Universal architectural concepts underlying protein folding patterns

Arthur M. Lesk[a,b], Ramanan Subramanian[c], Lloyd Allison[c], David Abramson[d], Peter J. Stuckey[c,e], Maria Garcia de la Banda[c], and Arun S. Konagurthu[c,*]

[a]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, U.K.; [b]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, U.S.A.; [c]Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia; [d]Research Computing Center, University of Queensland, Brisbane, QLD 4072, Australia; [e]School of Computing and Information Systems, University of Melbourne, VIC 3010, Australia

ABSTRACT

What is the architectural 'basis set' of the observed universe of protein structures? Using information-theoretic inference, we answer this question with a comprehensive dictionary of 1,493 substructural *concepts*. Each *concept* represents a topologically-conserved assembly of helices and strands that make contact. Any protein structure can be dissected into instances of concepts from this dictionary. We dissected the world-wide protein data bank and completely inventoried all concept instances. This yields an unprecedented source of biological insights. These include: correlations between concepts and catalytic activities or binding sites, useful for rational drug design; local amino-acid sequence–structure correlations, useful for *ab initio* structure prediction methods; and information supporting the recognition and exploration of evolutionary relationships, useful for structural studies. An interactive site, PROÇODIC, at http://lcb.infotech.monash.edu.au/prosodic (click) provides access to and navigation of the entire dictionary of concepts, and all associated information.

KEYWORDS: architectural concepts, protein building blocks, structural motifs, lossless compression, information theory

INTRODUCTION

The polypeptide chains of amino acids (primary structure) in most proteins fold into helices and strands of sheet (secondary structure), which in turn assemble to give proteins their intricate three-dimensional shapes and folding patterns (tertiary structure). Experimental methods have already provided over 140,000 entries in the world-wide Protein Data Bank (wwPDB), containing the three-dimensional coordinates of proteins and protein-nucleic acid complexes from a wide range of species.

Unravelling protein architecture and discovering the relationship among these three major levels of structural description provides the key to understanding how proteins function, how their 3D folding patterns form, and how they evolve (1). Investigations of protein folding patterns have revealed recurrent themes at all structural levels (2–8), which form the basis for widely-used hierarchical classifications of protein structures (9–11). Nevertheless, many aspects of the relationships across structural levels have remained unresolved.

Chothia and Lesk (6) introduced the idea of a *core* of the folding patterns of homologous proteins. This core comprises a maximal set of secondary structural elements that assemble in a common 3D topology, while withstanding a certain amount of distortion. The parts outside the core are structurally more variable.

Many related proteins contain some but not all of the same common substructures that form their cores. Therefore, it is of crucial interest to discover the nature of the substructures that contribute to the cores of

* To whom correspondence should be addressed. E-mail: arun.konagurthu@monash.edu

Conceptualization: ASK; Methodology: AML, LA, DA, PJS, MG, and ASK; Software: RS and ASK; Validation: AML, LA, and ASK; Analysis: AML, LA and ASK; Investigation: AML, RS, PJS, MG and ASK; Resources: AML and DA; Data Curation: ASK; Writing - Orginal Draft: AML and ASK; Writing - Review & Editing: RS, LA, DA, PJS and MG; Visualization: AML and ASK; Supervision: ASK; Project Administration: ASK; Funding Acquisition: AML, PJS, MG and ASK.

protein families. Some of these are *supersecondary structures* – small conserved combinations of *successive* elements of secondary structure, such as the β-α-β subunit. Supersecondary structures recur within many protein folds, and can be shared even by unrelated proteins. For example, the β-α-β subunit appears in NAD-binding domains, in TIM barrels, and in many other proteins.

Early definitions of supersecondary structures relied strongly on experts spotting and naming them (4, 12). With the steady growth of the `wwPDB`, several methods have been developed to identify automatically, with varying operational definitions, a *library* of substructures that form what can be considered as the 3D building blocks of protein structures (8, 13–25). However, these approaches have yielded limited libraries containing mostly short oligopeptide fragments, or assemblies of typically 2 to 4 secondary structural elements. It has been a challenge so far to go further than that and dissect protein structures into a more complete set that includes *larger* conserved substructures. Apart from the enormous computational challenge this problem poses, the attempts made so far have lacked a statistically-rigorous framework in which to describe, compute, identify and resolve a dictionary of conserved assemblies of secondary structures.

Here we address this problem and unravel protein folding patterns into a universal dictionary of architectural building blocks (*concepts*) containing topologically-conserved assemblies of helices and strands that make contact. This dictionary advances the current knowledge of these conserved patterns significantly beyond the classical supersecondary structures and other known patterns. Our approach relies on a rigorous information-theoretic framework that allows the inference of a dictionary that (a) avoids overfitting (i.e., inferring a dictionary that is more complex than necessary to explain the observed folding patterns) and (b) achieves an objective trade-off between the descriptive complexity of concepts in the dictionary and their fidelity (i.e., the amount of compression) gained when explaining the observed protein folding patterns. Thus, this work presents the 'basis set' of concepts underlying all observed protein folding patterns, and allows for any protein chain to be decomposed optimally into parts corresponding to concepts from this set. We put together a comprehensive online resource, PROÇODIC (click), that can contribute to: understanding the fundamental principles of protein structure, correlating concepts with ligand binding sites to suggest function, and applying sequence conservation observed within concepts to protein structure prediction.

RESULTS

Automatic identification of a dictionary of substructural concepts. This work uses the concise *tableau* representation of protein folding patterns introduced by Lesk (26), which is based on the idea that the essence of a protein folding pattern is captured by the order, contacts and geometry of the assembly of secondary structural elements along the amino-acid chain. A tableau corresponds to the 3D structure of a single protein domain (or sometimes chain), and has the form of a symmetric matrix (Fig. 1(a,c); **Supplementary §S1**). Importantly, in this representation supersecondary structures can be defined in a compact and computable way as subtableaux containing two or more *successive* secondary structure elements in contact (Fig. 1(d-e)).

We have constructed the universal dictionary reported here using our recently-developed method to infer, automatically, conserved assemblies of secondary structural elements within *any* given source collection of tableaux (27). The idea of a concept is constrained by the requirement that every secondary structural element in the concept must be in contact with at least one other secondary-structure element in that concept. Our concept inference approach (27) is based on the powerful minimum message length criterion for statistical inductive inference (28–30) and lossless data compression (**Supplementary §S2**). We have

applied this method to compress the source collection of Astral SCOP domains (9, 10, 31) (**Supplementary §S1**). This has allowed us to infer a dictionary of 1,493 substructural concepts that *most concisely* and *losslessly* describes the entire source collection, and does so without any prior knowledge or preconceived notions regarding these recurrent substructures.

The total computational effort required to identify this dictionary is equivalent to about 7 years of runtime on a modern computer. Therefore, we parallelised our method and ran it on a high-performance computing cluster using 240 cores to identify the PROÇODIC dictionary in 14 days (**Supplementary §S2**).

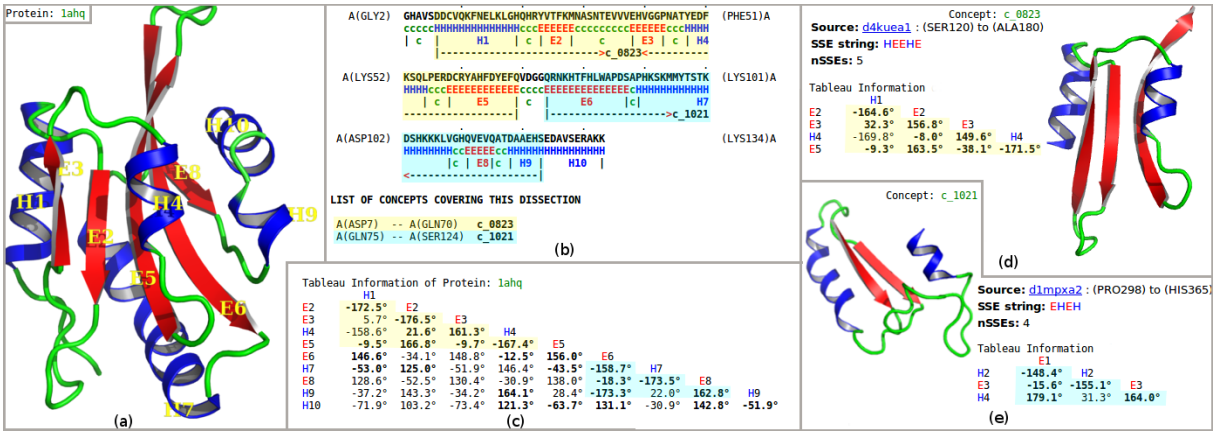

Fig. 1. (a) Secondary-structural cartoon representation of the crystal structure of the Actin-binding protein actophorin from *Acanthamoeba* (1AHQ) (32). (b) Secondary structural assignment (using SST (33); H = helix, E = strand of sheet) and the optimal dissection of the protein chain into non-overlapping regions, using the inferred concept dictionary. This information is shown with reference to the amino-acid sequence information in a marked-up format: the dissection of 1AHQ uses *concepts* (see text) c_0823 (highlighted in yellow) and c_1021 (highlighted in blue). (c) Tableau representation of the folding pattern of 1AHQ. The highlighted subtableaux correspond to concepts c_0823 and c_1021. Here, only the lower-triangle part of the tableau information is shown because the full tableau is a symmetric matrix. The rows and columns are indexed by secondary structure elements in order of appearance in the polypeptide chain. Off-diagonal elements record the angles between pairs of secondary structural elements; boldface indicates that there is a contact between the corresponding pair of secondary structural elements. (d-e) The concepts c_0823 and c_1021 are shown, together with their *archetypal* tableaux and corresponding secondary structural representation.

PROÇODIC: The dictionary of inferred concepts. Each of the 1,493 concepts in the dictionary is designated by an identifier of the form 'c_' followed by 4 digits: c_0001—c_1493. This order follows (1) the decreasing length in the number of secondary structural elements (nSSEs) defining each concept, and (2) the lexicographic order of their secondary structural strings, where we represent any helix by 'H' and any strand by 'E'.

Fig. 2 shows the top 100 concepts in the dictionary. The largest concept (c_0001) contains 28 secondary structural elements. The smallest concepts (c_1441—c_1493) – not shown in Fig. 2 – contain only 2 elements. (Note that a single helix or a single strand/extended region is not considered here as a concept.) The distribution of inferred concept sizes is shown in Fig. 3(a): 9 concepts (c_0001—c_0009) are composed of an assembly of $\geq$ 20 secondary structural elements (SSEs), 48 concepts (c_0010—c_0057) have between 15 and 19 SSEs, 217 concepts (c_0058—c_0274) contain between 10 and 14 SSEs, 217 concepts (c_0058—c_0274) contain between 10 and 14 SSEs, and 368 concepts (c_0275—c_0642) contain between 9 and 6 SSEs. The remaining concepts contain between 5 and 2 SSEs. The median concept size is 5 SSEs.

The complete inferred dictionary is available via the interactive website PROÇODIC (for Protein Concept dictionary – the cedilla allows the pronunciation as 'prosodic') at http://lcb.infotech.monash.edu.au/prosodic. As discussed below, this site allows the exploration of any structure the user provides as input, or specific concepts that are of motivating focus for the user, including: the usages of concepts in other structures,

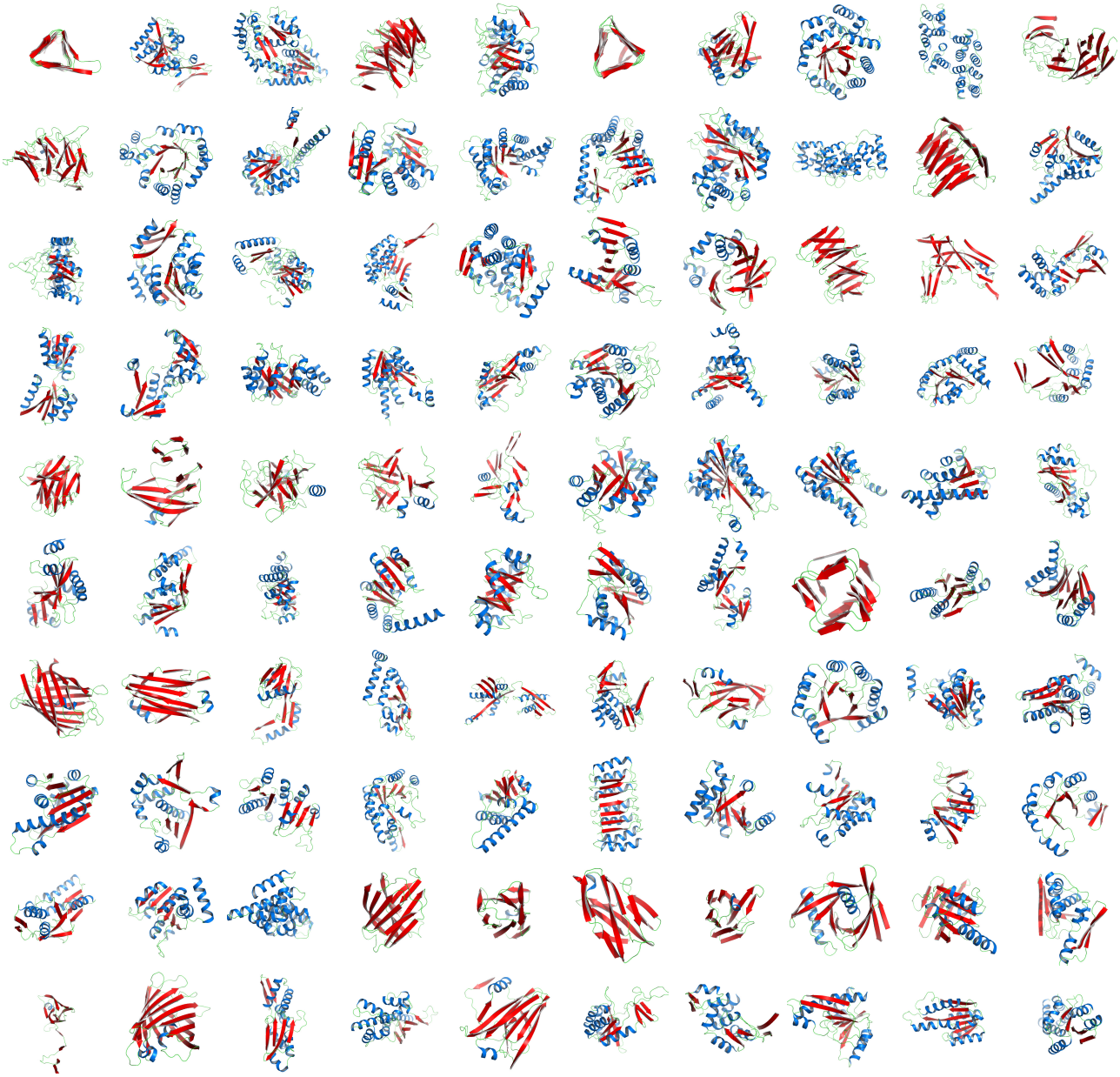

Fig. 2. Representative structural cartoons of the top 100 concepts from the inferred dictionary containing 1,493 concepts, ranked in decreasing order of number of secondary-structure elements (row-wise top-left to bottom-right: c_0001 to c_0100). Strands of sheet are shown in Red; helices in Blue. (See website for the full interactive listing.) The inference of the whole dictionary is automatic without any prior knowledge or preconceived notions of these recurrent themes. The inferred concepts subsume known patterns, for example: 'α-β Barrel' (c_0005), 'Armadillo repeat' (c_0083), 'β Barrel' (c_0061), 'β Propeller' (c_0004), 'Icosahedral (Virus)' (c_0067), Immunoglobulin (c_0062), 'Jellyroll architecture' (c_0084), 'Left-handed β-Helix' (c_0001), 'Leucine-rich repeat' (c_0076), 'Right-handed quadrilateral β-Helix' (c_0058) 'NAD-binding domain' (c_0002), 'TIM barrel' (c_0008) etc. Other classical supersecondary structures not shown in this figure such as β-hairpin (c_1442), α-hairpin (c_1484), β-α-β unit (c_1240) appear lower down in the dictionary of concepts, ordered from largest to smallest. See text (Page 5) where classical supersecondary structural motifs are discussed.

both homologous and non-homologous; or the inspection of frequently occurring keywords within the 'KEYWDS' records and the ligand-binding information from the 'HETATM' records extracted from the source wwPDB coordinate files (see **Discussion**).

