 2008. Domain tree-based analysis of protein architecture evolution. *Mol. Biol. Evol.* 25:254–64
37. Friedberg I, Godzik A. 2005. Connecting the protein structure universe by using sparse recurring fragments. *Structure* 13:1213–24
38. Friedberg I, Godzik A. 2005. Fragnostic: walking through protein structure space. *Nucleic Acids Res.* 33:W249–51
39. George RA, Spriggs RV, Thornton JM, Al-Lazikani B, Swindells MB. 2004. SCOPEC: a database of protein catalytic domains. *Bioinformatics* 20(Suppl. 1):i130–36

40. Getz G, Vendruscolo M, Sachs D, Domany E. 2002. Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. *Proteins Struct. Funct. Bioinforma.* 46:405–15
41. Glasner ME, Gerlt JA, Babbitt PC. 2006. Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.* 10:492–97
42. Goldstein RA. 2008. The structure of protein evolution and the evolution of protein structure. *Curr. Opin. Struct. Biol.* 18:170–77
43. Gough J. 2005. Convergent evolution of domain architectures (is rare). *Bioinformatics* 21:1464–71
44. Govindarajan S, Recabarren R, Goldstein RA. 1999. Estimating the total number of protein folds. *Proteins* 35:408–14
45. Grant A, Lee D, Orengo C. 2004. Progress towards mapping the universe of protein folds. *Genome Biol.* 5:107
46. Grigoriev IV, Zhang C, Kim SH. 2001. Sequence-based detection of distantly related proteins with the same fold. *Protein Eng.* 14:455–58
47. **Grishin NV. 2001. Fold change in evolution of protein structures. *J. Struct. Biol.* 134:167–85**
48. Gu J, Bourne PE. 2009. *Structural Bioinformatics*. Hoboken, NJ: Wiley-Blackwell. 1,035 pp.
49. Hadley C, Jones DT. 1999. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* 7:1099–112
50. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. 2002. Quantifying the similarities within fold space. *J. Mol. Biol.* 323:909–26
51. Holland TA, Veretnik S, Shindyalov IN, Bourne PE. 2006. Partitioning protein structures into domains: Why is it so difficult? *J. Mol. Biol.* 361:562–90
52. Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233:123–38
53. Holm L, Sander C. 1996. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* 24:206–9
54. Holm L, Sander C. 1998. Dictionary of recurrent domains in protein structures. *Proteins* 33:88–96
55. Hou Y, Hsu W, Lee ML, Bystroff C. 2003. Efficient remote homology detection using local structure. *Bioinformatics* 19:2294–301
56. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. 1999. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 27:254–56
57. James LC, Tawfik DS. 2003. Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* 28:361–68
58. Jefferson ER, Walsh TP, Roberts TJ, Barton GJ. 2007. SNAPPI-DB: a database and API of structures, interfaces and alignments for protein-protein interactions. *Nucleic Acids Res.* 35:D580–89
59. Joseph AP, Valadié H, Srinivasan N, de Brevern AG. 2012. Local structural differences in homologous proteins: specificities in different SCOP classes. *PLoS One* 7:e38805
60. Kihara D, Skolnick J. 2003. The PDB is a covering set of small protein structures. *J. Mol. Biol.* 334:793–802
61. Kinch LN, Grishin NV. 2002. Evolution of protein structures and functions. *Curr. Opin. Struct. Biol.* 12:400–8
62. Koehl P. 2001. Protein structure similarities. *Curr. Opin. Struct. Biol.* 11:348–53
63. Koehl P. 2006. Protein structure classification. *Rev. Comp. Chem.* 22:1–55
64. Kolodny R, Koehl P, Levitt M. 2005. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.* 346:1173–88
65. Kolodny R, Linial N. 2004. Approximate protein structural alignment in polynomial time. *Proc. Natl. Acad. Sci. USA* 101:12201–6
66. Kolodny R, Petrey D, Honig B. 2006. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Curr. Opin. Struct. Biol.* 16:393–98
67. Koonin EV, Tatusov RL, Galperin MY. 1998. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* 8:355–63
68. Koonin EV, Wolf YI, Karev GP. 2002. The structure of the protein universe and genome evolution. *Nature* 420:218–23

47. Reviews events in the course of evolution that can change a fold.

69. **Kosloff M, Kolodny R. 2008. Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins* 71:891–902**

69. Shows frequent pairs of proteins with similar sequence and dissimilar structure, which have an effect on hierarchical classifications based on similar sequences having similar structures.

70. Levitt M. 2007. Growth of novel protein structural data. *Proc. Natl. Acad. Sci. USA* 104:3183–88

71. **Levitt M, Chothia C. 1976. Structural patterns in globular proteins. *Nature* 261:552–58**

71. Identified the four classes of proteins, all-alpha, all-beta, alpha/beta, and alpha+beta, subsequently used in the SCOP classification.

