Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Searching protein space for ancient sub-domain segments

Rachel Kolodny

Evolutionary processes that formed the current protein universe left their traces, among them homologous segments that recur, or are 'reused,' in multiple proteins. These reused segments, called 'themes,' can be found at various scales, the best known of which is the domain. Yet, recent studies have begun to focus on the evolutionary insights that can be derived from sub-domain-scale themes, which are candidates for traces of more ancient events. Characterizing these may provide clues to the emergence of domains. Particularly interesting are themes that are reused across dissimilar contexts, that is, where the rest of the protein domain differs. We survey computational studies identifying reused themes within different contexts at the sub-domain level.

Address
Department of Computer Science, University of Haifa, 3498838 Haifa, Israel

Corresponding author: Kolodny, Rachel (trachel@cs.haifa.ac.il)

Current Opinion in Structural Biology 2021, **68**:105–112

This review comes from a themed issue on **Sequences and topology**

Edited by **Nir Ben-Tal** and **Andrei N Lupas**

For a complete overview see the Issue and the Editorial

Available online 18th January 2021

https://doi.org/10.1016/j.sbi.2020.11.006

Introduction

How did the protein universe emerge? The question is challenging because it calls for syllogizing about the early stages of the evolutionary process from its result: present-day proteins. We know quite a bit about the result, given the availability of significant amounts of protein data [1–8], coupled with tools to compare proteins along different facets: sequence, structure, function, and context. In particular, sequence similarities offer valuable clues into evolutionary processes, as such similarities are a hallmark of shared ancestry; that is, random sampling of a specific sequence is a very low-probability event, such that it is unlikely that two similar long sequence segments formed independently [9–14]. Indeed, sequence similarities embody traces of evolutionary events such as mutations, duplications, and recombinations (e.g. Refs. [15,16,17••,18,19,20]). Such events may be either recent or ancient, given that, as Eck and Dayhoff argued, recent events have not erased all traces of ancient ones, because natural selection inhibits change to ancient well-adapted parts on which other essential components depend [21].

The recurrence of similar sequence segments, or 'themes,' across proteins reflects the fact that the evolutionary processes of duplication and recombination happened and left their traces. Indeed, it is more efficient to 'reuse' segments [22] than to be invent them ab-initio [20,23]. These processes could have also acted on segments shorter than domains, even if they are too short to fold independently. Ancient segments (denoted Ancestral Domain Segments (ADS) by Lupas *et al.*) could have existed in an oligomeric state [12]. Smock *et al.* re-enacted such a scenario experimentally [24]. Alternatively, these segments could have been stabilized by the ribosome [25,26]. Also, themes that existed within full domains could have been re-used and grafted into other domains. Bharat *et al.* re-enacted this scenario experimentally [27].

Thus, one would expect the protein universe to include traces that are (divergent versions of) longer segments formed by duplication and recombination of shorter and more ancient ones [11,12,21,28,29]. According to this logic, shorter themes (with the exception of those produced by processes such as domain atrophy [30]) are candidates for traces of more ancient events. Notably, though classification of proteins according to similarity in longer segments — or domains — is a fundamental tenet of protein research, it is only relatively recently that researchers have begun to focus computational search efforts on identifying shorter, sub-domain-scale themes that recur across proteins and that may constitute traces of evolutionary processes.

In what follows, we review this emerging field of research. In particular, we survey recent computational studies that search for reuse across multiple proteins and relate the choice of computational procedure to the types of traces found. We note that we do not discuss the special case of intra-chain repetitions, or searching for duplications within a chain, as a recent review by Alva and Lupas [23] has covered the insightful studies addressing this case.

A note on terminology: usage of the terms 'global' versus 'local' in the study of protein space

