The continuity of protein structure space is an intrinsic property of proteins

Jeffrey Skolnick1, Adrian K. Arakaki, Seung Yup Lee, and Michal Bylinski

Center for the Study of Systems Biology, Georgia Institute of Technology, Atlanta, GA 30318

Edited by Barry H. Honig, Columbia University, New York, NY, and approved August 3, 2009 (received for review July 9, 2009)

The classical view of the space of protein structures is that it is populated by a discrete set of protein folds. For proteins up to 200 residues long, by using structural alignments and building upon ideas of the completeness and continuity of structure space, we show that nearly any structure is significantly related to any other using a transitive set of no more than 7 intermediate structurally related proteins. This result holds for all structures in the Protein Data Bank, even when structural relationships between evolutionary related proteins (as detected by threading or functional analyses) are excluded. A similar picture holds for an artificial library of compact, hydrogen-bonded, homopolymer peptide structures. The 3 sets share the global connectivity features of random graphs, in which the local connectivity of each node (i.e., the number of neighboring structures per protein) is preserved. This high connectivity supports the continuous view of single-domain protein structure space. More importantly, these results do not depend on evolution, rather just on the physics of protein structures. The fact that evolutionary divergence need not be invoked to explain the continuous nature of protein structure space has implications for how the universe of protein structures might have originated, and how function should be transferred between proteins of similar structure.

To compare a pair of protein structures, a structural alignment is done to identify their “optimal” structural similarity. In practice, structure alignment algorithms employ different structural similarity metrics and approaches to identify this “best” structural alignment (17–23). Especially when 2 proteins have subtle structural similarities, different comparison metrics will capture different structural features (24). One widely used structure comparison metric is the TM-score whose range is 0–1, with 1.0 indicating structurally identical proteins (25). The average TM-score of the best structural alignment between randomly related structures is 0.30, with a SD of 0.01 (16). The TM-score offers the advantage that, unlike many other metrics (26), the statistical significance of an alignment for a given TM-score is protein length-independent and no rigid distance cutoffs are introduced so that more subtle structural similarities can be detected. The TM-align structure alignment algorithm (25, 27), used later in this article, uses the TM-score, but any sensitive structural alignment algorithm could be used in the analysis that follows.

The continuity of fold space does not require that the library of solved protein structures in the current PDB be complete. Is there a limited, but large repertoire of single-domain topologies such that, at some point, the likelihood of discovering a new protein structure would be minimal? Or is protein fold space essentially infinite? Kihara and Skolnick (28) demonstrated for single-domain proteins that the PDB is likely already complete; however, this conclusion is not true for multi-domain proteins or proteins with many compartments. Early support for the continuity of protein structure space at least for approximately 130-residue-long substructures that transgress fold type came from Shindyalov and Bourne (11). Similarly, Harrison et al. concluded that fold space is a continuum for some topology types in the β or α/β secondary structure class (12). Yang and Honig (13) also detected structural similarities between different folds in SCOP (14). Consistent with these ideas, recent protein structure comparison studies suggest the alternative view that protein structure space is continuous, in the sense that “there are meaningful structural relationships between proteins that are classified very differently” (15), with many structural intermediates (16). However, protein structure space could be piece-wise continuous, with the space of α-helical, β-proteins, and α/β proteins disjoint from each other. This view is supported by the work of Kim et al., who found that protein structures associated with each type of secondary structure emerge from a common center (4), with sparse intervening regions that possibly arise because certain folds are unstable. Alternatively, this sparseness could arise because of the insensitivity of the structure comparison algorithms used (4) or because the library of solved structures is not yet complete.

Author contributions: J.S. and A.K.A. designed research; J.S., A.K.A., S.Y.L., and M.B. performed research; J.S., A.K.A., S.Y.L., and M.B. analyzed data; and J.S. wrote the paper.

The authors declare no conflict of interest.

1To whom correspondence should be addressed. E-mail: skolnick@gatech.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907683106/DCSupplemental.
multimeric protein structures. More recently, it was shown that, by analyzing randomly generated 100- and 200-residue compact conformations of generic homopolyptides in simplified and all-atom protein models, all have similar folds in the PDB, and conversely, all compact, single-domain protein structures in the PDB have structural matches to the set of compact homopolypeptide structures (16). Thus, both sets are quite likely complete, with the protein fold universe arising from compact conformations of hydrogen-bonded, secondary structures. As side chains are represented by Cαβ in both protein models, these results suggest that the observed protein folds are insensitive to chain packing details. Sequence specificity enters in fine-tuning the structure and stabilizing a given fold with respect to alternatives.

