

Less is more: towards an optimal universal description of protein folds

Joseph D. Szustakowski¹, Simon Kasif^{1,2} and Zhiping Weng^{1,2,*}¹Department of Biomedical Engineering and ²Bioinformatics Program, Boston University, 44 Cummington Street, Boston, MA 02215, USA

ABSTRACT

Motivation: Identification and characterization of protein structure regularities can reveal the mechanisms governing protein structure, function and evolution. Here we focus on an intermediate level of regularity. We have developed automated methods to systematically construct a dictionary of supersecondary structures that can be used as ‘protein parts’ to describe fold-sized structures.

Results: The dictionary was constructed by aligning representative structures of all known folds, clustering similar substructures and selecting the most descriptive substructures in a minimum description length fashion. We show that the dictionary is compact and descriptive, capable of describing a substantial fraction of all known protein folds. We performed simulations using independent sets of training and testing folds. Dictionaries generated using the training set had high coverage over the folds in the testing set, suggesting that dictionary entries reflect general features of protein structures and should be capable of describing novel protein folds.

Contact: zhiping@bu.edu

1 INTRODUCTION

Proteins are often represented as a hierarchy of increasingly complex structures, from primary sequences, to secondary structure elements (SSEs), to 3D structures. Identification and characterization of regularities at each of these levels is paramount in understanding the mechanisms that govern protein structure, function and evolution. It has long been recognized that proteins exhibit a modular architecture with domains serving as autonomous folding units. Several databases (Murzin *et al.*, 1995; Orengo *et al.*, 1999) classify protein domains according to their overall architecture or ‘fold’. Such schemes are instrumental in advancing the understanding of protein structure. Nevertheless, many questions regarding protein folds remain unanswered. What defines a protein fold—stereochemical rules or evolutionary constraints? How different must two structures be in order to be considered distinct folds? How many folds exist in nature? Are these the only viable folds? Are there evolutionary paths between different folds? Are new folds evolving today? Can we engineer new folds?

Many of these questions cannot be answered until we bridge the gap between secondary structures and 3D folds. Unrelated protein structures often share small arrangements of secondary structures. Such ‘super-secondary structures’ (SSSs) represent an intermediate level of protein structure between secondary structures and folds.

SSSs are typically defined as specific 3D arrangements of secondary structures that share a common sequential ordering (Efimov, 1984; Boutonnet *et al.*, 1998; Salem *et al.*, 1999). Presnell and Cohen thoroughly examined the four-helix bundle (Presnell and Cohen, 1989), which is found in many topological variations and often as part of large helical proteins. Holm and Sander (1993) described an anti-parallel β -barrel conserved among several folds. Boutonnet *et al.* (1998) performed an extensive analysis of a secondary structure motif (SSM) consisting of two consecutive hydrogen-bonded β -strands and one α -helix. Other well-known examples include the α - and β -hairpins, EF-Hand and β - α - β motif.

In this paper we propose a method to explore systematically this intermediate level of protein structure. Our automated techniques identified a rich, compact and descriptive set of SSSs. Many of these motifs resemble those previously identified manually, whereas others extend or build upon familiar structural themes. These motifs can be used to describe a substantial portion of protein structure space and may be thought of as protein legos—basic protein parts that can be assembled in a variety of ways to construct different protein domains (we shall use the terms SSS, motif, and lego interchangeably in this paper). We expect these protein legos will prove useful in the prediction of unknown structures from known sequences.

2 METHODS

The systematic construction of protein lego dictionaries presented here is performed in three basic steps. First, we align all pairs of domains in a representative set of proteins. We then cluster these alignments to identify recurrent structures. Finally, we select a subset of recurrent structures that can best describe protein fold space as represented by the input set.

2.1 Structure alignments

We selected one representative structure from each SCOP (Murzin *et al.*, 1995) fold as follows. We first listed all SCOP (version 53) domains from the α , β , $\alpha+\beta$ and α/β classes and then excluded those solved by NMR and those with fewer than three SSEs (Frishman and Argos, 1995). We then limited the sequence identity of any pairs of proteins to <95% to eliminate redundant structures (Brenner *et al.*, 2000), resulting in 3831 protein structures we shall refer to as the ‘full set’. We aligned each of the proteins in the full set against all other proteins in the full set from the same SCOP fold using the program K2 (Szustakowski and Weng, 2000, 2002). For each fold, the protein with the largest sum of K2 scores was selected as the representative.

Selection of one representative per SCOP fold creates a broad but shallow sampling of protein fold space, which should include examples of the major protein architectures and exclude proteins which share substantial structural similarities resulting from homology or convergence at the domain level. We aligned the 441 representative protein structures all-against-all with K2

*To whom correspondence should be addressed.