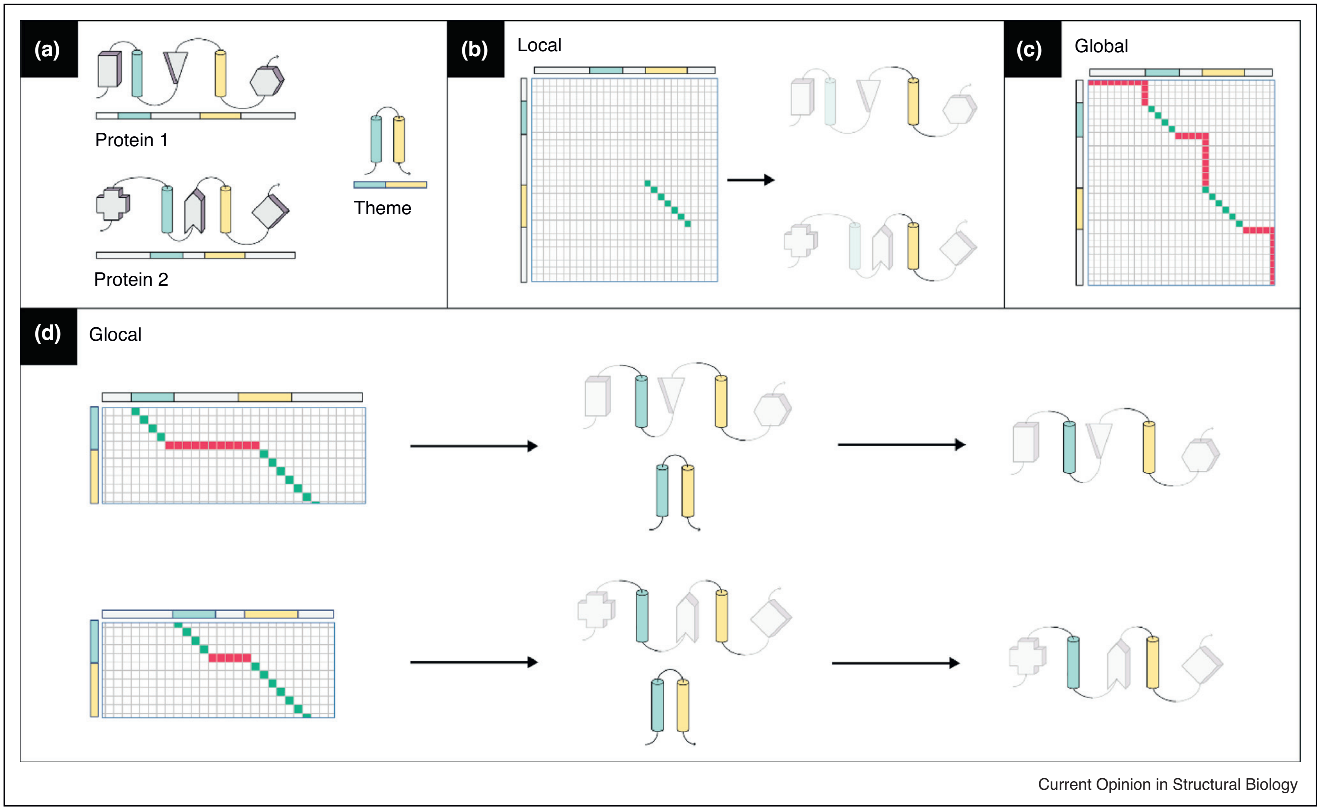
Before launching into our discussion, we clarify the terminology used to refer to different types of comparisons within protein space. In general, protein comparisons can be distinguished according to whether they search for 'global', 'local', or 'glocal' similarities. Yet, confusingly, comparison studies also use these terms in a different meaning, to describe the extent of the space studied — such that a 'global' refers to a comparison across all proteins (e.g. the global view of protein space [15,31]), whereas a 'local' refers to comparison within a smaller set (e.g. a set of homologous proteins [32]). Here, however, we use these terms in their other common usage, namely, as modifiers that describe the extent of similarity within compared objects (e.g. protein domains). In this case, a 'global' comparison between two objects identifies similarity across the entirety of those objects, whereas a local comparison identifies partial similarities. In sequence comparison, the distinction between global and local comparisons is that between the dynamic programming algorithms used: Needleman-Wunch for global comparisons, versus Smith-Waterman for local comparisons. The hybrid, 'glocal,' comparison finds similarities that span the entire (i.e. global) sequence of one object, but are only partial (i.e. local) in the other (see Figure 1(b–d)).

Global similarity implies local similarity, but the reverse is not true. Hence, a search for local similarities between objects is likely to identify more pairs of 'similar' objects compared with a search for global similarities. Transitive similarities are similarities in which, for any three proteins A, B, and C, the similarities (A, B), and (B, C), imply the similarity (A, C). Global similarities can (at least generally) satisfy this transitive property. In contrast, the partial nature of local similarities renders them non-transitive.

Reliance on global versus local similarities in protein classification

All meaningful sequence similarities (local ones included) reveal something about protein evolution. Focusing only

Figure 1



Alternative protein sequence comparisons: **(a)** protein1 and protein2 are globally different proteins that share local segments. The structures and sequences of the proteins are represented as cartoons and lines. The proteins share a green and an yellow segment; the unrelated segments are shown in gray. **(b)** The optimal local similarity matches only the more similar of the two possibly matching segments because optimizing the score of a local alignment (marked in green) avoids the negative contribution from the non-matching segments connecting the green and yellow ones. We show a cartoon of the aligned parts in the structures as non-dimmed. **(c)** Globally comparing the two proteins reveals they are not similar: although the green and yellow segments match with a positive contribution to the overall score (marked in green), the unrelated gray segments do not, and their gaps have a negative contribution to that score (marked in red). **(d)** A glocal comparison with a bait can identify local similarities within globally different contexts: Here, the bait theme is a concatenation of the green and yellow segments; comparing it glocally to both proteins identifies the two segments in both proteins because the bait is matched in its entirety. These parts in the two proteins can then be globally aligned to each other.

on global (and transitive) similarities offers the advantage of enabling protein space to be organized into groups, with each group containing proteins that are similar to one another. Because many chains share only one of their domains (i.e. are globally different yet locally similar) [33], the pioneers of protein classification built classification systems for domains. Once chains were parsed into domains, the focus could be shifted back to global similarities.