In this article, we further explore the issues of the continuity and completeness of protein fold space. We adopt and extend the view of continuity described by Kolodny et al. (15), according to which one can “navigate” fold space to link 2 arbitrarily selected structures, following a path of statistically significant similar structures. We show for proteins up to 200 residues long that nearly any protein structure can be related to any other protein structure using a transitive set of no more than 7 intermediate structurally related proteins. Thus, protein structure space is almost completely connected; viz. when fold space is conceived as a graph, a giant sub-graph exists wherein every protein pair is structurally connected. Although one cause of the connectivity and continuity of fold space is the process of fold evolution (9, 29), we show that this condition is not necessary by excluding structural relationships between proteins that have an evolutionary relationship as identified by threading and/or that share a common fold/function. More importantly, we show that the library of randomly generated, compact hydrogen bonded, homopolypeptide structures whose secondary structures match those in the PDB is also extremely connected. As the latter set of proteins have no evolutionary relationship whatsoever, this implies that the continuity of fold space is a fundamental property of protein structures and protein physics, which is then exploited during the course of protein evolution.

Results
Structural Relationships in the PDB. All-against-all structural alignments of compact proteins containing 40 to 300 residues that cover the PDB at no greater than 35% pair-wise sequence identity—the PDB300 set—were done. We consider here the PDB300 subset of proteins whose functions are either known or can confidently be predicted, so as to be able to exclude proteins with apparent functional relationships from the analysis (30). Proteins 200 to 300 residues long can act as bridges between proteins no greater than 200 residues long—PDB200holo—but their exclusion leaves the results essentially unchanged. To compare a pair of protein structures, a structural alignment is done to identify their “closest” structural similarity as assessed by a structure comparison metric. We define the template as the structure of the protein being aligned to the protein structure of interest, the target. We use the TM-score to compare structures (25); any protein pair with a TM-score greater than 0.40 is structurally related [e.g., Fig. 1A; 1gnyA (template)–1ekrA (target)].

At a finer-grained level, what happens when a helical protein is aligned to a β-protein? As in Fig. 1B, when a single helix is aligned to a β-strand, the β-strand aligns parallel to the principal axis of the helix, with half the helical residues omitted on average. As geometric objects, the aligned coordinates will be quite close in space. In this manner, the spatial proximity of secondary structural elements is maintained even when proteins of different secondary structure class are aligned. Because the TM-score depends on the number of aligned residues and half the helical residues are unaligned whereas all β-strand residues are aligned, for the single helix and β-strand considered here, the TM-score (helix→β) is 0.5 TM-score (β→helix). This effect introduces a subtle secondary structure dependence of the connectivity of protein space, with the set of helical proteins being most highly connected, followed by β-proteins, followed by mixed α/β proteins (see Discussion).

A further illustration as to how one can connect 2 apparently disparate protein structures of different secondary structure types is shown in Fig. 1A, in which we link the structure of the β-protein 1gnyA to the helical protein 101m [TM-score(1gnyA→101m) = 0.31] via an intermediate mixed α and β containing protein structure: 1gnyA→1ekrA→101m. The TM-score (1gnyA→1ekrA) is 0.43 and that of 1ekrA→101m is 0.41. Among other features, this gives the partly unaligned gray helix in 101m (Fig. 1A Upper Right). There are a sufficient number of aligned residues so that the 2 proteins (1ekrA and 101m) bear a significant structural relationship to each other. Thus, we have a transitive walk in structure space from a β- to a helical protein.

For template protein B aligned to target protein A, if the TM-score(B→A) ≥ d, then template B is a first neighbor of target protein A in structure space, at a TM-score cutoff d. If target and template structures A and B, and B and C, but not A and C, satisfy this criterion, then C is a second neighbor of A, with a transitive relationship C→B→A. More generally, template X is a kth neighbor of target Y if the length of the shortest path from X to Y is k. In Fig. 2A, the mean fraction of proteins, fs, in PDB200holo that are no more than k = first, second, fourth, eighth, 16th, and 32nd neighbors to another protein are shown versus d. The converged value of f∞(k∞→) is fmax. As shown in Fig. 2A (thick line), at d = 0.40, fs≤k = fmax = 98.5%; i.e., nearly
all proteins are no more than eighth structural neighbors. In fact, $f_{k=4} = 83.2\%$, indicating that the majority of proteins are just fourth structural neighbors at most.