Our dictionary subsumes known supersecondary structural motifs. Our automatically identified dictionary includes many concepts that match the known repertoire of supersecondary structural motifs (34). Matched motifs involving assemblies of a small number of helices and strands include: antiparallel (c_1442) and parallel (c_1443) β-β assemblies, α-α hairpin (c_1484) α-β/β-α assembly, (c_1459/c_1472), basic helix-loop-helix (c_1351), β-α-β motif (c_1240), EF-Hand (c_1342, c_1491), ϕ-motif (c_1178), helix-turn-helix motifs (c_0826 – winged type I, c_0870 – winged type II, c_1373 – plain), four-helix bundle (c_1101 – type I, c_1117 – type II), β-meander (c_1187), Greek key (c_0964), Zinc finger (c_1230), helix-hairpin-helix motif (c_1068), β-sandwich (c_0390), and $\alpha\beta$-sandwich (c_0603) among others.

Our dictionary also includes larger assemblies of helices and strands that match known *repeating* structural motifs. These include: three-sided left-handed β-helix (c_0001, c_0380), three-sided right-handed β-helix (c_0388), right-handed quadrilateral β-helix (c_0058), ankyrin repeat (c_0370, c_0632), armadillo repeat (c_0083, c_0888), kelch repeat (c_0395), α-solenoid (c_0270, c_0271), and Leucine rich repeat (c_0076) among others.

To understand the relationships between different concepts, we have clustered hierarchically the 1,493 PROÇODIC concepts. Since each concept archetype defines a (sub)tableau derived from a tableau of the domain in the source collection, we start by inferring the dictionary of *meta-concepts* that best explains all the PROÇODIC concept tableaux. This is achieved by using exactly the same unsupervised inference methodology that was used to infer PROÇODIC concepts. That is, we now treat the tableaux representing 1,493 archetypes from our inferred prosodic concept dictionary as the source collection, and rerun our inference method (**Supplementary §S2**). This yielded 34 meta-concepts that dissect (i.e., best explain) the inferred 1,493 concepts. The text file containing these meta-concepts, along with the corresponding list of PROÇODIC concepts that use each meta-concept within their dissections, is available in the supporting data file: metaConceptsAndUsageList.txt (click).

Once the meta-concepts are obtained, we dissect each of the 1,493 concepts using these meta-concepts. This permits a 34-dimensional feature vector representation of concepts, where each vector-component denotes the number of times the corresponding meta-concept is used in that concept dissection. We note that this representation is similar to the bag-of-words model (35) used in information retrieval and natural language processing. Using this feature vector representation, the 1,493 PROÇODIC concepts are clustered hierarchically by:

1. Constructing a $1,493 \times 1,493$ similarity matrix that compares all pairs of these 34-dimensional vectors, and

2. Using the resultant similarity matrix to cluster all PROÇODIC concepts hierarchically, based on the unweighted pair-groups method using arithmetic averages (36).

This procedure yielded a hierarchical tree of concept relationships, available in interactive format from: prosodicConceptClustering.html (click). This tree reveals similarities that are also detectable by examining the concept archetypes, their usages and keywords. For example, c_0009 and c_0018 are both helical bundles related to the architecture of Annexin proteins, with c_0009 having one extra helix compared to c_0018. Another example is the cluster containing c_0001, c_0006, c_0113, and c_0380, where all represent left-handed β-helical motifs composed of 28, 20, 12 and 7 β-strands, respectively.

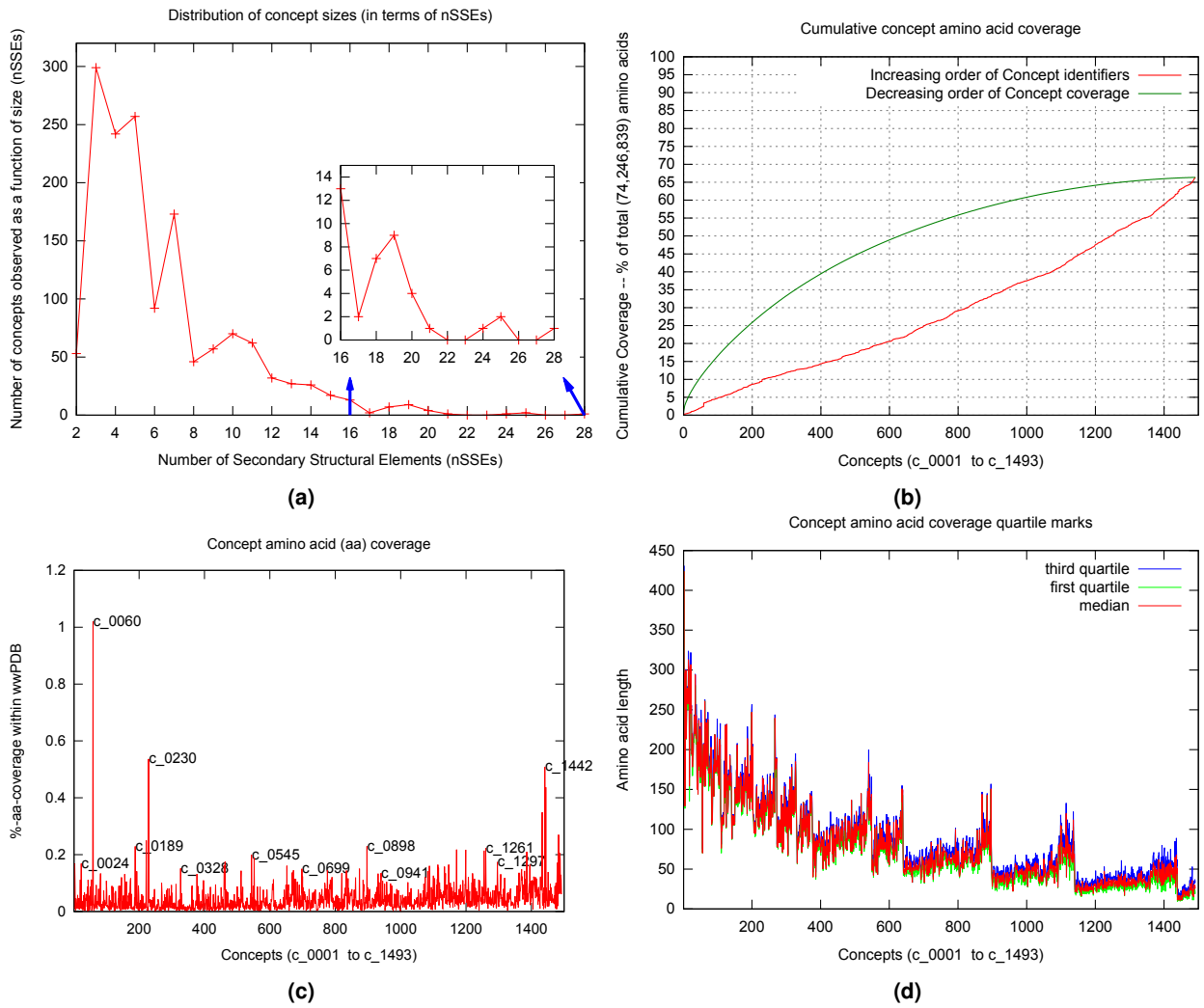

Fig. 3. (a) Distribution of concept lengths in terms of the number of secondary structural elements (nSSEs) they contain. The smallest concepts have 2 secondary structural elements; the largest has 28. (b) Cumulative amino acid coverage of concepts (as a percentage of the total 74,246,839 number of residues) after dissecting 275,014 protein chains using the inferred dictionary. The green curve gives the distribution in the decreasing order of individual concept amino acid coverage – i.e., the concept with largest coverage is listed first, that with second largest coverage is listed second, and so on. The red curve gives the same cumulative distribution in the serial order of concept identifiers – i.e., concept `c_0001` is shown first, concept `c_0002` second and so on. (c) Individual concept amino acid coverage (as a percentage of the total 74,246,836 residues) in the serial order of concept identifiers (with some concepts highlighted). (d) Dissecting the protein chains from the `wwPDB` allows us to catalogue the regions where each concept is used. Underlying each concept usage is an amino acid sequence of variable length (although the associated strings corresponding to the types and order of secondary structural elements match exactly). This graph plots the first, second and third quartile points in the distribution of amino acid lengths for each concept's set of usages in the `wwPDB`. Concepts are listed in the decreasing order of lengths, followed by the lexicographic order of their secondary structural strings. Since the average strand of sheet (denoted as 'E') has fewer amino acids than the average helix (denoted as 'H'), the lexicographic order creates in the plot the observed piecewise increasing trend among concepts with same number of secondary structural elements.

Although our average concept archetype is significantly smaller (with 47.6% of the number of SSEs) than its source protein domain, several concepts inferred in our dictionary describe conserved folding patterns at the level of domains. These include: NAD-binding domain (c_0173), β-grasp fold (*e.g.* c_729), β-propeller (c_0382), Swiss/Jelly roll fold (c_0406), Ferredoxin (plait) fold (c_0581), TIM barrel (c_0008), Immunoglobulin fold (c_0118, c_0121), Ubiquitin roll (c_0737), and large β-barrel (c_0061).

These results show that our dictionary encompasses a broader set of substructural invariants across the protein folding space than previous studies achieved (see section 'Comparison with previous related work'). This advantage is mainly due to our use of tableaux to capture concisely the essence of protein folding patterns, and our use of the information-theoretic criterion of minimum message length to yield an objective dictionary complexity-versus-fidelity trade-off.

Dissection of wwPDB and coverage of concepts across the protein folding space. The methods used for this work also permit the optimal *dissection*, within seconds, of any protein chain into non-overlapping regions that are explained (compressed) using the concepts from the inferred dictionary. Regions not assigned to any dictionary concept (notionally designated to the *null* concept, c_0000) remain uncompressed. These include the small set of proteins that have no secondary structure, for instance wheat-germ agglutinin (9WGA). Fig. 1 shows an example of the dissection of the crystal structure of the Actin-binding protein actophorin from *Acanthamoeba* (1AHQ). (See PROÇODIC website to dissect any protein structure of interest.)

We have dissected the entire wwPDB, which at the time of calculation resulted in tableaux corresponding to 275,014 protein chains containing 74,246,839 amino acid residues overall. (Note that the dictionary was constructed using an *unbiased* set of domains from ASTRAL; but the subsequent dissection of the entire wwPDB reflects the biases in the distribution of protein folding patterns in the full database.) The usages of the resulting concepts cover regions within proteins that account for 66.35% (49,262,577) of the total (74,246,839) amino acids in the wwPDB protein chains we dissected. The remaining 33.65% is dominated by single secondary structural elements, plus loops between successive concept assignments along a dissected chain. Fig. 3(b-d) show the distributions of amino acid coverage of concept usages within the wwPDB. Concept c_0060 (click) has the largest coverage in terms of the number of amino acids its usages cover. This concept is composed of 14 secondary structural elements (SSE string: EEEEHHEEEEHHEE) assembling into a four layer architecture, with its core containing two layers of closely-packed five-stranded β-sheets (37) that are sandwiched between two outer layers, containing two α-helices each. In total, this concept was used within 3,892 protein chains, with a median value of amino acid coverage equal to 194 residues. Examination of these usages reveals that they come from the protein chains of 285 proteasome complexes. At the other extreme is concept c_0568 (click), which has the smallest amino acid coverage: 561 residues over 13 protein chains related to plant and bacterial Ferredoxins (38). This concept is composed of 6 secondary structural elements (SSE string: EEHEEE).

Novel insights regarding the concepts can be gained from their usage information. For example, consider the concepts c_0060 and c_0568 mentioned above: the concept c_0060 covers the β5 subunit of a recently solved structure of the native human 20S proteasome at 1.8 Å resolution (5LE5) (39). This landmark study revealed a number of functionally important differences with respect to what was known from the previously published 20S proteasome structures. In particular, it identified chloride ions within all active sites, thus significantly revising the description of the proteasome active site, and providing new insights into peptide hydrolysis that underpin the '*development of next-generation proteasome-based cancer therapeutics*' (39). Examination of the usages of c_0060 within the dissection of 5LE5 (chain Y – β5 subunit), reveals that this concept is directly linked to proteolytic active sites (Fig. 4(a)). Analyses

of the human-annotated keywords used in the wwPDB coordinate files from these usages showed among its top 10 frequently used phrases terms such as 'Cancer (therapy)', 'Drug resistance' and 'Bortezomib' – an anti-cancer drug and the first therapeutic proteasome inhibitor to be used in humans (see PROÇODIC). This is strong evidence of the concept being linked to a proteolytic active site. A similar examination of the usage instances of the concept c_0568, directly links it to the Fe_2S_2-cluster binding Ferredoxins (see Fig. 4(b)), which mediate electron transfer (40).

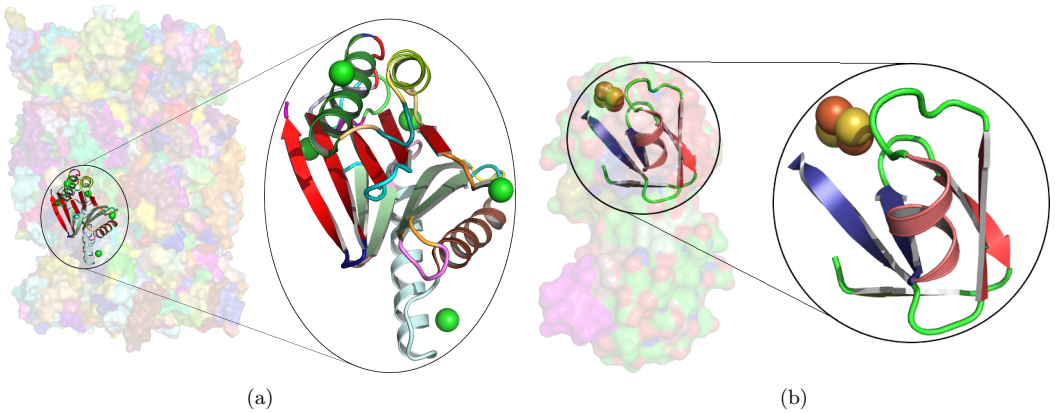

Fig. 4. (a) Transparent surface rendering of the native human 20S proteasome at 1.8 Å (5LE5), with the usage of concept c_0060 in the $\beta 5$ subunit (chain Y in the amino acid region THR1 to ASN191) shown in cartoon. The closeup of this region reveals a chloride ion in all active sites. Chloride ions are known to facilitate a proton shuttle catalytic mechanism (39). (b) Similar rendering as above for the usage of concept c_0568 in the 2.3 Å Ferredoxin structure from *Mastigocladus laminosus* (3P63 chain A in the amino acid region THR48 to GLU90). The closeup shows the region linked to the Fe_2S_2-cluster binding.

Comparison with previous related work. Many previous studies have attempted to identify a canonical set of recurrent patterns within the structures of proteins. Among the earliest such studies is the seminal work by Unger and colleagues (13), who demonstrated that most hexapeptide fragments in the (then) known proteins are structurally similar to a set of about 100 representative fragment types, using a normalised root mean square devision (RMSD) based clustering methodology. Much subsequent work have followed novel variations of this RMSD-based clustering strategy involving short oligopeptide fragments (of differing length) to produce different fragment libraries (8, 16–19, 41–43). Among other noteworthy approaches to roster oligopeptide fragments are the investigations of Baker and coworkers, who studied the distributions of local structures by clustering short amino acid sequence information (44, 45). Their fragment libraries, together with the inferred local sequence-structure relationships, now underpin successful and popular *ab initio* structure prediction methods (46). Further, Nepomnyachiy et al. (23) recently proposed a pipeline to explore reuse of regions in proteins based on their amino acid sequence relationships. This work reported reuse of sequence segments between 35 and 200 amino acids in length. However, relying on amino acid sequence relationships to identify reuse is rather limiting, as sequences diverge more drastically than structures in evolution (6).

Focusing on the methods that rely on structural information, Grishin and colleagues (20) recently proposed a method to enumerate constructively all idealised *parallel/antiparallel* arrangements of up to 5 SSEs. This work proposed a systematic enumeration of all possible parallel/antiparallel arrangements using a 3D lattice model. This allowed them to model theoretical arrangements of SSEs and use them to search for observed occurrences of each arrangement within the PDB. However, their idealised models are limited to parallel/antiparallel orientations, which poses a considerable restriction in exploring the full set of SSE arrangements observed in the PDB.

