

Tertiary alphabet for the observable protein structural universe

Craig O. Mackenzie^a, Jianfu Zhou^b, and Gevorg Grigoryan^{a,b,c,1}

^aInstitute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH 03755; ^bDepartment of Computer Science, Dartmouth College, Hanover, NH 03755; and ^cDepartment of Biological Sciences, Dartmouth College, Hanover, NH 03755

Edited by David Baker, University of Washington, Seattle, WA, and approved September 22, 2016 (received for review May 4, 2016)

Here, we systematically decompose the known protein structural universe into its basic elements, which we dub tertiary structural motifs (TERMs). A TERM is a compact backbone fragment that captures the secondary, tertiary, and quaternary environments around a given residue, comprising one or more disjoint segments (three on average). We seek the set of universal TERMS that capture all structure in the Protein Data Bank (PDB), finding remarkable degeneracy. Only ~600 TERMS are sufficient to describe 50% of the PDB at sub-Angstrom resolution. However, more rare geometries also exist, and the overall structural coverage grows logarithmically with the number of TERMS. We go on to show that universal TERMS provide an effective mapping between sequence and structure. We demonstrate that TERM-based statistics alone are sufficient to recapitulate close-to-native sequences given either NMR or X-ray backbones. Furthermore, sequence variability predicted from TERM data agrees closely with evolutionary variation. Finally, locations of TERMS in protein chains can be predicted from sequence alone based on sequence signatures emergent from TERM instances in the PDB. For multisegment motifs, this method identifies spatially adjacent fragments that are not contiguous in sequence—a major bottleneck in structure prediction. Although all TERMS recur in diverse proteins, some appear specialized for certain functions, such as interface formation, metal coordination, or even water binding. Structural biology has benefited greatly from previously observed degeneracies in structure. The decomposition of the known structural universe into a finite set of compact TERMS offers exciting opportunities toward better understanding, design, and prediction of protein structure.

tertiary motif | structural degeneracy | protein structural universe | sequence–structure relationships | structural modularity

In this work, we aim to decompose the protein structure space into its basic elements as a way of understanding its design principles and describing its limits. Reductionist representations of protein structure have been of long-standing interest (1), with many studies having shown degeneracy at various structural levels (2–5). Features ranging from backbone or side-chain dihedral angles (6, 7) to domains and folds (8, 9) have been classified, and structural basins of attraction have been found in select motifs (10–17), offering a glimpse of a modular space with frequently repeating elements. The reason behind this modularity appears to be a combination of evolutionary history and the fundamental physics of structure. In particular, degeneracy at the level of domains, folds, and functional modules is likely strongly influenced by evolution, whereby such elements recur in different proteins often because of a common ancestor (18, 19). On the other hand, statistics of more detailed structural features are better explained from the thermodynamic perspective. For example, observed Ramachandran backbone dihedral angle preferences are largely determined by local backbone energetics (6). Similarly, the frequency of amino acids in different secondary-structural environments is closely related to thermodynamic propensities (20–24).

Degeneracy at the secondary and supersecondary structural levels has been well studied (2, 3, 16), with emergent insights greatly benefiting protein design and structure prediction applications (4, 5, 25–30). Much work has focused on clustering short

contiguous backbone fragments of fixed length (5, 31–34). For example, Kolodny et al. (31) created libraries of four- to seven-residue fragment clusters, which were later used to enable rapid search for structural similarity (35). The BriX project created a thorough hierarchical library of contiguous backbone fragments clustered by length (4–14 residues) with an rmsd threshold ranging from 0.5 Å to 1.0 Å (5). This library has been used to model loops (5) and reconstruct protein backbones (3). In addition, using pairs of fragments, this database has been extended to characterize quaternary interactions (36) and incorporated into protein–peptide docking (4). In general, the discovery of modularity at the contiguous-backbone level, with emergent sequence/structure statistics, has laid some of the foundation for modern structure prediction and protein design methods (2, 26, 27, 37–41).

In early work on supersecondary structure, Thornton and coworkers (42) showed that some supersecondary motifs were over-represented in the most common protein folds. A more recent analysis by Fiser and coworkers (16) classified all instances of two consecutive regular secondary-structural elements (SSEs) connected by a loop based on four parameters defining the relative orientation of the two SSEs, showing considerable degeneracy and saturation of the Protein Data Bank (PDB). The library of these motifs (Smotifs) has been used in loop and structure prediction (28, 29). The modularity of Smotifs is consistent with emerging experimental evidence to suggest that supersecondary motifs can serve as standard building blocks of structure. For example, Kopec and Lupas (43) found that small blade-like motifs can give rise to highly diverse folds, and Tawfik and coworkers (44) experimentally demonstrated evolutionary trajectories for the emergence of β -propeller proteins from similar short motifs. Furthermore, Alva et al. (45) identified a set of supersecondary fragments, found across divergent folds, that could have served as ancestral peptides for a broad variety of domains through repetition, fusion, and recombination.

Significance

Proteins fold into intricate 3D structures, determined by their amino acid sequences. Different proteins can fold into drastically different structures, and the space of all possible structures appears hopelessly complex. However, this is precisely the space that needs to be described to understand how sequence encodes structure. In this paper, we decompose the set of known protein structures into standard reusable building blocks, which we call tertiary structural motifs (TERMs). Strikingly, we find that only ~600 TERMS describe 50% of the known protein structural universe at sub-Angstrom resolution. Furthermore, we find the natural utilization of TERMS gives us a means of uncovering sequence–structure relationships. These insights can be harnessed for protein structure prediction, protein design, and other applications.

Author contributions: G.G. designed research; C.O.M., J.Z., and G.G. performed research; C.O.M. and G.G. analyzed data; and C.O.M. and G.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹To whom correspondence should be addressed. Email: gevorg.grigoryan@dartmouth.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1607178113/-DCSupplemental.

In this study, we aim to characterize the degeneracy across all levels of the structural hierarchy, including tertiary and quaternary structure. We hypothesize that the universe of allowed local 3D structural environments is not continuous but better described as a collection of attractors corresponding to naturally recurrent motifs. Beyond the trivial effects of volume exclusion or electrostatics to prohibit some geometries, there is another fundamental reason to expect degeneracy—namely, the differential “designability” of protein structure. Designability quantifies how feasible it is to engineer a given structure using naturally occurring amino acids, which can be defined as the number of sequences that uniquely fold into the structure. Although systematic experimental validation of this concept has been difficult, differential designability has been clearly demonstrated *in silico* (46–48) and suggested to play a key role in the evolutionary selection of folds (49–51). Over- and underrepresentation of geometries within select structural motifs has also been associated with designability (10, 26). A simple example is the α -helical coiled coil, a ubiquitous protein domain in which two or more helices wrap around each other (52). Geometric parameters in natural coiled coils fall within well-defined ranges (10), also corresponding to structures designed either *de novo* or rationally (52–54). On the other hand, the majority of parameter space, much of it entirely plausible from the perspective of molecular mechanics, does not appear to contain folded states for natural or designed sequences. We argue that this concept should generalize beyond coiled coils to tertiary structure types in general—it should be harder to create productive amino acid interactions in the context of some backbone geometries than others, introducing strong biases in natural abundance.

Here, we have undertaken a substantial computational effort (over 25 processor years) toward describing the fundamental degeneracy of the known protein structural universe by identifying its most recurrent local 3D backbone geometries, which we call tertiary structural motifs (TERMs). Our framework intentionally avoided limiting the definition of a TERM to, for example, a motif composed of a fixed number of SSEs. Instead, we identified the optimal set of motifs automatically by minimizing the number of different ones needed to describe the structural database. The emergent TERMS thus represent an extremely compact (yet nearly complete) summary of protein structure. In fact, just 625 TERMS are sufficient to describe over 50% of the structural universe. We go on to show that TERMS capture fundamental sequence–structure relationships, presumably because sequence statistics associated with instances of a TERM are constrained by the physics relevant to the formation of the motif. We show, for example, that sequences designed purely based on the statistics of TERMS comprising a structure of interest are similar to native

sequences, whether NMR or X-ray backbones are used as input (~22% and ~29% sequence identity, respectively). TERM-based sequence profiles also agree closely with corresponding evolutionary profiles, with as much as 42% identity between computed and evolutionary consensus sequences. Furthermore, the presence and locations of specific TERMS in protein chains can be predicted from sequence alone, using preferences emergent from TERM instances in unrelated proteins. We find that TERMS often recur within entirely unrelated proteins and in different topological environments, suggesting that they may have reemerged in multiple contexts by convergent evolution. It thus appears that much of the observed degeneracy is not the result of insufficient sampling by either nature or the structural database, but may be a consequence of an underlying process that filters out unproductive (i.e., nondesignable) motifs. As structural data continue to accumulate, more accurate characterizations of structural degeneracies should provide increasingly informative principles to the benefit of protein design, structure prediction, and other problems of structural biology.

Results

Discovery of TERMS. The degeneracy of secondary structure can be revealed by finding a representative set of motifs (contiguous backbone fragments) that together describe the local backbone geometry around all residues in the database (3). By analogy, to capture the degeneracy at higher levels of the structural hierarchy, we must find motifs (not necessarily sequence-contiguous) that describe tertiary/quaternary structural environments around all residues. We define such an environment for each residue in the database via its surrounding structural motif—the residue’s TERM. Specifically, the TERM for residue i is defined as the union of its local backbone (residues $i-2$ to $i+2$) and the local backbones around all residues with which i forms “potential contacts” (PCs) (examples in Fig. 1A). A PC is a pair of protein positions that can accommodate contacting amino acids (*Materials and Methods*) (55).