(Szustakowski and Weng, 2000, 2002). These ‘fold–fold’ alignments were performed at the SSE level without regard to the ordering of the SSEs in the proteins’ primary sequences. We believe that the geometric packing of SSEs is primarily determined by stereochemistry, whereas their sequential ordering is more heavily influenced by evolution. It has been shown that similar β -sheet and four-helix bundle arrangements exist with many different sequential orderings (Richardson, 1977; Presnell and Cohen, 1989). There are also examples of proteins with the same function that have undergone sequential rearrangements (Lindqvist and Schneider, 1997). Thus, it is reasonable and necessary to ignore sequential ordering when searching for protein legos.

K2 employs a hierarchical approach to aligning protein structures that includes a rapid SSE alignment method used here. Briefly, K2 represents each SSE with a 3D vector corresponding to the linear least squares fit to its α carbon atom (C_α) positions. Secondary structure alignments are scored with a log-likelihood scoring function and optimized by a maximal weighted bipartite graph matching algorithm (Fredman and Tarjan, 1987) when sequential ordering is ignored as it is here. All alignments are then subjected to additional geometric constraints. The angle between aligned SSEs must be $<35^\circ$. Additionally, aligned SSEs must share at least three unique pairs of C_α atoms, one from each SSE, that are closer than some threshold distance (4.3 Å for strands, 5.0 Å for helices).

2.2 Identifying common subalignments and clustering structures

Proteins often share overlapping, but not identical, regions of structural similarity. This pervasive phenomenon has been dubbed the Russian doll effect (Swindells *et al.*, 1998). To address this, we state that two alignments share a common subalignment if they contain at least n SSEs from the same protein, where $3 \leq n < \min(l_1, l_2)$, and l_1 and l_2 are the lengths of the alignments. We extracted all common subalignments and added them to the list of fold–fold alignments.

We employed a graph-theoretic approach to cluster the resulting structure alignments. We represented each alignment (fold–fold and common subalignments) as a node on a graph. All pairs of alignments were compared to identify compatible alignments. Two alignments were deemed compatible if they contained the exact same set of SSEs from the same protein. Nodes corresponding to compatible alignments were connected with an edge.

Clustering was based on the detection of all maximal cliques on the graph using the sufficiently fast Bron and Kerbosch algorithm (Born and Kerbosch, 1973). We first selected the largest clique and then combined its constituent alignments into a multiple alignment. If more than one largest clique existed, the clique with the smallest sum of root mean squared deviations among the alignments was selected. We then pruned structures from the combined alignment deemed too dissimilar in 3D space from the other structures in the alignment. Following pruning, the clustered nodes were removed from the graph. This process of clique detection, combination and pruning was repeated until there were no cliques larger than two nodes remaining on the graph (Szustakowski, 2003).

2.3 Lego selection

The alignment and clustering procedures returned a very large number of clustered alignments—examination of the 7000 highest-scoring alignments from the 431 representative proteins resulted in 15 484 clusters. Ideally, we would like to identify a small subset of these clusters that acts as a minimal dictionary of protein legos and provides maximum coverage of protein fold space. To construct such a dictionary, we developed a target function based on the minimum description length principle (MDLP). The MDLP states that the best model to infer from a set of data is the model that minimizes the size of the model plus the size of the data when described in terms of the model (Rissanen, 1978, 1986, 1987; Quinlan and Rivest, 1989). More formally, the optimal model M^* is selected according to the equation

$$M^* = \min_i (l(M_i) + l(D|M_i)) \quad (1)$$

where $l(M_i)$ is the length of the i -th model and $l(D|M_i)$ is the length of the data when described in terms of M_i . Typically all lengths are measured in bits and are calculated based on problem-specific encoding schemes. The MDLP may be thought of as a formalization of Occam’s razor, favoring simpler (shorter) explanations over complex (longer) explanations. The MDLP requires no prior knowledge of underlying functional distributions and can be applied to non-continuous data. The MDLP balances the competing and conflicting goals of a small but descriptive model by considering not only the size of a model but also how effectively it describes the initial data.