Furthermore, two new types of motif libraries have also been recently proposed: the Smotif library (21) and the TERMs library (22). An Smotif is designated by the arrangement of a *pair* of SSEs (of one of the following types: EE, EH, HE, and HH). A library of Smotifs is a collection of such SSE-pairs with different geometries. Their work utilises an RMSD theshold of 2.5Å to cluster 11,068 observed pairs of SSEs from a collection of 1,200 protein structures (i.e., one randomly-chosen protein domain per SCOP fold). These fragments serve in their work as the representatives of the protein structural space. Thus, any consecutive pair of secondary structures within a protein chain is assigned to the closest (based on RMSD) representative Smotif.

Beyond these basic pairwise assemblies identified by the Smotif library, the tertiary motif (TERMs) library (22) was able to find bigger assemblies of short oligopeptide fragments using the following approach. For each amino acid residue i in the non-redundant collection of 29,000 residues, a candidate TERM is defined using one or more oligopeptide fragments formed by the union of the residues $i-2, \ldots, i+2$ together with all penta-peptide regions around residues that form a 'potential contact' with the residue i. For each candidate TERM, the method finds matching tertiary fragments using an RMSD-based search method. A subset of candidate TERMs is realised by posing it as the classical set cover problem and realising the minimal cover using a greedy approximation method that iteratively identifies the TERMs (based on their coverage) that match proteins in the considered set. This iterative procedure yields about *half a million* (458,251) TERMs. The minimum TERM has 1 oligopeptide fragment containing 5 amino acids, while the maximum TERM has 10 fragments with 52 amino acids. Importantly, an average TERM in their library is composed of 3 oligopeptide fragments covering 19 amino acids (i.e., 6 amino acids per fragment). Furthermore, inspecting the TERMs that cover 50% of their proteins in their considered collection of 29,000 protein structures, we find that each TERM averages 2 fragments with 12 amino acids. Moreover, inspecting the top 24 TERMs (see Fig 2A of (22)), we find many repetitions of short helices and antiparallel strands.

In comparison, our work results in only 1,493 architectural concepts (two orders of magnitude more concise than TERMs), where our minimum concept is composed of 2 SSEs whose median usage in the PDB covers 19 amino acids, while the maximum is composed of 28 SSEs whose median usage covers 171 amino acids. An average prosodic concept in our dictionary is composed of 6 SSEs covering 75 amino acids. Considering the prosodic concepts that cover 50% of the PDB, an average concept has 5 SSEs covering 66 amino acids. Thus, using this framework, our dictionary yields concepts that are a substantially larger than TERMs, and define a significantly more economical dictionary that explains the entire PDB. Moreover, our methodology defines a direct and efficient (dynamic programming based) method to dissect any given protein structure using the inferred PROÇODIC dictionary.

These results are achieved due to the expressive power of tableaux to compactly represent the essence of protein folding patterns. This tableau-representation together with the statistically rigorous minimum message length inference methodology, provides a significantly more powerful framework to losslessly compress and identify redundancies in the protein folding space.

Discussion

Many concepts are linked to ligand-binding sites. The molecular function of proteins is often mediated via interactions with chemical components such as metal ions, coenzymes, metabolic substrates, and nucleic acids, amongst others. Knowledge of such interactions is central to annotate protein function (47, 48), engineer new proteins (49), and design novel drugs (50, 51). These functionally critical interactions impose structural constraints on protein structures, as their domains evolve from a common ancestor. As noted by Lesk and Chothia (52), in many cases active sites are the best-conserved regions within a family of protein structures (as seen in Fig. 4).

We have analysed our dictionary and systematically identified concepts directly related to protein-ligand interactions. To achieve this, we mined and catalogued frequent-ligand information (from 'HETATM' records) derived from the source wwPDB entries of each concept usage (i.e., each instance in the wwPDB where the concept is used in the dissection of that protein's tableau). Our definition of a *ligand* comes from the inventory of 23,258 chemical components specified by the LigandExpo (53) database. We note that this inventory does *not* exclude simple monovalent ions (such as Na^+, $K^+$and Cl^-) or those that are often not biologically functional (such as sulphate SO_4^{2-} ions). To complement this information, we also mined and catalogued keywords (from 'KEYWDS' records) derived again from the source wwPDB entries of these concept usages. We used the observed frequencies of the bound ligands within the regions of concept usages, to narrow the initial set down to the 463 (31%) concepts that stand out in terms of recurrent patterns of interactions with the same set of ligand(s). These encompass interactions with monovalent ions, di-/tri-/tetra-valent ionic species, small molecules (including nucleotides), and macromolecular compounds, among others.

The full annotated list of concepts with observed interactions with ligands/chemical components is available in the supporting data file: conceptsWithLigandInteractions.txt (click).

Fig. 5 shows the distribution of 69 distinct chemical interactions observed within the shortlisted set of 463 concepts. Fig. 6(a-g) shows examples of concept usages for a random selection of 8 concepts associated with metal-binding activity. Table 1 shows a partial list of concepts for which all (100% of) their usages show binding to the specified ligand/chemical components. Also shown are the extracted high-frequency keywords associated with usages of that concept, providing useful insights to impute functional roles. Among the shortlisted set of 463 concepts are those that demonstrably show binding specificity linked with target recognition, reception and signalling (see Table 2).

The full list of inferred concepts putatively linked to molecular reception, recognition, and signalling is available in the supporting data file: receptorConcepts.pdf (click).

Table 1. A partial list of concepts for which all (100% of) their usages show interactions with ligand/chemical components. This is derived by inspecting the ligand ('HETATM') records within the source coordinate files of each concept usage. The bound ligands are shown (in the second column) using their standardized abbreviations, along with their observed frequency within the usages in parentheses. Also shown (in the third column) are the top keyword terms (from 'KEYWDS' records specified by the structures' authors) recurring within the usage coordinate files with their associated frequencies. (Note: CA = Calcium ion)

Concept ID	Ligand/Chemical component (freq)	Keyword (freq)
c_0011	PQQ (100%), CA (100%)	OXIDOREDUCTASE (90%), QUINOPROTEIN (27%)
c_0036	ZN (100%)	HYDROLASE (85%), EXOPEPTIDASE (46%), CARBOXYPEPTIDASE B (46%)
c_0065	FES (100%)	OXIDOREDUCTASE (96%), XANTHINE OXIDASE (32%), IRON SULFUR (30%)
c_0096	FMN (100%)	OXIDOREDUCTASE (100%), ROSSMANN FOLD (55%)
c_0108	HEM (100%), CA (100%)	OXIDOREDUCTASE (85%), PEROXIDASE (63%)
c_0110	HEM (100%)	OXIDOREDUCTASE (82%), MONOOXYGENASE (43%), CYTOCHROME P450 (34%)
c_0124	SF4 (100%), MG (100%)	OXIDOREDUCTASE (91%), [NIFE]HYDROGENASE (26%)
c_0144	CA (100%)	TRANSFERASE (81%), CGTASE (36%), ACARBOSE (33%)
c_0156	ZN (100%)	TRANSFERASE (90%), SET DOMAIN (39%), EPIGENETICS (28%)
c_0159	SF4 (100%)	OXIDOREDUCTASE (96%), NIFE HYDROGENASE (17%)
c_0208	CU (100%)	OXIDOREDUCTASE (97%), BETA BARREL (34%), LACCASE (32%)
c_0374	HEM (100%)	OXYGEN TRANSPORT (56%), HEMOGLOBIN (26%)
c_0397	ZN (100%)	OXIDOREDUCTASE (94%), SUPEROXIDE DISMUTASE (27%)
c_0424	PCA (100%)	HYDROLASE (95%), GLYCOSIDASE (35%), CELLULOSE DEGRADATION (32%)
c_0546	ZN (100%)	HYDROLASE (88%), PHOSPHODIESTERASE (32%), PDE (28%)
c_0568	FES (100%)	ELECTRON TRANSPORT (77%), FERREDOXIN (38%)
c_0604	HEM (100%)	ELECTRON TRANSPORT (100%), HEME (57%), CYTOCHROME B5 (40%)
c_0624	ZN (100%)	APOPTOSIS (47%), ZINC FINGER (44%), METAL BINDING (30%)
c_0714	NAG (100%)	VIRAL PROTEIN (84%), HEMAGGLUTININ (39%), GLYCOPROTEIN (22%)

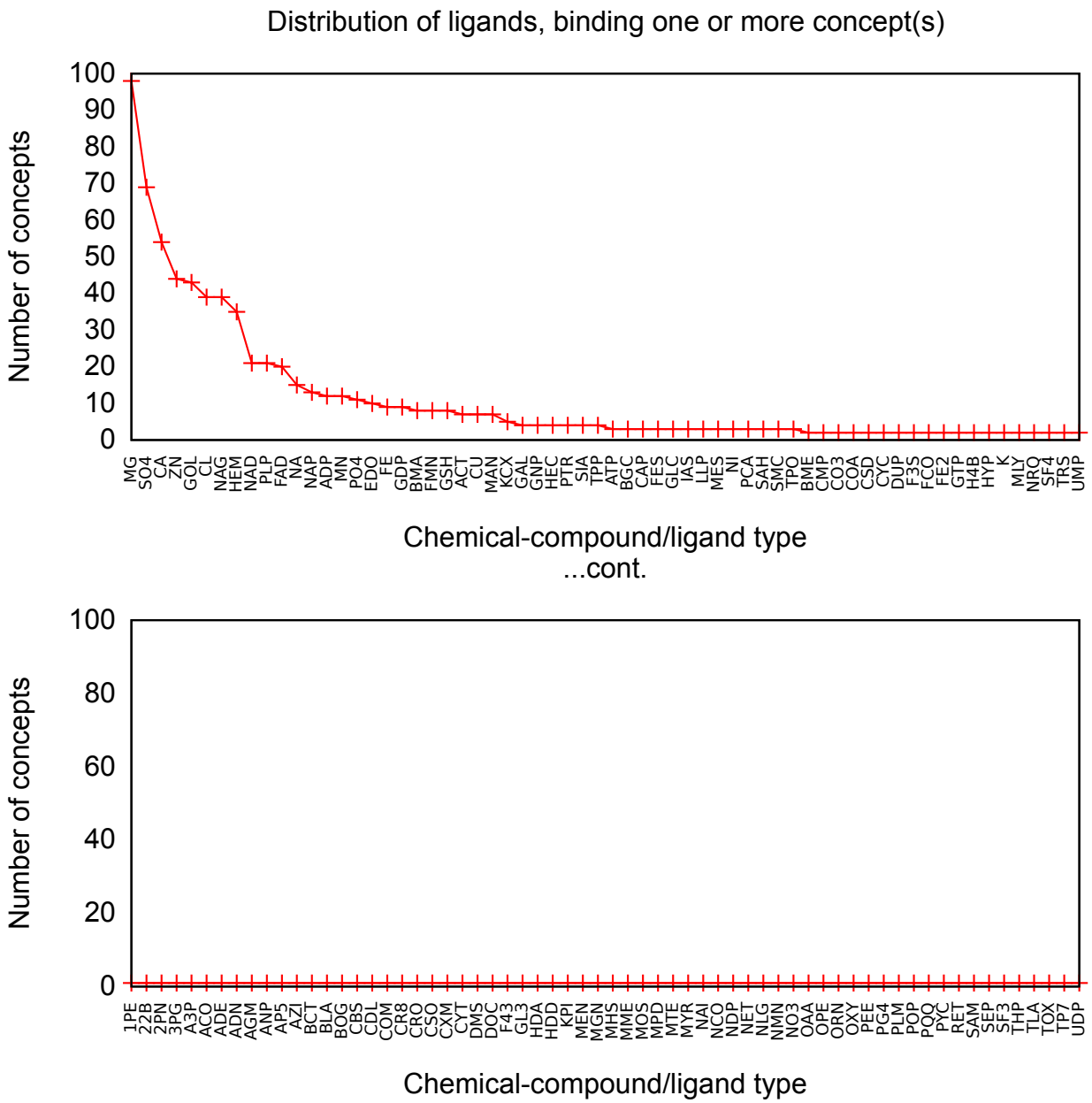

Fig. 5. Distribution of 128 distinct ligands binding a shortlisted set of 463 concepts. A concept from the dictionary is shortlisted to be correlated to a binding site if $> 30\%$ of its usages in the `wwPDB` bind to a common ligand. The full details are available in `conceptsWithLigandInteractions.txt` (click). For readability, this distribution is split across 2 plots in the decreasing order of the number of shortlisted concepts that each ligand binds. In the top figure, the distribution varies between 98 (far left: `MG`) and 2 (far right: `UMP`). In the bottom figure, all ligands appear exactly in one of the shortlisted 463 concepts.

Table 2. A partial list of concepts putatively linked to molecular reception, recognition, and signalling.

Concept ID	Ligand/Chemical component (freq)	Frequent Keywords (freq)
c_0062	NAG (96%), BMA (70%)	IMMUNE RECOGNITION (21%)
c_0133	ZN (35%)	AMPA RECEPTOR (26%), NEUROTRANSMITTER RECEPTOR (20%)
c_0205	GAL (36%)	CARBOHYDRATE RECOGNITION (11%)
c_0252	MYR (40%)	RHINOVIRUS COAT PROTEIN (20%), RECEPTOR (17%), ANTIVIRAL COMPOUND (10%)
c_0304	CA (60%)	ANTIBODY RECEPTOR (18%), CARBOHYDRATE RECOGNITION DOMAIN (15%)
c_0335	NAG (34%)	CELL ADHESION (29%), RECEPTOR (16%)GLYCOPROTEIN (11%)
c_0352	GOL (67%)	PEPTIDOGLYCAN RECOGNITION PROTEIN (10%)
c_0423	NAG (63%)	IMMUNE SYSTEM (87%),ANTIGEN PRESENTATION (26%), T CELL RECEPTOR (12%)
c_0572	FMN (67%)	PHOTORECEPTOR (36%), LIGHT INDUCED SIGNAL TRANSDUCTION (13%)
c_0819	ZN (58%)	SIGNALING PROTEIN (19%), PHOTORECEPTOR (13%)

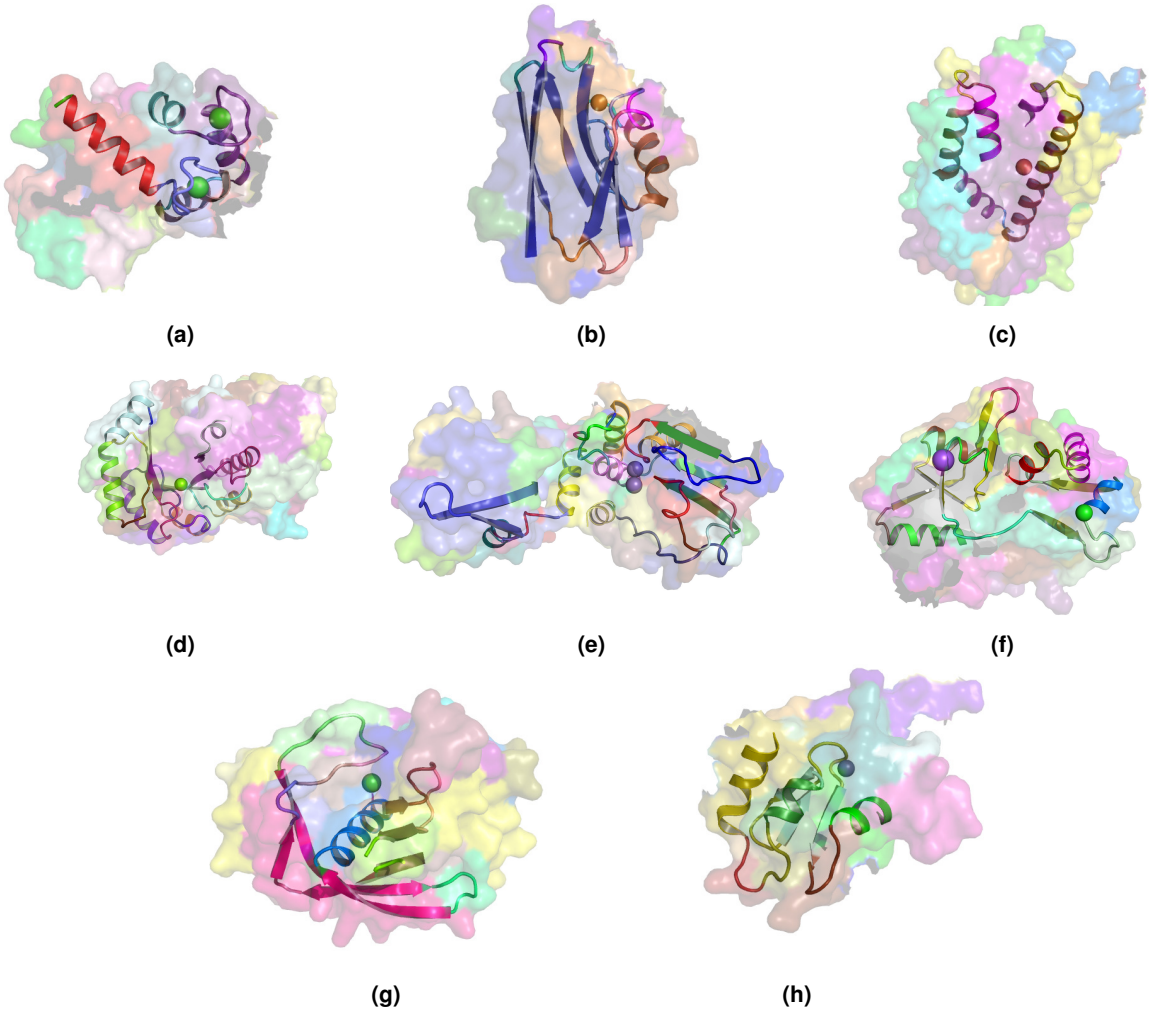

Fig. 6. Exemplars of usages of eight concepts linked to metal binding activity. The region of concept usage is shown in cartoon in the context of the surface rendering of the source protein chain. (a) Usage of concept c_1099 within the calcium-bound calmodulin (1CDL (54)). (b) Usage of concept c_432 within the coppper-bound electron transfer protein (1A4B (55)). (c) Usage of concept c_885 within the iron-bound oxidoreductase (2VUX). (d) Usage of concept c_139 within the magnesium-bound lyase (3TTE). (e) Usage of concept c_186 within the manganese-bound hydrolase (1K23 (56)). (f) Usage of concept c_133 within the sodium cation-bound Kainate and AMPA receptors (3G3G (57)). (g) Usage of concept c_280 within the nickel-bound peptide deformylase (2AIA). (h) Usage of concept c_624 within the zinc-bound Melanoma inhibiting anti-apoptotic protein (1OY7 (58)).