From Fig. 2B, the fraction of proteins in the largest strongly connected component [LSCC; the largest sub-graph of a directed graph, where a path exists from every vertex in the sub-graph to every other vertex in the sub-graph, i.e., every pair of vertices is connected in both directions; see supporting information (SI) Fig. S1], is $S = 0.986$. Here, the LSCC at a given $k^{th}$ cutoff includes the largest subset of proteins such that every possible pair of proteins in the subset are no more than $k^{th}$ neighbors to each other, with $k \neq 0$. Thus, for $k = 1$, the LSCC comprises the maximal subset of proteins where the TM-score (A→B) $\geq d$ and the TM-score (B→A) $\geq d$ for every pair of proteins A and B that are members of the subset (i.e., the largest clique in the corresponding digraph). This figure clearly shows that protein structure space is almost completely connected (i.e., most proteins belong to the LSCC) and is continuous ($15$) in that one can link $2$ arbitrarily selected structures, following a path of statistically significant similar structures. The results are very close to what happens when random digraphs with the same distribution of first neighbors are generated (Fig. 2B, thin line; $S = 0.992$).

At the transition midpoint of $f_{\text{max}}$, $d = 0.49$, and the strongly connected members are no more than 32nd neighbors, with a significant fraction of structure pairs ($0.44 \%$) are 16th neighbors. As protein structures are highly similar at this TM-score threshold (their structural alignment Z-score is 19), this further reinforces the conclusion that protein structure space is globally continuous and highly inter-linked. However, as $d$ increases further, we recover the traditional discrete view of protein structure space ($31$).

Similar results are shown in Fig. S2. When the full PDB300 structures up to 300 residues is used, at $d = 0.4$, then $f_{\text{max}} = 0.95$ and the transition midpoint moves to $d = 0.515$. Thus, the results are remarkably insensitive to protein length or the size of the database used, with the transition midpoint occurring at a $d$ well above the regime where statistically significant structural similarities are found.

Fig. 3 shows the length distribution of proteins in PDB200$holo$ not belonging to the LSCC; these proteins are either very small or very large. Because the TM-score is not commutative, there are large proteins aligned to the smallest proteins, but it takes at least second neighbors of intermediate length for a small protein to be aligned to a large protein. In practice, a tiny subset of the smallest proteins are not acceptable templates for the largest proteins. Moreover, just because of their size, some large proteins fail to have a sufficient number of neighbors to belong to the LSCC. They are effective templates for the smaller proteins, but do not have many structures aligned to them with a TM-score $\geq d$. As the strongly connected component requires reciprocity, they are excluded. Which protein folds that are not part of the LSCC is somewhat anecdotal; when the entire PDB300 is used, all proteins excluded from the LSCC for PDB200$holo$ become part of the LSCC.

In Fig. 2C, for PDB200$x$, i.e., PDB200$holo$ where functional relationships as detected by threading ($32$) and FINDSITE ($30$) are excluded, we plot the fraction of proteins that are no more than $k = \text{first, second, fourth, 16th, and 32nd}$ neighbors as a function of the TM-score cutoff $d$. The goal here is to remove

![Fig. 2](image-url) Mean fraction of proteins in PDB200$holo$ (A), PDB200$x$ (C), and the homopolyptide library (E) that are no more than $k^{th}$ neighbors ($f_k$), whose first neighbors have a TM-score $\geq d$. Relative size of the LSCC in PDB200$holo$ (B), PDB200$x$ (D), and the homopolyptide library (F), as a function of the $k^{th}$ neighbor cutoff, at $d = 0.40$. The thick line with diamonds corresponds to the values in the original set; the thin line indicates the median values obtained from 2,000 randomly generated digraphs with the same number of nodes and first-order local connectivity per node as in the original set (error bars indicate the minimum and maximum values from the 2,000 random graphs).

