Received: 6 August 2019 | Revised: 30 September 2019 | Accepted: 1 October 2019

DOI: 10.1002/pro.3742

FOR THE RECORD

The *Urfold*: Structural similarity just above the superfold level?

Cameron Mura[1] | **Stella Veretnik**[1] | **Philip E. Bourne**[1,2]

[1]Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia

[2]School of Data Science, University of Virginia, Charlottesville, Virginia

Correspondence
Philip E. Bourne, Department of Biomedical Engineering, University of Virginia; Charlottesville, VA 22908.
Email: peb6a@virginia.edu

Funding information
Division of Molecular and Cellular Biosciences, Grant/Award Number: 1350957; University of Virginia

Abstract

We suspect that there is a level of granularity of protein structure intermediate between the classical levels of "architecture" and "topology," as reflected in such phenomena as extensive three-dimensional structural similarity above the level of (super)folds. Here, we examine this notion of architectural identity despite topological variability, starting with a concept that we call the "*Urfold*." We believe that this model could offer a new conceptual approach for protein structural analysis and classification: indeed, the Urfold concept may help reconcile various phenomena that have been frequently recognized or debated for years, such as the precise meaning of "significant" structural overlap and the degree of continuity of fold space. More broadly, the role of structural similarity in sequence↔structure↔function evolution has been studied via many models over the years; by addressing a conceptual gap that we believe exists between the architecture and topology levels of structural classification schemes, the Urfold eventually may help synthesize these models into a generalized, consistent framework. Here, we begin by qualitatively introducing the concept.

KEYWORDS

architecture, fold space, molecular evolution, protein structure classification, secondary structure, superfold, topology, β-sheet

1 | INTRODUCTION

A deep challenge in molecular evolution concerns the development of a robust, quantitative, and lucid model for protein structural evolution, capable of affording insight into both the physicochemical and biological (functional) facets underlying various evolutionary mechanisms and processes.[1,2] A central pillar in this area is the concept of a protein "fold." Though widely invoked, the notion of a fold does not have a clear quantitative foundation,[3] and often a given protein cannot be unambiguously assigned to one fold versus another.[4] Here, we follow Orengo and colleagues[5] in considering a fold to be the "*global arrangement of the main secondary structural elements (SSEs), in terms of their relative orientations (architecture) and patterns of connectivity (topology)*." The space of all folds (known and unknown) can be conceptually organized in at least three distinct ways: (a) using discrete, hierarchical classification schemes, with greater levels of similarity between entities (folds or individual three-dimensional [3D] structures within a given fold class) that occupy lower (more finely detailed) classification levels[6]; (b) as acyclic graphs, with vertices denoting folds and edges representing structural similarity between two folds[7]; and (c) as dendrograms, wherein proteins with similar SSEs are neighboring leaves in these taxonomic trees.[8] The first approach is taken by the well-known structure classification schemes FSSP,[9] SCOP,[10] CATH,[11] and ECOD.[12] While these various systems differ in their methodological approaches and underlying assumptions, their top levels always consist of very generic classes (e.g., all-α, α/β) and, nearer the bottom levels, folds

Abbreviations: FS, fold space; PSS, protein structure space; SBB, small β-barrel; SSE, secondary structural element.

Cameron Mura and Stella Veretnik contributed equally.

become partitioned into families that exhibit sufficiently strong sequence similarity to indicate homology within the family (i.e., clear evolutionary relatedness).

It has been noted multiple times that hierarchical classification schemes—while useful in conceptualizing and organizing protein structure space (PSS)*—unavoidably miss significant relationships between disparate folds (e.g., ref [13]) and also depend on whether the continuity of fold space is considered.[14] Interfold similarities, including those which are missed, stem from geometric similarities of structural motifs within the folds.[13–17] Claims as to (a) the extent of structural overlap between two otherwise disparate folds (i.e., the characteristic size of the structural motifs), (b) any conclusions regarding their origin (e.g., convergent vs. divergent evolution), and (c) the potential functional significance of such motifs vary greatly in the literature.[13,18–20] While a detailed and comprehensive treatment of that topic is beyond the scope of this Note, inter-fold relationships clearly exist, and fold space (FS) can be viewed as rather continuous.[13–19,21,22] In the network view of FS, the degree of connectivity between folds varies, often depending on the precise computational methods. For example, the α/β region of FS appears to be highly interlinked,[4,7,22] and the all–α-helical region may show more connections than other regions[21]; simultaneously, others have found similar levels of interconnectivity within FS.[16] Though not always the case, in many instances, one can reach fold $\mathcal{B}$ from fold $\mathcal{A}$ by a sequence of smooth, continuous deformations, $\mathcal{A} \to \mathcal{A}' \to \mathcal{A}'' \to \cdots \to \mathcal{B}$.[23] Thus, a more accurate model will not binarily classify folds $\mathcal{A}$ and $\mathcal{B}$ as either identical or nonidentical but rather by their degree of similarity, as one can almost always find a structural relationship between two distinct folds; a similar point has been made by Sippl.[24] Though there may seem to be a natural tension between the continuous versus mostly-discrete views of FS (the latter of which is implicitly taken by all the predominant classification approaches), this need not be the case: as lucidly described in Sadreyev et al.,[25] these are two sides of the same coin, and discrepancies and distinctions chiefly arise from the application of fixed numerical thresholds (of similarity).