Representative secondary and supersecondary motifs have been identified by clustering (3, 56), but this method would be complicated for TERMS by the ambiguity of comparing different-sized motifs comprising different numbers of disjoint segments. To mitigate this problem, we adopted an approach that avoids comparing motifs to each other, and instead directly seeks the smallest set of motifs that jointly describe the structural universe. Each TERM is characterized by the subset of the universe it describes, which enables the smallest subset of TERMS covering the entire universe to be found by solving the classical set cover problem (57). More specifically, the universe was defined as the set of all nonredundant residues and PCs (“universe elements”) in our

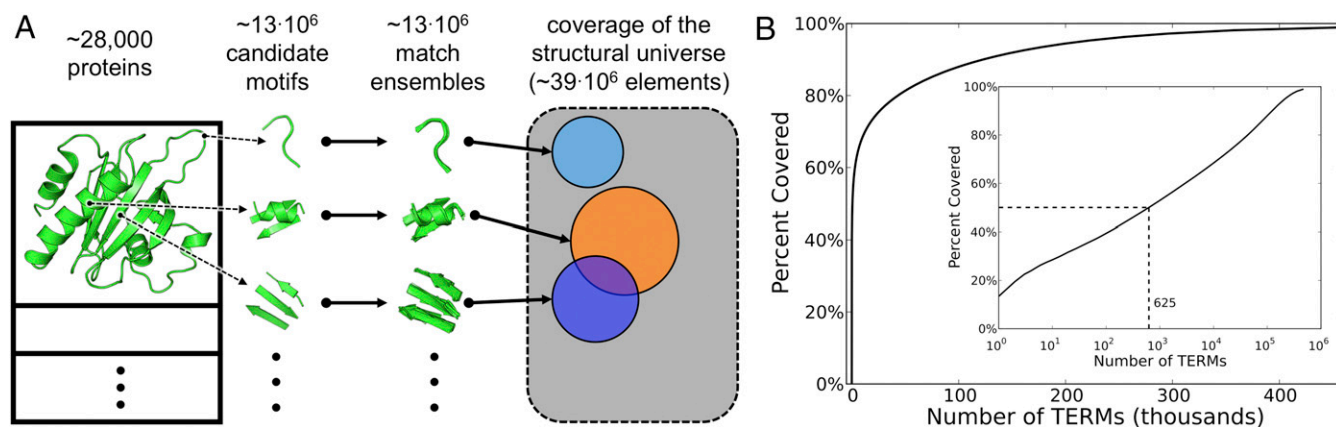


Fig. 1. Discovering TERMS that optimally describe the protein structural universe. (A) A candidate motif is defined around each residue in the database, structural matches (from within the database) to each motif are identified using MASTER (58), and these matches are used in defining the coverage of every motif. Next, the set cover problem is solved to find the minimal set of motifs that jointly cover the structural universe. (B) Coverage of the universe as a function of the number of TERMS, in the order discovered by the greedy algorithm (inset uses logarithmic scale along the x axis).

structural database of ~29,000 proteins and ~67,000 chains (redundancy due to homology removed; *Materials and Methods*). Then, for each candidate motif (one defined around each residue in the database), we used our search engine MASTER (58) to identify all matching substructures in the database, with universe elements within these matches said to be covered by the candidate. Finally, the minimal subset of candidate motifs covering the entire universe, the universal TERMS, was found (Fig. 1A). The candidate motif that gave rise to each universal TERM was referred to as the TERM's centroid and all other matching substructures as its instances.

The set cover problem is NP-complete (59), but a simple greedy solution guarantees a close approximation of the optimum (57). In our case, this greedy approach involved keeping track of universe elements already covered and iteratively choosing the candidate motif that covers the most currently uncovered elements. The set of universal TERMS resulting from this analysis was thus an approximation of the smallest set of motifs that jointly described the structural universe. Moreover, motifs emerged from this procedure in the order of their relative importance for describing the universe, hereafter referred to as "priority" (higher rank order corresponds to lower priority). It was thus trivial to create subsets of universal TERMS sufficient to cover any fraction of the universe by considering the highest-priority TERMS adding up to the necessary coverage level.

Because we did not prespecify a fixed number of disjoint segments for universal TERMS, motifs are discovered automatically, dictated by the natural patterns within the structural database. Complex multisegment motifs have the advantage that each match covers a large number of universe elements. On the other hand, simpler motifs (e.g., short single-segment ones) are likely to have many more matches, although each may cover few elements. A balance between these two considerations establishes the most parsimonious representation of the universe, which is precisely what is needed to understand the extent of its degeneracy. Furthermore, because we defined the universe to include both residues and PCs, and we included unique protein-protein interfaces in our database (*Materials and Methods*), emergent universal TERMS describe secondary, tertiary, as well as quaternary structural levels, to the extent that this information is represented in the PDB.

Definition of a Structural Match. The above procedure needs to find for each candidate motif all matching substructures in the database. The rmsd computed over backbone atoms is a convenient metric of structural similarity, but one needs to define a cutoff for admitting an alignment as a match. A constant rmsd cutoff, applied across motifs of different sizes and complexities, would unfairly reward smaller motifs. We thus propose an empirical rmsd cutoff for a motif t , as a function of its size and complexity (see *SI Appendix, SI Methods* for a detailed derivation):

$$c(t) = \sigma_{\max} \sqrt{\left(1 - \frac{2}{N(N-1)} \sum_k \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} e^{(i-j)/L}\right)}. \quad [1]$$

Here, σ_{\max} is a "resolution" parameter (i.e., the maximum possible rmsd cutoff that is imposed in the limit of large/complex structures), the first sum under the square root extends over disjoint segments of the motif, n_k is the length of the k th segment, N is the total length of the motif (i.e., $N = \sum_k n_k$), and L is a correlation length between residues in protein chains (on the order of ten residues; *SI Appendix, Fig. S1*). Here, we chose σ_{\max} to be around 1.0 Å to provide atomistic resolution, which meant that for most TERMS, the rmsd required for a match was far below an angstrom and could get as high as ~1 Å for the largest/most complex motifs (*SI Appendix, Fig. S2A*). We have verified, using motifs of varying sizes and complexities, that this universal rmsd cutoff definition produces match ensembles consistent with our intuition of structural similarity and does not appear to unreasonably reward either small or large motifs (*SI Appendix, Fig. S2B*). Based on our exper-

imentation, σ_{\max} and L values in the ranges of 0.9–1.1 Å and 10–20 residues, respectively, give reasonable results. Unless otherwise specified, data presented here were generated with the 1.0 Å/20 parameter combination, although results were generally quite similar with different cutoffs in this parameter range.

A Small Number of TERMS Describe Most of the Structural Universe.

Universal TERMS emergent from the set cover procedure reveal substantial degeneracy, with just 625 TERMS describing over half of the structural universe (or ~39 · 10⁶ universal elements; Fig. 1B). In fact, just 3.5% of the ~13 · 10⁶ candidate motifs considered (or ~458,000) describe 99% of all structure in our database. The coverage appears to increase logarithmically with the number of motifs, showing that a few TERMS are enough to describe most structural space, but many are needed to describe everything (Fig. 1B, *Inset*).

Universal TERMS vary in size, ranging from 5 to 56 residues and 1 to 10 segments (*SI Appendix, Fig. S3 B and C*). β -Strand content increases with the number of segments, whereas helical content decreases (*SI Appendix, Fig. S3A*). Three-segment TERMS are the most frequent, representing 30% of all motifs. As seen in Fig. 2B, high-priority TERMS exhibit considerable diversity, ranging from simple single-segment to more complex multisegment motifs, representing prototypical structural patterns. Out of the top 24 TERMS (together covering a third of the universe), there are 4 major categories: helices, two-strand β -sheets, turns, and helix-helix motifs (Fig. 2A). There are important structural differences between TERMS in each category. For example, the second and fifth ranked TERMS are both two-segment antiparallel β -sheets and may appear to be redundant. However, the average rmsd between instances of these two TERMS is 2.0 Å, so they cannot substitute for one another. Closer inspection reveals that the two motifs are centered on two topologically distinct sites in antiparallel β -sheets (*SI Appendix, Fig. S4*) (60). Because such a distinction does not occur in parallel sheets, we may expect fewer parallel than antiparallel two-strand TERMS. Indeed, out of 114 two-strand β -sheets among the top universal TERMS (up to 50% coverage) only 29 are parallel. Note that this recognition of distinct environments in anti-parallel β -sheets, described previously as the alternation of "small" and "large" H-bonded rings in the main chain (60), arose automatically in motif generation. Delineation of other subtle but important structural details is also apparent with helical TERMS, where individual motifs represent α - and 3₁₀-helices along with turn-helix and helix-turn geometries (e.g., TERMS 1, 20, 15, and 16, respectively).