Inferring biological function from concept usage information. Many proteins are deposited into the wwPDB with unspecified functional annotation, especially those coming from structural genomic initiatives. Functional characterization of such proteins is of crucial importance to the structural biology community. Its importance can be evidenced by the community-wide Critical Assessment of protein Function Annotation programme (CAFA, biofunctionprediction.org/cafa/), which assesses methods dedicated to predicting protein function from amino-acid sequence.

As previously shown (Fig. 4(a)), the rich source of information within this concept dictionary is useful to investigate and impute biological function. More evidence of this is shown by another case study involving the haze-forming thaumatin-like protein in white wines made from *Vitis vinifera* (4JRU containing 201 residues). Fig. 7 gives the dissection of 4JRU composed of two concepts c_0111 and c_1442.

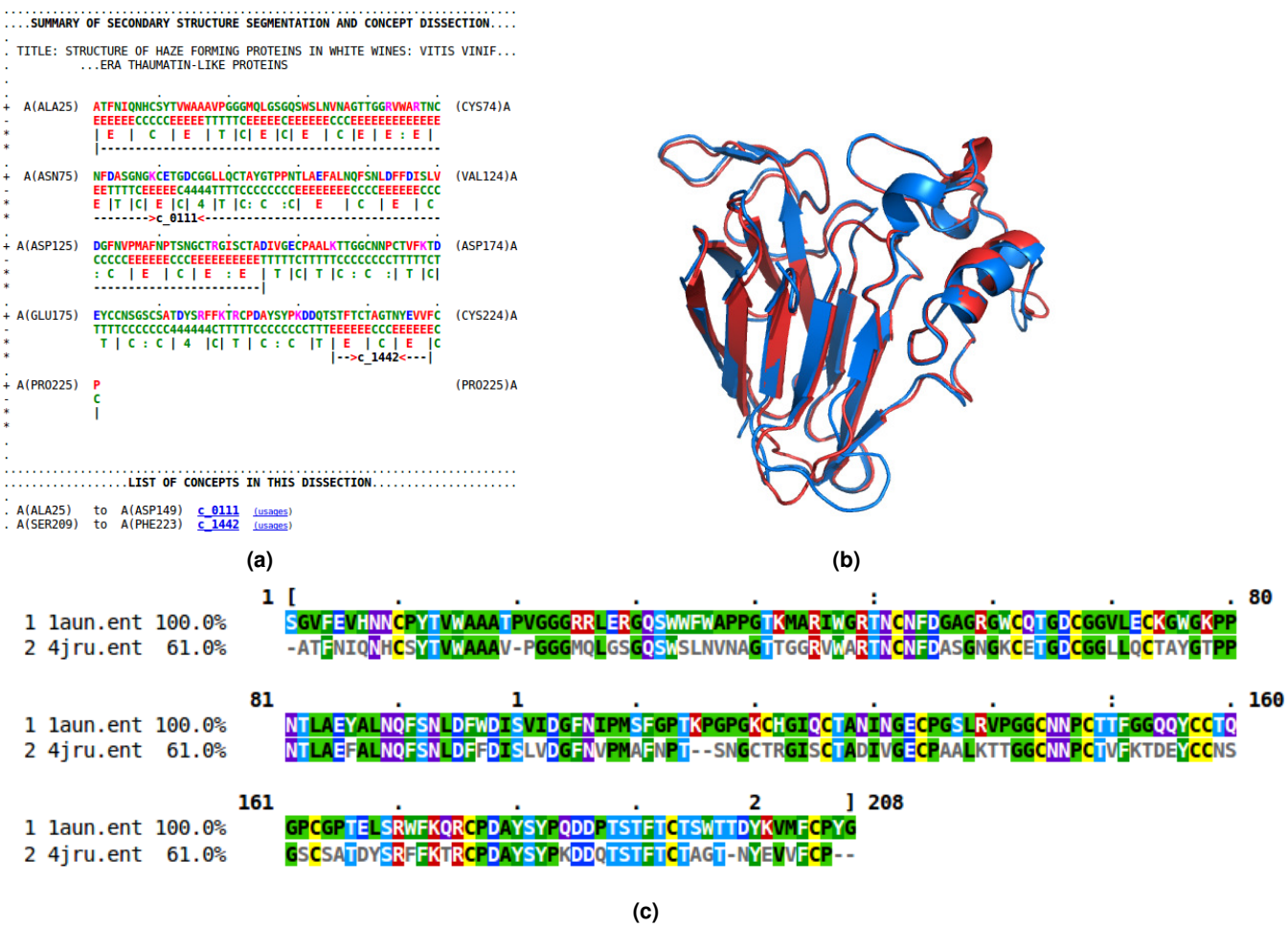

Fig. 7. (a) Dissection output from PROÇODIC of the haze-forming thaumatin-like protein in white wines from *Vitis vinifera* (4JRU). (b) Superposition of this haze-forming protein and the pathogenesis-related PR-5d protein of tobacco (*Nicotiana tabacum*; 1AUN). 4JR is shown in blue; 1AUN is shown in red. This superposition was based on the structural alignment produced by MMLigner (59), which is shown in (c).

Concept c_1442 is of less functional interest as it defines a common β-hairpin unit consisting of two antiparallel β-strands. On the other hand, c_0111 contains 12 strands that assemble to form mainly two face-to-face packed antiparallel β-sheets with an extended β-ribbon connected by an Ω-loop (60). This multi-stranded motif is characteristic of thaumatin-like proteins (61). Examining the usages of this concept within the wwPDB via our PROÇODIC web site, we find it is used at 15 other loci, most of them thaumatin/osmatin-like proteins, with their top two keywords displaying 'antifungal protein

(53.3%)' and 'plant protein (46.7%)', respectively. Fig. 7(b-c) show the structural alignment of 4JRU with the usage in the pathogenesis-related PR-5D protein of tobacco (*Nicotiana tabacum*; 1AUN with 208 residues) that results in a superposition with 1.47 Å root-mean-square deviation (r.m.s.d.) over 201 amino acid residues between the C_α coordinates of the two structures. This specific PR-5D protein is classified functionally as an antifungal protein and, in general, proteins of this class have known pathogenesis-related antifungal activity. This suggests that the haze-forming protein might exhibit the same biological function.

In some cases, the information provided by this dictionary can lead to a reliable but less-specific functional classification prediction, for example putatively identifying a general type of function such as 'Oxidoreductase' or 'Lyase'. Such generic functional classification can be useful, as it may provide guidance for laboratory experiments aimed at defining the function more precisely, especially if clues about a ligand-binding site are available. For example, consider the crystal structure of dihydrodipicolinate synthase (DapA) from *Agrobacterium tumefaciens* (2HMC). The dissection of this DapA structure shows the usage of concept c_0008 covering its entire chain A. About 90% of c_0008's 118 usages show the functional classification as 'Lyase'. DapA belong to the family of amine-lyases that catalyze the cleaving of carbon-nitrogen bonds, playing an important role in lysine biosynthesis in prokaryotes, phycomycetes, and plants (62). A similar example would be the identification of HI0073 from the *Haemophilus influenzae* structural genomics project as a nucleotide-binding protein (63).

Local sequence-structure correlation within concept usages. The identification of structural features that have strong amino-acid sequence preferences is central to structure prediction (45). Therefore, we studied the concept usages within the wwPDB to explore the conformational preferences of local sequences. To achieve this, for each concept, the amino acid sequences in the regions of concept usages within the wwPDB were extracted, and the sequences in each set were aligned and clustered (64).

Almost 20% of the concepts in our dictionary (288 out of 1,493) have associated amino acid sequences that cluster into a single group. Fig. 8(a) shows the number of clusters produced for each set of concept usage sequences – in general, the fewer the clusters, the stronger the local sequence-structure relationship. When considering the (normalized) ratio of clusters over the number of *non-identical* amino acid sequences of concept usages (Fig. 8(b)), almost 30% of the concepts (441) have a ratio smaller than 0.05, while almost 50% (738) have a ratio between 0.05 and 0.1. Together, this indicates that for almost 80% of the concepts in our dictionary, their usages of amino-acid sequences cluster into a small number of groups ($< 10\%$ of their total unique amino acid sequences).

This strong sequence dependence is expected, particularly for concepts linked to ligand binding or other functional units. For example, Fig. 9 shows the sequence logo obtained from the multiple sequence alignment of the usage sequences of the concept c_0397. This concept is related to the Cu-Zn type I (SODI) superoxide dismutase, which has a $\beta-$barrel-like subunit with copper and zinc ions bound at the active site. This is common in many Gram-negative bacterial pathogens (amongst others) to counteract a burst of toxic superoxide radicals under oxidative stress (66).

There is a potential application of the observed sequence-structure correlations to structure prediction. We downloaded the coordinate files of 33 wwPDB structures specified in the description field of the CASP12 target list available at http://predictioncenter.org/casp12/targetlist.cgi. Each chain from these 33 structures was independently dissected using the PROÇODIC dictionary of concepts. The dissection of protein chains defines non-overlapping regions assigned either to one of the dictionary concepts (c_0001 – c_1493), or a *null* concept (c_0000). For each region assigned to a dictionary concept, we extracted the associated target amino-acid sequence and performed a pairwise *sequence* alignment with each of the local amino acid sequences defined by the concept usages. This exercise identified a subset of concept usages in the wwPDB whose local amino acid sequences have a detectable similarity with the target. Table 3 quantifies the extent

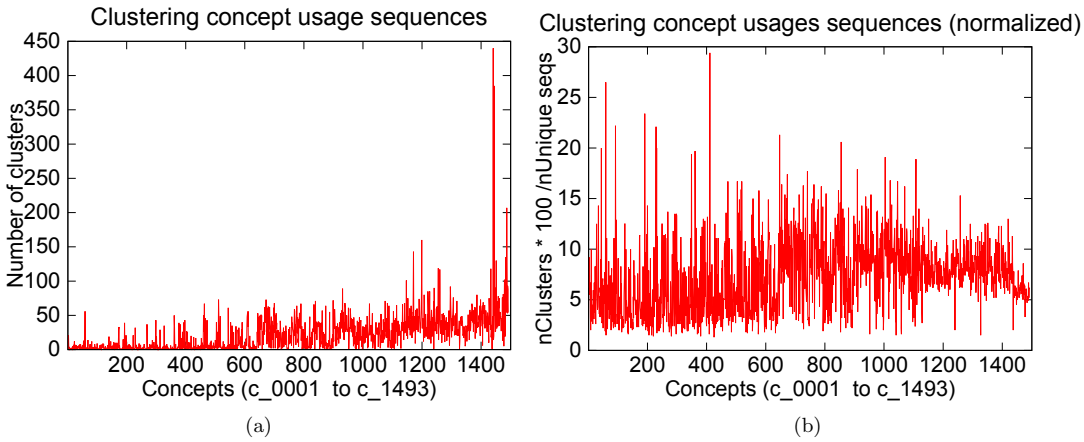

Fig. 8. (a) Number of clusters produced by `Clustal-Omega` (64) based on its computation of the multiple sequence alignment of the usage-amino-acids-sequences for each concept. `Clustal-Omega` uses the `mBed` algorithm to cluster sequences (65). (b) Normalised plot to account for the differences in the number of usages per concept. Normalization involves dividing the number of clusters by the total number of *unique* amino acid sequences observed in the set of usages per concept.

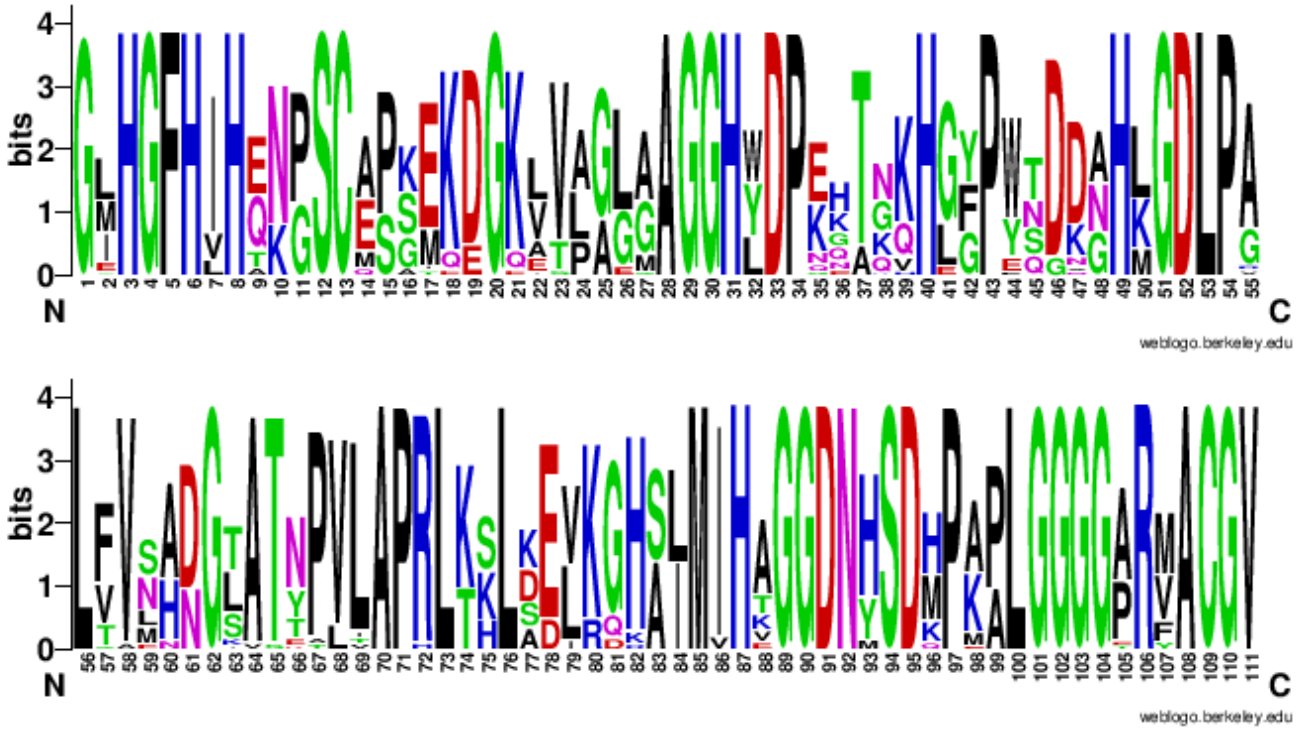

Fig. 9. Amino acid sequence logo (in two parts: columns 1-55 and 56-111) showing the sequence consensus across the usages of a randomly chosen concept `c_0397` directly related to the Cu-Zn binding superoxide dismutase. Of the 111 columns in the multiple sequence alignment (of the `c_0397`'s 33 usage sequences) corresponding to this logo, 46 aligned columns show a consensus of 100%.