2 | THE URFOLD CONCEPT

Several properties of FS, such as the above continuous/discrete dichotomy, motivate us to propose the existence of a level of structural organization that we term the *Urfold*. First, network representations of FS feature highly interconnected nodes that are bridges or hubs. Such hubs have been proposed to contain (sub)structures that are common to many different folds.[4,16,17] Depending on the threshold of structural overlap, the degree of interconnectedness between distinct folds can range from dense to sparse. Second, a highly skewed distribution of folds—in terms of their population by known 3D structures—was first observed long ago,[6,26] and a power-law trend has persisted after many more observations (e.g., post-structural genomics): more than 1,300 folds (as defined by CATH) are currently known, and 10 of these accounts for 50% of all known domain structures. These enriched folds, termed *superfolds*,[26,27] can be viewed as dense "attractors"[28] in FS. The 3D structural arrangements of SSEs in such superfolds are thought to be particularly stable (thermodynamically) and mechanistically readily accessible (vis-à-vis folding kinetics), leading to an unusually broad sequence space capable of adopting these folds; these features, in turn, account for the vast functional diversification within superfolds. Third, a striking jump in the populations of two adjacent layers of structural granularity (Figure 1a,b) has been consistently observed in hierarchical classification schemes, whereby relatively few groups expand into a disproportionately large number of entities at the next-finer level (Figure 1d). In CATH, the jump occurs between the *Architecture* and *Topology* levels (41 Architectures ↢ 1,391 Topologies); in SCOP, it occurs between *Classes* and *Folds* (4 Classes ↢ 1,232 Folds); and in ECOD, the jump is between *Architectures* and *X-groups* (20 Architectures ↢ 2,247 X-groups).†

We suspect that the three phenomena outlined above are interrelated, pointing to the existence of a bonafide new grouping that lies above the topological level of structural organization but below the architectural level; this is a level of structural granularity that we believe has been hitherto neglected. We introduce the term Urfold‡ to describe such an entity—an aggregation, collection, or "grouping" near the architectural level (Figure 1a,b). The Urfold can be viewed as capturing 3D architectural similarity despite topological variability (Figure 2). As such, it is a coherent, topology-independent structural unit that likely reflects 3D arrangements of SSEs that are particularly favorable (likely for geometric or physicochemical reasons). In other words,

[*] We generally use the phrase PSS to refer to the set of all protein 3D structures, both known and unknown. We do not consider this strictly equivalent to the somewhat less precise phrase "fold space," though we do occasionally use them interchangeably. In such instances, we do so knowingly—i.e., our usage of PSS and fold space (FS) as synonyms, in certain cases, means that we do not intend to distinguish between these two subtly different concepts for the purposes of the argument at hand.

[†] These statistics were gathered in early 2019 from the respective website of each structural database.

[‡] We chose the term *Urfold* because the prefix "*ur-*" indicates "*primitive*," "*ancestral*," or "*one step higher in scope*" (http://en.wiktionary.org/wiki/ur-). Prior to broad adoption of the term "*domain*" to refer to the Archaeal, Bacterial, and Eukaryal domains of life, Woese and Fox referred to the highest level taxonomic rank as an "*urkingdom*." In terms of the granularity of protein structural classification levels, the Urfold is one "step" above the fold, and yet, it is distinct from the concept of a superfold (Figure 1).

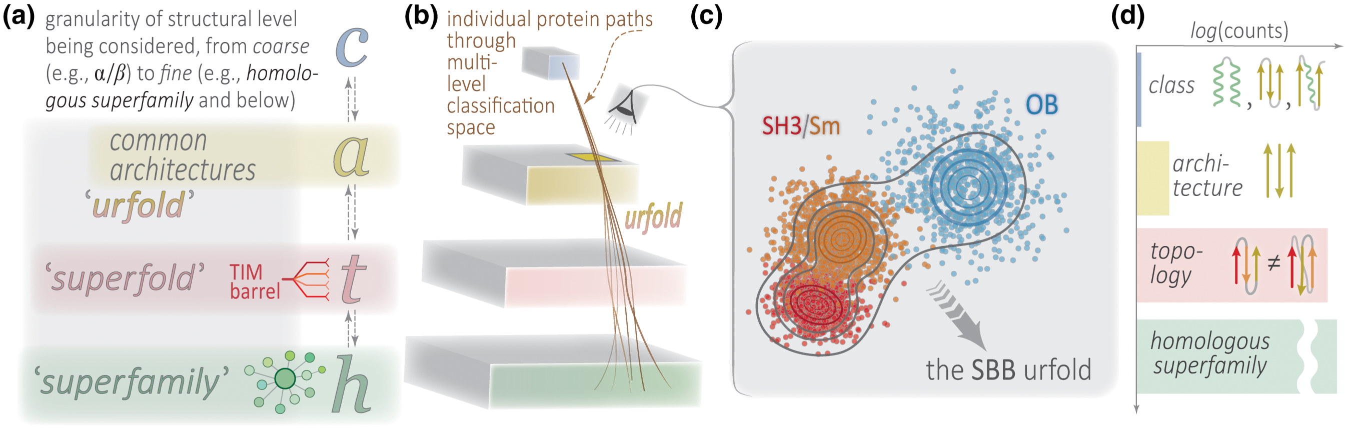
(a) granularity of structural level being considered, from coarse (e.g., α/β) to fine (e.g., homologous superfamily and below)
c
common architectures
a
'urfold'
'superfold'
TIM barrel
t
'superfamily'
h
(b) individual protein paths through multi-level classification space
urfold
(c)
SH3/Sm
OB
the SBB urfold
(d)
log(counts)
class
archi-tecture
topo-logy
≠
homologous superfamily