Total description length calculations for the lego dictionaries (model) and representative proteins (data) were based on a simple scheme for describing protein structures as unordered 3D arrangements of SSEs. Each SSE was represented as a vector, described by the x , y and z coordinates of its head and tail and a label indicating whether it is a strand or helix. Description of a single SSE in this scheme requires $C_{\text{SSE}} = 6^*C_c + C_s$ bits, where C_c is the length of a single 3D coordinate (21 bits) and C_s is the length needed to label a SSE as a β -strand or α -helix (1 bit). The total description length for a set of protein structures with S_i total SSEs is therefore $S_i^*C_{\text{SSE}}$. Each lego added to the dictionary increases the model length $l(M_i)$ because its structure must be described once in the dictionary. Once a lego has been added to the dictionary it can be used to describe a portion of a representative protein by specifying which lego to use (C_l , 9 bits) and where to place it within the representative protein (6^*C_c). The most descriptive legos can be used like this in many representative proteins, offsetting the fixed costs of adding them to the dictionary and reducing the description length of the data $l(D|M_i)$. The difference in total description length resulting from the inclusion of a lego in the dictionary can be computed as

$$\Delta C_i = S_i^*(6^*C_c + C_s) - V_i^*(6^*C_c + C_s) + P_i^*(C_l + 6^*C_c) \quad (2)$$

The first term in Equation (2) describes the number of bits needed to enter lego i with S_i SSEs into the dictionary. The second term computes the number of bits no longer needed to describe the representative protein structures, where V_i is the coverage of lego i and is defined as the number of previously uncovered SSEs in the representative structures that are now covered by lego i . The third term describes the number of bits needed to use lego i to describe the P_i proteins it covers. For each protein, C_l bits are required to specify which lego to use and 6^*C_c bits are used to specify its 3D location within the protein.

Identifying the optimal set of legos that minimizes the total description length is a difficult combinatorial problem; even a simplified version of this problem is equivalent to the NP-hard set-covering problem. Minimizing description length generalizes the set-covering problem by assigning a score to each lego [Equation (2)]. Moreover, these scores change as each selected lego may decrease the coverage V_i of some of the remaining legos. We implemented a greedy algorithm to optimize dictionary construction. The greedy algorithm has a reasonable ratio bound for the related set-cover problem (Cormen, 2001) and is an appropriate choice here. Dictionaries constructed with the greedy algorithm are ‘top-heavy’, with the best legos added first. In each iteration the greedy algorithm selects the lego with the most negative value for ΔC_i [Equation (2)]. Once a lego is added to the dictionary, the ΔC_i values for the remaining unselected legos are updated. The algorithm exits when there are no unselected legos with $\Delta C_i < 0$. This exit condition corresponds to the point at which the cost of adding the best remaining lego to the dictionary is greater than the savings it generates when used to describe the representative protein structures.

3 RESULTS

3.1 Overview of the dictionaries

Example protein legos from a dictionary constructed from the 7000 top-scoring alignments from all 431 fold representatives are depicted in Figure 1. The automated methods used here identified structural motifs that are generally consistent with manually curated SSSs. These legos are small (mean = 3.4 SSEs) and compact, largely

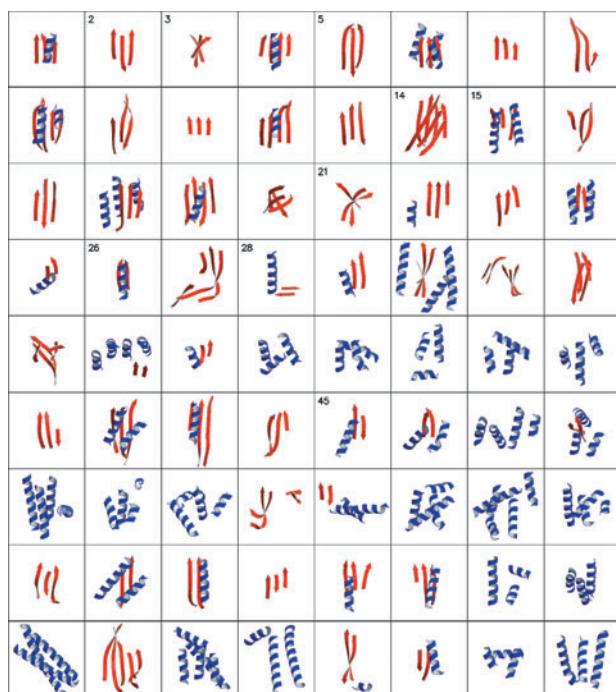


Fig. 1. Example legos from the 431-fold, 7000-alignment dictionary. Of the top 100 structures, 72 are depicted here, from top left to bottom right in order of their entry into the dictionary.