That the hierarchical domain classifications SCOP [34], CATH [35], and ECOD [4] capture meaningful global similarities is evident from the grouping of domains into disjoint sets [36]. For example, SCOP groups domains of closely related sequences into families, families into superfamilies, and superfamilies of similar structures into folds [34,37]. The classifications differ [4,38,39], each offering its own perspective on meaningful similarities: For example, the less conservative ECOD includes more remote global homologies [4]. The structural levels also have a subjective component in that similarity is based on core-features, rather than on all residues [11,13].

Classification systems that capture the global similarities between domains portray only a partial picture. Missing from it are local similarities between globally different domains, including similarities across folds [40]. It is unclear whether local structural similarities that have been identified across CATH and SCOP folds [19,41] are due to homology or to biophysical constraints (i.e. convergent evolution) [42,43]. Notably, however, some of these structural similarities are accompanied by significant cross-fold local (i.e. sub-domain-level) sequence similarities, and these suggest homology: for example, those found by Alva *et al.* [18], Nepomnyachiy *et al.* [15], and Ferruz *et al.* [44••].

Together, these observations imply that if we were to model protein similarities as edges in a graph (or network) connecting nodes that represent domains, the hierarchical classifications would form graphs with many disconnected sub-graphs (e.g. one per fold); adding the cross-fold local similarities would connect these subgraphs [45]. Several studies have tried to explore these connections, shifting the curation of evolutionary relationships among domains to include local similarities, for example, Nepomnyachiy *et al.* [15], SISYPHUS [46], and most recently SCOP2 [47,48•]. Nevertheless, evolutionary studies continue to rely primarily on global domain-level similarities, because the analysis of such similarities is simpler.

We believe that finding ways to include local similarities at the sub-domain level is critical to studying protein evolution. First, an inherent assumption of domain-based classification is that the boundaries of the domains are correct. Yet, as Bourne argued, it is difficult to identify domain boundaries correctly, and indeed, the definitions of domains vary across classifications [49]. The problem is not just an issue of mistaken boundaries: Local similarities reveal complex patterns of homology among different, often overlapping parts, which hint at the evolutionary processes that formed them [16,46,47,48•].

The discussion of whether global similarities across proteins suffice when studying evolutionary processes versus the necessity of adding local similarities echoes an old scientific debate regarding whether protein domain space is discrete or continuous [15,36,50,51]. In the discrete model, hierarchical classifications (i.e. only global similarities) are an adequate description of the protein space; in the continuous model, the many cross-fold, or local, similarities must also be considered (e.g. those relating the ancient alpha/beta domains [15]).

The case of domains and locally similar but globally different chains

In general, it is difficult to derive insights from local patterns; several studies suggest borrowing methods and inspiration from linguistics [52,53,54••]. Fortunately, however, given that shared domains are local similarities within globally different chains, we can derive insights and methodologies from the vast field of study of domain architectures (see Refs. [33,53,55] for recent reviews).

In general, domain architectures are calculated by glocally aligning sequences of a pre-curated domain set to non-overlapping parts of protein chains [53]. Using the domains in the pre-curated set as 'baits' reveals local (domain-size) similarities in globally different chains, namely, when a bait matches multiple chains [20]. Note that using baits is a different search from locally comparing all versus all chain sequences (see Figure 1 for an example).

Analysis of domain architectures has led to evolutionary insights. For example, studies have shown that insertions are more common than deletions, especially at the termini [56,57]. Domain architectures were used in functional comparative genome analysis [58], and domain ages were estimated from those of the genomes at the root of the sub-phylogenetic trees where they occur [56,59,60]. Interestingly, not all domain combinations exist, and it is possible to quantify, using relative entropy, the difference between the limited repertoire of pairs in the protein universe and the complete repertoire of domain pairs that could be associated at random [54••]. Nonetheless, because domains are considered independently folding units, it is believed that generally all combinations can be formed in the lab. Important exceptions to this rule are outer membrane beta barrel domains, for which it was assumed that there is at most one domain per chain [23,61] — until recently, when the Ben-Tal lab discovered chains with multiple barrel domains (unpublished results).

The case of sub-domains and locally similar but globally different domains

Similar analysis methodologies can be used to investigate local, sub-domain-level similarities shared among globally different domains. Yet, this level of analysis presents additional challenges compared to the above-domain level. Curating a bait set can be challenging, as there are no pre-curated databases. When analyzing the genomic history of domain architectures, one can realistically assume that for most current architectures, ancestral architectures are also found, unchanged, in other proteins [33]. Finally, using shorter baits finds fewer reliable glocal sequence alignments. Several studies took the first step and identified traces of homologous segments at the sub-domain level between domains that are not overall homologous, including the set of Alva *et al.* [52], which extends the ADS set [12], the 'Fuzzle' dataset by Ferruz *et al.* [44••], and the recent set of bridging themes by Kolodny *et al.* [62••].