FIGURE 1 Schematic representation of the Urfold concept, with respect to protein structure space. This diagram sketches the granularity of structural levels that are typically considered (a), ranging from coarsest (e.g., "α/β class") to finer levels (e.g., "homologous superfamily" and below). Note that the terms used here (*class*, *architecture*, etc.) closely align with the usage in systems such as CATH, but they are not necessarily identical (the "*c*," "*a*," etc. in panel a are lowercase for this reason—we do not mean to imply, simply by using these terms, that the present work strictly adheres to any particular classification scheme). The exact position of the Urfold, between the topology (red) and architecture (yellow) levels, is currently indeterminate. These conceptual terms are elaborated in (b) and (c). Panel (b) shows the relationships, in terms of a hierarchical concept map or ontology, between (a) the various conceptual levels of protein structural entities found in most hierarchical classification systems (class, architecture, topology, etc.), in the vertical direction, and (b) the grouping or "aggregation" function served by such terms as "superfamily" and "superfold" (and, now, "urfold") represented in the mostly horizontal direction (semitransparent slabs, color matched to panel a). The "eye" icon in (b) gazes down (and through) the yellow slab, representing entities at the *architecture* level, whereupon we see a set of architecturally identical protein folds (SH3/Sm, OB, etc.) that can be grouped into the small β-barrel (SBB) Urfold in (c); here, contour lines represent different thresholds, or stringencies, of clustering discrete entities at that given level along the structural classification hierarchy (the concept planes/slabs). In a sense, the Urfold concept is to the architecture level as the superfold concept is to the topology(/fold) level. The histogram in (d) roughly indicates the relative populations of these structural levels. A noticeable jump occurs between the upper levels in most classification schemes (CATH, SCOP, ECOD), and we suggest that the Urfold corresponds to structural entities lying within the architecture ↭ topology gap

the same arrangement of SSEs in 3D space can be readily achieved via different arrangements of SSEs along a protein sequence. Belonging to a given Urfold neither requires sequential contiguity or identical order of structural elements (see, e.g., the OB vs. SH3/Sm topologies in fig. 3 of Reference 29), nor does it preclude strand reversal,[23] as illustrated here by the K Homology (KH) domain (Figure 2b). Taken even further, some degree of "mismatch" between the types of aligned SSEs may be allowed[§]: such variation has been detected in the fold change of homologous proteins[23] and presumably stems from the capacity to achieve similar packings of compact, hydrogen-bonded SSEs.[30]

3 | EXAMPLES OF PUTATIVE URFOLDS

Relatively simple and more intricate examples of putative urfolds are illustrated by the P-loop NTPases and the KH domain, shown in Figures 2a and 2b, respectively. The adenylate kinase and RecA catalytic domains in Figure 2a are architecturally similar, and they are also topologically equivalent under a simple strand re-ordering (in 3D, within the sheet). Thus, this is a conceptually straightforward example of the "same architecture, different topology" principle. Next, if we allow (a) strand reorderings, (b) more severe reordering of SSEs (sequence-level swapping of α ⇄ β elements), and (c) reversal of strand directions in 3D (so ↿↾ and ↿⇂ are taken as equivalent), then the KH domains of hnRNP K and ribosomal protein S3 (RPS3) coalesce into a single Urfold, shown in Figure 2b. An intriguing example wherein greater topological variation does not correspond to more 3D architectural variation is shown by the series of proteins in Figure 2c, all of which build upon a fully antiparallel four-stranded β-sheet: 1M5Q contains the C-term domain of an archaeal Sm protein (SmAP3),[31] 1XXA is the C-term region of the DNA-binding arginine repressor,[32] 1VHH is the N-term signaling domain of Sonic hedgehog,[33] and 1ELO is a domain from the elongation factor G translocase.[34] The 1M5Q → {1XXA, 1VHH} → 1ELO progression, schematized in the rightmost panel of Figure 2c, shows that the same architecture can persist despite increasingly severe topological changes (apart from swapping the location [in sequence] of the helices in the ⌇$_{1/2}$ pair, the 1XXA and 1VHH structures are topologically equivalent).

[§] A mild mismatch would be, for example, not distinguishing between a 3_{10}– and α-helix; more severe would be to treat a helix and strand as interchangeable.