consisting of SSEs that are hydrogen bonded or otherwise packed against each other. Several of these structures are similar to motifs reported elsewhere in the literature, including examples of the β - α - β motif, Rossmann fold, immunoglobulin-like β -barrel and orthogonal β - β - α motif (see below). The SSE composition of the legos is fairly representative of the SCOP database, with 38.4% mixed, 37.5% all- α and 24.1% all- β structures. Based on the SCOP distribution of fold types and a simplifying assumption about alignment outcomes, one would expect $\sim 46\%$ mixed alignments, $\sim 32\%$ all- α alignments and $\sim 22\%$ all- β alignments. The all- β legos are dominated by three-stranded sheets in parallel, anti-parallel and mixed anti-parallel arrangements, which exhibit a spectrum of strand lengths and twists (Fig. 1.2, 1.3, 1.5, 1.21). One of the largest legos identified consists of two anti-parallel β -sheets arranged in a barrel configuration reminiscent of the immunoglobulin fold (Fig. 1.14). This motif has previously been identified as common to a number of folds (Holm and Sander, 1993), and we identified examples of this motif in at least 33 different SCOP folds. This immunoglobulin-like lego contains seven SSEs, making it the largest lego contained in this dictionary. The mixed protein legos (those containing α -helices and β -strands) are dominated by variations on the β - α - β motif. Mixed legos were identified with two to four or more strands and one to three or more helices. Figure 1.26 depicts a typical β - α - β unit with the β -strands in parallel orientation, whereas Figure 1.45 depicts a unit with anti-parallel strands. The canonical Rossmann fold (Fig. 1.15) is also included in this dictionary. Several variations on these basic mixed arrangements are found: parallel, anti-parallel, and mixed anti-parallel sheets flanked by one or more helices on one or both sides.

The lego depicted in Figure 1.28 closely resembles the orthogonal β - β - α motif described by Boutonnet *et al.* (1998). Their

definition of this motif required two consecutive hydrogen-bonded β -strands that either immediately precede (β - β - α) or follow (α - β - β) an α -helix in the primary sequence. We found this same basic geometric arrangement of SSEs in several different sequential configurations (Fig. 2). Ribonuclease T1 contains the canonical β - β - α motif (Fig. 2A). A similar arrangement of SSEs is found in Chaperonin GroEL (Fig. 2B). Although the relative ordering of the SSEs is still β - β - α , this example falls outside the purview of Boutonnet's definition because the β -strands are separated in the primary sequence by 81 positions, which include three helices and two strands. This lego is also present in Staphylococcal enterotoxin C2 (Fig. 2C), with the SSEs ordered as β - α - β , where the α - β SSEs are separated by 51 positions, which include one helix and two strands. Yet another variation on this lego is present in tyrosine hydroxylase, in which neither β -strand is proximal to the helix in the primary sequence. Consequently, the helix is slightly displaced relative to the two strands (Fig. 2D). Nevertheless, these SSEs exhibit the same basic 3D arrangement and the helix and strands pack against each other both directly and via the first two positions in the turn at the carboxy end of the helix.

The β - α - β motif depicted in Figure 1.26 is known to prefer a left-handed topology (β_1 - α - β_2). This particular SSM was constructed from 31 of the representative fold structures. Although this motif demonstrates a preference for maintaining the ordering of the left-handed topology (β_1 - α - β_2) ($n = 21$), it was also present in four of the five other possible orderings ($n = 10$). The only ordering not present was β_1 - β_2 - α , which would require a topologically tricky overhanded connection between the beta strands.

3.2 Estimating dictionary coverage

The minimum description length methods used here aim to create lego dictionaries that can describe not only the training structures used in their creation but new protein structures as well. The ability to generalize is an important characteristic of any model and is crucial when considering protein structures because we know that the current sampling of protein structure space is incomplete (Murzin *et al.*, 1995; Zhang and DeLisi, 1998; Orengo *et al.*, 1999; Pearl *et al.*, 2000; Wolf *et al.*, 2000). Consequently any model of protein structures will be routinely challenged by the determination of novel protein structures and folds in the future.

To test the ability of the lego dictionaries to generalize we conducted several bootstrapping simulations in which we divided the 3831 protein structures into mutually exclusive training and test sets. For each simulation, some number of SCOP folds ($j = \{30, 50, \dots, 200\}$) were randomly designated as training folds, and the remaining $441 - j$ folds were designated as test folds. The 3831 structures were then separated into training and test sets based on these fold designations. For each simulation, we constructed a lego dictionary from the j representative protein structures corresponding to the training folds. We repeated this process five times for each value of j .