72. Lo W-C, Lee C-C, Lee C-Y, Lyu P-C. 2009. CPDB: a database of circular permutation in proteins. *Nucleic Acids Res.* 37:D328–32

73. **Lupas AN, Ponting CP, Russell RB. 2001. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* 134:191–203**

73. Discusses the evolution of protein folds and how proteins may have evolved from peptide precursors.

74. McLachlan AD. 1987. Gene duplication and the origin of repetitive protein structures. *Cold Spring Harb. Symp. Quant. Biol.* 52:411–20

75. Melo F, Marti-Renom MA. 2006. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. *Proteins* 63:986–95

76. Menke M, Berger B, Cowen L. 2008. Matt: Local flexibility aids protein multiple structure alignment. *PLoS Comput. Biol.* 4:e10

77. Moreno-Hernández S, Levitt M. 2012. Comparative modeling and protein-like features of hydrophobic-polar models on a two-dimensional lattice. *Proteins* 80:1683–93

78. Murzin AG. 1998. How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* 8:380–87

79. Murzin AG. 2008. Metamorphic proteins. *Science* 320:1725–26

80. Murzin AG, Bateman A. 1997. Distant homology recognition using structural classification of proteins. *Proteins Suppl.* 1:105–12

81. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–40

82. Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–53

83. Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–34

84. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—a hierarchic classification of protein domain structures. *Structure* 5:1093–108

85. Orengo CA, Taylor WR. 1996. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* 266:617–35

86. **Osadchy M, Kolodny R. 2011. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proc. Natl. Acad. Sci. USA* 108:12301–6**

86. Characterizes the fundamental relationship between protein structure and protein function by considering all structural similarities between domains, instead of relying on a hierarchical classification.

87. Panchenko AR, Wolf YI, Panchenko LA, Madej T. 2005. Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins* 61:535–44

88. Paoli M. 2001. Protein folds propelled by diversity. *Prog. Biophys. Mol. Biol.* 76:103–30

89. **Pascual-García A, Abia D, Ortiz ÁR, Bastolla U. 2009. Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput. Biol.* 5:e1000331**

89. Shows how clustering techniques used in SCOP and CATH (single versus average linkage) influence classification.

90. Pawlowski K, Godzik A. 2001. Surface map comparison: studying function diversity of homologous proteins. *J. Mol. Biol.* 309:793–806

91. Pearson WR. 1996. Effective protein sequence comparison. *Methods Enzymol.* 266:227–58

92. Peng J, Xu J. 2011. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* 79(Suppl. 10):161–71

93. Ponting CP, Russell RR. 2002. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* 31:45–71

94. Przytycka T, Aurora R, Rose GD. 1999. A protein taxonomy based on secondary structure. *Nat. Struct. Biol.* 6:672–82

95. Qian J, Luscombe NM, Gerstein M. 2001. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* 313:673–81

96. Ranea JA, Sillero A, Thornton JM, Orengo CA. 2006. Protein superfamily evolution and the last universal common ancestor (LUCA). *J. Mol. Evol.* 63:513–25

97. Redfern O, Bennett C, Orengo C. 2004. Structure comparison and protein structure classifications. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, ed. MJ Dunn, LB Jorde, PFR Little, S Subramaniam. New York: Wiley

98. Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA. 2007. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput. Biol.* 3:e232

99. Rogen P, Fain B. 2003. Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci. USA* 100:119–24

100. Sadowski MI, Taylor WR. 2010. On the evolutionary origins of "Fold Space Continuity": a study of topological convergence and divergence in mixed alpha-beta domains. *J. Struct. Biol.* 172:244–52

101. Sadreyev RI, Kim BH, Grishin NV. 2009. Discrete-continuous duality of protein structure space. *Curr. Opin. Struct. Biol.* 19:321–28

102. Sam V, Tai C-H, Garnier J, Gibrat J-F, Lee B, Munson P. 2008. Towards an automatic classification of protein structural domains based on structural similarity. *BMC Bioinforma.* 9:74

103. Samson AO, Levitt M. 2011. Normal modes of prion proteins: from native to infectious particle. *Biochemistry* 50:2243–48

104. Schaeffer RD, Daggett V. 2011. Protein folds and protein folding. *Protein Eng. Des. Sel.* 24:11–19

105. Schaeffer RD, Jonsson AL, Simms AM, Daggett V. 2011. Generation of a consensus protein domain dictionary. *Bioinformatics* 27:46–54

106. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. 2011. Protein disorder—a breakthrough invention of evolution? *Curr. Opin. Struct. Biol.* 21:412–18

107. Selkoe DJ. 2003. Folding proteins in fatal ways. *Nature* 426:900–4

108. Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI. 2003. Functional fingerprints of folds: evidence for correlated structure-function evolution. *J. Mol. Biol.* 326:1–9