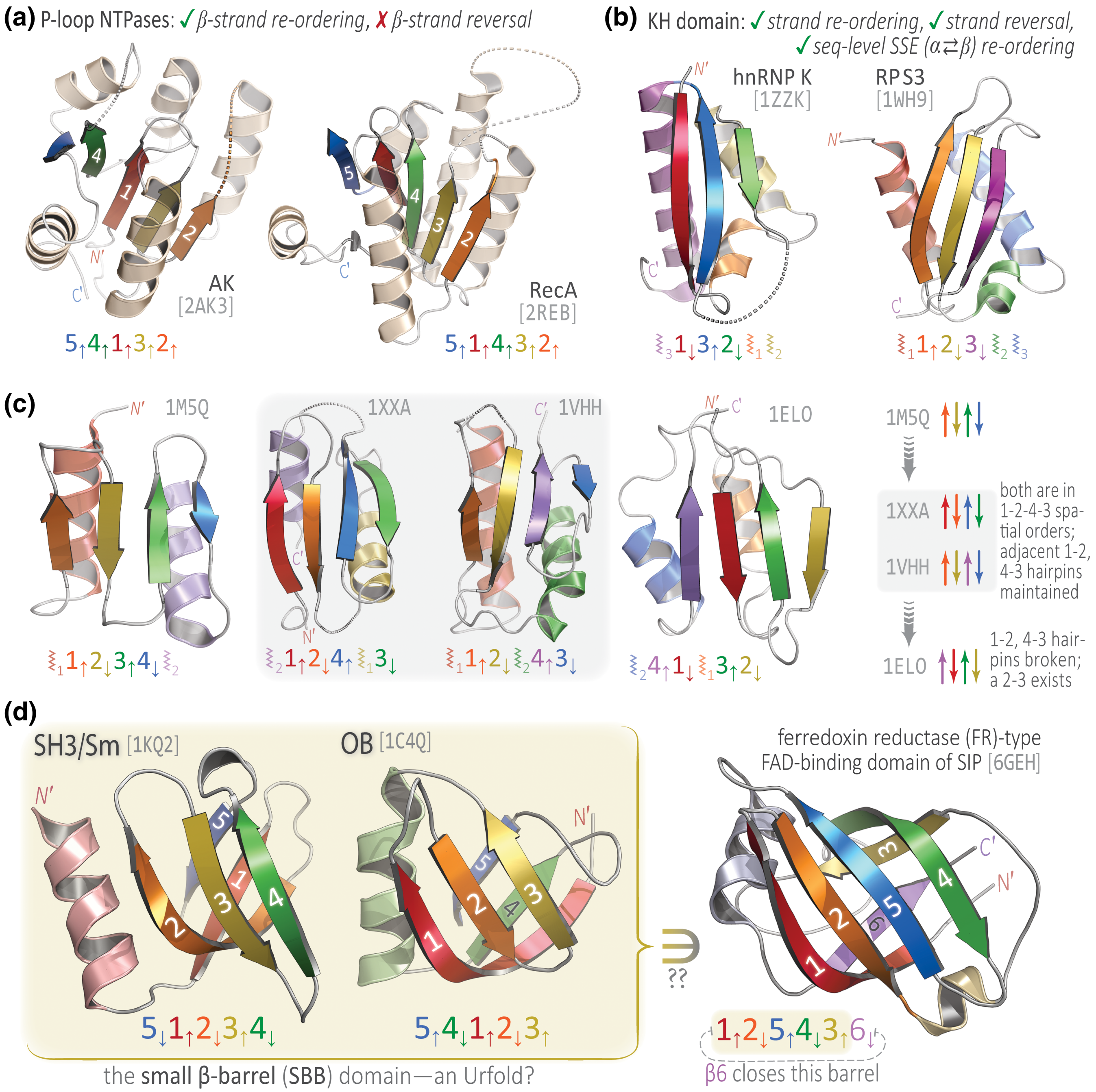
(a) P-loop NTPases: ✓ β-strand re-ordering, ✗ β-strand reversal
AK
[2AK3]
N′
C′
1
2
4
5↑4↑1↑3↑2↑
RecA
[2REB]
5
4
3
2
C′
5↑1↑4↑3↑2↑
(b) KH domain: ✓ strand re-ordering, ✓ strand reversal,
✓ seq-level SSE (α⇄β) re-ordering
hnRNP K
[1ZZK]
N′
C′
≀3 1↓3↑2↓ ≀1 ≀2
RPS3
[1WH9]
N′
C′
≀1 1↑2↓3↓ ≀2 ≀3
(c)
1M5Q
N′
≀1 1↑2↓3↑4↓ ≀2
1XXA
N′
C′
≀2 1↑2↓4↑ ≀1 3↓
1VHH
C′
≀1 1↑2↓ ≀2 4↑3↓
1ELO
N′ C′
≀2 4↑1↓ ≀1 3↑2↓
1M5Q
1XXA
1VHH
both are in 1-2-4-3 spatial orders; adjacent 1-2, 4-3 hairpins maintained
1ELO
1-2, 4-3 hairpins broken; a 2-3 exists
(d)
SH3/Sm [1KQ2]
N′
5
1
2
3
4
5↓1↑2↓3↑4↓
OB [1C4Q]
N′
5
1
2
4
3
5↑4↓1↑2↓3↑
the small β-barrel (SBB) domain—an Urfold?
⊇
??
ferredoxin reductase (FR)-type
FAD-binding domain of SIP [6GEH]
C′
N′
1
2
3
4
5
6
1↑2↓5↑4↓3↑6↓
β6 closes this barrel

FIGURE 2 Some examples of putative Urfolds and analyses thereof. Many protein structures exhibit architectural similarity despite topological variability, irrespective of considerations of homology—a principle we term the *Urfold*. This concept is illustrated here using (within each panel) two or more examples of distinct folds that adopt equivalent architectures, suggesting them as putative Urfolds. All 3D structures are shown as cartoon ribbon diagrams, and PDB codes are indicated near each structure (light-gray). The N' and C'-termini are marked in most cases (space permitting), and individual SSEs are color ramped from $N' \rightarrow C'$ along the visible spectrum (red → orange → yellow → ···). The helices are of secondary importance for the immediate purposes of (a) and (d); so in those two panels their color is either light-tan (a) or a hue that is intermediate between the adjoining strands (d). Also in (a) and (d), individual β-strand numbers appear on the cartoons. The strand layout for each β-sheet is diagrammed underneath each representation, for example, as $5_{\downarrow}1_{\uparrow}2_{\downarrow}3_{\uparrow}4_{\downarrow}$, for the SH3/Sm superfold in (d). For cases wherein we consider the helices to have a pivotal role in defining a particular Urfold (i.e., panels b and c), these schematic diagrams are used to also indicate the approximate location of each helix, for example, the "$\wr_3 1_{\downarrow} 3_{\uparrow} \cdots$," for the KH domain of hnRNP K in (b). In general, the coloring and diagrammatic schemes are intended to expose the nature of the equivalencies and other mappings between the salient SSEs. Further descriptions of these putative Urfold examples are provided in the text