We then aligned the legos from each dictionary against the 3831 protein structures. A lego was considered to hit a protein if an alignment was found that met the geometric parameters described above and if all of the SSEs in the lego were included in the alignment. Dictionary coverage was measured as the fraction of SSEs in the test and training protein structures that were involved in at least one such lego hit. Dictionary coverage increased with the number of training folds used in constructing the dictionary (Fig. 3). The

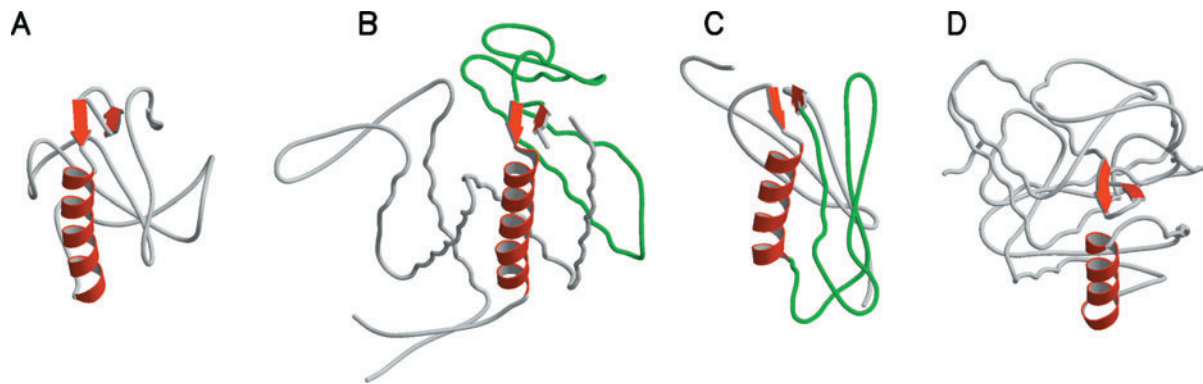


Fig. 2. Four examples of the orthogonal β - β - α motif, depicted here in red, in various configurations. (A) The canonical β - β - α configuration (SCOP d9rnt_). (B) In Chaperonin GroEL (SCOP d1oela1) the β -strands are separated in the primary sequence by 81 positions (green). (C) This same SSE arrangement is also present in a β - α - β arrangement, with 51 positions (green) separating the α - β SSEs (SCOP d1ste_2). (D) An example of the same geometric arrangement in which neither strand is proximal to the helix in the primary sequence (SCOP d1toh_).

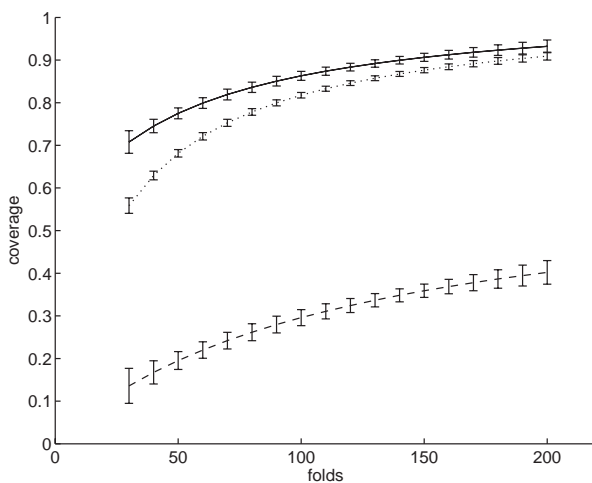


Fig. 3. Dictionary coverage for bootstrapping simulations. This figure shows dictionary coverage as a function of the number of training folds used in the bootstrapping simulations. Lines indicate fit to the rational function described in the text for dictionary coverage against the training set (solid) and test set (dotted). Coverage for the negative control set is indicated by the dashed line. 95% confidence intervals are indicated by vertical bars.

coverage–fold relationship was consistent with a rational function of the form $f(x) = (\beta_1 x + \beta_2)/(x + \beta_3)$ with a horizontal asymptote $\lim_{x \rightarrow \infty} f(x) = \beta_1$. As expected, the coverage curve is higher for proteins in the training set than for those in the test set (Fig. 3). Nevertheless, coverage of the test set proteins is only slightly lower than that of the training sets, and the β_1 values for both curves were effectively the same (training = 1.03, test = 1.02), suggesting that as the number of training folds increases, dictionary coverage should approach 100% for both data sets.

As a negative control we created a number of randomized dictionaries for the sake of comparison. For each of the dictionary simulations described above we constructed a matching negative control dictionary with the same number of legos and SSEs per lego by randomly selecting SSEs from a randomly selected training protein. Coverage for the negative control dictionaries was substantially lower than for the real dictionaries (Fig. 3).