Because the coverage curve slows down considerably after its initial rapid increase (Fig. 1B), low-priority TERMS must be describing less common geometric scenarios. As shown in *SI Appendix, Fig. S5*, incidences of high B-factors and low occupancies increase with decreasing TERM priority but are overall quite low even toward the tail end of coverage. Thus, although some low-priority TERMS arise due to poorly modeled structural regions, many appear to represent legitimate infrequent conformations. To determine whether these motifs are variations on high-priority TERMS or represent entirely different geometries, we measured the amount of structural "novelty" represented by late-arising motifs, which was defined as the ratio between the portion of the universe covered exclusively by low-priority TERMS and the total portion of the universe covered by these TERMS. As shown in *SI Appendix, Fig. S6*, after the first few thousand TERMS are added, the majority of what is to be described by lower-priority TERMS is already captured by the high-priority ones. Thus, we can think of low-priority TERMS as mainly representing relatively small deviations from their high-priority counterparts, which nevertheless put them past the defined similarity cutoff. Consistent with this finding, low- and high-priority TERMS also show similar distributions of size and number of segments (*SI Appendix, Fig. S7*).

The PDB is Close to Saturated in TERMS. Protein structural data will continue to accumulate. We thus wondered whether newly characterized proteins, especially those considerably different from any in the current database, would still be composed of the same universal TERMS and would exhibit similar degeneracy. To

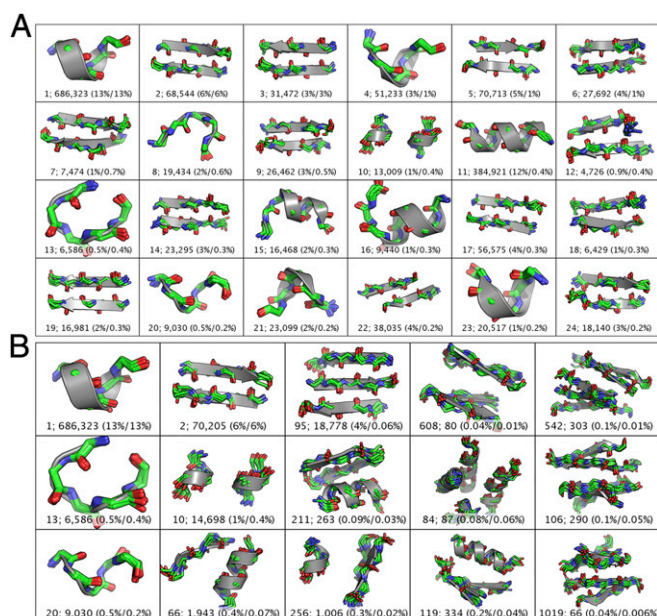


Fig. 2. Universal TERMS. (A) Top 24 TERMS ranked by the number of elements covered in the set cover procedure; jointly these cover roughly a third of the universe elements. (B) A diverse selection of high-priority TERMS that span from one- to five-segment motifs, shown in the first to fifth columns, respectively. Shown in each column are representatives from the three most common secondary-structure classes for the given number of segments (*SI Appendix, SI Methods*). In both A and B, each TERM is represented with ten randomly chosen matches along with its centroid. The text underneath each TERM is formatted as follows: r ; n (s/c) where r is the rank of the TERM in the set cover (lower rank corresponds to higher priority), n is the number of unique matches, s is the total fraction of universe elements covered by the TERM, and c is the marginal fraction of the universe elements covered by the TERM (i.e., fractional coverage of those elements not already covered by preceding TERMS in the set cover).

answer this question, we identified all protein structures deposited into the PDB since the start of our project and selected only those that shared less than 35% sequence identity to any structure in our original database (also applying the same quality filters as for the original set; *Materials and Methods*). This procedure yielded 1,095 novel structures, and we analyzed the extent to which these were covered by universal TERMS derived from the initial database. Fig. 3 compares coverage of the original database with that of novel structures, using a growing subset of universal TERMS considered in the order of their priority. Clearly, coverage is roughly identical up to ~70%, which is achieved with the first ~9,000 TERMS. Thereafter, coverage of the novel dataset slows, with a total of 82% of the new structural data eventually covered. This finding argues that the most important TERMS discovered by our analysis, those that arise early in the set cover process and cover the majority of the universe, are indeed universal and will likely continue to represent the majority of designable structure space. On the other hand, new rare motifs, such as those represented in the tail end of the coverage curve in Fig. 1B, are likely to continue arising. Because our novel dataset here comprised of only proteins highly divergent from any in the current database, this experiment constitutes a stringent test of TERM universality, so that coverage of an “average” future structural dataset is expected to be much higher. This observation was robust to small differences in the rmsd cutoff function. The plot shown in Fig. 3 corresponds to the rmsd cutoff function with $L = 15$ and $\sigma = 1.1$ Å, whereas the combination with $L = 20$ and $\sigma_{\max} = 1.0$ Å showed a similar behavior (*SI Appendix, Fig. S8*; in each case, the same cutoff definition is used for deriving universal TERMS and measuring coverage within novel structures).

TERM Instances Rediscovered by Nature. The fact that TERMS recur in unrelated proteins suggests that they may represent design-

able attractors in the protein structure space, which may have reemerged multiple times throughout evolution. To further support this idea, we sought to characterize whether TERMS recur in similar or different topological contexts. In particular, we asked whether in multisegment TERMS the order of segments (in sequence) is generally preserved or divergent. Given that TERM segments are often quite far away in sequence, it appears unlikely that alternative segment permutations would recur as a result of close evolutionary homology. Having analyzed two-, three-, and four-segment TERMS, we see that topology is indeed typically sampled broadly among TERM instances. For example, among the frequently recurring three-segment TERMS (i.e., those with at least 20 nonredundant matches; *Materials and Methods*), 33% have all six possible segment permutations within their instances, with at least two different permutations occurring in 80% of cases (Fig. 4B). A similar topological diversity exists with two- and four-segment motifs (Fig. 4A and C), although it appears to decrease slightly with the number of segments. Furthermore, on average 41% of three-segment TERM matches have a topology different from that in the centroid motif (Fig. 4D). Interestingly, there does appear to be some bias toward specific segment orders (e.g., Fig. 4D), which may be a consequence of weak residual homology in our universe, but it could also reflect underlying kinetic or energetic preferences. Baker and coworkers (26) have shown that for certain supersecondary structural motifs, segment order does affect the energetics of folding, and distributions expected from simulation generally agree with those observed natively.

Specialization of TERMS. Native protein structures are evolutionarily filtered not only for designability, but also function. Thus, some universal TERMS may recur in a variety of contexts by virtue of being associated with a ubiquitous function. Such function-linked TERMS may be somewhat less “generic” than their purely structural counterparts, in that they may occur only within a functionally biased subset of proteins. However, these motifs would still need to be used broadly across diverse proteins within such subsets to emerge from our set cover procedure with high priority. We looked for TERMS specialized for metal binding by identifying motifs whose instances are enriched in contacts with metal ions (*SI Appendix, SI Methods*). *SI Appendix, Fig. S9* shows the most enriched TERMS for several metals. These motifs originate from diverse proteins, and some of their instances are unrelated to metal coordination. For example, we found the ubiquitous calcium-binding domain EF hand (61) to be clearly reflected in universal TERMS, with a representative motif shown in Fig. 5A–C. Note that metals did not directly participate in our universe database, so this motif (like other metal-specialized TERMS) was identified purely on the basis of recurrent backbone geometry. In accordance, not all of the TERM’s instances correspond to EF hands (Fig. 5C), and the ones that do originate from diverse proteins (*SI Appendix, Table S1*) and in a highly variable sequence contexts (Fig. 5D). This result agrees well with prior findings on the independent evolutionary past of EF-hand subfamilies (62). It is also interesting that the above TERM appears to correspond to the most structurally conserved portion of the EF hand and excludes the binding loop and much of the N-terminal helix (Fig. 5A–C), both of which are known to exhibit substantial variation (63). Additional examples of functionally linked TERMS, including a modular metal-coordinating motif and water-binding TERMS, are presented in *SI Appendix, SI Results* and Figs. S10–S13.

Quaternary Structure. We also looked for TERMS that specialize in forming interfaces. To this end, we ranked TERMS by the number of covered quaternary PCs (i.e., PCs that involve residues on different chains). The top 25 TERMS, shown in *SI Appendix, Fig. S14*, fall into two anticipated classes—helical-bundles and β -sheets. Although the helix-helix TERMS in *SI Appendix, Fig. S14A* appear similar by eye, they correspond to distinct interhelix geometric parameters (10, 64). Whereas these TERMS denote some of the most common interfacial binding modes, most are not exclusive to interfaces and the majority of their instances originate from

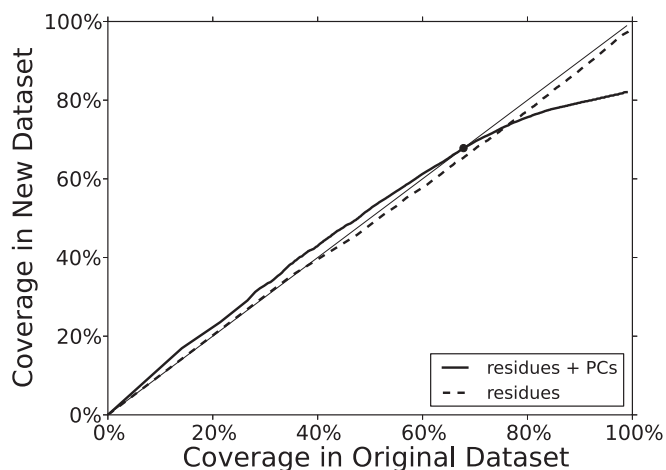


Fig. 3. Coverage in a test dataset of 1,095 proteins highly divergent from those used to create universal TERMS. The thick solid and dashed lines represent coverage of all universe elements and just residues, respectively. The thin line designates $x=y$. Up to 68% (indicated by a bold dot) the two sets are covered roughly identically.

noninterfacial contexts. This finding supports previous suggestions that interfacial and tertiary geometries largely resemble one another (36). On the other hand, TERMS in *SI Appendix, Fig. S144* jointly describe only 4% of all quaternary PCs in the universe. For comparison, top 25 TERMS chosen for tertiary coverage jointly describe 15% of tertiary PCs in the universe. Thus, although the most common geometries may be shared between tertiary and quaternary structure, there are significant differences between the two, with the latter appearing to sample more diversity. In agreement with this observation, coverage of quaternary contacts generally lags behind that of tertiary ones as the set cover progresses (*SI Appendix, Fig. S15*).