How much can a pair of structures vary and still be part of the same Urfold? (How stringently do we delimit folds from one another, when collecting them into Urfolds?) The type of progression shown in Figure 2c helps elucidate these questions by showing the relationships (equivalencies, alterations) between individual folds within a single putative Urfold. In a decision tree–based approach to systematizing β-structures,[29] these four proteins form a natural progression, with 1ELO more distant from 1M5Q than are 1XXA and 1VHH. Somewhat similar in spirit, toy models could be used together with a machine learning–based fold classifier to examine the relationship between classification results and systematically varied geometric descriptors, such as the crossover angle between helices (e.g., as part of a helix-turn-helix motif or, more distant in sequence, as tertiary contact sites).

Finally, Figure 2d illustrates the small β-barrel (SBB) domain, which we propose is an Urfold that aggregates the SH3/Sm and OB superfolds.[29] The SBB spurs the question of whether a particular Urfold might be part of a larger structural unit (e.g., a large-sized domain)? For example, the ferredoxin reductase (FR)–like fold, found within a recent structure of a siderophore-interacting protein (SIP; 6GEH), bears "a certain resemblance"[23] to the SBB, as shown in Figure 2d. Did the FR-like fold evolve by being built incrementally from an SBB Urfold core, via addition of the β6 strand (an idea bolstered by the fact that other examples exist of an SBB augmented with a sixth strand, such as the RNase P subunit of Rpp29 mentioned in Reference 29)? At this stage, alluring possibilities such as this are intended as more predictive and conjectural, not conclusive.

4 | THE URFOLD IN CONTEXT: DOMAINS AND GREGARIOUSNESS

In formulating the Urfold, the size of the structural unit being considered for comparison, grouping, and so forth is crucial, as it defines the extent of the similarity,[24] and hence the extent of connectivity among folds (viz., the discrete ↔ continuous FS extrema). Folds are generally viewed as corresponding to the level of structural domains,[6,35] though even for the smallest of folds many subtle and intertwined signals can be detected, such as covariation of amino acid residues that are distant in sequence but near in space.[17] These signals are presumably evolutionary echoes of the physicochemical interactions that stabilize a fold, integrated over millions to billions of years; thus, it may be feasible to detect subtle similarities in patterns within covariance matrices for subsets of proteins lying within a given Urfold (via, e.g., the evolutionary couplings approach). As envisaged here, the Urfold can be a full domain, most likely of relatively small size (e.g., the SBB of Reference 29), or it may comprise a significant fraction of the structural "core" of a larger sized domain (e.g., the β-grasp in the work of Shi et al.[36]).

The Urfold concept closely relates to the "gregariousness" quantity, defined by Harrison et al.[4] to measure the structural overlap amongst different folds. While gregariousness is a property that can be computed for any type of fold, its utility in defining what *is* a fold (characteristic sizes, recurring spatial patterns of SSEs) has not been systematically explored across the PSS. We suspect that highly gregarious folds are archetypal Urfolds. Given that, an Urfold differs from a highly gregarious fold insofar as the structural entity is defined less rigidly—we allow for strand reversals, rearrangements in the order of SSEs, and even some level of mismatch between SSEs (see above and Figure 2). At one extreme, a free-standing helix or β-strand (or even β-hairpin) is too small to be an Urfold, and in the other limit, a two-domain protein is too large. Between these two extremes, there are "motifs" of SSEs that have been found to recur in certain folds, and many of these are rather more "gregarious" than others. The key point is that any two entities within the same Urfold have a shared 3D architecture. In terms of minimal size requirements, note that we define an Urfold as larger than typical "structural motifs" (ref 37 is an early example of this terminology), which range from several residues (e.g., P-loop, Zn-finger, Asp box[38]) to two or three SSEs (e.g., a helix-turn-helix motif[39]). When part of a larger domain, we require an Urfold to be central to the structural core (versus, e.g., a peripheral element or other "decoration," in the sense of examples in the work of Youkharibache et al.[29]).