Table 3. Statistics showing the extent of detectable sequence similarity on each of the 33 CASP12 targets with their wwPDBIDs specified at http://predictioncenter.org/casp12/targetlist.cgi. First column: wwPDBID of the 3D experimental structure of each CASP12 target. Second column: The coverage statistics in terms of the total number of amino acids (#a.a.) within the amino acid (sub-)sequences defined by the dissected regions of the target protein with detectable sequence similarity with amino acid (sub-)sequences of their corresponding concept usage instances (see main text). Third column: The total number of amino acids in the target protein, cumulative over all chains. Fourth column: Percentage coverage = Second column*100/Third column.

Target's wwPDBID	#a.a.'s in regions with usage seq hits	Total #a.a.'s in all chains	%Covered
3JB5	1046	2076	50.4
4YMP	202	215	94.0
5A7D	4468	5065	88.2
5AOT	91	102	89.2
5AOZ	125	141	88.7
5D9G	154	502	30.7
5ERE	417	540	77.2
5FHY	155	458	33.8
5FJL	88	136	64.7
5G3Q	100	168	59.5
5G5N	580	1022	56.8
5HKQ	160	263	60.8
5IDJ	63	242	26.0
5J4A	339	440	77.0
5J5V	815	1065	76.5
5JMB	103	182	56.6
5JMU	172	219	78.5

Target's wwPDBID	#a.a.'s in regions with usage seq hits	Total #a.a.'s in all chains	%Covered
5JO9	215	239	90.0
5JZR	203	262	77.5
5KKP	166	509	32.6
5KO9	73	253	28.9
5LEV	323	375	86.1
5M2O	171	211	81.0
5MQP	2674	4801	55.7
5NSJ	150	284	52.8
5NV4	713	1377	51.8
5SY1	786	1458	53.9
5T87	444	745	59.6
5TF2	331	338	97.9
5TJ4	2640	5462	48.3
5UNB	378	681	55.5
5UVN	954	2496	38.2
5UW2	211	332	63.6

of coverage of these regions for each of the 33 CASP12 targets. This table shows that in 26 of 33 cases, more than 50% of the target amino-acid chain has detectable sequence similarity that can be derived from the usage information.

It should be noted that we used the structural information of CASP12 targets to dissect the protein chains, before identifying the sequence relationships of the target sequence and those within the concept usages. However, for the proper application to structure prediction, the identification of sequence hits with concept usages should be carried out using only the target sequence. In principle, this can be done by sliding along the target sequence with varying window sizes, and exploring the sequence similarity with the sequences across all usages of every concept in the dictionary. Nevertheless, this preliminary analysis can be used to hypothesise reasonably that these local sequence-structure relationships provide a strong potential to support structure prediction efforts, especially since an average concept usage spans significantly longer stretches along the protein chain than the currently considered oligopeptide fragment libraries used by fragment-based *ab initio* protein modeling approaches. Thus, this information can be potentially utilized to model several non-overlapping regions in the target protein chains by the current state-of-the-art structure prediction servers (67–69).

The amino acid subsequences of non-overlapping regions dissected using the PROÇODIC dictionary of concepts is available at: casp12_prosodic_dissections.tgz (click). The information of dissected target region followed by other subsequences in the usages of the corresponding (assigned) concept with demonstrable sequence similarity (under pairwise sequence alignment with the target subsequence) is available at: casp12_concept_usage_hits.tgz (click). The multiple sequence alignments (using MUSCLE (70) with default options) of the identified sequence hits is available at: casp12_concept_usage_hits_msa.tgz (click).

Exploration of substructures and structural relationships. In addition to the applications explored above, the dictionary can be used to complement standard protein structural studies. Researchers can approach the dictionary with a particular structure or family of structures in mind. For example, dissecting

the human haemoglobin (1HHO, chain A) at the PROÇODIC web site identifies the concepts c_0375, c_0894 and c_1410. Choosing one of the concepts, for example c_0894, its archetype is found in d1x9fd, a globin from the annelid *Lumbricus terrestris*. Note that related proteins can present dissections into different concepts. However, these concepts can still be related (see discussion on hierarchical clustering of concepts on Page 5). For example, c_0375 and c_0894 are related concepts linked to globins, with the former being more elaborate (with three extra helices) than the latter. Examining the corresponding concept 'usages' link on the PROÇODIC web site reveals that many usages of these related concepts appear in other globins. **Supplementary §S3** contains several examples of use of PROÇODIC to explore protein substructural similarities.

CONCLUSION

Most protein domains fold into geometrical assemblies of helices and strands of sheet. Our work has analysed the domains of known structures and inferred a 'basis set' containing 1,493 folding concepts into which folding patterns of domains can be dissected. The discovery of the dictionary was completely automatic, and unbiased by any previous structural analysis of folding patterns, or by any hidden sequence or structure-based patterns. The effectiveness of our inference method is validated by the discovered concepts, which subsume classic supersecondary structures (α-hairpin, β-hairpin, and β-α-β unit etc.), known repeat patterns (ankyrin, armadillo, kelch repeat etc.) and many other known patterns (β-propeller, Jellyroll, and Immunoglobulin architecture, among others). We note that the scope of the concepts in our dictionary far exceeds these known patterns.

The discovery of this dictionary allows us to dissect optimally the structures of any protein domain in seconds, and map dictionary concepts onto non-overlapping regions along its chain. Importantly, the dissections of domains into concepts provide a plethora of useful biological insights, including:

1. Understanding the fundamental components of protein folding patterns. Our dictionary of concepts will support innovative projects aimed at the analysis of protein structures.

2. Correlation, in many cases, of concepts with functions directly or indirectly, via ligand binding sites. This provides useful predictions in the case of proteins with known structure but unknown function.

3. Many concepts show amino-acid sequence correlation; that is, some conservation of sequence patterns. These results are applicable to protein structure prediction by suggesting conformations of local regions.

The results of dissecting all structures in the current wwPDB, or of dissecting a user-supplied set of protein coordinates, are accessible from the PROÇODIC web site: http://lcb.infotech.monash.edu.au/prosodic (click). This site supports the interactive exploration of protein structures and their relationships.

SUPPLEMENTARY MATERIAL

Supplementary information, including description of material and methods supporting this work, is available as a separate PDF.

ACKNOWLEDGEMENTS

This research is funded by Australian Research Council (ARC) Discovery Project grant (DP150100894). We thank Research Computing Centre, University of Queensland for the High-Performance Cluster Infrastructure that supported this project over the last 3 years. AML thanks the Medical Research Council

Laboratory of Molecular Biology for their hospitality during his sabbatical year. We thank Sureshkumar Balasubramanian for proofreading this work.

REFERENCES

1. Lesk A (2016) *Introduction to Protein Science: Architecture, Function, and Genomics*. (Oxford University Press), 3rd edition.
2. Pauling L, Corey RB, Branson HR (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* 37(4):205–211.
3. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences* 37(5):251.
4. Rao ST, Rossmann MG (1973) Comparison of super-secondary structures in proteins. *Journal of Molecular Biology* 76(2):241–256.
5. Lesk AM, Rose GD (1981) Folding units in globular proteins. *Proceedings of the National Academy of Sciences* 78(7):4304–4308.
6. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *The EMBO Journal* 5(4):823.
7. Richards FM, Kundrot CE (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins: Structure, Function, and Bioinformatics* 3(2):71–84.
8. Camproux A, Tuffery P, Chevrolat J, Boisvieux J, Hazout S (1999) Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Engineering* 12(12):1063–1073.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* 247(4):536–540.
10. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2013) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research* 42(D1):D310–D314.
11. Orengo CA, et al. (1997) CATH – a hierarchic classification of protein domain structures. *Structure* 5(8):1093–1109.
12. Kister AE, ed. (2013) *Protein Supersecondary Structures*. (Springer-Humana Press).
13. Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins: Structure, Function, and Bioinformatics* 5(4):355–373.
14. Rooman MJ, Rodriguez J, Wodak SJ (1990) Automatic definition of recurrent local structure motifs in proteins. *Journal of Molecular Biology* 213(2):327–336.
15. Unger R, Sussman JL (1993) The importance of short structural motifs in protein structure analysis. *Journal of Computer-Aided Molecular Design* 7(4):457–472.
16. Micheletti C, Seno F, Maritan A (2000) Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins: Structure, Function, and Bioinformatics* 40(4):662–674.
17. Kolodny R, Koehl P, Guibas L, Levitt M (2002) Small libraries of protein fragments model native protein structures accurately. *Journal of Molecular Biology* 323(2):297–307.
18. Friedberg I, Godzik A (2005) Connecting the protein structure universe by using sparse recurring fragments. *Structure* 13(8):1213–1224.
19. Joseph AP, et al. (2010) A short survey on protein blocks. *Biophysical Reviews* 2(3):137–145.
20. Chitturi B, Shi S, Kinch LN, Grishin NV (2016) Compact structure patterns in proteins. *Journal of Molecular Biology* 428(21):4392–4412.
21. Dybas JM, Fiser A (2016) Development of a motif-based topology-independent structure comparison method to identify evolutionarily related folds. *Proteins: Structure, Function, and Bioinformatics* 84(12):1859–1874.
22. Mackenzie CO, Zhou J, Grigoryan G (2016) Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences* 113(47):E7438–E7447.
23. Nepomnyachiy S, Ben-Tal N, Kolodny R (2017) Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proceedings of the National Academy of Sciences* 114(44):11703–11708.
24. de Oliveira SH, Deane CM, Valencia A (2018) Combining co-evolution and secondary structure prediction to improve fragment library generation. *Bioinformatics* 1:9.

25. Joshi RR (2018) Diversity and motif conservation in protein 3d structural landscape: exploration by a new multivariate simulation method. *Journal of Molecular Modeling* 24(4):76.
26. Lesk AM (1995) Systematic representation of protein folding patterns. *Journal of Molecular Graphics* 13(3):159–164.
27. Subramanian R, et al. (2017) Statistical compression of protein folding patterns for inference of recurrent substructural themes in *Data Compression Conference (DCC), 2017*. (IEEE), pp. 340–349.
28. Wallace CS, Boulton DM (1968) An information measure for classification. *The Computer Journal* 11(2):185–194.
29. Wallace C (2005) *Statistical and Inductive Inference by Minimum Message Length*. (SpringerVerlag).
30. Allison L (2018) *Coding Ockham's Razor*. (Springer).
31. Chandonia JM, Fox NK, Brenner SE (2017) SCOPe: Manual curation and artifact removal in the structural classification of proteins–extended database. *Journal of Molecular Biology* 429(3):348–355.
32. Leonard SA, Gittis AG, Petrella EC, Pollard TD, Lattman EE (1997) Crystal structure of the actin-binding protein actophorin from *Acanthamoeba*. *Nature Structural Biology* 4(5):369–373.
33. Konagurthu AS, Lesk AM, Allison L (2012) Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics* 28(12):i97–i105.
34. Efimov AV (2013) Super-secondary structures and modeling of protein folds in *Protein Supersecondary Structures*. (Springer), pp. 177–189.
35. Harris ZS (1954) Distributional structure. *Word* 10(2-3):146–162.
36. Sokal RR (1958) A statistical method for evaluating systematic relationship. *University of Kansas Science Bulletin* 28:1409–1438.
37. Chothia C, Levitt M, Richardson D (1977) Structure of proteins: packing of α-helices and pleated sheets. *Proceedings of the National Academy of Sciences* 74(10):4130–4134.
38. Tagawa K, Arnon DI (1962) Ferredoxins as electron carriers in photosynthesis and in the biological production and consumption of hydrogen gas. *Nature* 195(4841):537–543.
39. Schrader J, et al. (2016) The inhibition mechanism of human 20S proteasomes enables next-generation inhibitor design. *Science* 353(6299):594–598.
40. Nechushtai R, et al. (2011) Allostery in the ferredoxin protein motif does not involve a conformational switch. *Proceedings of the National Academy of Sciences* 108(6):2240–2245.
41. Tramontano A, Chothia C, Lesk AM (1989) Structural determinants of the conformations of medium-sized loops in proteins. *Proteins: Structure, Function, and Bioinformatics* 6(4):382–394.
42. Hutchinson EG, Thornton JM (1996) Promotif—a program to identify and analyze structural motifs in proteins. *Protein Science* 5(2):212–220.
43. Kihara D, Skolnick J (2003) The PDB is a covering set of small protein structures. *Journal of Molecular Biology* 334(4):793–802.
44. Bystroff C, Simons KT, Han KF, Baker D (1996) Local sequence-structure correlations in proteins. *Current Opinion in Biotechnology* 7(4):417–421.
45. Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of Molecular Biology* 281(3):565–577.
46. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta in *Methods in Enzymology*. (Elsevier) Vol. 383, pp. 66–93.
47. Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics* 36(3):307–340.
48. Goldstein RA (2008) The structure of protein evolution and the evolution of protein structure. *Current Opinion in Structural Biology* 18(2):170–177.
49. Gutteridge A, Thornton JM (2005) Understanding nature's catalytic toolkit. *Trends in Biochemical Sciences* 30(11):622–629.
50. Rognan D (2007) Chemogenomic approaches to rational drug design. *British Journal of Pharmacology* 152(1):38–52.
51. Kinjo AR, Nakamura H (2009) Comprehensive structural classification of ligand-binding motifs in proteins. *Structure* 17(2):234–246.
52. Lesk AM, Chothia C (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of Molecular Biology* 136(3):225–270.
53. Feng Z, et al. (2004) Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* 20(13):2153–2155.
54. Meador WE, Means AR, Quiocho FA (1992) Target enzyme recognition by calmodulin: 2.4 Angstrom struc-

ture of a calmodulin-peptide complex. *Science* 257(5074):1251–1256.
55. Messerschmidt A, et al. (1998) Rack-induced metal binding vs. flexibility: Met121His azurin crystal structures at different pH. *Proceedings of the National Academy of Sciences* 95(7):3443–3448.
56. Ahn S, et al. (2001) The "open" and "closed" structures of the type-c inorganic pyrophosphatases from *Bacillus subtilis* and *Streptococcus gordonii*. *Journal of Molecular Biology* 313(4):797–811.
57. Chaudhry C, Weston MC, Schuck P, Rosenmund C, Mayer ML (2009) Stability of ligand-binding domain dimer assembly controls kainate receptor desensitization. *The EMBO Journal* 28(10):1518–1530.
58. Franklin MC, et al. (2003) Structure and function analysis of peptide antagonists of melanoma inhibitor of apoptosis (ml-iap). *Biochemistry* 42(27):8223–8231.
59. Collier JH, et al. (2017) Statistical inference of protein structural alignments using information and compression. *Bioinformatics* 33(7):1005–1013.
60. Leszczynski JF, Rose GD (1986) Loops in globular proteins: a novel category of secondary structure. *Science* 234(4778):849–855.
61. Ogata CM, Gordon PF, de Vos AM, Kim SH (1992) Crystal structure of a sweet tasting protein thaumatin I, at 1.65 Å resolution. *Journal of Molecular Biology* 228(3):893–908.
62. Mirwaldt C, Korndorfer I, Huber R (1995) The crystal structure of dihydrodipicolinate synthase from *Escherichia coli* at 2.5 Å resolution. *Journal of Molecular Biology* 246(1):227–239.
63. Lehmann C, et al. (2005) Structure of HI0073 from *Haemophilus influenzae*, the nucleotide-binding domain of a two-protein nucleotidyl transferase. *Proteins: Structure, Function, and Bioinformatics* 60(4):807–811.
64. Sievers F, Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Multiple Sequence Alignment Methods* pp. 105–116.
65. Blackshields G, Sievers F, Shi W, Wilm A, Higgins DG (2010) Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms for Molecular Biology* 5(1):21.
66. Forest KT, Langford PR, Kroll JS, Getzoff ED (2000) Cu, Zn superoxide dismutase structure from a microbial pathogen establishes a class with a conserved dimer interface. *Journal of Molecular Biology* 296(1):145–153.
67. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research* 32(suppl_2):W526–W531.
68. Källberg M, et al. (2012) Template-based protein structure modeling using the RaptorX web server. *Nature Protocols* 7(8):1511.
69. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PDB Viewer: an environment for comparative protein modeling. *Electrophoresis* 18(15):2714–2723.
70. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5):1792–1797.