TERM Sequence Statistics Enable Design. Next, we tested the hypothesis that amino acid statistics from TERM instances represent, to some degree, fundamental sequence–structure relationships. To this end, we asked whether TERM statistics would predict sequences optimally compatible with native backbones to be close to the corresponding native sequences. This experiment, known as “native sequence recovery,” is a common means of evaluating scoring functions in computational protein design (65).

Our sequence design procedure for a target backbone consisted of three major steps (*Materials and Methods* and *SI Appendix*). First, we found all instances of universal TERMS within the target

structure using MASTER. Second, positional (self) and pairwise pseudoenergies were calculated from the sequences of the matching TERMS (*SI Appendix, Eqs. S3 and S9*). These pseudoenergies effectively captured amino acid distribution biases at positions or pairs of positions from corresponding TERM matches. Lastly, we used integer linear programming to find the sequence that minimized the total pseudoenergy for each target structure. Note that in all of these operations we removed any homology to the target protein from TERM statistics (*Materials and Methods*).

We performed the native sequence recovery experiment on four different datasets: two with X-ray and two with NMR backbones (26, 66, 67) (*Materials and Methods* and *SI Appendix, Table S3*). On average, predicted sequences were 29% and 22% identical to the native with X-ray or NMR backbones, respectively (Table 1). To put these results in perspective, we also evaluated the performance of the state-of-the-art protein design suite Rosetta Design (68), with the atomistic scoring function talaris2013 (65), on the same datasets. Rosetta achieved higher sequence identities than the TERM-based approach (Table 1), by 5% on average (only 2.5% for NMR structures). The small difference is notable because, whereas fixed-backbone design benefits greatly from strong steric constraints (i.e., the “backbone-memory” effect), the TERM-based approach interprets the backbone much more loosely, seeking sequences consistent with structural ensembles of constituent TERM instances. Accordingly, fixed-backbone design exhibits a larger difference between NMR and X-ray backbones than the TERM-based approach (Table 1). This result is especially clear with X-ray-2 and NMR-2 sets, which contain X-ray and NMR structures, respectively, of the same 11 proteins, verified to be devoid of any major structural differences (67). Here Rosetta easily outperforms TERMS with X-ray structures (32% versus 27%), whereas the two perform similarly on NMR structures (23% versus 21%). Furthermore, corresponding sequences designed for X-ray-2 and NMR-2 sets by the TERM-based method were 46% identical to each other (on average), whereas this similarity was only 26% for Rosetta. The two structure sets represent the same folded ensembles, so ideally similar sequences should be obtained when designing with either. TERM-based backbone decomposition thus appears to strike a balance between interpreting backbone coordinates loosely enough to recognize similar conformations as representing related ensembles, and yet precisely enough to suggest native-like sequences.

As stated above, we aimed to carefully remove influence from direct evolutionary homology to the redesigned protein in our TERM-based approach. However, to test the possibility that remote residual homology may still influence our results, we considered the set of five NMR backbones of de novo-designed proteins recently published by Baker and coworkers (26), the NMR-1 set. Unsurprisingly, these proteins are not close in

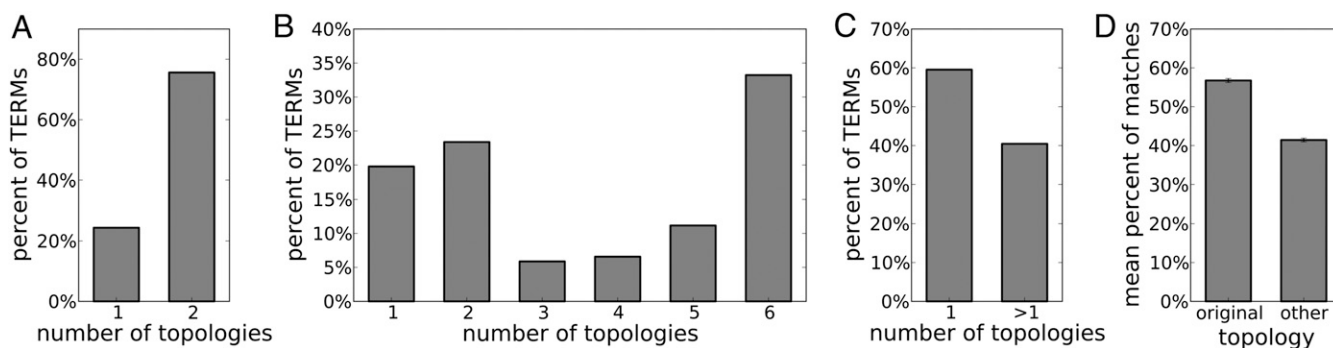


Fig. 4. The distribution of the number of topologies found in high-priority TERMS with two (A), three (B), and four (C) segments. All four-segment TERMS with more than one topology are placed into a single bin (>1 representing TERMS containing 2–24 topologies). (D) The topology of TERM matches compared with the centroid motif for three-segment TERMS. Bin “original” represents matches with the same topology as the centroid, whereas “other” corresponds to any of the other five possible topologies. Shown is the mean percentage of matches (over all TERMS) in either category, with error bars designating SE.

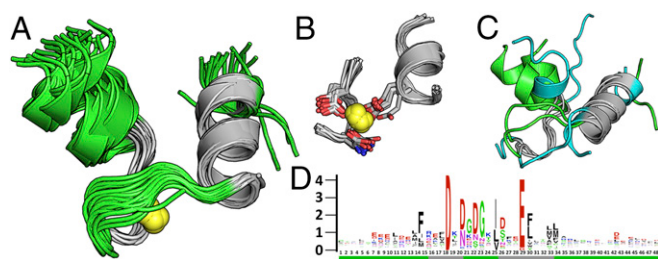


Fig. 5. An EF-hand TERM. (A) The 31 nonredundant EF hand-containing instances of the TERM (gray) with adjacent structure (green). Calcium atoms from TERM instances are shown as yellow spheres. (B) TERM instances alone with calcium-contacting side chains shown with sticks. (C) Variability among TERM instances. Four instances are shown in gray: two EF-hand examples with varying loop geometries (surrounding structure in green) and two non-EF-hand instances (from PDB ID codes 3HNO and 1CB7, surrounding structure in cyan), including one with TERM segments belonging to different chains. (D) Sequence logo of nonredundant EF hand-containing matches of the TERM. Position 18 corresponds to the canonical EF hand loop position 1 (61).

sequence to any other proteins in the PDB and do not, by construction, have any homologs. Nevertheless, our TERM-based approach performs equally well on this set as on other NMR structures and also quite similarly to Rosetta which was used to design these proteins (Table 1).

Sequences designed using TERMS were only 24% identical to the corresponding ones from Rosetta, suggesting that the two approaches are considerably orthogonal and may be complementary. To test the latter, we used positional TERM-based pseudoenergies to limit the choice of amino acids at each position to an average of ~ 10 possibilities (*Materials and Methods*). Doing so improved Rosetta's already high native-sequence recovery rates in all cases but especially with NMR structures. This result suggests that TERM-based information may be an effective means of capturing ensemble preferences in the context of standard fixed-backbone design.

TERMs Explain Evolutionary Variation. As a further test of the hypothesis that TERMS capture fundamental sequence–structure relationships, we asked whether evolutionary sequence variation could also be rationalized on the basis of TERM statistics. To this end, we used the pseudoenergies computed above to perform Metropolis Monte Carlo (MC) simulations in sequence space for all proteins in the benchmark sets (*SI Appendix, Table S3*). Final sequences from each of 100 independent simulations constituted the predicted sequence variation for the given protein. Evolutionarily variation was measured by clustering results of a BLAST search (69) against the nonredundant protein database (*Materials and Methods*).

For each protein, we compared the positional frequencies predicted through the MC simulation with frequencies emergent from the evolutionary multiple-sequence alignments (MSA). Both were normalized by the background amino acid frequency, producing enrichment ratios, to remove the trivial effect of a non-uniform genomic distribution present in both cases. The overall correlation between TERM-predicted and evolutionary enrichment ratios (log-transformed) was $R = 0.51$, with 30% of the positions having correlations of $R = 0.8$ or higher (*SI Appendix, Figs. S16 and S17*). Furthermore, the predicted enrichment ratios were within a factor of two of the evolutionary ones for 46% of position/amino acid combinations, and within a factor of three for 71% of these.