![Fig. 3](image-url) Length distribution of proteins not belonging to the LSCC at $d = 0.40$ relative to all proteins in the PDB200$holo$ set, the PDB200$x$ set, and the homopolyptide library. Relative abundance is the fraction of the total number of proteins excluded from the LSCC that fall in a given interval of protein length divided by the fraction of the total number of proteins in the set in the same protein length interval.
evolutionary relationships between proteins so that, as far as is practical, the underlying structural relationships are explored. At \( d = 0.4 \), \( f_{2:8} = f_{max} = 0.947 \), with approximately 74% of proteins no more than fourth neighbors (\( f_{4:4} = 0.737 \)). The transition midpoint shifts to a TM-score of 0.47 from 0.49 for PDB200\text{holo}. Fig. 2D shows for PDB200\text{thick}, at \( d = 0.40 \), the fraction of proteins in the LSCC as a function of the \( k^{th} \) neighbor cutoff. For \( k \geq 8, S = 0.947 \), i.e., a rather minor diminution in the size of the LSCC compared with PDB200\text{holo}. The thin line shows the random digraph results given the same distribution of first neighbors for each protein as in PDB200\text{holo}. Again, the asymptotic behavior is essentially indistinguishable from a random digraph (\( f_1 = 0.005 \)). However, real protein structures see the entire space of the LSCC at a larger number of neighbors than for the corresponding randomly generated graph; this is a result of a protein length effect that is entirely ignored in the random digraph. As in PDB200\text{holo}, those proteins excluded from the LSCC lie at the extremes of protein size; see Fig. 3. Thus, even when we try our best to remove evolutionary relationships between proteins, protein structure space is still almost completely connected.

We then considered the subset of PDB200\text{holm} comprised only of proteins of identical secondary structure class. Interestingly, for structurally significant alignments among the 229 helical proteins, at \( d = 0.4 \), \( S = 0.952 \), while \( f_1 = 0.172 \). For structural alignments among the 209 purely \( \beta \)-proteins, at \( d = 0.4 \), \( S = 0.874 \), while \( f_1 = 0.174 \). Finally, for structural alignments among the 388 mixed motif proteins, at \( d = 0.4 \), \( S = 0.895 \), while \( f_1 = 0.087 \). The lower \( f_1 \) for \( \alpha \beta \) proteins reflects the TM-score reduction effect discussed earlier, when helices are aligned to \( \beta \)-strands, resulting in a smaller average number of neighbor structures. Nevertheless, in all 3 cases, the majority of structures within a given class belong to the LSCC. The fact that \( S \) is considerably larger for helical proteins than \( \beta \)-proteins reflects the fact that helices are longer than strands, so the average number of secondary structural elements in a helical protein of a given length is less than for \( \beta \)-proteins. Thus, the space of helical structures is effectively more compact (33). Interestingly, the distributions of the length of the shortest path \( k \), linking protein pairs of identical secondary structure class are remarkably similar to those for protein pairs of different secondary structure classes (Fig. S3), indicating that if \( k \) would be used as a metric of protein similarity, protein structure space would be less segregated by secondary structure class than previously reported (4).

**Structure Space of Compact, Sticky Homopolypeptides.** Despite our best efforts in PDB200\text{holm} to remove evolutionary relationships among proteins to expose the purely structural characteristics of protein fold space, we still cannot guarantee that all such relationships have been excised. Previously, by examining the relationship between a randomly generated set of compact sticky, hydrogen-bonded homopolypeptides and real proteins, we demonstrated that all such structures were in the PDB and that the converse was also true (16). This suggested that the PDB is likely complete and that the completeness arises from the packing of hydrogen-bonded, compact arrangements of secondary structural elements, and nothing more. Here, in a similar spirit, for a set of polyvaline homopolypeptides, each with the same secondary structure assignment per amino acid position as one of the members of PDB300\text{holm}, we further explore the nature of the structure space of the library of folds generated by TASSER in the ab initio limit (33).

A number of interesting results were found. Using the original hydrogen bond scheme of TASSER (33), for \( \beta \)-sheet-containing proteins, protein structure space was not highly connected. We did find that, for the subspace of helical proteins, at \( d = 0.4 \), all helical protein structures are no more than eighth neighbors and the subspace of helical structures is almost completely connected. Examination of \( \beta \)-strand containing proteins above approximately 100 residues revealed that they failed to form hydrogen-bonded sheets, an effect exacerbated with increasing length. This is an echo of our previous work (16), where we showed that, if the hydrogen bonding is turned off, most generated compact structures are not in the PDB. Thus, we concluded that hydrogen bonding is necessary to generate protein-like structures. Without well formed \( \beta \)-sheets, there are many more geometric arrangements of the strands, and as a consequence, the resulting structure space is not so well connected.