Supporting material for '*Universal architectural concepts underlying protein folding patterns*'

Arthur M. Lesk[a,b], Ramanan Subramanian[c], Lloyd Allison[c], David Abramson[d], Peter J. Stuckey[c,e], Maria Garcia de la Banda[c], and Arun S. Konagurthu[c,1]

[a]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, U.K.; [b]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, U.S.A.; [c]Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia; [d]Research Computing Center, University of Queensland, Brisbane, QLD 4072, Australia; [e]School of Computing and Information Systems, University of Melbourne, VIC 3010, Australia

PROÇODIC website: An interactive website describing the inferred concepts and associated information is available at: http://lcb.infotech.monash.edu.au/prosodic/

S1. Material and Terminology

Tableau representation of protein folding patterns. As mentioned in the main text, the essence of any protein folding pattern can be captured by the order, geometry and contact information of its secondary structural elements (1). Lesk (2) developed the *tableau* representation to capture this information in the form of a concise symmetric matrix (see Fig. 1 in main text). A tableau encapsulates: (1) the order in which helices and strands-of-sheet appear in the protein chain, represented by a string, $\mathbf{S}$, of length $|\mathbf{S}|$ over the $\{\text{H(for helix)}, \text{E(for strand)}\}$ alphabet; (2) the geometry of each pair of secondary structural elements, represented by a square-symmetric matrix of angles, $\mathbf{\Omega}$, of order $|\mathbf{S}| \times |\mathbf{S}|$, where angles are in the range $(-180°, 180°]$; (3) the corresponding interactions between pairs of secondary structural elements, represented by a contact matrix, $\mathbf{\Xi}$, of 0/1 values and order $|\mathbf{S}| \times |\mathbf{S}|$, where 1 represents contact and 0 otherwise (3, 4). Formally, any tableau τ is a three-tuple of the form $(\mathbf{S}, \mathbf{\Omega}, \mathbf{\Xi})$.

Source collection used for the inference of PROÇODIC dictionary of concepts. A *source collection* is a collection of (source) tableaux, denoted by the set $\mathcal{T} = \left\{\tau_1, \tau_2, \ldots, \tau_{|\mathcal{T}|}\right\}$.

Since the full `wwPDB` has a lot of redundancy in terms of entries with similar structures, to infer the dictionary of concepts, we use the ASTRAL SCOP-95 (5–7) (v2.05) dataset which has been produced to remove bias due to over-represented structures, while explicitly incorporating structure quality at each step of the domain selection. (8). This data set used is composed of 26,949 domains, representing only 12% of the full SCOPe (v2.05) domain dataset. Of these, 13,365 domains have $< 40\%$ sequence similarity. Although the maximum sequence similarity two proteins can share is 95%, the average sequence similarity is significantly lower $(< 53\%)$. The full list of ASTRAL SCOP-95 domains used to infer the reported dictionary is available in the supporting data file: `prosodicInferenceList.txt` (click).

Subtableaux and the dictionary of concepts. PROÇODIC concepts are represented as contiguous subtableau. A candidate concept, denoted by c, can be instantiated by selecting a source tableau $\tau_\ell \in \mathcal{T}$ and specifying a continuous range of indices $[i, i+1, \ldots, j-1, j]$, such that $1 \leq i < j \leq |\tau_\ell|$, provided it satisfies the constraint that the graph defined by the corresponding contact matrix is connected – any undirected graph is said to be connected if there exists a path between any two vertices in the graph.

[*] To whom correspondence should be addressed. E-mail: arun.konagurthu@monash. edu

Formally, any concept $c \in \tau_\ell^{[i\ldots j]}$ defines a three-tuple $\left(\mathbf{S}_{\tau_\ell}^{[i\ldots j]}, \mathbf{\Omega}_{\tau_\ell}^{[i\ldots j]}, \mathbf{\Xi}_{\tau_\ell}^{[i\ldots j]}\right) \subseteq \left(\mathbf{S}_{\tau_\ell}^{[1\ldots|\tau_\ell|]}, \mathbf{\Omega}_{\tau_\ell}^{[1\ldots|\tau_\ell|]}, \mathbf{\Xi}_{\tau_\ell}^{[1\ldots|\tau_\ell|]}\right)$. Here, $\mathbf{S}_\ell{}^{[i\ldots j]}$ is the substring of $\mathbf{S}_\ell$, defined by the range $[i, \ldots, j]$, while $\mathbf{\Omega}_\ell{}^{[i\ldots j]} \subseteq \mathbf{\Omega}_\ell$ and $\mathbf{\Xi}_\ell{}^{[i\ldots j]} \subseteq \mathbf{\Xi}_\ell$ are the corresponding geometry and contact submatrices respectively. We define a *dictionary* as a set of concepts, denoted by the set $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$. A dictionary can contain an arbitrary number of concepts ($|\mathcal{C}|$), with each concept $c_y \in \mathcal{C}$ containing an arbitrary number of secondary structural elements ($|c_y| \geq 2$). Any possible dictionary is a potential candidate to compress the source collection of tableaux, $\mathcal{T}$.

Associated with each concept $c_y \in \mathcal{C}$ is a concentration parameter, κ_y, corresponding to a von Mises circular (angular) probability distribution (9). This parameter controls the assignment of probabilities used to estimate the encoding length of entries in $\mathbf{\Omega}$ when compressing regions of the source tableaux. That is, κ_y controls the *flexibility* of an inferred concept. A smaller/larger κ_y yields greater/lesser flexibility of the concept's usages for compressing source tableaux regions. In this work, all $\{\kappa_y\}_{\forall 1 \leq y \leq |\mathcal{C}|}$ lie in the range $[\kappa_{\min} = 10, \kappa_{\max} = 100]$, with their precise value inferred (to a precision of $\epsilon_\kappa = 0.5$) as a part of the dictionary search (see below).

Collection of `wwPDB` files dissected using this dictionary. At the time of writing this article, PROÇODIC includes, collectively within the usages of concepts, the dissections of 113,724 protein coordinates files from the `wwPDB` (10). The specific list is available in the supporting data file: `prosodicDissectedWWPDBList.txt` (click). In addition to these dissections, the PROÇODIC website allows users interactively to dissect any protein structure on demand.

S2. Methods

Secondary structure assignment and construction of tableaux. Secondary structure is assigned to any given protein coordinate data using our algorithm, SST (11). SST assigns secondary structures using solely the C_α coordinates of protein structures. The web version of SST is available from: http://lcb.infotech.monash.edu.au/sst. Using SST we can delineate any protein structure into its standard secondary structural elements (SSEs) and thence generate its tableau representation.

Inference of dictionary of concepts. In the proceedings of 2017 IEEE Data Compression Conference, we described the compression-based methodology of inference of recurrent subtableaux on any source collection of tableaux using the Bayesian method of Minimum Message Length (MML) inference (12). The dictionary we report and analyse in this work has been inferred using this methodology with minor modifications. For convenience of the reader, the description of our methodology is reproduced in the Appendix (see Page 9).

S3. Supplementary Notes for 'Exploration of substructures and structural relationships'

Globins. As seen in the main text, different related proteins can present dissections into different related concepts; this is the result of the calculation to optimize the representation of the whole set of proteins.

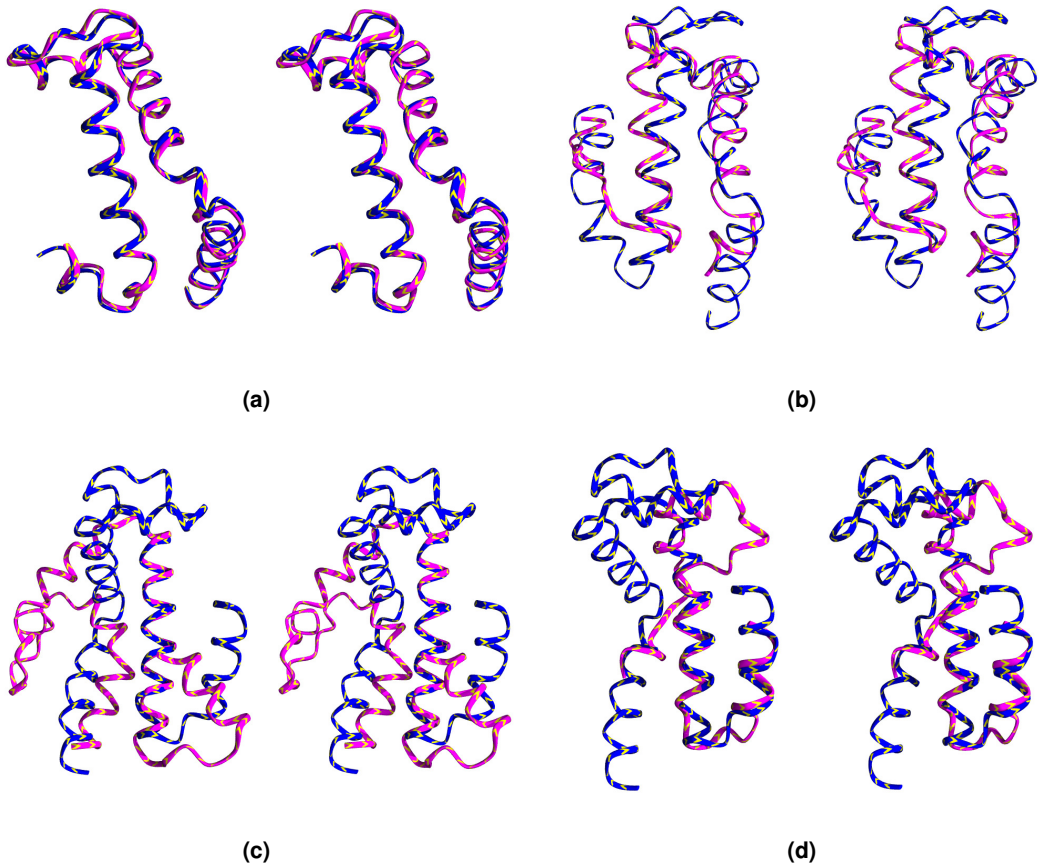

Fig. SF1. Superposition of the usage instances of concept c_0894 (shown in stereo). (a) Superposition of instances from the α chains of human deoxyhaemoglobin (blue) and oxyhaemoglobin from common pigeon *(Columba livia)* 2R80 (pink). (b) Superposition of instances from human oxyhaemoglobin (1hho) and the truncated globin of *Tetrahymena pyriformis* 3AQ5. (c) Superposition of instances from human haemoglobin (1HHO chain A) with the corresponding usages in an *unrelated* protein, complex II (succinate dehydrogenase) from *E. coli* (1NEN chain B). (d) Superposition of 1HHO with unrelated human squalene synthase (3VJ8).

For closely-related proteins, the three-dimensional structures of the usage instances of a given concept are superposable. Fig. SF1(a) shows the superposition of the instances of c_0894 from the α subunits of human deoxyhaemoglobin (2DN2), and oxyhaemoglobin from the common pigeon *(Columba livia)* (2R80). In the structural superposition, 78 Cα atoms from these regions fit to an r.m.s.d. of 0.83 Å.

More distantly-related globins may also share the concept but the regions are not so precisely superposable. A structurally-diverged class of globins are the truncated globins, substantially shorter than sperm-whale myoglobin and showing substantial structural changes. Fig. SF1(b) shows the superposition of the instances of c_0894 from human oxyhaemoglobin (1HHO) and the truncated globin from the ciliate *Tetrahymena pyriformis* (3AQ5). Note that the helix lengths are much more variable than in the superposition shown in Fig. SF1(a). The loop region at the top of this figure does not superpose well between the two structures, and indeed does not even have the same length. This emphasises that our representation of folding patterns captured via the subtableaux of concepts is at the level of geometry of secondary structure elements.

The list of wwPDB entries reported from the 'usage' link contains proteins identified as non-globins, for instance complex II (succinate dehydrogenase) from *Escherichia coli,* a membrane protein, (1NEN chain B residues ASP144–LEU197). Superposing the residues in this region with the instance in 1HHO gives the result shown in Fig. SF1(c). The helices from the N and C termini of these regions fit well. The intervening region shares the secondary structure with some conformational differences.

Could it be that *Escherichia coli* complex II (succinate dehydrogenase) is really homologous to the globins? Examination of 1NEN shows that this protein includes strands of β-sheet, ruling out the possibility of similar topologies of their overall folding patterns.

Another example that shares a concept with human haemoglobin, is the human squalene synthase (3VJ8). (Fig. SF1(c)) Comparing the superpositions within the globin family with the superpositions involving globins and non-globins, closely-related globins show a well-fitting superposition of all secondary structures using the same overall rotation and translation, but globin–non-globin superpositions do not. This is because preserving the angles between successive secondary structures does not fix the global structure, although it does constrain it.

TIM-barrels. When probing the web site for a standard type of structure, the TIM barrel, we enter into the keyword window the string EHEHEHEHEHEHEHEH – or its regular expression: (EH){8} – signifying an eight-fold repeat of a β-α unit.

The web site returned two concepts, c_0008 and c_0032 containing the pattern EHEHEHEHEHEHEHEH. The first, c_0008, contains TIM barrels. The structure of the concept comprises the canonical 8-fold β-α barrel plus four additional C-terminal helices. As the reader is encouraged to try, clicking on the image produces a large 'still' high-quality graphic display; clicking on 'view interactively' produces an image rotatable under mouse control. Clicking for 'full details' gives the full secondary-structure assignment for the 'fold archetype' of c_0008, SCOP domain d3flua_, and the tableau computed for this domain. (Our methodology enables each concept in the dictionary to converge to an archetype that can viewed as the topological *median* over all usages of that concepts in the source collection the dictionary compresses. Varying ideas defining such topological medians representing supersecondary structural motifs were previously explored, especially as 'attractors in fold space' to enable protein structural classification efforts (13, 14).

Clicking on 'usages' reports other instances of this concept in the wwPDB. There are 118 other usages, all TIM barrels. However, they are not the only TIM barrels in the wwPDB. Many structures that do contain the 8-fold barrel but lack the four C-terminal helices are dissected into smaller units, some containing β-α

subsets of the barrel. Others proteins have helices inserted at different points, deviating from the specified secondary structural pattern. Therefore, searching using patterns alone does not return all the TIM barrels in the wwPDB.

It is possible to type 'TIM barrel' into the keyword field. The web site will then return many concepts, some of which do correspond to TIM barrels, but others of which do not. For instance, in response to the keyword query 'TIM barrel' PROÇODIC returns c_0004 which contains 24 consecutive β-strands but no helices. The reason is that one of the wwPDB entries in which c_0004 appears is the human PRMT5:MEP50 complex (4QGB): this entry protein does contain a TIM-barrel domain (in which c_0004 does not appear), and the entry file contains TIM BARREL in its wwPDB KEYWDS record line, which triggers a hit on this concept. In summary, to find TIM barrels in the wwPDB, a PROÇODIC search for (EH){8} returns too little, a search for 'TIM BARREL' returns too much, and there is no Goldilocks compromise. In any event, this is a solved problem. For this particular question, other tools such as SCOP are more convenient and appropriate.

The other concept returned for the query EHEHEHEHEHEHEHEH, contains two domains, each with four β-α units, but not closed into a barrel. This appears because a sequence of secondary structure elements does not uniquely define three-dimensional structure.

Uncompressed regions in dissections and unusual structural components. Suppose a region in a protein contains an unusual conformation. It may require a shorter message to send the subtableau information of this region raw (without compressing that region) than to include a representative within the dictionary and compress it. This is because, given its rarity, the overhead of adding it in the dictionary does not justify its inclusion as a separate concept. Overall, in the dissections of $\sim$114 000 wwPDB entries, $\sim$66% of the residues are covered by dictionary concepts.

An example of an uncompressed region appears in the dissection of *Chironomus* erythrocruorin, a globin from an insect. Although the overall structure of this molecule is similar to that of globins, there are deviations from the usual structure in the region corresponding to the D helix of sperm whale myoglobin. This results in different dissections of sperm whale myoglobin (1MBD) and *Chironomus* erythrocruorin (1ECD):

Sperm whale myoglobin (1MBD):		*Chironomus* erythrocruorin (1ECD):	
SER3 – PRO37	c_1483	ALA3 – ASP31	c_1368
PRO37 – LYS79	c_1141	ALA53 – MET136	c_1140
HIS82 – LEU149	c_1433		

Observe that the region in *Chironomus* erythrocruorin from residues 31–53 is not part of the dissection. This region is explained directly without compression, using what we call the *null* concept.