Factors that influence the sensitivity of searches for ancient sub-domain themes

When a search of a domain dataset fails to find novel local sub-domain-level similarities, it may be because such similarities are rare and the passage of time has eroded their traces, or, alternatively, because the search was not sensitive enough.

There are two main factors that can influence the sensitivity of a search. The first is the dataset of searched domains. A straightforward choice for the domain dataset is all domains classified by the various established hierarchies; for these, we know which are globally similar. We suggest that, instead of considering a small (e.g. 30%), non-redundant subset of these domains, one should search in a somewhat redundant set (70%), and cluster the results. The reason is that we seek a representative set of globally different (e.g. ECOD X-group) domains with the strongest local similarity signals [63], and the best representatives may be non-standard exemplars within their groups (see, for example, the locally similar P-loop and Rossmann domains described in Ref. [64••]).

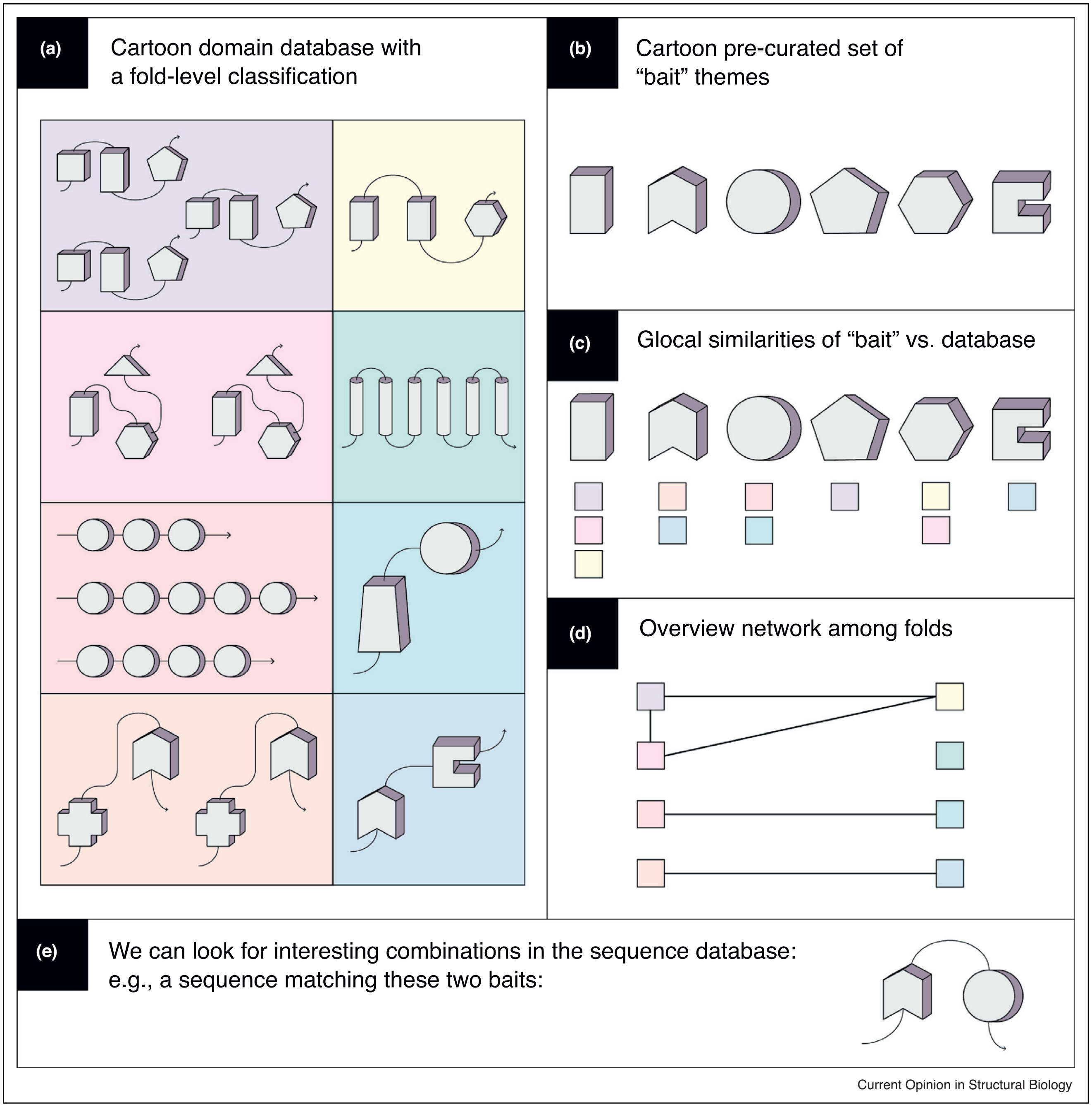
The second factor influencing search sensitivity is the mode of comparison adopted. Figure 1 highlights ways to identify local segments shared among globally different domains. The domains can be compared directly with a local aligner (e.g. as in Refs. [17••,44••,52], Panel b). Alternatively, one can rely on a pre-curated set of baits, and compare these to the domains, using a glocal aligner (e.g. as in Ref. [62••], Panel d). While a direct comparison does not require a bait set, the alternative of curating such a set is closer in spirit to how domain architectures are identified [53]. Indeed, using bait themes focuses the search on the relevant parts within the domains. Figure 2 (panels a-d) shows a cartoon overview of using baits to search for all similarities in a database of classified domains, where the results are organized as a graph (panel d).