Using the improved hydrogen bond scheme described in Methods (Fig. 2E), for the homopolypeptide structures, we plot \( f_1 \) as a function of \( d \). The transition midpoint is at \( d = 0.435 \). When \( d = 0.40 \), \( f_{max} \) is 0.888 and \( S \) is 0.887. As in the case of real proteins, the space of homopolypeptide structures is extremely connected. Consistent with Fig. 2E, at \( d = 0.4 \), Fig. 2F shows that all members of the LSCC are no more than eighth structural neighbors. Interestingly, the random digraph with the same distribution of first neighbors for each node has a somewhat larger size of the LSCC (\( S = 0.946 \)).

Focusing on the subset of helical proteins, we find that at \( d = 0.40 \) nearly every protein is no more than an eighth neighbor, with \( f_1 \) equal to 0.18 (essentially, the same as in PDB200\text{holm}), and \( S \) equal to 0.991. The transition midpoint occurs at \( d = 0.49 \). In contrast, for the structural subspace comprised only of \( \beta \)-proteins, at \( d = 0.40 \), \( f_1 \) is 0.019 and \( S \) is 0.518, with all proteins in the LSCC no more than 16th neighbors. For the structural subspace of \( \alpha \beta \) proteins, at \( d = 0.40 \), \( f_1 \) is 0.011 and \( S \) is 0.404, again with all proteins in the LSCC no more than 16th neighbors. The qualitative trends are the same as in PDB200\text{holm}, but the nonhelical structural subspaces are more diffuse in the homopolypeptide library.

The most striking difference between the homopolypeptide structural library and the real PDB wherein detectible evolutionary relationships are excluded is in the size of the LSCC. This is a consequence of the difference in the average fraction of structures that are first neighbors, with an \( f_1 \) of 0.024 compared with 0.051 for real proteins. Note that, for the homopolypeptide structure library, if we consider the top 2 and 5 clusters for each pattern of secondary structure, \( f_1 \) values are 0.027 and 0.025, respectively. Thus, \( f_1 \) is quite insensitive to the size of the homopolypeptide structural library. To reproduce the qualitative behavior of the size of the LSCC in real PDB structures, we need to include the top 8 clusters per secondary structure arrangement; that is, the structure space covered by the homopolypeptide library is less connected than that of PDB200\text{holm}.

If we randomly delete connections in PDB200\text{holm} so that \( f_1 \) is 0.024 as in the homopolypeptide structural library, then the size of the LSCC is \( S = 0.881 \) at \( d = 0.4 \), i.e., essentially the same as in the homopolypeptide library, where \( S \) is 0.887 at \( d = 0.4 \). In other words, for the size of the LSCC, the real PDB200\text{holm} set excluding detectible functional relationships, i.e., PDB200\text{holm}, behaves the same as the homopolypeptide structure library. The size of the LSCC is also comparable for the random digraphs that preserve the first-order local connectivity, with \( S = 0.965 \) for the random digraphs corresponding to the first neighbor depleted PDB200\text{holm} and \( S = 0.946 \) for random digraphs corresponding to the homopolypeptide library. That is, the continuity of protein structure space is largely a generic property of randomly connected nodes that reflects the intrinsic structural similarities of proteins.

**Discussion**

This study builds on previous work that strongly suggested that the library of folds of compact single-domain proteins found in the PDB is already likely complete and that the set of protein structures arises from the packing of compact, hydrogen-bonded secondary structural elements (16). This does not require that
protein structure space be continuous in the sense defined by Kolodny et al. (15), which is unrelated to the mathematical concept of graph continuity, nor that it be highly connected (e.g., if the LSCC were small in the real PDB library, then the set of randomly generated homopolypeptide structures corresponding to them would also have a very small LSCC). Here, we further argue that protein structure space is nearly completely connected, whereby essentially all protein structures can be linked from an arbitrary starting structure using a transitive set of 7 structurally related neighbors or less. The fact that protein structure space is almost completely connected suggests a means by which the observed universe of protein folds as generated by evolution (9, 10) could have arisen. During evolution, sequences that adopt any arbitrary fold that is at least marginally stable can give rise to sequences whose structures eventually filled all of fold space. Our homopolypeptide structure library results suggest that there are many ways that this could happen, and nature took advantage of at least one scenario.