The architectural similarity of SSEs that is the hallmark of an Urfold ultimately stems from the purely physicochemical properties of a given protein sequence, subject to statistical mechanical sampling.[21] From this perspective, the spatial arrangement of SSEs that defines a particular Urfold also governs the overall (thermodynamic) stability of any of the particular folds that belong to that Urfold. Because the Urfold is agnostic of the specific connectivity of SSEs (i.e., is topology independent), in general, there would exist a range of thermodynamic stabilities ($\Delta G^{\circ}_{\mathrm{fold}}$) among the individual folds that comprise an Urfold. In terms of folding kinetics, note that efficient folding of a 3D structure correlates with the sequential proximity of SSEs (at least for the folding nucleus[40,41]); however, even the folding nucleus can consist of SSEs that are non-contiguous in sequence.[42]

5 | THE URFOLD AND STRUCTURAL CLASSIFICATION SYSTEMS

The Urfold relaxes the constraint of identical topologies (at least partially) while still requiring the spatial arrangements of SSEs between two folds (that are members of the same urfold) to at least roughly match (Figure 2). Thus, in terms of structural hierarchies, it lies above topology (i.e., fold) but somewhat below the level of architecture, at least as usually defined. Closely related to this, note that the "architecture," at least as operationally defined in structural classification systems, is rather generic. For this reason, we find low numbers of such entities in CATH (46 architectures) and ECOD (20 architectures), relative to the number of distinct topologies (1,391 in CATH and 2,247 in ECOD); the "architecture" concept does not explicitly appear in SCOP.

We propose that the number of entities at the Urfold level smoothly bridges the jump that can be empirically seen in the populations of the architecture and topology/fold levels (Figure 1d). In terms of network representations of fold space, we suspect that Urfolds will generally correspond to "hub" regions, with high degrees of connectivity linking them to numerous discrete folds that are one level lower (Figure 1b; "lower" in an analogous sense as reticulated networks being a generalization of phylogenetic trees[43]). From the perspective of structural classification systems, we suspect that applying the Urfold concept would yield a reorganization of population distributions in existing classification levels (in CATH and SCOP). This might occur in a manner similar to ECOD, where disparate folds (or superfamilies) often coalesce for reasons related to an underlying sequence similarity, yielding new categories (groupings) not observed in other classification schemes.[12] However, note that the conceptual underpinning of the Urfold is actually disjoint from that of ECOD: while inferred homology is central to ECOD's classification scheme, the Urfold is agnostic of homology. Rather, an Urfold is inferred mostly on the basis of recurrent (and thus presumably favorable) spatial arrangements of SSEs, which, in turn, are governed by physicochemical principles (and evolutionary principles only implicitly, over far longer timescales, as captured by approaches such as evolutionary couplings[44]).

New levels of protein structural classification have been suggested before. For example, a "metafold"[45] was proposed to address clear cases of homology among disparate folds (a motivation shared by the ECOD system as shown in Reference 12). Interestingly, the Urfold concept does relate to that of the metafold, but the Urfold is more generic, as it does not rely upon inferred evolutionary relationships among structures. The concept of "hyperfamilies," representing yet another level of protein structural classification, was proposed[4] to account for possibly significant structural overlap between Homologous superfamilies that belong to different Topologies in CATH (i.e., the gregariousness concept). The Urfold relates to, but is not identical to, these other conceptualizations of protein folds and structural classes.

The Urfold concept was initially motivated by our discovery[29] that two distinct superfolds, namely, the SH3 and OB, exhibit extensive structural and functional similarities, yet have distinct topologies that are not equivalent under circular permutations or other rearrangements (strand invasion, strand swaps, deletions) that have been described as permissible for homologous proteins.[23,45] In fact, in the CATH system, the SH3 and OB domains even belong to two distinct architectures (2.40.50 [OB] and 2.30.30 [SH3]). The striking 3D structural similarity among these seemingly unrelated proteins was initially detected visually, by multiple independent human experts (see also Reference 46). Along with 10 additional folds that have similar overall architectures, we recently termed these superfolds the "small β-barrel" (SBB) domain.[29] The sequence similarities among members of each fold within the SBB urfold (as well as between the SH3 and OB folds) are often minimal (below the twilight zone), perhaps because of both homologous and analogous relationships between the individual entities. Indeed, such a confounding mixture of effects—one largely evolutionary (homology, divergent) and the other more physicochemical (analogy, convergent)—might hold even within the SH3 superfold itself.[47] As presented here (Figure 2d), the SBB is an archetypal Urfold: a grouping of folds with (a) the same architecture, *broadly defined* (i.e., not necessarily or strictly mapping to identical Architecture levels in CATH), (b) potentially differing topologies, and (c) perhaps some telling functional similarities (*potentially* indicative of homology). For instance, the SH3/Sm and OB folds both function extensively in nucleic acid metabolic pathways.[29]

Cases similar to that described above for the SBB can be found with other folds. For example, we posit that the various topological organizations of barrels that have been grouped under the umbrella term "cradle-loop barrel" metafold[45] comprise an Urfold, the members of which span 13 different topologies, four architectures and even two different classes in CATH (see table 1 in Reference 45). Other notable examples (Figure 2) involve (a) the KH domains, which occur as two different topologies[48]; (b) the β-grasp domain, which exists as a separate domain or embedded within a larger context[36,49]; and (c) the P-loop NTPases and Rossmann-like motif, which is detected in over 20% of all structures and even in multiple different folds.[50]

6 | CONCLUSIONS, OUTLOOK

Most known cases of topologically permuted folds have been discovered via sequence similarity.[23,45,51] Such instances of different folds—with similar architectures and clear evidence of homology, yet distinct topologies—can serve as helpful starting points in developing approaches to identify cases of similar architecture which do *not* show clear sequence or topological relationships (essentially, they could serve as true positives). In formulating such an approach, some conceivable parameters to consider include (a) the minimal size of an Urfold (number of

SSEs, total number of residues); (b) stringency levels for alignment of SSEs/backbones (e.g., related to the above example of a helical crossing angle); (c) the extent of topological variability allowed among the folds that comprise a single, well-defined Urfold (SSEs that belong to the folding nucleus likely will be contiguous in sequence, as noted for the SBB,[29] although the rest of the architecture for a given Urfold might be arranged around that core in topologically different ways); (d) the degree to which different types of SSEs are allowed to count as a "match" (a hallmark of "homologous fold change"[23,47]); and (e) any further thresholds that might be imposed on the minimal structural contribution to the core.