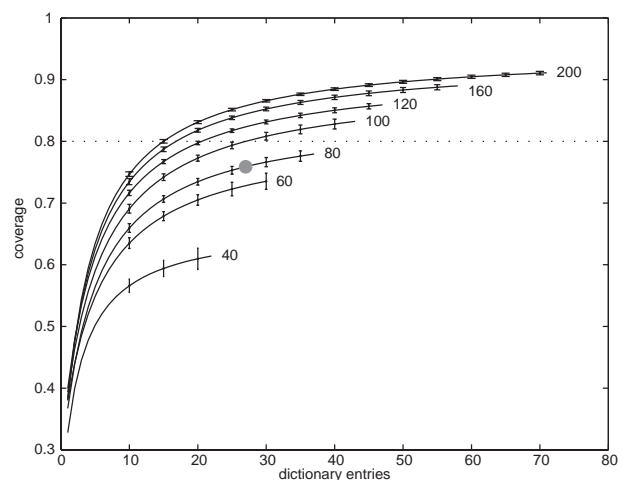


Fig. 4. Cumulative dictionary coverage for dictionaries constructed from 40, 60, 80, 100, 120, 160 and 200 training folds. The abscissa indicates the number of SSMs used from the dictionaries of interest. The ordinate indicates the coverage achieved by a particular dictionary and number of SSMs used. Solid lines indicate fit to the rational function described in the text, and vertical bars indicate 95% confidence intervals. As an example, the dictionary from 80 training folds provides coverage of 0.76 from the first 27 SSMs in the dictionary, indicated here by the gray dot. The dotted horizontal line indicates coverage equal to 0.8. The dictionaries from 60, 80, 100, 120, 160 and 200 training folds achieve coverage equal to 0.8 with 228, 56, 27, 21, 17 and 15 SSMs, respectively.

The greedy algorithm used to construct the dictionaries results in top-heavy dictionaries with the most descriptive legos listed first. For each dictionary we computed its coverage as a cumulative function of the number of legos used, i.e. for the first $i = \{1, 2, \dots, N\}$ entries in the dictionary. The effects of the top-heavy construction are evident in Figure 4. For all dictionaries, coverage increases rapidly for the first $\sim 30\%$ of the entries and then continues to increase more slowly over the remaining entries. It is clear that most of the descriptive power of the dictionaries resides in the first entries. Therefore, it should be possible to safely ignore all but the best dictionary legos and still maintain excellent protein coverage. As an example, the



Fig. 5. SCOP domain d1cyx_(gray) can be largely reconstructed from five supersecondary structures (red, yellow, green, blue and purple).

dictionary from 80 training folds provides coverage of 0.76 from the first 27 SSMs in the dictionary, indicated by the gray dot in Figure 4.

It is possible to reconstruct full-sized protein domains by selecting a set of protein legos that cover a substantial portion of its structure. One such reconstruction is depicted in Figure 5. In this example, 15 of 17 SSEs are described by 5 different protein legos. The reconstructed structure exhibits good overall similarity to the original domain, with only minor loss of details concerning the precise orientations of the SSEs.

4 DISCUSSION

Our results indicate that there is substantial regularity among protein structures at the SSS level. Through the application of specific geometric criteria and the minimum description length selection process we were able to infer a compact, descriptive model of SSSs that captures much of the complexity at this level of structure.

What is the origin of these protein legos? Common evolutionary histories seem unlikely in the majority of cases because the legos are present across a diverse set of (presumably) unrelated proteins. A more plausible explanation extends the idea of structural attractors put forth by Holm and Sander (1996). Specifically, these legos may correspond to localized attractors in fold space; that is to say, stable, kinetically accessible structures that act as evolutionary sinks. In this scenario, protein sequences would evolve to take on these localized motifs because they acted as folding nuclei, added stability to the protein structure and/or served as scaffolding for constructing larger, more complex structures.

The confluence of increasing computer power, growth in the number of solved protein structures and maturing structure alignment methods has fueled a number of large-scale structure comparison efforts (Holm and Sander, 1996; Orengo *et al.*, 1997; Gerstein and Levitt, 1998; Orengo *et al.*, 1999; Shindyalov and Bourne, 2000; Yang and Honig, 2000; Dietmann and Holm, 2001). Shindyalov and Bourne (2000) demonstrated that structural similarities are abundant

across a wide spectrum of sizes. Recent work by Harrison *et al.* (2002) indicates that statistically significant similarities are present even among unrelated, generally dissimilar structures. Taylor (2002) reported success in matching protein domains to many idealized model structures. Day *et al.* (2003) analysis of several structure classifications identified common metafolds. Results from these and other similar projects have led to an emerging view of protein fold space as a sort of continuum in which most or all protein structures, homologous and non-homologous, share varying degrees of similarity over a broad range of sizes.