Further support for the view that the high connectivity of fold space is an intrinsic structural property of proteins emerges from the fact that the size of the LSCC of real single-domain proteins in the PDB and the homopolypeptide structure library are quite consistent with the properties of a random directed graph (whose nodes represent structures) with the same number of nodes and first neighbors per node. The major difference between the space of real protein structures and homopolypeptide structures is that, in the latter, there are fewer first neighbors, the number of which is essential for dictating the size of the LSCC. Whether this reflects the residual influence of evolution or is caused by inadequacies in the potential used to fold the homopolypeptides is uncertain at this time.

Our results further emphasize the importance of hydrogen bonding. In an earlier report (16), we showed that, if hydrogen bonding were removed, the library of compact homopolypeptide structures do not resemble PDB structures. Rather, the average TM-score of the closest related homopolypeptide structure corresponds to the mean value of the best structural alignment of a pair of randomly related structures. Here, we show that it is the inclusion of a reasonable H-bond scheme that restricts the conformational space of the compact homopolypeptides so that their structural space has a large LSCC. In other words, hydrogen bonding acts at the level of individual protein structures as well as dictating the size and connectivity of the structural space of single domain proteins. Thus, the ability to reproduce the features of the structure space of real proteins can be another design criterion used to optimize the hydrogen bond potential used to fold proteins.

This work augments and supports the idea that fold space is discrete at high structural similarity and continuous at lower but still significant structural similarity (31). Moreover, we show that most protein pairs are separated by just 3 structural neighbors or less, irrespective of their secondary structure class. There are other biological ramifications of this study as well. Considerable effort has been expended over the years in the development of fold classification schemes that study the interrelationships of protein structures, not only because they might provide functional insights but also because they are of fundamental interest (5, 34). The fact that one need not invoke evolution to explain the structural interrelationships of almost all protein structures provides potentially important insights into how structure space is globally organized. Furthermore, because such interrelationships arise in the homopolypeptide library, wherein proteins with similar structures are not evolutionary related, care must be exercised in transferring function on the basis of structural similarity alone, without additional local structure and sequence-based filters. Thus, this analysis provides a foundation for the study of the interplay of evolution and protein physics in the nature of protein structure space.

Methods

The PD8300 Set. The PD8300 set is a representative set of 5,906 compact PDB proteins with pair-wise sequence identity no greater than 35% and containing between 40 and 300 residues; many are in a ligand-free state. As our goal is to understand their underlying structural relationships, we wish to remove all detectable evolutionary relations between protein pairs. We therefore constructed the subset of PD8300, in a procedure detailed later, for which we had observed/predicted binding sites and associated binding ligands, the PD8300holo set comprised of 1,932 proteins. The subset of 1,186 proteins whose length is no greater than 200 residues is the PD8200holo set.

Detection of Distant Evolutionary Relationships Among Proteins. For a given protein in PD8300holo, we thread against the entire structural template library using PROSPECTOR.3.5 (32) and exclude structural relationships between all protein partners whose z-score is ≤ 4 or higher. As PROSPECTOR.3.5 is driven by a strong sequence profile component, it can detect distant evolutionary relationships among proteins. To detect even more evolutionary distant pairs of proteins, we examined proteins of similar structure that share at least one common binding site and which are predicted to bind a correlated set of ligands using FINDSITE (30). The set of structural relationships after these exclusions is the PD8300set. PD8200holo is defined analogously from PD8200holo.

Additional details are in the SI Appendix.

TM-Score and the Structural Alignment Program TM-Align. The TM-score between the structure of template protein B with respect to target protein A, of lengths \( N_b \) and \( N_a \), respectively, is:

\[
\text{TM-score}(B \rightarrow A) = \frac{1}{N_a} \sum_{i=1}^{N_{align}} \frac{1}{1 + (d(d_i(N_A))^{1/3} - 1.8)}
\]

where \( N_{align} \) is the number of aligned residues, \( d_i \) is the distance between the \( i^{th} \) pair of aligned residues (1 ≤ \( i \) ≤ \( N_{align} \)) and \( d_i(N_A) \) is the average distance between a pair of residues in a randomly related structure pair (25). For unequal-length protein pairs, from Eq. 1a, \( \text{TM-score}(B \rightarrow A) = \text{TM-score}(A \rightarrow B) \), i.e., the TM-score is non-symmetric. To perform structural alignments, we employ an improved version of TM-align, fr-TM-align (27).