Erythrocruorins. Typing 'erythrocruorin' in the keyword field returns 26 concepts. The top one, concept c_0547, is an assembly of seven helices, corresponding to helices A-B-C-E-F-G-H in the canonical globin fold. There are 268 usage instances of this concept, which include erythrocuorins and globins. (The seven helix pattern fits the $\alpha-$chain of mammalian haemoglobins, which lack a D helix, but not the β-chain which contains a D helix.) The distinction between similar structures that differ in some detail which breaks a pattern, can be seen as both a strength and as a weakness. We saw a similar phenomenon with the TIM barrels and with globins.

Substructures of the 'globin fold' containing 6 helices, include c_0640. This corresponds to globin helices A-B-E-F-G-H. Examples include phycocyanins and phycoerythrins, colicin, and certain globins (15). To be a proper instance of c_0640, a globin must *lack* both C and D helices. In these cases the region corresponding to the D helix is not helical, but the region corresponding to the C-helix is nevertheless quite

close to the expected 3_{10} helix. This is because the region is distorted enough to drag at least one hydrogen bond outside the thresholds of acceptance in distance and/or angle. As a result, the region is assigned as a coil, and only six helices are attributed to the chain. In other cases, the compression criterion encodes a regular globin structure as more than a single concept. Thus, human oxyhaemoglobin (1HHO), chain A, is decomposed into:

SER3 – ARG92	c_0894
VAL96 – ARG141	c_1410

Note that this chain does have a 3_{10} C helix, but *not* a D helix.

In addition to helical concepts that are substructures of the canonical globin fold, querying for erythrocruorins (as a keyword on the web site) returns concepts containing purely β-sheet concepts. These are known to appear in large, multimeric, extracellular invertebrate erythrocruorins, as linker regions between globin-like all-helical domains (16). Our dictionary has therefore called for attention to these additional substructures, not customary structure components from the familiar globin family.

As a β-sheet structure is rare in the globins and erythrocruorins, it is of interest to explore the relationships of these substructures to other families. Are there homologues, of which the erythrocruorin linker domain might be part of a chain of evolutionary relationships?

Consider the concept c_0559, comprising six β-strands. Its *usages* provides a list of chains in which this concept occurs. Checking the list of chains against the SCOP classification gives the following results:

#Instances	SCOP classification	#Instances	SCOP classification
39	b.60.1.2	1	d.129.3.8
5	e.7.1.1	1	d.129.3.5
3	h.1.2.1	1	b.97.1.1
3	b.8.1.1	1	b.82.1.23
3	b.30.5.4	1	b.82.1.11
2	d.85.1.1	1	b.61.7.1
2	d.25.1.1	1	b.61.1.1
2	b.60.1.1	1	b.60.1.8
1	h.1.32.1	1	b.60.1.7
1	g.12.1.1	1	b.23.1.1
1	f.1.1.1	1	b.1.9.2
1	d.169.1.8	1	b.163.1.1

It is no surprise that most of these are from SCOP all-β class, as they are assemblies of β-strands, with class b.60.1.2, 'Fatty acid binding protein-like' as the exception. (However, because the *usage* option gives results for the entire wwPDB, the frequency may well reveal the experimental bias of solved structures within that family.)

Fig. SF2b shows the structural superposition of the regions from the linker region of the multimeric erythrocruorin from the earthworm, *Lumbricus terrestris,* (2GTL, chain o) and Human cellular retinol binding protein III (1GGL, chain a).

Are the erythrocruorin linker domains and the retinol-binding proteins homologues? Structural comparison shows that only the regions of these proteins share secondary structural elements and the geometry of their assembly (as shown in Fig. SF2b). The rest of the domains are quite different. The regions are not homologues. However, the dictionary has identified a substructure which they, and other domains, share. Like other supersecondary structures, their appearance in unrelated proteins show that they are shared pieces of protein folds but not signs of homology. Indeed, an appeal to SCOP shows that the erythrocruorin linker domains are in a superfamily of their own, sharing a folding topology with 'Streptavidin-like' domains.

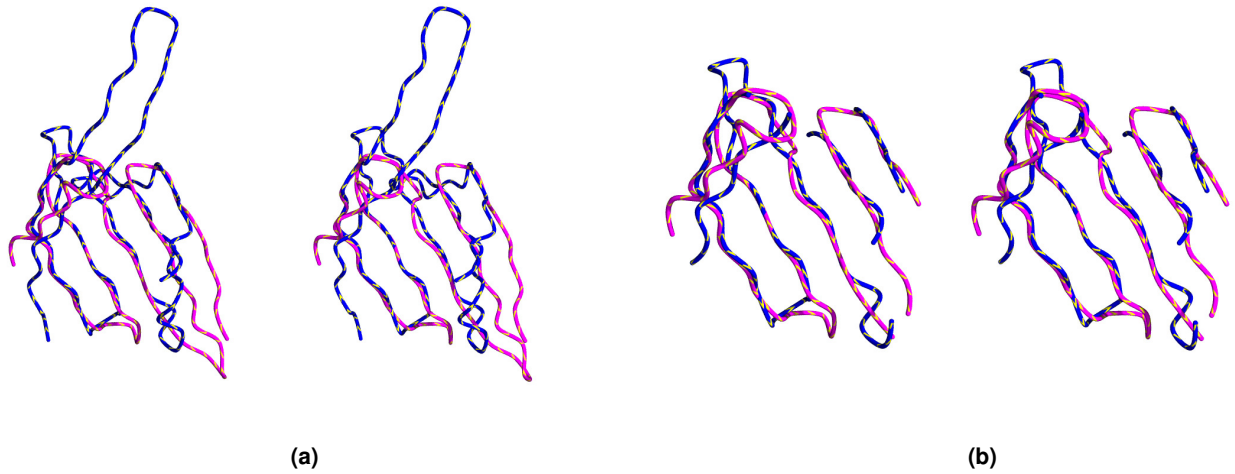

Fig. SF2. Superposition (in stereo) of region containing concept from (blue) linker region of earthworm multimeric erythrocruorin and (pink) Human cellular retinol binding protein III. (a) Entire region comprising concept c_0559. (b) Restriction to well-fitting residues from region comprising concept c_0559.

References

1. Chothia C, Finkelstein AV (1990) The classification and origins of protein folding patterns. *Annual Review of Biochemistry* 59(1):1007–1035.
2. Lesk AM (1995) Systematic representation of protein folding patterns. *Journal of Molecular Graphics* 13(3):159–164.
3. Konagurthu AS, Stuckey PJ, Lesk AM (2008) Structural search and retrieval using a tableau representation of protein folding patterns. *Bioinformatics* 24(5):645–651.
4. Konagurthu AS, Lesk AM (2010) Cataloging topologies of protein folding patterns. *Journal of Molecular Recognition: An Interdisciplinary Journal* 23(2):253–257.
5. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* 247(4):536–540.
6. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2013) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research* 42(D1):D310–D314.
7. Chandonia JM, Fox NK, Brenner SE (2017) SCOPe: Manual curation and artifact removal in the structural classification of proteins–extended database. *Journal of Molecular Biology* 429(3):348–355.
8. Brenner SE, Koehl P, Levitt M (2000) The astral compendium for protein structure and sequence analysis. *Nucleic acids research* 28(1):254–256.
9. Mardia KV, Jupp PE (2009) *Directional statistics*. (John Wiley & Sons) Vol. 494.
10. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research* 35(suppl 1):D301–D303.
11. Konagurthu AS, Lesk AM, Allison L (2012) Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics* 28(12):i97–i105.
12. Subramanian R, et al. (2017) Statistical compression of protein folding patterns for inference of recurrent substructural themes in *Data Compression Conference (DCC), 2017*. (IEEE), pp. 340–349.
13. Dietmann S, et al. (2001) A fully automatic evolutionary classification of protein folds: Dali domain dictionary version 3. *Nucleic acids research* 29(1):55–57.
14. Orengo C, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. *Nature* 372(6507):631.
15. Holm L, Sander C (1993) Structural alignment of globins, phycocyanins and colicin A. *FEBS Letters* 315(3):301–306.
16. Ruggiero Bachega JF, et al. (2015) The structure of the giant haemoglobin from *Glossoscolex paulistus*. *Acta Crystallographica. Section D, Biological Crystallography* 71(6):1257–1271.
17. Wallace CS, Boulton DM (1968) An information measure for classification. *Computer Journal* 11(2):185–194.
18. Wallace CS (2005) *Statistical and Inductive Inference using Minimum Message Length*, Information Science and Statistics. (SpringerVerlag).

19. Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27:379–423.
20. Wallace CS, Patrick JD (1993) Coding decision trees. *Machine Learning* 11(1):7–22.

Funding This research is funded by Australian Research Council (ARC) Discovery Project grant (DP150100894).

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to Arun Konagurthu. (email: arun.konagurthu@monash.edu).

Appendix

The construction of PROÇODIC dictionary of concepts relies on the compression-based methodology we developed and presented at the 2017 IEEE Data Compression Conference (12). This methodology (with minor changes and additional details specific to this work) is reproduced below with permission for the convenience of the reader.

Our goal is to learn the static dictionary $\mathcal{C}$ (i.e., hypothesis) that offers the best compression of the source collection $\mathcal{T}$ (i.e, observed data). The general statistical framework used to achieve this relies on the criterion of minimum message length (MML) inference (17, 18). MML inference is best understood as a lossless two-part communication between an *imaginary transmitter-receiver pair*. In the first part the transmitter encodes and communicates the hypothesis to the receiver, while in the second part it communicates the observed data *given* the stated hypothesis. The best hypothesis in this framework is the one that yields the shortest two-part lossless message to communicate the observed data. Formally, for any static dictionary $\mathcal{C}$ and source collection $\mathcal{T}$, the two-part message length (in bits) is denoted by the terms:

$$\mathcal{I}\Big(\mathcal{C}\&\mathcal{T}\Big) = \underbrace{\mathcal{I}\Big(\mathcal{C}\Big)}_{\text{first part}} + \underbrace{\mathcal{I}\Big(\mathcal{T}|\mathcal{C}\Big)}_{\text{second part}} \qquad \text{bits,} \qquad [1]$$

where $\mathcal{I}\Big(\cdot\Big) = -\log_2(\Pr(\cdot))$ is the Shannon's measure of information content (19).

The two-part message shown in Eqn. 1 is contrasted with the (single-part) *null model* message, that is, the encoding of the observed data *as is*, without the support of any hypothesis. The null model message length is denoted as $\mathcal{I}_{\text{null}}(\cdot)$. Thus, the quality of an inferred dictionary $\mathcal{C}$ is measured as the compression obtained by encoding the source collection $\mathcal{T}$ using $\mathcal{C}$, i.e., as $\mathcal{I}_{\text{null}}(\mathcal{T}) - \mathcal{I}\Big(\mathcal{C}\&\mathcal{T}\Big)$. This yields an inference problem with the following objective: $\max_{\mathcal{C}} \mathcal{I}_{\text{null}}(\mathcal{T}) - \mathcal{I}\Big(\mathcal{C}\&\mathcal{T}\Big)$.

Addressing this inference problem requires the following: (1) A method to estimate the null model encoding length, $\mathcal{I}_{\text{null}}(\mathcal{T})$, for any given collection $\mathcal{T}$; (2) A method to estimate the dictionary model encoding length, $\mathcal{I}\Big(\mathcal{C}\&\mathcal{T}\Big)$ for any given dictionary $\mathcal{C}$ and collection $\mathcal{T}$; (3) A search method for an *optimal* dictionary (one that maximizes compression, as per the stated objective). These methods are presented below.

Estimation of $\mathcal{I}_{\text{null}}(\mathcal{T})$. The null encoding of the source collection $\mathcal{T}$ involves the encoding of the number of tableaux over an integer code, followed by the null encoding of each tableau $\tau_\ell \in \mathcal{T}$:

$$\mathcal{I}_{\text{null}}(\mathcal{T}) = I_{\text{integer}}\Big(|\mathcal{T}|\Big) + \sum_{\ell=1}^{|\mathcal{T}|} I_{\text{null}}\Big(\tau_\ell\Big) \qquad \text{bits,} \qquad [2]$$

where $I_{\text{integer}}(.)$ is the message length of encoding any positive integer over a $\log^*$ distribution (20). Further, the estimation of each $I_{\text{null}}\Big(\tau_\ell\Big)$ term is carried out by encoding the number of secondary structural elements using the same integer code, followed by encoding the three-tuples $(\mathbf{S}_{\tau_\ell}, \mathbf{\Omega}_{\tau_\ell}, \mathbf{\Xi}_{\tau_\ell})$ using uniform probability distributions on their respective supports. This implies that each character in the $\mathbf{S}_\ell$ string takes one bit to encode, each contact state in $\mathbf{\Xi}_\ell$ also takes one bit, and each angle in $\mathbf{\Omega}_\ell$, specified to a precision of

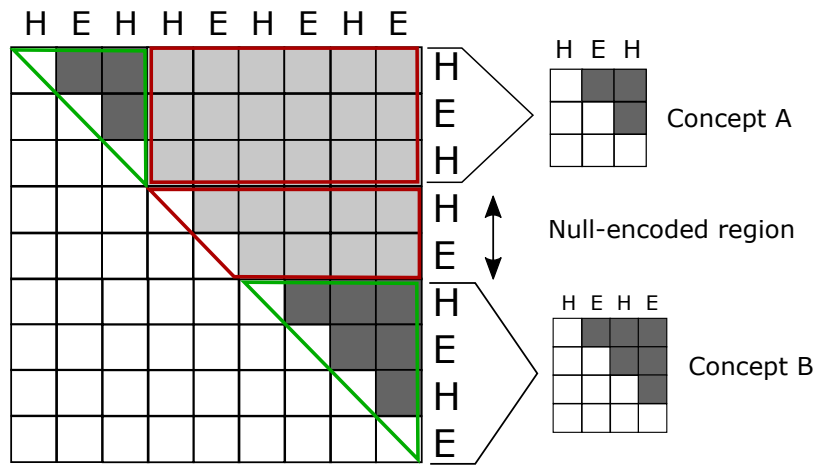

Fig. SF3. An illustration of a partition of a tableau.

0.1° in the range $(-180^\circ, +180^\circ]$, takes $\log_2(360/0.1) = \log_2 3600$ bits. Thus, the null message length for communicating each tableau τ_ℓ is given by:

$$I_{\text{null}}\big(\tau_\ell\big) = \underbrace{I_{\text{integer}}\big(|\mathbf{S}_\ell|\big) + |\mathbf{S}_\ell|}_{\mathcal{I}_{\text{null}}(\mathbf{S}_\ell)} + \underbrace{\binom{|\mathbf{S}_\ell|}{2}\log_2(3600)}_{\mathcal{I}_{\text{null}}(\mathbf{\Omega}_\ell|\mathbf{S}_\ell)} + \underbrace{\binom{|\mathbf{S}_\ell|}{2}}_{\mathcal{I}_{\text{null}}(\mathbf{\Xi}_\ell|\mathbf{S}_\ell)}. \qquad [3]$$

Estimation of the first part, $\mathcal{I}\big(\mathcal{C}\big)$, term in Eqn. 1. Each concept c_y in any given dictionary is a (sub-)tableau. Therefore, c_y can be encoded using the null model as shown in Section S3, using $\mathcal{I}_{\text{null}}(c_y)$ bits. In addition, its associated κ_y parameter also needs to be encoded. As seen in Section S1, each κ_y lies in the range $[\kappa_{\min}, \kappa_{\max}]$ specified to a precision of ϵ_κ, and can be encoded using a uniform probability distribution over this support. Using these component terms, the resulting encoding length for the full dictionary takes (in bits):

$$\mathcal{I}\big(\mathcal{C}\big) = I_{\text{integer}}\big(|\mathcal{C}|\big) + \sum_{j=1}^{|\mathcal{C}|} I_{\text{null}}\big(c_j\big) + |\mathcal{C}|\log_2\Big(\frac{\kappa_{\max} - \kappa_{\min}}{\epsilon_\kappa} + 1\Big). \qquad [4]$$

Estimation of the second part, $\mathcal{I}\big(\mathcal{T}|\mathcal{C}\big)$, term in Eqn. 1. The encoding length of the source collection $\mathcal{T}$ given the dictionary $\mathcal{C}$ is computed as the sum of the code lengths required to encode each tableau $\tau_\ell \in \mathcal{T}$ using $\mathcal{C}$. To compute this code length, $\mathcal{I}\big(\tau_\ell|\mathcal{C}\big)$, each tableau τ_ℓ is *partitioned* into *non-overlapping* regions of variable sizes (see Fig. SF3). Any *partition* of τ_ℓ, $p(\tau_\ell)$, is specified by an increasing sequence of integer indices $1 \equiv z_0 < z_1 < z_2 < \ldots < z_{|p(\tau_\ell)|} \equiv |\tau_\ell| + 1$. The set of possible 2-grams from this partition sequence, $\{\langle z_{k-1}, z_k\rangle\}_{\forall 1 \le k \le |p(\tau_\ell)|}$, define $|p(\tau_\ell)|$ consecutive, non-overlapping regions of τ_ℓ of the form $\tau_\ell^{[z_{k-1}\ldots z_k-1]} \subseteq \tau_\ell$. The total number of possible partitions for a given τ_ℓ is $2^{|\tau_\ell|-1}$.