In cases where the evolutionary ensemble could be determined with good confidence (i.e., final MSA had at least 1,000 sequences), we also directly compared the emergent sequence logos with those from MC simulations. The 20 cases that fell into this category were ranked by the average per-structure correlation between TERM-predicted and observed enrichment ratios, with the best, worst, and median correlations being $R = 0.74$, $R = 0.41$,

and $R = 0.58$, respectively. In *SI Appendix, Fig. S18*, we compare predicted and observed sequence logos for the three proteins representing these extreme cases. Overall, there is a striking similarity between prediction and observation, even in the worst case, despite the fact that the influence of direct homology to the protein being analyzed was removed in calculating TERM-based energies.

Finally, we observed that in 35% of positions, the most frequent amino acid in the MC-generated ensemble was also the most common in the evolutionary ensemble, and this value increases to 42% for the 20 cases with most confident evolutionary MSAs (*Materials and Methods*). Interestingly, these fractions are higher than the native sequence recovery rates in Table 1, suggesting that TERM-based pseudoenergies are more reflective of sequences compatible with the ensemble of states to which the given backbone belongs rather than the precisely specified conformation.

TERMs Map Sequence to Structure. In the final test of TERM statistics, we asked whether they could also enable the prediction of structure from sequence alone. Specifically, we tested whether sequence preferences emergent from a given TERM's instances are sufficient to predict its likely occurrences in nonhomologous protein chains. Such prediction is challenging because a TERM encompasses only some of the determinants behind its formation in a given structure, the rest contributed by the surrounding environment and interactions. Thus, the sequence signature of an isolated motif, even if it reflects the underlying physics of structure exactly, may not be sufficient to identify whether and where the motif occurs within a protein sequence. Adding to the challenge is the fact that the number of possible alignments of a TERM onto a given sequence grows exponentially with the number of disjoint segments in the TERM, with only a handful of these, if any, being correct.

Using the weak coupling framework reported by Weigt and coworkers (70), we built a two-body statistical sequence model for each TERM from the MSA of its PDB instances (*SI Appendix, SI Methods*). With this model, we scored all possible alignments for each of the top 4,000 highest-priority universal TERMS (with up to three segments) onto each protein sequence from the above X-ray-1 and NMR-1 sets. The best scoring alignments for each protein were then predicted to form the corresponding TERM. Fig. 6A shows the fraction of these predictions corresponding to correct structural alignments as a function of the number of predictions made, with native and de novo-designed proteins separated (*SI Appendix, Fig. S19* shows examples of correctly predicted

Table 1. Sequence recovery results

Dataset*	Method	SID, % [†]	B/E, % [‡]	Cons., % [§]	Top 3, %	Cov., % [¶]
X-ray-1 (66), 64	TERMs	29.3	30/27	49.5	51.8	96
	Rosetta	35.5	39/29	50.6		
	Combined	36.1	39/32	51.1		
X-ray-2 (67), 11	TERMs	26.7	30/22	50.8	49.9	98
	Rosetta	31.9	38/23	49.1		
	Combined	32.6	38/26	48.6		
NMR-1 (26), 5	TERMs	25.4	25/26	60.0	45.3	93
	Rosetta	28.3	32/22	54.6		
	Combined	32.0	33/29	55.3		
NMR-2 (67), 11	TERMs	20.6	23/17	44.8	54.6	90
	Rosetta	22.9	25/19	41.6		
	Combined	24.5	27/20	42.3		

*Dataset name, source citation (in parentheses), and number of proteins.

[†]Sequence identity (SID) between designed and native sequences.

[‡]SID among buried/exposed (B/E) positions.

[§]The degree of conservation (cons.) between designed and native sequences (by physicochemical class; see *Materials and Methods*).

^{||}The frequency of the native amino acid being in the top three residues by TERM self pseudo-energy.

[¶]Average fraction of residues in the corresponding dataset covered (cov.) by TERMS.

TERMs from a representative protein; *SI Appendix, SI Methods*). Despite the challenges outlined above, there is a clear tendency for best-scoring alignments to be correct. Interestingly, performance on de novo proteins is markedly higher than on native ones, which is consistent with a key design principle behind the de novo structures—complete agreement between local and global interactions (i.e., lack of conformational frustration) (26). On the other hand, such high performance on de novo-designed proteins, which by definition do not have any homologs, demonstrates that success of TERM statistics here is not due to any residual homology that may escape our filtering procedure.

The success rates for two- and three-segment TERMs are isolated in Fig. 6B and C, respectively. Because the number of possible alignment associated with these motifs is very high (e.g., on the order of 10^4 and 10^6 alignments for a two- and three-segment TERM, respectively, onto a 100-residue protein), one would expect

a near-zero success rate at random, whereas we see considerably higher rates in practice. For example, for two-segment motifs, over 20% of predicted alignments are correct for de novo proteins and 5–8% for native ones. However, are multisegment TERMs contributing unique information on the compatibility between their segments, or are they merely having the effect of filtering for secondary structure? To answer this question, we estimated the expected rate of successful alignments in a scenario where we have full knowledge of the local backbone conformation. That is, we know all of the structurally matching alignments for each individual segment of a multisegment TERM, but we do not know which combinations correspond to correct alignments of the entire TERM (if any) and predict at random. Under this scenario, we find that correct alignments for two- and three-segment TERMs would be discovered at rates of 0.2 and 0.01%, respectively, in our benchmark proteins. Thus, data from multisegment TERMs do, in fact, provide considerable additional information beyond secondary structure. This is not to say that secondary structural information is not helpful. In fact, when predicted secondary structure is used to bias TERM alignments (*SI Appendix, SI Methods*), the performance increases considerably in all cases, with top alignments for two-segment TERMs in native proteins now correct in ~10% of cases (*SI Appendix, Fig. S20*). Furthermore, even if not all top-predicted alignments are correct, most correct alignments are, in fact, found toward the top of the list ordered by the statistical energy, as shown in Fig. 6D and *SI Appendix, Fig. S20D*.

Discussion

The goal of this study was to develop a systematic decomposition of the known protein structure space that is compact, universal, and detailed enough to provide insight into structure–sequence relationships. The collection of TERMs we synthesize here constitutes just such a decomposition, covering secondary, tertiary, and even quaternary levels of the structural hierarchy. Furthermore, the method by which we extract TERMs, via the set cover formalism, is general and can be used to develop decompositions with alternative structural databases and definitions of coverage.

We find that the protein structural universe is highly degenerate, which is clear from the rapid increase of structural coverage as a function of the number of TERMs used (Fig. 1B). On the other hand, to go past 70–80% of coverage, tens of thousands of individual TERMs are required, and the overall coverage curve follows a power law (Fig. 1B, *Inset*). It thus appears that despite being highly degenerate and repetitive, the universe nevertheless continually innovates. Motifs toward the tail end of the coverage curve in Fig. 1B represent infrequent geometries, some of which arise from inaccuracies of structural determination (*SI Appendix, Fig. S5*), but most represent genuine (albeit relatively small) departures from the more canonical universal TERMs (*SI Appendix, Figs. S6 and S7*).

Structural thermodynamic considerations must influence the occurrence of TERMs as they do constrain the evolution of protein structure. Other physical constraints likely also contribute, including designability, structural specificity, solubility, etc. Thus, we can think of TERMs, with their pattern of recurrence and sequence biases, as encoding a certain sequence–structure mapping, driven by a complex metric that incorporates the above properties (among others). We interrogated this mapping by using TERM-based statistics to propose likely sequences given native backbones. This procedure generated sequences similar to native ones (Table 1), which is notable given that the TERM-based approach does not explicitly consider atomistic details. TERM-based pseudoenergies are even more successful at predicting evolutionary sequence variation (*SI Appendix, Fig. S18*), producing the correct consensus amino acid in 35% of positions. Together with the high relative performance on NMR backbones (Table 1), these results suggest that TERM statistics may reflect preferences of the structural ensemble represented by the given backbone, and not just the specific conformation provided.

We emphasize that our design procedure is highly simplistic and our goal in developing it was merely to probe the apparent