Calculation of Kth Neighbors in Protein Structure Space. For a given TM-score value of the structural similarity cutoff \( d \), the mean fraction of structures that are Kth neighbors, \( f_k \), and the fraction of proteins that are part of the LSCC, \( S \), are calculated using a standard-depth first algorithm from graph theory (35). The implementation details are in the SI Appendix.

Random Directed Digraphs. We represent the network of structural relationships between protein pairs as a 2-colored directed graph or digraph, whose nodes are protein structures and whose edges have a direction. The number of template proteins aligned to protein i is the in-degree and the number of proteins that protein i is aligned to is the out-degree of i. Each node can have one of 2 possible colors that correspond to small proteins (up to 200 residues) or large proteins (200–300 residues) that act as structural bridges. Random digraphs are generated by a procedure described in the SI Text that conserves the number of nodes, their colors, and first-order local connectivity. Fig. S1 illustrates how we represent structural relationships between proteins in a graph.

Polyvaline Simulations. For each protein in the PD8300holo library, we extract its secondary structure (helix, strand, and coil) and transfer this bias to a polyvaline sequence of the same length. For helices, we use the highly accurate helix extraction subroutine in TM-align (25). For strands, we use the high-accuracy strand assignment algorithm our group described earlier (36). Residues assigned as helices are unchanged in the simulation, whereas strands and loops/turns experience a conformational bias and can dissolve and reform simulation. Folding is done using the ab initio version of TASSER (33), with modifications to the hydrogen bond potential described in the SI Text. The resulting structures are clustered and the structural properties of the space comprised of the top 8 clusters per secondary structure arrangement in PD8300holo are reported.

ACKNOWLEDGMENTS. This work was supported by grants GM-44835 and GM-37408 of the Division of General Medical Sciences of the National Institutes of Health. We thank Dr. Jose Borreguero for suggesting threading to remove evolutionary relationships between proteins and Dr. Shashi Pandit and Jake Bogdan for useful discussions.
SI METHODS

Graph representation of the structural relationships between proteins

Figure SI 1 illustrates how we represent structural relationships between proteins in a graph at two TM-score cutoff values \( d=0.4 \) and \( d=0.60 \). Figures SI 1A and SI 1D show the corresponding TM-score matrices for a set of 5 protein structures identified as target and template respectively, at these two TM-score cutoffs. Figure SI 1B and SI 1E show the corresponding graphs derived from the TM-score matrices, with the set of proteins belonging to the LSCC located inside the blue curve. Finally, Figures SI 1C and SI 1F show the matrices representing the neighboring order \( k \) between each pair of template-target proteins.

Algorithm for the calculation of \( k^{\text{th}} \) neighbors in protein structure space

Template protein \( j \) is defined to be a structural first neighbor of target protein \( i \) at a TM-score cutoff \( d \) if the TM-score(\( j \to i \)) \( \geq d \). We define a nonsymmetric matrix whose matrix elements \( t_1(i,j)=1 \) if \( j \) is a structural neighbor of \( i \); otherwise, \( t_1(i,j)=0 \). For such protein pairs, we consider structure \( j \) to be a first neighbor, i.e. \( k=1 \), of protein structure \( i \). If there are \( N \) structures in the PDB library, then the total number of \( k=1 \) neighbors of protein structure \( i \) is given by

\[
n_1(i) = \sum_{j=1}^{N} t_1(i,j)
\]

For computational efficiency, we want the list of all \( n(i,k-1) \) neighbors of \( i \), the \( m^{\text{th}} \) member of which is \( jj \). This can be compactly represented by

\[
nl_1(m,i) = jj \text{  for  } m=1,\ldots, n_1(i).
\]
If target and template structures $i$ and $j$ and $j$ and $l$ are first neighbors (i.e. $t_1(i,j)=1$ and $t_1(j,l)=1$), but $i$ and $l$ are not ($t_1(i,l)=0$), then $i$ and $l$ are second neighbors. That is, $t_2(i,l)=1$. If structures $i$ and $l$ are not second neighbors, then $t_2(i,l)=0$.

More generally, we construct the neighbor matrix,

$$neib(i, j, k') = k'$$

(3a) that is to say, structures, $i$ and $j$ are $k'^{th}$ structural neighbors. The value of $k'$ is determined by the minimum $k$ such that $t_{k-1}(i,j)=0$ while $t_k(i,j)=1$. Otherwise,

$$neib(i, j, kk') = 0$$

(3b) for all $kk' < k'$ and all $kk' > k'$.