The bait set can vary: In a recent study [62••], we used a set of previously identified themes [16]. Other potential bait sets include: (1) unit segments in repeating proteins, as these segments are believed to be duplications of ancient segments [23]; there are tools to identify these, for example, [65]; (2) the elementary functional loops (EFLs) identified by Berezovsky *et al.* [66]. The EFL database includes ligand-binding segments of approximately 30 residues [67]; (3) Similarly, the 'Fuzzle' fragments curated by Hocker and co-workers [44••] and (4) structural units (e.g. the beta-blades set used in Ref. [68]) or functional units (e.g. adenine-binding themes used in Ref. [69••]). Notably, and in contrast to the above-domain setting, in which globally similar domains are grouped together, theme baits may have more complicated similarity patterns (e.g. a series of expanding segments, akin to a pattern of Russian dolls [16]). Thus, although the glocal aligner matches the entire theme, redundancy among themes can lead to finding redundant instances.

Constructing a comprehensive and meaningful catalogue of recurrent sub-domain themes

The computational efforts initiated to identify locally similar segments across globally different domains [12,44••,52,62••] can contribute to a comprehensive catalogue of such themes. This catalogue will allow scholars to characterize these themes and their co-occurrence patterns. Focal characteristics might include, for example, biophysical features such as polar/hydrophobic profile; tendency to be core or surface elements; or tendency to be located toward the termini or middle of the chain. For some themes, it may be possible to identify functional roles — for example, binding — on the basis of known structures. Such characterizations could extend the report of Alva et al., that their set of recurrent protein fragments is enriched with nucleotide-binding, nucleic-acid-binding, and metal-binding motifs [52], or the metal-binding themes described in Refs. [62••,70]. The catalogue can also be used to reveal novel co-occurrences of themes. Even though the catalogue is based on PDB data, because the search is sequence-based, one can search for novel co-occurrences in the far larger sequence databases [71]. Each catalogue theme is found in multiple global contexts, suggesting that the themes are ancestral within their domains. The notion that the same themes exist together in distinct global contexts (Figure 2e) is interesting because it implies that evolution joined these contexts together, or created a non-monophyletic domain [12]. Characterization of the themes can then be used to study the evolution of protein function – computationally (e.g. Ref. [69••]) and experimentally (e.g. Ref. [72]), and perhaps even in protein engineering [73].

Figure 2


(a)
Cartoon domain database with a fold-level classification
(b)
Cartoon pre-curated set of "bait" themes
(c)
Glocal similarities of "bait" vs. database
(d)
Overview network among folds
(e)
We can look for interesting combinations in the sequence database: e.g., a sequence matching these two baits:
Current Opinion in Structural Biology

Overview describing how to use a pre-curated set of baits to characterize local similarities among globally different proteins at the sub-domain level. **(a)** A domain classification clusters domains into groups: in our cartoon representation, each fold is boxed in a different color. **(b**,**c)** We can glocally align a pre-curated set of themes, to identify cases of different folds (marked by their color) that share a theme. **(d)** The similarities among the folds can be represented as a network: each node is a fold, and edges connect folds that share a theme. **(e)** We can then look for novel combinations of bait themes that were found in at least two folds within the sequence databases.

We note that previous studies by us [15] and others [52] searching for locally similar sequences within globally different domains added a structure-based post-filter. Specifically, to focus on cases with supporting structural evidence of homology, those studies did not suffice with sequence similarity but instead also required structural similarity, such that segments with different conformations were not considered to be similar. Here, we plead in favor of these discarded cases. The structures of mutating sequences are generally robust [74]. However, as pointed out by Grishin [11], there is no strict correlation between homology and structural similarity. The environment of a sequence can change the sequence's structure, as observed in short chameleon segments (up to 10 residues), sub-domain sections in different oligomeric states, and even complete domains and proteins [75–77]. Indeed, there are many cases in naturally occurring or designed proteins where a few mutations have a dramatic structural impact; for recent reviews see Refs. [78–80]. Thus, not only can these cases be bona-fide instances of a shared ancestry, they may be particularly interesting ones. The capacity of a sequence's structure to change in different environments, or through a few mutations, is evolutionarily advantageous for exploring structure space [63] and evolving new functions [81].