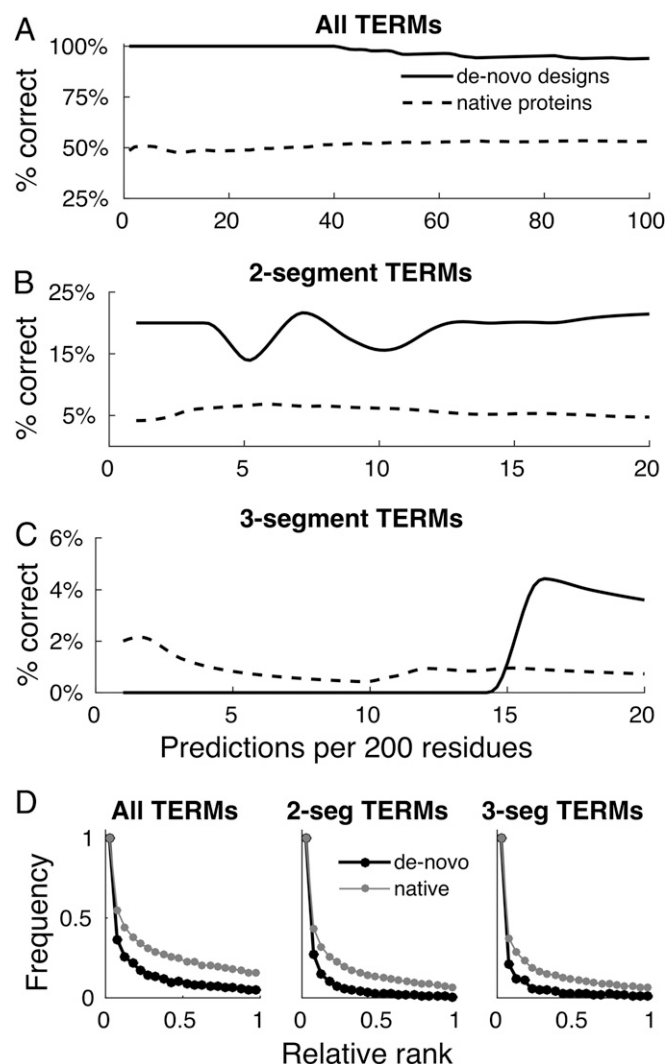


Fig. 6. TERMs enable going from sequence to structure. The benchmark set (X-ray-1 and NMR-1; Table 1) comprises de novo-designed (PDB ID codes 2KL8, 2LN3, 2LTA, 2LV8, and 2LVB) and native proteins. A shows the overall success rate of identifying structurally correct alignments, while B and C isolate the performance for two- and three-segment TERMs. Shown in D are histograms of the relative ranks of structurally correct alignments (out of all possible alignments). Relative rank for an alignment is computed as its rank based on score (with 1 corresponding to the best-scoring alignment) divided by the total number of possible alignments for the motif onto the sequence of a given benchmark protein.

sequence–structure relationships encoded by TERMS. Whether a variant of such a method could be used for a robust approach to protein design in general remains to be explored. On the other hand, our results suggest that TERM-based insight may be of utility to protein design methods. As a very first step toward exploring this possibility, we used TERM-based pseudoenergies to automatically restrict the amino acid alphabet in Rosetta-based sequence redesign, resulting in even higher native sequence recovery rates than by Rosetta alone (Table 1). A possible interpretation of this result is that requiring agreement between Rosetta's detailed atomistic scoring function and the looser ensemble-based TERM pseudoenergy, both of which are only partially accurate, enriches the remaining sequence space for good solutions (possibly at the expense of reducing the sequence space more than necessary).

Having interrogated the ability of TERMS to predict sequence from structure, we next showed that the opposite is also possible—the use of TERM statistics to predict local structural motifs from sequence alone (Fig. 6 and *SI Appendix*, Figs. S19 and S20). This capability is especially consequential for multisegment TERMS, because it means that segments distant in sequence can be predicted to be adjacent in space—a major challenge in structure prediction. Significant strides toward addressing this challenge have recently emerged from prediction of contacts based on evolutionary covariation (70, 71), but, importantly, such prediction is only applicable to native proteins, and especially those with available deep MSAs. On the other hand, TERM-based mining appears to be quite applicable to de novo proteins and requires no homology (Fig. 6), providing further evidence for the generality of TERM-encoded statistics. These results compliment our prior findings showing that sequence statistics from TERM-like motifs are sufficient, on their own, to discriminate between good and poor structure-prediction models on par with or better than leading scoring functions (55). The availability of prebuilt universal TERMS, each with its own statistical model, should enable a multitude of novel uses toward improving structure prediction.

An important fundamental question is why TERMS recur. Is it primarily due to the biophysics of protein structure and designability, or is much of the degeneracy due to, for example, functional constraints of evolution? Although the answer is difficult to determine definitively, it is likely a combination of these two factors. Intuition suggests that high-priority TERMS, which occur across an extremely diverse set of proteins, and which are not associated with any specific cellular role, localization, or host species, likely recur as a result of fundamental biophysical principles. On the other hand, TERMS that occur within proteins that are functionally biased (although still quite diverse) are likely influenced by functional constraints and evolutionary history. We have specifically sought examples of TERMS that recur in the context of a function, such as metal or water binding (Fig. 5 and *SI Appendix*, Figs. S9–S13). Importantly, however, not every function will necessarily impact TERM selection. This bias will only emerge if: (i) the function is associated with relatively well-defined structural motifs and (ii) the corresponding geometries are either otherwise ubiquitous or the function itself is common (among diverse proteins). In either case, we can regard the resulting motifs as true modules of protein structure. Thus, the universality of TERMS discovered by the set cover procedure should hold generally, whether they have an associated function or not.

Materials and Methods

Structural Data. Protein-containing biological units of X-ray entries with R-free below 0.3 and resolution better than 2.6 Å were downloaded from the PDB on 8/16/2014 resulting in 69,319 structures. All nonprotein chains were eliminated as were residues with missing backbone atoms and nonstandard names (except for MSE, HSC, HSD, HSE, and HSP). Chains or chain segments of less than 5 residues were also removed. Structures with no chains or chain segments over 10 residues were eliminated. Thus, 67,199 entries remained in the database, with 133,057 chains and ~34 million residues.

Intrastructure Redundancy Removal. Some biological units in the PDB have considerable internal redundancy (e.g., viral capsids). We thus used the

createPDS program from the MASTER package (58) to identify the minimally nonredundant subset of chains within each entry that preserved all unique quaternary interfaces. Briefly, the procedure begins by enumerating all chain “neighborhoods”—i.e., a central chain and all chains in contact with it. Two neighborhoods are considered redundant if: (i) the two central chains are at least 90% identical in sequence and superimpose to within 1.0 Å; (ii) every pair of corresponding contacting chains pass the same filter; and (iii) the entire neighborhoods superimpose to within 2.0-Å rmsd. A chain neighborhood is said to cover all chains and interchain interfaces contained with it and all neighborhoods redundant to it. Thus, having identified all redundant chain neighborhoods, *createPDS* proceeds to solve the greedy set cover problem to arrive at a subset of chain neighborhoods that together cover all unique chains and interfaces in the entry. The union of these subsets is then output, significantly reducing structure size for cases with considerable internal redundancy. Any remaining redundancy is removed in subsequent steps described below.

Gross Interbiounit Redundancy Removal. We further removed obvious redundancy from our database while maintaining unique interfaces between chains. To this end, we used BLASTclust (69) to cluster surviving chains from above at 80% sequence identity (95% coverage). Each biounit was then assigned a set of clusters associated with its chains. Two biounits were considered equivalent if their sets of associated clusters and the number of chains associated with each cluster were equivalent. Thus, unique quaternary interfaces were preserved at the expense of allowing occasional redundancy between chains in nonequivalent biounits, which was removed at a lower sequence identity cutoff in later stages. The resulting dataset, DB80, contained one representative for each class of equivalent biounits, comprising 28,747 PDB entries and 62,556 chains.

Motif Creation. A candidate motif was generated around each residue in the database (the central residue of the motif) as described in the *Results*. To find residue pairs capable of making contacts (PCs) we introduced the measure of “contact degree.” For a given pair of positions i and j , this value is calculated by first finding all possible rotamers (of all amino acids) at both positions that do not clash with the backbone [the Richardson penultimate rotamer library was used (72)]. Contact degree is then computed as the weighted fraction of rotamer combinations at i and j that have closely approaching nonhydrogen atoms:

$$\tau(i, j) = \frac{\sum_{a=1}^{20} \sum_{b=1}^{20} \sum_{r_i \in R_i(a)} \sum_{r_j \in R_j(b)} C_{ij}(r_i, r_j) Pr(a) Pr(b) p(r_i) p(r_j)}{\sum_{a=1}^{20} \sum_{b=1}^{20} \sum_{r_i \in R_i(a)} \sum_{r_j \in R_j(b)} Pr(a) Pr(b) p(r_i) p(r_j)}, \quad [2]$$

where $R_i(a)$ is the set of nonclashing rotamers of amino acid a at position i , $C_{ij}(i, j)$ is a logical variable indicating whether rotamers r_i and r_j at positions i and j , respectively, have nonhydrogen atom pairs within 3 Å, $Pr(a)$ is the frequency of amino acid a in the structural database, and $p(r_i)$ is the probability of rotamer r_i from the rotamer library. Contact degree varies from 0 to 1, with higher values corresponding to position pairs more likely to influence each other's amino acid identities. If $\tau(i, j)$ was above 0.05, residues i and j were said to form a PC (55).

For each candidate motif (query), we obtained all similar substructures in DB80 (instances of the motif) using MASTER (58). Specifically, MASTER was asked to return all matches with both C α and full-backbone rmsds below the cutoff specified by our rmsd cutoff function (Eq. 1 and *SI Appendix*, *SI Methods*). If the candidate becomes a universal TERM via the set cover procedure (see below), the query motif used to create the TERM is referred to as the centroid of the TERM and the matches (TERM instances) are sorted in ascending order by their rmsd to the centroid.

Minimum Set Cover. The universal set to cover consisted of all unique residues and PCs in DB80, representing secondary and tertiary/quaternary information, respectively. The manner in which DB80 was prepared ensured no obvious redundancy (i.e., no repeating proteins or close variants), but homologous proteins could still remain in the database. It was, of course, possible to further filter DB80 such that no pair of chains would have a detectable evolutionary relationship, but we reasoned that such filtering would unnecessarily discard unrelated regions/domains from proteins that may also have related regions. We therefore adopted an approach that accounted for local redundancy on the level of individual residues and PCs. To this end, all pairs of chains in DB80 were aligned using the Needleman–Wunsch global alignment algorithm in the FASTA package (73). Pairs of sequences were considered “neighbors” if the sequence identity over the entire alignment was above 30% and the ratio of the smaller to larger chain lengths was greater than 0.75. The sequence identity threshold was adjusted for chains below 30 residues, linearly increasing from 30 to 60% as chain length decreased from 30 to 5 residues.