Assuming the above plan is realized—i.e., that effective parameter sets are found—we can then ask: Does the Urfold concept enable exploration and discovery of any new features of protein structure space? For example, (a) how frequently does an Urfold constitute an entire domain, and how often is an Urfold embedded in a larger structure (i.e., below the level of structural domain)? And, are there any recurrent characteristics of an Urfold in the context of larger domains? (b) Are there prevalent 3D spatial arrangements of protein backbones in Urfolds? If so, do these arise mostly from interactions among SSEs and super-SSEs that are local in sequence, as has been detected in earlier studies[26,52,53] or are such SSEs equally likely to come from noncontiguous regions[42]? (c) Are Urfolds more often associated with known superfolds than with other folds? (d) What are the connectivity properties of fold space, assuming distinct Urfolds? (e) Where precisely do Urfolds sit, in terms of granularity level (Figure 1b) in classification schemes such as CATH, SCOP, and ECOD? A key issue that relates to each of the above questions will be how robust are the characteristics and properties of FS (points a → e), under varying definitions of the Urfold (points a → e of the preceding paragraph).

We propose the Urfold as a distinct type of entity, akin to "*the fold*," but capturing more general (and basic) physicochemical principles that underlie protein structure and function. Computationally detecting and systematically identifying urfolds will enable a new approach to explore the organization of protein structure space, particularly at the relatively coarse and intermediate levels of architecture and topology/fold. Such studies could, in turn, offer a new conceptual platform for deepening our understanding of protein structure, in terms of fundamental physical principles as well as potential evolutionary relationships—and, most significantly, the interplay between these two fundamentally different approaches in protein science.[1,2]

Finally, note that the Urfold raises some deep questions regarding our conceptual models of PSS, including (a) the development of a more precise, quantitative, and computable definition of the Urfold; (b) implementation of this definition and systematic application to all known 3D structures; and (c) elucidation of the impact of Urfold-level entities on the relationships among these known structures—for example, are classification schemes such as CATH, SCOP, and ECOD altered by allowing for an Urfold entity? (If so, how?) These basic problems offer intriguing directions and quantitative challenges for further investigation.

ACKNOWLEDGMENTS

We thank J. Bushweller, E. Draizen and G. Lipkowitz for feedback, and K. Stanek for bringing the SIP structure to our attention. Portions of this work were supported by the University of Virginia and NSF Career award MCB-1350957.

ORCID

Cameron Mura https://orcid.org/0000-0001-7985-2561
Stella Veretnik https://orcid.org/0000-0002-6222-7281
Philip E. Bourne https://orcid.org/0000-0002-7618-7292

REFERENCES

1. Liberles DA, Teichmann SA, Bahar I, et al. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 2012;21:769–785.
2. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. J Royal Soc Interface. 2014;11:20140419.
3. Kolodny R, Petrey D, Honig B. Protein structure comparison: Implications for the nature of 'fold space', and structure and function prediction. Curr Opin Struct Biol. 2006;16:393–398.
4. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. Quantifying the similarities within fold space. J Mol Biol. 2002;323:909–926.
5. Dessailly BH, Dawson NL, Das S, Orengo CA. Function diversity within folds and superfamilies. In: Rigden DJ, editor. From protein structure to function with bioinformatics. Dordrecht, The Netherlands: Springer, 2017; p. 295–325.
6. Brenner SE, Chothia C, Hubbard TJ. Population statistics of protein structures: Lessons from structural classifications. Curr Opin Struct Biol. 1997;7:369–376.
7. Nepomnyachiy S, Ben-Tal N, Kolodny R. Global view of the protein universe. Proc Natl Acad Sci U S A. 2014;111:11691–11696.
8. Przytycka T, Aurora R, Rose GD. A protein taxonomy based on secondary structure. Nat Struct Biol. 1999;6:672–682.
9. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res. 1994;22:3600–3609.
10. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural classification of proteins—Extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014;42:D304–D309.
11. Dawson NL, Lewis TE, Das S, et al. CATH: An expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2017;45:D289–D295.
12. Cheng H, Schaeffer RD, Liao Y, et al. ECOD: An evolutionary classification of protein domains. PLoS Comput Biol. 2014;10:e1003926.
13. Cuff A, Redfern OC, Greene L, et al. The CATH hierarchy revisited—Structural divergence in domain superfamilies and the continuity of fold space. Structure. 2009;17:1051–1062.