Concluding remarks

Computational and experimental studies of protein evolution are complementary approaches to decipher the historical record and the mechanisms that govern it. Traces of many ancient evolutionary events are long gone, having left only scant marks on the current protein universe. Indeed, the only way to study the parts no longer in view is to re-enact them experimentally (e.g. Refs. [24,27,82,83]). Nonetheless, it is worthwhile to find the rare traces, embodied in sub-domain-level themes: not only because these traces are part of the historical record, but also because their persistence throughout time suggests they are important. To characterize these remnants, we must identify them in the current universe [64••,69••], and sifting through the vast amounts of data necessitates specialized computational procedures. Then, their ancestral sequences can be reconstructed, and these may be studied experimentally [24,83], towards deciphering how proteins emerged and continue to evolve.

Conflict of interest statement

Nothing declared.

Acknowledgements

We thank Dan Latovicz for his help with the figures. This research was funded by ISF Grant 450/16.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Levitt M: **Nature of the protein universe**. *Proc Natl Acad Sci U S A* 2009, **106**:11079-11084.

2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The protein data bank**. *Nucleic Acids Res* 2000, **28**:235-242.

3. Hubbard TJ, Murzin AG, Brenner SE, Chothia C: **SCOP: a structural classification of proteins database**. *Nucleic Acids Res* 1997, **25**:236-239.

4. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim B-H, Grishin NV: **ECOD: an evolutionary classification of protein domains**. *PLoS Comput Biol* 2014, **10**:e1003926.

5. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

6. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR: **CDD: a Conserved Domain Database for the functional annotation of proteins**. *Nucleic Acids Res* 2010, **39**:D225-D229.

7. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A *et al.*: **The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution**. *Nucleic Acids Res* 2007, **35**:D291-D297.

8. Consortium TU: **The universal protein resource (UniProt)**. *Nucleic Acids Res* 2008, **36**:D190-D195.

9. Doolittle RF: **Similar amino acid sequences: chance or common ancestry?** *Science* 1981, **214**:149-159.

10. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches**. *J Mol Biol* 1999, **287**:1023-1040.

11. Grishin NV: **Fold change in evolution of protein structures**. *J Struct Biol* 2001, **134**:167-185.

12. 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.

13. Lupas A, Koretke K: **Evolution of protein folds**. *Computational Structural Biology: Methods and Applications*. World Scientific Hackensack, NJ; 2008:131-152.

14. Fetrow JS, Godzik A: **Function driven protein evolution. A possible proto-protein for the RNA-binding proteins**. *Pac Symp Biocomput*. 1998:485-496.

15. Nepomnyachiy S, Ben-Tal N, Kolodny R: **Global view of the protein universe**. *Proc Natl Acad Sci U S A* 2014 http://dx.doi.org/10.1073/pnas.1403395111.

16. 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.

17. Franklin MW, Nepomnyachyi S, Feehan R, Ben-Tal N, Kolodny R,
•• Slusky JS: **Evolutionary pathways of repeat protein topology in bacterial outer membrane proteins**. *eLife* 2018, **7**:e40308

A detailed analysis of the evolutionary relationships among outer membrane beta-barrels (OMBBs) to describe evolutionary paths leading to the OMBBs found today (including a loop to strand conversion).

18. Alva V, Remmert M, Biegert A, Lupas AN, Söding J: **A galaxy of folds**. *Protein Sci* 2010, **19**:124-130.

19. Edwards H, Deane CM: **Structural bridges through fold space**. *PLoS Comput Biol* 2015, **11**:e1004466.

20. Chothia C, Gough J, Vogel C, Teichmann SA: **Evolution of the protein repertoire**. *Science* 2003, **300**:1701-1703.

21. Eck RV, Dayhoff MO: **Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences**. *Science* 1966, **152**:363-366.

22. Jacobs T, Williams B, Williams T, Xu X, Eletsky A, Federizon J, Szyperski T, Kuhlman B: **Design of structurally distinct proteins using strategies inspired by evolution**. *Science* 2016, **352**:687-690.

23. Alva V, Lupas AN: **From ancestral peptides to designed proteins**. *Curr Opin Struct Biol* 2018, **48**:103-109.

24. Smock RG, Yadid I, Dym O, Clarke J, Tawfik DS: **De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints**. *Cell* 2016, **164**:476-486.

25. Lupas AN, Alva V: **Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins**. *J Struct Biol* 2017, **198**:74-81.

26. Kovacs NA, Petrov AS, Lanier KA, Williams LD: **Frozen in time: the history of proteins**. *Mol Biol Evol* 2017, **34**:1252-1260.

27. Bharat TA, Eisenbeis S, Zeth K, Höcker B: **A βα-barrel built by the combination of fragments from different folds**. *Proc Natl Acad Sci U S A* 2008, **105**:9942-9947.

28. Brenner S: **The molecular evolution of genes and proteins: a tale of two serines**. *Nature* 1988, **334**:528-530.

29. Koonin EV, Wolf YI, Karev GP: **The structure of the protein universe and genome evolution**. *Nature* 2002, **420**:218-223.

30. Prakash A, Bateman A: **Domain atrophy creates rare cases of functional partial protein domains**. *Genome Biol* 2015, **16**:1.