Chains were then greedily clustered based on this similarity criterion. Specifically, at each iteration, the chain with the highest number of neighbors not already assigned to a cluster was identified. This chain and its neighbors were then assigned to a new cluster. This process was repeated until all chains were assigned to clusters, resulting in 14,515 clusters.

Within each cluster, we wished to identify equivalent positions and PCs and assign them matching identifiers. To this end, for the residue at position i_A in chain A , we looked through the Needleman–Wunsch alignments of A with all of its cluster neighbors, in the order of decreasing sequence identity, to find the first in which i_A is aligned against a position with an already assigned identifier. If such a neighboring chain existed, say B with position i_B aligned to i_A , then i_A was assigned the identifier of i_B . Otherwise, i_A received a new identifier representing a unique position not yet seen in the cluster. To identify equivalent PCs (corresponding to pairs of positions in alignments) within a cluster, we followed a similar procedure except that the simultaneous alignment of two residues was required. That is, for a PC between positions i_A and j_A in chain A , we sought an alignment of A with one of its neighbors, say B , with corresponding positions i_B and j_B , such that: (i_B, j_B) is a PC with a previously assigned identifier, i_B and i_A share the same position identifier, and j_B and j_A share the same position identifier. The latter two conditions were imposed to make sure that PC identifiers were assigned based on chains that were close enough for both positions to be equivalent individually. If such a chain was found, then PC (i_A, j_A) was assigned the same identifier as PC (i_B, j_B), whereas if no such chain was found, then PC (i_A, j_A) received a new unique identifier, reflecting its status as a new type of contact not already described. The process for assigning identifiers to quaternary PCs was the same, except the aligned PC (i_B, j_B) had to additionally also be quaternary. [SI Appendix, Table S2](#) summarizes the resulting number of residues and PCs before and after this redundancy removal. The nonredundant residues and PCs resulting from this procedure were referred to as the set of universe elements, U .

Each candidate motif was said to cover all universe elements that occurred within its matches. Our goal was to find the smallest set of TERMS, C , which collectively covered all of U . This is an instance of the NP-complete minimum set cover problem (59). We applied a simple greedy solution, which guaranteed roughly an $\ln(n)$ approximation to the optimum in terms of the number of TERMS needed, where n is the number of universe elements (the approximation is typically much better in practice) (57). In our case, because n is $\sim 13 \cdot 10^6$, the greedy approximation will need at most 16 times as many TERMS as the optimal solution, and likely many fewer.

The greedy procedure works by keeping track of two sets: M —the set of motifs not yet chosen as TERMS (starts as the set of all candidate motifs) and U' —the set of universe elements not already covered by any chosen TERMS (starts as U). At each iteration, the TERM, $m \in M$, that covers the most structural units in U' is removed from M and added to C , and universal elements covered by m are then subtracted from U' . We repeated this process until at least 99% of all elements in U were covered. The emergent motifs were referred to as universal TERMS with their order of discovery defining TERM priority.

Nonredundant TERM Instances. The above ensured that TERMS were selected based on nonredundant coverage of universe elements. For some analyses, however (e.g., sequence recovery experiments), it was also useful to have a set of nonredundant instances for each TERM. To this end, we used the chain clusters described above to label two instances of a TERM as redundant if: (i) corresponding segments between the two instances originated from chains form the same clusters, and (ii) for any two segments originating from a single chain in one instance, the corresponding segments in the other instance also originated from a single chain. The list of nonredundant instances of a given TERM was then built by iterating through all instances, by increasing rmsd, and accepting those not redundant to previously accepted ones.

Validation of Coverage. All X-ray PDB entries deposited since the generation of DB80, resolved to 2.6 Å or better, and with R-free below 0.3, were downloaded and processed in the same manner as when building DB80. This set of 5,218 structures was further filtered by removing any entries with chains having more than 35% sequence identity to any chain in DB80, resulting in 1,095 structures (test set). Next, all residues and PCs were identified within this database using the same procedure as for DB80. Finally, MASTER was used to search the test set for structural matches to each of our universal TERMS, using the TERM centroid as the query and the same rmsd cutoff function as used in the set cover procedure. A TERM was said to cover any elements contained in its matches from the test set. To compare coverage of the test set and DB80, we considered subsets of N top-priority TERMS from the set cover procedure, with N varying from 1 to the total number of TERMS, measuring the joint coverage in both databases for each subsets (Fig. 3).

Topology Determination. The topology of a TERM match originating from a single chain was defined as the order of segments within that chain, in the N-to-C direction (there are $n!$ possible topologies for a TERM with n segments). Topologies of two TERM instances were considered the same if corresponding segments appeared in the same order. This analysis was applied only to instances originating from single chains due to the order ambiguity that arises with multiple chains.

Design Procedure. Given a target backbone, all instances of universal TERMS within it were found with MASTER, using the same rmsd cutoff as in the set cover procedure. A TERM was said to cover all positions and PCs contained within any of its matching regions in the target. We then proceeded to compute amino acid self pseudoenergies for each position, using all TERMS covering that position, and amino acid pair pseudoenergies for each PC, using all TERMS covering that PC. These pseudoenergies took the form of log-transformed weighted self and pair frequencies, respectively. The weighting was needed for two reasons: (i) to consider the relative contributions of different TERMS that cover a given position or PC and (ii) to weigh TERM instances higher if they originated from structural environments similar to the corresponding structural context in the target. The last consideration, for example, would assure that when deriving self pseudoenergies for a surface position, covering TERM matches originating from surfaces of other proteins would be weighted higher. The procedure for computing pseudoenergies is detailed in [SI Appendix, SI Methods](#).

Sequence Recovery. Four datasets were used to test sequence recovery (Table 1). Set X-ray-1 consisted of 64 X-ray monomer structures used in a Rosetta sequence recovery/design study (66). Set NMR-1 was made up of five NMR structures de novo-designed by Koga et al. (26). Lastly, we used 11 proteins with both X-ray and NMR structures, deposited by the Northeast Structural Genomics consortium (sets X-ray-2 and NMR-2, respectively), taken from a study on NMR structure refinement with Rosetta (67). Out of the 40 original (unrefined) NMR structures from that study, we selected those with MolProbity scores less than -2.0 , backbone rmsds less than 2.0 Å to the corresponding X-ray structure, and backbone rmsds less than 1.0 Å between ordered regions NMR structures and corresponding regions in X-ray structures (67).

Because amino acids were designed only for positions in the target covered by TERMS, sequence identity to the native protein was calculated over those positions only. The talaris2013 scoring function was used for Rosetta (65). For the combined Rosetta-TERMs design procedure, we limited the amino acids that Rosetta could use at position r to the smallest set that accounted for at least 80% cumulative probability according to TERM-based self-energy (i.e., $\{a_1, \dots, a_N\} : \sum_{i=1}^N P(a_i, r) \geq 0.8$; [SI Appendix, Eq. S3](#)). Across the sequence recovery datasets, this approach limited choices to an average 9.8 aa per position.

Sequence conservation was measured as the fraction of positions assigned amino acids within the same physiochemical class as the corresponding native amino acids. The following eight physiochemical classes were used in both sequence recovery and evolutionary analysis: Ala, Val, Ile, Leu, Met (hydrophobic); Phe, Tyr, Trp (aromatic); Lys, Arg, Glu, Asp (charged); Gln, Asn, His (polar); Ser, Thr (small polar with hydroxyl); Gly (special case); Pro (special case); and Cys (special case).

Evolutionary and Predicted Sequence Profiles. For each native structures in the sequence recovery set ([SI Appendix, Table S3](#)), the amino acid sequence was extracted and used in a BLAST search against the nonredundant database with an E value cutoff of 10 (69). The list of matches was filtered to remove alignments that had less than 35% sequence identity or those covering less than 70% of the query. Surviving match sequences were combined to produce a single MSA using Clustal Omega (74), with positions corresponding to gaps in the query or query positions not covered by TERMS subsequently removed. Sequence logos were generated from the resulting MSA with Seq2Logo, using the Hobohm clustering algorithm to correct for redundancy (75). Predicted sequences were generated via 100 independent MC trajectories (100,000 steps in each) with the kT parameter set to 0.5 pseudoenergy units (chosen empirically to produce distributions of reasonable sequence entropy). Corresponding sequence logos were created using the same settings of Seq2Logo as above. Enrichment ratios were calculated by dividing positional frequencies from Seq2Logo (i.e., the same frequencies as used for generating the sequence logos) by background genomic frequencies of corresponding amino acid.

ACKNOWLEDGMENTS. We thank Dr. David Jewell, Fan Zheng, Jack E. Holland, and Dr. William F. DeGrado for the careful reading and comments on the manuscript. This work was funded by National Science Foundation (NSF) Award DMR1534246 (to G.G.), NSF Award MCB1517032 (to G.G.), NSF Infrastructure Award CNS-1205521, National Institutes of Health Award P20-GM113132, and an award from the Neukom Institute at Dartmouth College (to G.G.).