109. Shi J, Blundell TL, Mizuguchi K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310:243–57

110. Shindyalov IN, Bourne PE. 2000. An alternative view of protein fold space. *Proteins* 38:247–60

111. Shortle D. 2009. One sequence plus one mutation equals two folds. *Proc. Natl. Acad. Sci. USA* 106:21011–12

112. Sierk M, Kleywegt G. 2004. Deja vu all over again: finding and analyzing protein structure similarities. *Structure* 12:2103–11

113. Sikosek T, Bornberg-Bauer E, Chan HS. 2012. Evolutionary dynamics on protein bi-stability landscapes can potentially resolve adaptive conflicts. *PLOS Comput. Biol.* 8(9):e1002659

114. Sippl MJ. 2008. On distance and similarity in fold space. *Bioinformatics* 24:872–73

115. Sippl MJ. 2009. Fold space unlimited. *Curr. Opin. Struct. Biol.* 19:312–20

116. Sippl MJ, Suhrer SJ, Gruber M, Wiederstein M. 2008. A discrete view on fold space. *Bioinformatics* 24:870–71

117. Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195–97

118. Söding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–60

119. Söding J, Lupas AN. 2003. More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays* 25:837–46

120. Stebbings LA, Mizuguchi K. 2004. HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res.* 32:D203–7

121. Subbiah S, Laurents DV, Levitt M. 1993. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* 3:141–48

122. Suhrer SJ, Wiederstein M, Gruber M, Sippl MJ. 2009. COPS—a novel workbench for explorations in fold space. *Nucleic Acids Res.* 37:W539–44

123. Taylor WR. 2006. Topological accessibility shows a distinct asymmetry in the folds of βαproteins. *FEBS Lett.* 580:5263–67

124. Taylor WR. 2007. Evolutionary transitions in protein fold space. *Curr. Opin. Struct. Biol.* 17:354–61

125. Thangudu RR, Sharma P, Srinivasan N, Offmann B. 2007. Analycys: a database for conservation and conformation of disulphide bonds in homologous protein domains. *Proteins* 67:255–61

114. Suggests a measure for distance between protein structures that is metric and thus well-suited in clustering of all protein structures.

126. Thornton JM. 2001. From genome to function. *Science* 292:2095–97

127. Tress ML, Ezkurdia I, Richardson JS. 2009. Target domain definition and classification in CASP8. *Proteins: Struct. Funct. Bioinforma.* 77:10–17

128. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN. 2004. Toward consistent assignment of structural domains in proteins. *J. Mol. Biol.* 339:647–78

129. Vogel C, Morea V. 2006. Duplication, divergence and formation of novel protein topologies. *BioEssays* 28:973–78

130. Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, et al. 2007. Towards fully automated structure-based function prediction in structural genomics: a case study. *J. Mol. Biol.* 367:1511–22

131. Wernisch L, Wodak SJ. 2005. Identifying structural domains in proteins. In *Structural Bioinformatics*, pp. 365–85. Hoboken, NJ: Wiley

132. Winstanley HF, Abeln S, Deane CM. 2005. How old is your fold? *Bioinformatics* 21:449–58

133. Winter C, Henschel A, Kim WK, Schroeder M. 2006. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.* 34:D310–14

134. Xia Y, Levitt M. 2004. Simulating protein evolution in sequence and structure space. *Curr. Opin. Struct. Biol.* 14:202–7

135. Xu Y, Xu D, Gabow HN. 2000. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* 16:1091–104

136. Yadid I, Kirshenbaum N, Sharon M, Dym O, Tawfik DS. 2010. Metamorphic proteins mediate evolutionary transitions of structure. *Proc. Natl. Acad. Sci. USA* 107:7287–92

137. Yang AS, Honig B. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* 301:665–78

138. Yang S, Bourne PE. 2009. The evolutionary history of protein domains viewed by species phylogeny. *PLoS One* 4:e8378

139. Ye Y, Godzik A. 2004. Comparative analysis of protein domain organization. *Genome Res.* 14:343–53

140. Yeats C, Redfern OC, Orengo C. 2010. A fast and automated solution for accurately resolving protein domain architectures. *Bioinformatics* 26:745–51

141. Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2006. Physical origins of protein superfamilies. *J. Mol. Biol.* 357:1335–43

142. Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33:2302–9

143. Zhou H, Xue B, Zhou Y. 2007. DDOMAIN: Dividing structures into domains using a normalized domain–domain interaction profile. *Protein Sci.* 16:947–55

Annual Review of
Biophysics

Volume 42, 2013

Contents

Index

Errata

An online log of corrections to *Annual Review of Biophysics* articles may be found at http://biophys.annualreviews.org/errata.shtml