31. Hou J, Sims GE, Zhang C, Kim SH: **A global representation of the protein fold space**. *Proc Natl Acad Sci U S A* 2003, **100**:2386-2390.

32. Narunsky A, Ben-Tal N, Kolodny R: **Navigating among known structures in protein space**. *Computational Methods in Protein Evolution*. Springer; 2019:233-249.

33. Forslund SK, Kaduk M, Sonnhammer ELL: **Evolution of protein domain architectures**. In *Evolutionary Genomics: Statistical and Computational Methods.* Edited by Anisimova M. New York: Springer; 2019:469-504 http://dx.doi.org/10.1007/978-1-4939-9074-0_15.

34. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536-540.

35. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH–a hierarchic classification of protein domain structures**. *Structure* 1997, **5**:1093-1108.

36. 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.

37. Brenner SE, Chothia C, Hubbard TJ, Murzin AG: **Understanding protein structure: using scop for fold interpretation**. *Methods Enzymol* 1996, **266**:635-643.

38. Kelley LA, Sternberg MJ: **Partial protein domains: evolutionary insights and bioinformatics challenges**. *Genome Biol* 2015, **16**:1-3.

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

40. Sippl MJ: **Fold space unlimited**. *Curr Opin Struct Biol* 2009, **19**:312-320.

41. Harrison A, Pearl F, Mott R, Thornton J, Orengo C: **Quantifying the similarities within fold space**. *J Mol Biol* 2002, **323**:909-926.

42. Tian P, Best RB: **How many protein sequences fold to a given structure? A coevolutionary analysis**. *Biophys J* 2017, **113**:1719-1730.

43. Deeds EJ, Shakhnovich EI: **A structure – centric view of protein evolution, design, and adaptation**. *Adv Enzymol Relat Areas Mol Biol* 2007, **75**:133.

44. •• Ferruz N, Lobos F, Lemm D, Toledo-Patino S, Farías-Rico JA, Schmidt S, Höcker B: **Identification and analysis of natural building blocks for evolution-guided fragment-based protein design**. *J Mol Biol* 2020 http://dx.doi.org/10.1016/j.jmb.2020.04.013

The 'Fuzzle' database of more than 1000 protein homologous fragments shared in different SCOP folds.

45. Ben-Tal N, Kolodny R: **Representation of the Protein universe using classifications, maps, and networks**. *Israel J Chem* 2014.

46. Andreeva A, Prlić A, Hubbard TJP, Murzin AG: **SISYPHUS—structural alignments for proteins with non-trivial relationships**. *Nucleic Acids Res* 2007, **35**:D253-D259.

47. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG: **SCOP2 prototype: a new approach to protein structure mining**. *Nucleic Acids Res* 2013 http://dx.doi.org/10.1093/nar/gkt1242.

48. • Andreeva A, Kulesha E, Gough J, Murzin AG: **The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures**. *Nucleic Acids Res* 2019, **48**:D376-D382

The new SCOP2 hierarchy which includes more complex relationships among domains. Notably, the domain boundaries are now defined for both the family and superfamily levels.

49. Holland TA, Veretnik S, Shindyalov IN, Bourne PE: **Partitioning protein structures into domains: why is it so difficult?** *J Mol Biol* 2006, **361**:562-590.

50. Sadreyev RI, Kim B-H, Grishin NV: **Discrete–continuous duality of protein structure space**. *Curr Opin Struct Biol* 2009, **19**:321-328.

51. 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.

52. Alva V, Söding J, Lupas AN: **A vocabulary of ancient peptides at the origin of folded proteins**. *eLife* 2015, **4**:e09410.

53. Scaiewicz A, Levitt M: **The language of the protein universe**. *Curr Opin Genet Dev* 2015, **35**:50-56.

54. •• Yu L, Tanwar DK, Penha EDS, Wolf YI, Koonin EV, Basu MK: **Grammar of protein domain architectures**. *Proc Natl Acad Sci U S A* 2019, **116**:3636

The authors highlight that not all domain combinations are found in protein space, and further quantify, using information theory tools, the difference between what is found and randomly associated domains.

55. Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E, Elofsson A: **Arrangements in the modular evolution of proteins**. *Trends Biochem Sci* 2008, **33**:444-451.

56. Nasir A, Kim KM, Caetano-Anollés GJPCB: **Global patterns of protein domain gain and loss in superkingdoms**. *PLoS Comput Biol* 2014, **10**:e1003452.

57. Björklund ÅK, Ekman D, Light S, Frey-Skött J, Elofsson A: **Domain rearrangements in protein evolution**. *J Mol Biol* 2005, **353**:911-923.

58. Koehorst JJ, Saccenti E, Schaap PJ, dos Santos VAM, Suarez-Diez MJF: **Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics**. *F1000Res* 2016, **5**.