Assigned to each non-overlapping region $\tau_\ell^{[z_{k-1}\ldots z_k-1]}$ in a specified partition $p(\tau_\ell)$ is one of the concepts $c_{y_k} \in \mathcal{C}$ (provided $\mathbf{S}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]} \equiv \mathbf{S}_{c_{y_k}}$), where $1 \le y_k \le |\mathcal{C}|$, or a *null concept* c_0. The key idea here is that the burden of explanation of the data in each non-overlapping region $\tau_\ell^{[z_{k-1}\ldots z_k-1]} \subseteq \tau_\ell$ defined by $p(\tau_\ell)$ is borne by the corresponding data in its assigned concept. That is, the data in $\Big(\mathbf{S}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}, \mathbf{\Omega}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}, \mathbf{\Xi}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}\Big)$ is communicated using the corresponding data $\Big(\mathbf{S}_{c_{y_k}}, \mathbf{\Omega}_{c_{y_k}}, \mathbf{\Xi}_{c_{y_k}}\Big)$. When the region is assigned to the null concept c_0, the encoding of the data in that region is identical to null model encoding described in Section S3. In the remaining text, $p(\tau_\ell)$ specifies not only the non-overlapping regions, but also their corresponding concept assignments.

Thus, given $p(\tau_\ell)$, the data within each $\tau_\ell \in \mathcal{T}$ can be communicated as follows. First, the tableau size is communicated with an integer code using $I_{\text{integer}}\Big(|\tau_\ell|\Big)$ bits. Second, the details specified by $p(\tau_\ell)$ are communicated using $\mathcal{I}\Big(p(\tau_\ell)\Big)$ bits. Third, each non-overlapping region $\tau_\ell^{[z_{k-1}\ldots z_k-1]}$ is explained by its assigned concept c_{y_k} (shown as the grey coloured regions in Fig. SF3), using $\mathcal{I}\Big(\tau_\ell^{[z_{k-1}\ldots z_k-1]}|c_{y_k}\Big)$ bits. Finally, the remaining data of τ_ℓ (the light-grey cells in Fig. SF3) are communicated using the null model (Section S3). We denote this data (using set notations) as $(\tau_\ell|p(\tau_\ell))^{\complement} = \tau_\ell \setminus \cup_{k=1}^{|p(\tau_\ell)|}\{\tau_\ell^{[z_{k-1}\ldots z_k-1]}\}$, with code length denoted by $\mathcal{I}_{\text{null}}((\tau_\ell|p(\tau_\ell))^{\complement})$.

Based on the above details, the *shortest* lossless encoding length of any tableau $\tau_\ell \in \mathcal{T}$ given a static dictionary $\mathcal{C}$, is one that that minimizes the following objective:

$$\mathcal{I}\Big(\tau_\ell|\mathcal{C}\Big) = I_{\text{integer}}\Big(|\tau_\ell|\Big) + \min_{\forall p(\tau_\ell)}(\mathcal{I}\Big(p(\tau_\ell)\Big) + \sum_{k=1}^{|p(\tau_\ell)|}\mathcal{I}\Big(\tau_\ell{}^{[z_{k-1}\ldots z_k-1]}|c_{y_k}\Big)) + \mathcal{I}_{\text{null}}((\tau_\ell|p(\tau_\ell))^{\complement}).$$

Combining the above term with Eqn. 4, gives us $\mathcal{I}\Big(\mathcal{T}|\mathcal{C}\Big)$ as:

$$\mathcal{I}\Big(\mathcal{T}|\mathcal{C}\Big) = I_{\text{integer}}\Big(|\mathcal{T}|\Big) + \sum_{\ell=1}^{|\mathcal{T}|}\mathcal{I}\Big(\tau_\ell|\mathcal{C}\Big). \qquad [5]$$

Below we describe the details of computing the code length terms involved in Eqn. 5.

Computation of $\mathcal{I}\Big(p(\tau_\ell)\Big)$: A partition $p(\tau_\ell)$ is encoded as follows. Since the tableau size $|\tau_\ell|$ has already been communicated, the size of the partition $|p(\tau_\ell)|$ is encoded in $\log_2|\tau_\ell|$ bits. Each index in the corresponding set of concept assignments $\{y_k\}_{\forall 1\leq k\leq |p(\tau_\ell)|}$ can take the values $0 \leq y_k \leq |\mathcal{C}|$, and, thus, can be encoded in $\log_2(|\mathcal{C}|+1)$ bits. Given this information, the values of z_0=1, $z_{|p(\tau_\ell)|} = |\tau_\ell|$+1, and the subset of $\{z_k\}$'s $(2 \leq k < |p(\tau_\ell)|)$ associated with regions *not* assigned to the null concept c_0 are already decipherable based on the assigned concept sizes. The remaining ones, associated with c_0, each take $\log_2(|\tau_\ell|-z_{k-1}+1)$ bits to state.

Computation of $\mathcal{I}\Big(\tau_\ell^{[z_{k-1}\ldots z_k-1]}|c_{y_k}\Big)$: A region $\tau_\ell^{[z_{k-1}\ldots z_k-1]}$ can be encoded using a null concept c_0, or using any concept $c_{y_k} \in \mathcal{C}$. The computation of the null concept encoding of a region follows the same scheme as the null-model encoding of a tableau described in Section S3. On the other hand, the encoding of $\tau_\ell^{[z_{k-1}\ldots z_k-1]}$ using a concept $c_{y_k} \in \mathcal{C}$ is permitted only when the corresponding secondary structural strings are identical, and when the corresponding contact information between pairs of secondary structural elements differ in no more than 10% $\left(= \lfloor \frac{1}{10} \cdot \binom{|c_{y_k}|}{2} \rfloor\right)$ places.

The details of encoding $\tau_\ell^{[z_{k-1}\ldots z_k-1]} \equiv \left(\mathbf{S}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}, \mathbf{\Omega}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}, \mathbf{\Xi}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}\right)$ with $c_{y_k} \equiv \left(\mathbf{S}_{c_{y_k}}, \mathbf{\Omega}_{c_{y_k}}, \mathbf{\Xi}_{c_{y_k}}\right)$ are now considered. Since $\mathbf{S}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}$ is identical to $\mathbf{S}_{c_{y_k}}$, it is only necessary to encode $\mathbf{\Xi}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}$ and $\mathbf{\Omega}_{\tau_\ell}^{[z_{k-1}\ldots z_k-1]}$ using the corresponding concept's matrices $\mathbf{\Xi}_{c_{y_k}}$ and $\mathbf{\Omega}_{c_{y_k}}$, respectively. Thus, the encoding N_m requires $\log_2\left(1+\lfloor \frac{1}{10}\binom{|c_{y_k}|}{2}\rfloor\right)$ bits, where $N_{\text{m}} \in [0, \lfloor \frac{1}{10}\binom{|c_{y_k}|}{2}\rfloor]$ is the total number of mismatches in this assigned region where $\mathbf{\Xi}_{c_{y_k}}$ and $\mathbf{\Xi}_{\tau_\ell}{}^{[z_{k-1}\ldots z_k-1]}$ differ. The locations of the mismatches are then encoded using a uniform distribution over the number of ways of identifying N_m locations out of $\binom{|c_{y_k}|}{2}$ cells. The resulting code length to identify each mismatched entry takes the logarithm of the corresponding binomial coefficient. Once this information is communicated, the corresponding entries of $\mathbf{\Omega}_{\tau_\ell}{}^{[z_{k-1}\ldots z_k-1]}$ are encoded using the null model, each using $\log_2 3600$ bits.

The matched angles are now transmitted. Each signed angle $\theta \in \mathbf{\Omega}_{\tau_\ell}{}^{[z_{k-1}\ldots z_k-1]}$ is encoded using the corresponding angle $\theta_\mu \in \mathbf{\Omega}_{c_{y_k}}$ and a $(90\%, 10\%)$ mixture model of von-Mises circular distribution (18)

(parameterized on θ_μ and concept's κ) and a uniform distribution over the support $(-180°, +180°]$.

Computation of $\mathcal{I}_{\text{null}}((\tau_\ell|p(\tau_\ell))^{\complement})$: This refers to the encoding of the off-diagonal $\mathbf{\Omega}_{\tau_\ell}$ and $\mathbf{\Xi}_{\tau_\ell}$ entries, shown as the light-grey coloured cells in Fig. SF3, and denoted here as $\mathbf{\Omega}_{(\tau_\ell|p(\tau_\ell))^{\complement}}$ and $\mathbf{\Xi}_{(\tau_\ell|p(\tau_\ell))^{\complement}}$. The total number of entries in this off-diagonal area in each of these matrices is $\binom{|\tau_\ell|}{2} - \sum_{k=1}^{|p(\tau_\ell)|}\binom{z_k - z_{k-1}}{2}$. Each angle in $\mathbf{\Omega}_{(\tau_\ell|p(\tau_\ell))^{\complement}}$ is encoded using the null model in $\log_2(3600)$ bits. Each contact in $\mathbf{\Xi}_{(\tau_\ell|p(\tau_\ell))^{\complement}}$ requires 1 bit.

Optimal partition of τ_ℓ given $\mathcal{C}$ via dynamic programming: The computation of $\mathcal{I}\Big(\tau_\ell|\mathcal{C}\Big)$ is carried out on the optimal partition of τ_ℓ given the concepts in $\mathcal{C}$. The identification of the optimal $p(\tau_\ell)$, the one that minimizes $\mathcal{I}\Big(\tau_\ell|\mathcal{C}\Big)$, is achieved using a one-dimensional dynamic programming algorithm, similar to the one devised in our earlier work for a *completely different* problem from the same domain (11).

To compute the optimal partition $p(\tau_\ell)$ of a tableau τ_ℓ under the dictionary $\mathcal{C}$, we first construct a set $\Big\{H^0, H^1, \ldots, H^{|\mathcal{C}|}\Big\}$ of $|\mathcal{C}|+1$ matrices called the "code length matrices". The (i,j)-th entry of code-length matrix H^t stores the message length for stating the subtableau region $\tau_\ell{}^{[i\ldots j]}$ using the concept $c_t \in \mathcal{C}$ plus the corresponding off-diagonal entries in that region $\mathcal{I}_{\text{null}}((\tau_\ell|p(\tau_\ell))^{\complement})^{[i\ldots j]}$. (If the region is stated using the null concept, this entry also adds the cost of stating the size of the null region.)

The manner in which partitioning has been formulated allows us to investigate the problem of determining the optimal partition by solving smaller independent sub-problems, thereby allowing us to propose a dynamic-programming (DP) approach. We first consider a cost function $M(j)$ defined using the following recurrence relationship, for every integer j such that $0 \leq j \leq |\tau_\ell|$:

$$M(j) = \begin{cases} 0, & \text{if } j = 0, \\ \min\limits_{0 \leq i < j} \left[M(i) + \min\limits_{0 \leq t \leq |\mathcal{C}|} H^t(i+1, j) + \log_2\Big(|\mathcal{C}|+1\Big) \right], & \text{otherwise.} \end{cases} \quad [6]$$

The recurrence relation in Eqn 6 implies that $M(j)$ is the optimal cost of stating the τ_ℓ up to the SSE index j. This can be verified as follows. For $j = 0$, $M(j)$ is interpreted as the cost of stating nothing, which is zero. For any other value j, we find the optimal dissection point $i < j$ such that the tableau up to SSE j can be stated incrementally as the tableau up to SSE i using some dissection, plus the rest of the tableau up to SSE j using a concept that best explains the region [i...j]. We can hence compute the optimal statement cost of the dissection of the entire tableau, $M\Big(j = |\tau_\ell|\Big)$. Adding to it the cost of stating the number of SSEs in the tableau and the cost of stating the number of segments in the optimal segmentation, we will have the expansion for $\mathcal{I}\Big(\tau_\ell \mid \mathcal{C}\Big)$ term in Eqn. 5 as:

$$\mathcal{I}\Big(\tau_\ell \mid \mathcal{C}\Big) = I_{\text{integer}}\Big(|\tau_\ell|\Big) + \log_2|\tau_\ell| + M\Big(|\tau_\ell|\Big). \quad [7]$$

The optimal segmentation can then be determined by tracing back on the cached optimal derivations of M, starting from $M\Big(j = |\tau_\ell|\Big)$ back to $M(j = 0)$.

Searching for an optimal dictionary. Our goal is to address the problem of inferring an optimal static dictionary, i.e., one that minimizes the two-part message length given by Eqn. 1. Finding a provably optimal dictionary is computationally intractable due to the enormous search space. Hence, a simulated annealing (SA) heuristic is devised to address this problem. Algorithms based on SA require an aperiodic irreducible Markov chain defined on a certain state space, and a *cooling schedule* to push iteratively the solution towards the optimum. In our case, the state space is the set of all possible dictionaries. The desired Markov chain is generated by defining a neighbourhood and the corresponding transition probabilities for every

state $\mathcal{C}$. A local neighbourhood for every state is explored through the following perturbation primitives: (1) `Add concept`: Creates a concept randomly from the source collection and adds it to $\mathcal{C}$. (2) `Remove element`: Chooses a concept randomly from the dictionary and deletes it. (3) `Perturb concept length`: Chooses a concept randomly from the dictionary, and extends/shortens it, in reference to its original source. (4) `Perturb concept kappa`: Increments/decrements the current value of κ associated with a randomly chosen concept. (5) `Swap concept with usage`: Chooses a concept randomly from the dictionary, and swaps it with a region in the collection that is currently encoded by it. (Note, the usage swapped-in as the perturbed concept could weakly violate the connectivity constraint in terms of its contact map, unless strict connectivity is also imposed on that chosen usage.)

At each iteration, one of the above five perturbations is chosen uniformly at random. The transition probability to the neighbour is computed as follows. If the two-part message length given by Eqn. 1 decreases, the transition probability to the perturbed state is 1. If the two-part message length increases by ΔI bits, the probability is $2^{-\Delta I/T}$, where T is the *temperature* parameter of the system controlled by the following cooling schedule: Start with a temperature of 5,000 and decrease it by a factor of 0.88. At each temperature step, perform 50,000 random perturbations unless the temperature is below 10, where the number of perturbations is increased to 500,000 per temperature step. When the temperature reaches below 0.1, the search stops and the current state of the dictionary is reported. See `pseudocode.pdf` (click).

Parallel Implementation. The methodology of inference described above was implemented in the `C++` programming language. To tackle the enormous amount of computational effort that was required to identify the PROÇODIC dictionary reported here, this program was parallelised and deployed on a large computing cluster. This was achieved using the `OpenMPI C++` library, under a Message-Passing Interface framework.

Specifically, this implementation was parallelized and was executed over 240 computing cores (requiring about 2 Giga Bytes of main memory per core) on the high performance computing cluster, `Tinaroo`, at the University of Queensland's Research Computer Center. It took roughly 11 days for our simulated annealing approach to converge on the identified PROÇODIC dictionary of 1493 concepts – this runtime is tantamount to about 7 years of computing over a sequential run.

The source collection gets sub-divided into roughly equal subsets of tableaux, and allocated to individual computing nodes. This data-parallel implementation exploits the ability for each computing node to generate *independently* the same sequence of uniform (pseudo-)random numbers. At each node, and operating on the allocated subset of tableaux from the collection that is being compressed, all random numbers required for performing the perturbations involved in the simulated annealing approach are ensured to work in tandem across the cores. This results in every node evolving the same dictionary independently, although working on different partitions of the source collection. This minimises the overhead required for communication between the compute nodes, by avoiding the overhead involved in broadcasting the current state of the dictionary and the details of the perturbations being performed at each perturbation step of the simulated annealing approach. The only communication required between nodes is the '*Allreduced*' total two-part message length, *summed* over all data partitions, before perturbations are synchronously accepted/rejected (as per the Metropolis criterion). The total speedup as a result of this parallelisation was measured to about 200 times over 240 nodes.