In contrast to the methods we used to infer a compact dictionary of SSSs, previous research has employed the inverse approach by applying predefined SSS definitions to a set of proteins (Boutonnet *et al.*, 1998; Salem *et al.*, 1999). As one might expect, our methods and less stringent criteria yielded a larger set of SSSs with greater coverage than previously reported (Salem *et al.*, 1999). Nevertheless, these results are in general agreement and indicate that there is substantial reuse of small arrangements of SSEs across a broad array of protein structures.

The difference in coverage between the real dictionaries and the negative control dictionaries (Fig. 3) underscores the non-random nature of proteins at the SSM level. Complexity theory describes an intimate relationship between complexity or entropy and minimum description length. Intuitively, non-random systems can be highly compacted or compressed, whereas random systems are fairly incompressible. Our method provides an approximate indication of the non-randomness of protein structure space by identifying recurrent patterns at the SSM level.

The orthogonal β - β - α motif described above helps to illustrate a larger point. Most of the motifs identified here share similar SSE geometry despite different orderings in their primary sequences. Our strategy in comparing protein structures differs from many previous efforts because we ignore the sequential ordering of the SSEs in the initial alignment, clustering and selection procedures. Each lego with N SSEs may exist in as many as $N!$ different primary sequence orderings. We examined the legos and lego hits for the simulation dictionaries constructed from 200 representatives and found that 93% of the three-SSE legos were present in all 6 possible orderings, and 33% of the four-SSE legos were present in all 24 possible orderings. Although none of the larger legos was found in all possible orderings, all of the five-SSE legos were found in 34 or more orderings, and 80% of the six-SSE legos were found in 51 or more orderings. These results underscore the utility of ignoring sequential ordering when searching for geometric similarities among protein structures. It would appear that for small motifs, nature and crystallographers have had sufficient time to sample most of the possible orderings. For larger legos, a smaller fraction (but sizable number) of orderings has been sampled. As new structures are determined, one might expect the discovery of motifs similar to those reported here with still new orderings.

One potential and immediate application for these legos is in the prediction of protein structures. The most successful *ab initio* prediction methods, threading (Dunbrack *et al.*, 1997; Jones, 1997; Marchler-Bauer and Bryant, 1997; Smith *et al.*, 1997) and fragment assembly (Simons *et al.*, 1997, 1999), attack this problem from opposite ends of the size spectrum. The compact, descriptive lego dictionaries described here provide a natural starting point for an intermediate structure prediction tactic. Accurately predicting the existence and structure of even a single protein lego within a

modeled structure would impose significant constraints on the model that would expedite prediction of the remainder of the structure. Moreover, results from the bootstrapping simulations suggest these legos may effectively describe even novel protein structures. Another potential application is to assist in the prediction of biochemical function. By describing proteins as collections of protein legos (e.g. as a bit vector that denotes the inclusion/exclusion of each dictionary entry in a particular protein) it would be possible to apply standard statistical and machine learning techniques to identify combinations of legos that correlate with specific functions.

Certainly, no single study can comprehensively address the open questions about protein folds. Our results do, however, represent concrete progress in bridging the secondary structure/fold gap. How different must structures be to be considered distinct folds? Clearly there are substantial similarities among proteins from different folds. Although any pair of structures may share only small regions of similarity, most structures may be effectively described as mosaics of small motifs ‘borrowed’ from other folds. How many folds exist in nature? Are these the only viable folds? Can we engineer new folds? The dictionaries described here represent a kind of vocabulary for protein structures. We hope to build upon these results by learning a motif grammar, i.e. the rules that govern the interactions among these motifs. Taken together, a structural vocabulary and grammar would allow us to predict or even design the basic structures of novel protein folds. Are there evolutionary paths between different folds? Are new folds evolving today? Recently, several mechanisms have been proposed to describe the evolution of new protein folds from existing structures or short polypeptides (Grishin, 2001; Lupas *et al.*, 2001). We believe the SSSs described here may enhance these models of protein evolution by providing candidate structures for insertion/deletion events as well as frequently occurring, stable structures that may act as evolutionary way stations.

ACKNOWLEDGEMENTS

We thank Prof. Temple Smith and ZLAB members for fruitful discussions. We also thank the anonymous reviewers for their comments and questions, which helped to improve this manuscript. This work was supported by NSF grants DBI-0078194, DBI-0239435 and ITR-048715. Calculations were performed on a 256-processor Linux cluster purchased with NSF Major Research Instrument grant DBI-0116574.

Conflict of Interest: none declared.