59. Winstanley HF, Abeln S, Deane CM: **How old is your fold?** *Bioinformatics* 2005, **21**:449-458.

60. Wang M, Jiang Y-Y, Kim KM, Qu G, Ji H-F, Mittenthal JE, Zhang H-Y, Caetano-Anollés G: **A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation**. *Mol Biol Evol* 2010, **28**:567-582.

61. Arnold T, Poynor M, Nussberger S, Lupas AN, Linke D: **Gene duplication of the eight-stranded β-barrel OmpX produces a functional pore: a scenario for the evolution of transmembrane β-barrels**. *J Mol Biol* 2007, **366**:1174-1184.

62. •• Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N: **Bridging themes: short protein segments found in different architectures**. *bioRxiv* 2020

The authors describe the results of a comprehensive search (using baits) of sub-domain themes that appear in different ECOD folds. They denote these bridging themes and highlight that in many cases the structures of these themes vary, depending on the context.

63. Bornberg-Bauer E, Huylmans A-K, Sikosek T: **How do new proteins arise?** *Curr Opin Struct Biol* 2010, **20**:390-396.

64. •• Longo LM, Jablonska J, Vyas P, Kolodny R, Ben-Tal N, Tawfik DS: **On the emergence of P-Loop NTPase and Rossmann enzymes from a beta-alpha-beta ancestral fragment**. *Elife* 2020, **9**:e64415

A detailed analysis of a specific sub-domain element that is found in two ancient folds: a P-loop and a Rossmann.

65. Biegert A, Söding J: **De novo identification of highly diverged protein repeats by probabilistic consistency**. *Bioinformatics* 2008, **24**:807-814.

66. Berezovsky IN, Guarnera E, Zheng Z: **Basic units of protein structure, folding, and function**. *Progr Biophys Mol Biol* 2017, **128**:85-99.

67. Zheng Z, Goncearenco A, Berezovsky IN: **Nucleotide binding database NBDB – a collection of sequence motifs with specific protein-ligand interactions**. *Nucleic Acids Res* 2015, **44**:D301-D307.

68. Kopec KO, Lupas AN: **β-Propeller blades as ancestral peptides in protein evolution**. *PLoS One* 2013, **8**:e77074.

69. •• Narunsky A, Kessel A, Solan R, Alva V, Kolodny R, Ben-Tal N: **On the evolution of protein–adenine binding**. *Proc Natl Acad Sci U S A* 2020, **117**:4701-4709

A survey of adenine-binding patterns in the protein universe.

70. Krishna SS, Sadreyev RI, Grishin NV: **A tale of two ferredoxins: sequence similarity and structural differences**. *BMC Struct Biol* 2006, **6**:8.

71. Farías-Rico JA, Schmidt S, Höcker B: **Evolutionary relationship of two ancient protein superfolds**. *Nat Chem Biol* 2014, **10**:710-715.

72. Romero Romero ML, Yang F, Lin YR, Toth-Petroczy A, Berezovsky IN, Goncearenco A, Yang W, Wellner A, Kumar-Deshmukh F, Sharon M *et al.*: **Simple yet functional phosphate-loop proteins**. *Proc Natl Acad Sci U S A* 2018, **115**:E11943-E11950.

73. Khersonsky O, Fleishman SJ: **Why reinvent the wheel? Building new proteins based on ready-made parts**. *Protein Sci* 2016, **25**:1179-1187.

74. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins**. *EMBO J* 1986, **5**:823-826.

75. Yadid I, Kirshenbaum N, Sharon M, Dym O, Tawfik DS: **Metamorphic proteins mediate evolutionary transitions of structure**. *Proc Natl Acad Sci U S A* 2010, **107**:7287-7292.

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

77. Narunsky A, Nepomnyachiy S, Ashkenazy H, Kolodny R, Ben-Tal N: **ConTemplate suggests possible alternative conformations for a query protein of known structure**. *Structure* 2015, **23**:2162-2170.

78. Davidson AR: **A folding space odyssey**. *Proc Natl Acad Sci U S A* 2008, **105**:2759-2760.

79. Lella M, Mahalakshmi R: **Metamorphic proteins: emergence of dual protein folds from one primary sequence**. *Biochemistry* 2017, **56**:2971-2984.

80. Zamora-Carreras H, Maestro B, Sanz JM, Jiménez MA: **Turncoat polypeptides: we adapt to our environment**. *ChemBioChem* 2020, **21**:432-441.

81. James LC, Tawfik DS: **Conformational diversity and protein evolution a 60-year-old hypothesis revisited**. *Trends Biochem Sci* 2003, **28**:361-368.

82. Studer S, Hansen DA, Pianowski ZL, Mittl PR, Debon A, Guffy SL, Der BS, Kuhlman B, Hilvert D: **Evolution of a highly active and enantiospecific metalloenzyme from short peptides**. *Science* 2018, **362**:1285-1288.

83. Longo LM, Despotović D, Weil-Ktorza O, Walker MJ, Jabłońska J, Fridmann-Sirkis Y, Varani G, Metanis N, Tawfik DS: **Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion**. *Proc Natl Acad Sci U S A* 2020, **117**:15731-15739.