1. Hou J, Jun SR, Zhang C, Kim SH (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc Natl Acad Sci USA* 102(10):3651–3656.
2. Han KF, Baker D (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci USA* 93(12):5814–5818.
3. Vanhee P, et al. (2011) BriX: A database of protein building blocks for structural analysis, modeling and design. *Nucleic Acids Res* 39(Database issue):D435–D442.
4. Verschueren E, Vanhee P, Rousseau F, Schymkowitz J, Serrano L (2013) Protein-peptide complex prediction through fragment interaction patterns. *Structure* 21(5):789–797.
5. Baeten L, et al. (2008) Reconstruction of protein backbones from the BriX collection of canonical protein fragments. *PLOS Comput Biol* 4(5):e1000083.
6. Porter LL, Rose GD (2011) Redrawing the Ramachandran plot after inclusion of hydrogen-bonding constraints. *Proc Natl Acad Sci USA* 108(1):109–113.
7. Shapovalov MV, Dunbrack RL, Jr (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 19(6):844–858.
8. Sillitoe I, et al. (2015) CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res* 43(Database issue):D376–D381.
9. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2014) SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Res* 42(Database issue):D310–D314.
10. Grigoryan G, Degradó WF (2011) Probing designability via a generalized model of helical bundle geometry. *J Mol Biol* 405(4):1079–1100.
11. Zhang SQ, et al. (2015) The membrane- and soluble-protein helix-helix interactome: Similar geometry via different interactions. *Structure* 23(3):527–541.
12. Hu C, Koehl P (2010) Helix-sheet packing in proteins. *Proteins* 78(7):1736–1747.
13. Ho BK, Curmi PM (2002) Twist and shear in beta-sheets and beta-ribbons. *J Mol Biol* 317(2):291–308.
14. Engel DE, DeGrado WF (2005) Alpha-alpha linking motifs and interhelical orientations. *Proteins* 61(2):325–337.
15. Platt DE, Guerra C, Zanotti G, Rigoutsos I (2003) Global secondary structure packing angle bias in proteins. *Proteins* 53(2):252–261.
16. Fernandez-Fuentes N, Dybas JM, Fiser A (2010) Structural characteristics of novel protein folds. *PLOS Comput Biol* 6(4):e1000750.
17. Feng X, Barth P (2016) A topological and conformational stability alphabet for multipass membrane proteins. *Nat Chem Biol* 12(3):167–173.
18. Sakarya O, et al. (2010) Evolutionary expansion and specialization of the PDZ domains. *Mol Biol Evol* 27(5):1058–1069.
19. Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* 27(10):514–520.
20. O'Neil KT, DeGrado WF (1990) A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* 250(4981):646–651.
21. Muñoz V, Serrano L (1995) Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J Mol Biol* 245(3):275–296.
22. Muñoz V, Serrano L (1994) Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: Comparison with experimental scales. *Proteins* 20(4):301–311.
23. Street AG, Mayo SL (1999) Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc Natl Acad Sci USA* 96(16):9074–9076.
24. Hsu HJ, et al. (2006) Assessing computational amino acid beta-turn propensities with a phage-displayed combinatorial library and directed evolution. *Structure* 14(10):1499–1510.
25. Kuhlman B, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302(5649):1364–1368.
26. Koga N, et al. (2012) Principles for designing ideal protein structures. *Nature* 491(7423):222–227.
27. Verschueren E, et al. (2011) Protein design with fragment databases. *Curr Opin Struct Biol* 21(4):452–459.
28. Vallat B, Madrid-Aliste C, Fiser A (2015) Modularity of protein folds as a tool for template-free modeling of structures. *PLOS Comput Biol* 11(8):e1004419.
29. Fernandez-Fuentes N, Fiser A (2013) A modular perspective of protein structures: Application to fragment based loop modeling. *Methods Mol Biol* 932:141–158.
30. Menon V, Vallat BK, Dybas JM, Fiser A (2013) Modeling proteins using a super-secondary structure library and NMR chemical shift information. *Structure* 21(6):891–899.
31. Kolodny R, Koehl P, Guibas L, Levitt M (2002) Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* 323(2):297–307.
32. Hunter CG, Subramaniam S (2003) Protein fragment clustering and canonical local shapes. *Proteins* 50(4):580–588.
33. Camproux AC, Tufféry P (2005) Hidden Markov model-derived structural alphabet for proteins: The learning of protein local shapes captures sequence specificity. *Biochim Biophys Acta* 1724(3):394–403.
34. 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* 40(4):662–674.
35. Budowsky-Tal I, Nov Y, Kolodny R (2010) FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proc Natl Acad Sci USA* 107(8):3481–3486.
36. Vanhee P, et al. (2009) Protein-peptide interactions adopt the same structural motifs as monomeric protein folds. *Structure* 17(8):1128–1136.
37. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 80(7):1715–1735.
38. Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* 281(3):565–577.
39. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268(1):209–225.
40. Bowie JU, Eisenberg D (1994) An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc Natl Acad Sci USA* 91(10):4436–4440.
41. Grigoryan G, et al. (2011) Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science* 332(6033):1071–1076.
42. Salem GM, Hutchinson EG, Orengo CA, Thornton JM (1999) Correlation of observed fold frequency with the occurrence of local structural motifs. *J Mol Biol* 287(5):969–981.
43. Kopec KO, Lupas AN (2013) β -Propeller blades as ancestral peptides in protein evolution. *PLoS One* 8(10):e77074.
44. Smock RG, Yadid I, Dym O, Clarke J, Tawfik DS (2016) De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints. *Cell* 164(3):476–486.
45. Alva V, Söding J, Lupas AN (2015) A vocabulary of ancient peptides at the origin of folded proteins. *eLife* 4:e09410.
46. Li H, Helling R, Tang C, Wingreen N (1996) Emergence of preferred structures in a simple model of protein folding. *Science* 273(5275):666–669.
47. Helling R, et al. (2001) The designability of protein structures. *J Mol Graph Model* 19(1):157–167.
48. Koehl P, Levitt M (2002) Protein topology and stability define the space of allowed sequences. *Proc Natl Acad Sci USA* 99(3):1280–1285.
49. Govindarajan S, Goldstein RA (1996) Why are some proteins structures so common? *Proc Natl Acad Sci USA* 93(8):3341–3345.
50. England JL, Shakhnovich BE, Shakhnovich EI (2003) Natural selection of more designable folds: A mechanism for thermophilic adaptation. *Proc Natl Acad Sci USA* 100(15):8727–8731.
51. Wingreen NS, Li H, Tang C (2004) Designability and thermal stability of protein structures. *Polymer (Guildf)* 45(2):699–705.
52. Grigoryan G, Keating AE (2008) Structural specificity in coiled-coil interactions. *Curr Opin Struct Biol* 18(4):477–483.
53. Huang PS, et al. (2014) High thermodynamic stability of parametrically designed helical bundles. *Science* 346(6208):481–485.
54. Thomson AR, et al. (2014) Computational design of water-soluble α -helical barrels. *Science* 346(6208):485–488.
55. Zheng F, Zhang J, Grigoryan G (2015) Tertiary structural propensities reveal fundamental sequence/structure relationships. *Structure* 23(5):961–971.
56. Fernandez-Fuentes N, Oliva B, Fiser A (2006) A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res* 34(7):2085–2097.
57. Cormen TH, Leiserson CE, Rivest RL (1990) The set-covering problem. *Introduction to Algorithms* (The MIT Press, Cambridge, MA), pp 974–978.
58. Zhou J, Grigoryan G (2015) Rapid search for tertiary fragments reveals protein sequence-structure relationships. *Protein Sci* 24(4):508–524.
59. Karp RM (1972) Reducibility among combinatorial problems. *Complexity of Computer Computations*, eds Miller RE, Thatcher JW (Springer, New York), pp 85–103.
60. Salem FR (1983) Structural properties of protein beta-sheets. *Prog Biophys Mol Biol* 42(2-3):95–133.
61. Gifford JL, Walsh MP, Vogel HJ (2007) Structures and metal-ion-binding properties of the Ca²⁺-binding helix-loop-helix EF-hand motifs. *Biochem J* 405(2):199–221.
62. Nakayama S, Moncrief ND, Kretsinger RH (1992) Evolution of EF-hand calcium-modulated proteins. II. Domains of several subfamilies have diverse evolutionary histories. *J Mol Evol* 34(5):416–448.
63. Zhou Y, et al. (2006) Prediction of EF-hand calcium-binding proteins and analysis of bacterial EF-hand proteins. *Proteins* 65(3):643–655.
64. Crick FH (1953) The Fourier Transform of a Coiled Coil. *Acta Crystallogr* 6:685–689.
65. O'Meara MJ, et al. (2015) Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J Chem Theory Comput* 11(2):609–622.
66. Jacak R, Leaver-Fay A, Kuhlman B (2012) Computational protein design with explicit consideration of surface hydrophobic patches. *Proteins* 80(3):825–838.
67. Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *J Am Chem Soc* 136(5):1893–1906.
68. Leaver-Fay A, et al. (2011) ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 487:545–574.
69. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
70. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108(49):E1293–E1301.
71. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* 110(39):15674–15679.
72. Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. *Proteins* 40(3):389–408.
73. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85(8):2444–2448.
74. Sievers F, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.
75. Thomsen MC, Nielsen M (2012) Seq2Logo: A method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res* 40(Web Server issue):W281–W287.