REFERENCES

- Born, C. and Kerbosch, J. (1973) Finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.
- Boutonnet, N.S. *et al.* (1998) Structural classification of alphabeta and betabetaalpha supersecondary structure units in proteins. *Proteins*, **30**, 193–212.
- Brenner, S.E. *et al.* (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.
- Cormen, T.H. (2001) Introduction to Algorithms. MIT Press, Cambridge, MA.
- Day, R. *et al.* (2003) A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci.*, **12**, 2150–2160.
- Dietmann, S. and Holm, L. (2001) Identification of homology in protein structure classification. *Nat. Struct. Biol.*, **8**, 953–957.
- Dunbrack, R.L., Jr., Gerloff, D.L., Bower, M., Chen, X., Lichtarge, O. and Cohen, F.E. (1997) Meeting review: the second meeting on the critical assessment of techniques for protein structure prediction (CASP2), Asilomar, California, December 13–16, 1996. *Fold. Design*, **2**, R27–R42.
- Efimov, A.V. (1984) A novel super-secondary structure of proteins and the relation between the structure and amino acid sequence. *FEBS Lett.*, **166**, 33–38.
- Fredman, M. and Tarjan, R. (1987) Fibonacci heaps and their uses in improving network optimization algorithms. *J. Assoc. Comput. Mach.*, **34**, 596–615.
- Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.
- Gerstein, M. and Levitt, M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.*, **7**, 445–456.
- Grishin, N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
- Harrison, A. *et al.* (2002) Quantifying the similarities within fold space. *J. Mol. Biol.*, **323**, 909–926.
- Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
- Holm, L. and Sander, C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
- Jones, D.T. (1997) Progress in protein structure prediction. *Curr. Opin. Struct. Biol.*, **7**, 377–387.
- Lindqvist, Y. and Schneider, G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.*, **7**, 422–427.
- Lupas, A.N. *et al.* (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.*, **134**, 191–203.
- Marchler-Bauer, A. and Bryant, S.H. (1997) A measure of success in fold recognition. *Trends Biochem. Sci.*, **22**, 236–240.
- Murzin, A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Orengo, C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
- Orengo, C.A. *et al.* (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279.
- Pearl, F.M. *et al.* (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.
- Presnell, S.R. and Cohen, F.E. (1989) Topological distribution of four-alpha-helix bundles. *Proc. Natl Acad. Sci. USA*, **86**, 6592–6596.
- Quinlan, J.R. and Rivest, R.L. (1989) Inferring decision trees using the minimum description length principle. *Inform. Comput.*, **80**, 227–248.
- Richardson, J.S. (1977) β -Sheet topology and the relatedness of proteins. *Nature*, **268**, 495–500.
- Rissanen, J. (1978) Modeling by shortest data description. *Automatic*, **14**, 37–38.
- Rissanen, J. (1986) Stochastic complexity and modeling. *Ann. Stat.*, **14**, 1080–1100.
- Rissanen, J. (1987) Stochastic complexity. *J. R. Statist. Soc. Ser. B*, **49**, 223–239 and 253–265.
- Salem, G.M. *et al.* (1999) Correlation of observed fold frequency with the occurrence of local structural motifs. *J. Mol. Biol.*, **287**, 969–981.
- Shindyalov, I.N. and Bourne, P.E. (2000) An alternative view of protein fold space. *Proteins*, **38**, 247–260.
- Simons, K.T. *et al.* (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
- Simons, K.T. *et al.* (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* (suppl.), 171–176.
- Smith, T.F. *et al.* (1997) Current limitations to protein threading approaches. *J. Comput. Biol.*, **4**, 217–225.
- Swindells, M.B. *et al.* (1998) Contemporary approaches to protein structure classification. *Bioessays*, **20**, 884–891.
- Szustakowski, J.D. and Weng, Z. (2000) Protein structure alignment using a genetic algorithm. *Proteins*, **38**, 428–440.
- Szustakowski, J.D. and Weng, Z. (2002) K2: protein structure comparisons and their statistical significance. In Fogel, G. and Corne, D. (eds), *Evolutionary Computation in Bioinformatics*. Morgan Kaufmann.
- Szustakowski, J.D. (2003) A protein structure alignment method and application to the discovery of recurrent protein structure motifs. Doctoral dissertation, Boston University, Boston, MA.
- Taylor, W.R. (2002) A ‘periodic table’ for protein structures. *Nature*, **416**, 657–660.
- Wolf, Y.I. *et al.* (2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.*, **299**, 897–905.
- Yang, A.S. and Honig, B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.*, **301**, 679–689.
- Zhang, C. and DeLisi, C. (1998) Estimating the number of protein folds. *J. Mol. Biol.*, **284**, 1301–1305.