14. Xu J, Zhang J. Impact of structure space continuity on protein fold classification. Sci Rep. 2016;6:23263.
15. Sadowski MI, Taylor WR. On the evolutionary origins of "fold space continuity": A study of topological convergence and divergence in mixed alpha-beta domains. J Struct Biol. 2010;172: 244–252.
16. Edwards H, Deane CM. Structural bridges through fold space. PLoS Comput Biol. 2015;11:e1004466.
17. Nepomnyachiy S, Ben-Tal N, Kolodny R. Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. Proc Natl Acad Sci U S A. 2017;114:11703–11708.
18. Shindyalov IN, Bourne PE. An alternative view of protein fold space. Proteins. 2000;38:247–260.
19. Friedberg I, Godzik A. Connecting the protein structure universe by using sparse recurring fragments. Structure. 2005;13:1213–1224.
20. Alva V, Dunin-Horkawicz S, Habeck M, Coles M, Lupas AN. The GD box: A widespread noncontiguous supersecondary structural element. Protein Sci. 2009;18:1961–1966.
21. Skolnick J, Arakaki AK, Lee SY, Brylinski M. The continuity of protein structure space is an intrinsic property of proteins. Proc Natl Acad Sci U S A. 2009;106:15690–15695.
22. Taylor WR, Chelliah V, Hollup SM, MacDonald JT, Jonassen I. Probing the "dark matter" of protein fold space. Structure. 2009; 17:1244–1252.
23. Grishin NV. Fold change in evolution of protein structures. J Struct Biol. 2001;134:167–185.
24. Sippl MJ. Fold space unlimited. Curr Opin Struct Biol. 2009;19: 312–320.
25. Sadreyev RI, Kim BH, Grishin NV. Discrete-continuous duality of protein structure space. Curr Opin Struct Biol. 2009;19:321–328.
26. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature. 1994;372:631–634.
27. Sillitoe I, Dawson N, Thornton J, Orengo C. The history of the CATH structural classification of protein domains. Biochimie. 2015;119:209–217.
28. Holm L, Sander C. Mapping the protein universe. Science. 1996; 273:595–603.
29. Youkharibache P, Veretnik S, Li Q, Stanek KA, Mura C, Bourne PE. The small β-barrel domain: A survey-based structural analysis. Structure. 2019;27:6–26.
30. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci U S A. 2006;103: 2605–2610.
31. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D. Structure and assembly of an augmented Sm-like archaeal protein 14-mer. Proc Natl Acad Sci U S A. 2003;100:4539–4544.
32. Van Duyne GD, Ghosh G, Maas WK, Sigler PB. Structure of the oligomerization and L-arginine binding domain of the arginine repressor of *Escherichia coli*. J Mol Biol. 1996;256:377–391.
33. Hall TM, Porter JA, Beachy PA, Leahy DJ. A potential catalytic site revealed by the 1.7-a crystal structure of the amino-terminal signalling domain of sonic hedgehog. Nature. 1995;378:212–216.
34. Aevarsson A, Brazhnikov E, Garber M, et al. Three-dimensional structure of the ribosomal translocase: Elongation factor G from *Thermus thermophilus*. EMBO J. 1994;13:3669–3677.
35. Day R, Beck DA, Armen RS, Daggett V. A consensus view of fold space: Combining SCOP, CATH, and the Dali domain dictionary. Protein Sci. 2003;12:2150–2160.
36. Shi S, Zhong Y, Majumdar I, Sri Krishna S, Grishin NV. Searching for three-dimensional secondary structural patterns in proteins with ProSMoS. Bioinformatics. 2007;23:1331–1338.
37. Berger B. Algorithms for protein structural motif recognition. J Comput Biol. 1995;2:125–138.
38. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol. 2001;134:191–203.
39. Iyer LM, Koonin EV, Leipe DD, Aravind L. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: Structural insights and new members. Nucleic Acids Res. 2005;33:3875–3896.
40. Dill KA, Fiebig KM, Chan HS. Cooperativity in protein-folding kinetics. Proc Natl Acad Sci U S A. 1993;90:1942–1946.
41. Wathen B, Jia Z. Protein β-sheet nucleation is driven by local modular formation. J Biol Chem. 2010;285:18376–18384.
42. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function. J Mol Biol. 1999;291:177–196.
43. Makarenkov V, Legendre P. From a phylogenetic tree to a reticulated network. J Comput Biol. 2004;11:195–212.
44. Hopf TA, Marks DS. Protein structures, interactions and function from evolutionary couplings. In: Rigden DJ, editor. From protein structure to function with bioinformatics. Dordrecht, The Netherlands: Springer, 2017; p. 37–58.
45. Alva V, Koretke KK, Coles M, Lupas AN. Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. Curr Opin Struct Biol. 2008;18:358–365.
46. Agrawal V, Kishan RK. Functional evolution of two subtly different (similar) folds. BMC Struct Biol. 2001;1:5.
47. Alva V, Remmert M, Biegert A, Lupas AN, Soding J. A galaxy of folds. Protein Sci. 2010;19:124–130.
48. Grishin NV. KH domain: One motif, two folds. Nucleic Acids Res. 2001;29:638–643.
49. Krishna SS, Grishin NV. Structural drift: A possible path to protein fold change. Bioinformatics. 2005;21:1308–1310.
50. Medvedev KE, Kinch LN, Grishin NV. Functional and evolutionary analysis of viral proteins containing a Rossmann-like fold. Protein Sci. 2018;27:1450–1463.
51. Murzin AG, Bateman A. Distant homology recognition using structural classification of proteins. Proteins Suppl. 1997;1:105–112.
52. Goldenberg DP. Finding the right fold. Nat Struct Biol. 1999;6:987–990.
53. Martinez JC, Serrano L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. Nat Struct Biol. 1999;6:1010–1016.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Mura C, Veretnik S, Bourne PE. The *Urfold*: Structural similarity just above the superfold level? *Protein Science*. 2019;28: 2119–2126. https://doi.org/10.1002/pro.3742