We are now in a position to construct the recursive relationships to ascertain the subset of structures that are no more than $k^{th}$ neighbors.

Consider for the $i^{th}$ structure, for all $m=1, 2, .. n_k(i)$ neighbors, the $jj^{th}$ structure that is a $k^{th}$ neighbor

$$jj = nl_k(m,i)$$

(4a) Consider now the neighbors of structure $jj$, the $m'^{th}$ of which is

$$ll = nl_{k'}(m', jj)$$

(4b) for all $m'$ neighbors $=1, 2, .. n(j,k')$. Then, structure $i$ and $ll$ are $k''$ structural neighbors, provided that all $t_k(i,ll)=0$ for $k<k''$ where

$$k'' = neib(i, jj, k) + neib(jj, ll, k')$$

(5a) and we set

$$neib(i, ll, k'') = k''$$

(5b) and
where of course for all $i$ and $ll$, we chose the minimum $k''$ obtained from eq. 5b for all intermediate structures $jj$.

From $t_{k''}(i,ll)$, we can calculate the $k''$ neighbor list for structure $i$ as follows

$$n_{k''}(i) = \sum_{j=1}^{N} t_{k''}(i,j)$$

The $m$th structure that is the $k''$th neighbor of structure $i$ immediately follows from eq. 5d, by just counting the number of structures for which $t(i,ll,k'')$ is non zero.

In practice, the recursion relationships embodied in eq. 2-5, identify all $2^{s-1}$ neighbors for $s=1, 2, 3,$ etc. The first round calculates all $k=1$ neighbors. Then, in the second round, $s=2$, we identify all structures that are at most second neighbors, viz. $k=2$. Let $i,j$ and $j,l$ be first neighbors and $i,l$ are second neighbors, schematically depicted as $l \rightarrow j \rightarrow i$. In the third round of the iteration, $s=3$, we will identify at most the following sets of structure neighbors, $q \rightarrow p \rightarrow l \rightarrow j \rightarrow i$; thus, structure $q$ is the fourth neighbor to $i$ and second neighbor to $l$. We will also identify structures $p \rightarrow l \rightarrow j \rightarrow i$, where $q$ is the third neighbor to $i$ and first neighbor to $l$.

In the next round ($s=4$), we select all structures between 5th and 8th neighbors. The reason for this is as follows: Consider a set of protein structures $v,q$ and where $v \rightarrow u \rightarrow t \rightarrow r \rightarrow q \rightarrow p \rightarrow l \rightarrow j \rightarrow i$, since structure $v$ is a fourth neighbor to structure $q$ and thus an eighth neighbor to structure $i$. In a similar fashion, structures $u,t$ and $r$ are seventh, sixth and fifth neighbors to structure $i$. Thus, this is a rapid way of identifying all $k=2^{s-1}$ neighbors in structure space after $s$ iterations.

The average fraction of proteins, $f_k$, with no more than $k^{th}$ neighbors is readily obtained from eq. 5d as follows
\[
    f_k = \frac{\sum_{k=1}^{N} \sum_{i=1}^{N} t_k(i,j)}{N^2}
\]  

(6)

In practice, we explore all \(s\) iterations until the results from iteration \(s\) and iteration \(s-1\) are the same. This defines the maximum average fraction of structures that are related, i.e. the converged value, \(f_{\text{max}}\). We further wish to identify the set of all strongly connected clusters all of whose members satisfy

\[
    t_s(i, j) = t_s(j, i)
\]

(7)

That is, both the target and template structure pairs \(i\) and \(j\) are structural neighbors. We identify LSCC, the largest strongly connected component, which is the largest strongly connected cluster of a directed graph, all of whose members satisfy eq. 7.

**FINDSITE based approach to establish functional relationships between proteins**

For each protein, binding sites were detected using FINDSITE (1), an evolution/structure-based approach for ligand-binding site prediction and functional annotation that has been demonstrated to provide accurate functional annotations by detecting common binding sites in evolutionarily related proteins. We employ a set of closely related template structures to assign highly confident binding pockets to the dataset proteins. First, for each protein, ligand-bound structures with the sequence identity >35% were selected from the PDB (Oct-08). Then, FINDSITE was used to transfer template-bounds ligands into a target protein upon the global target-template structure alignments generated by TM-align. Binding pockets were identified by the spatial clustering of ligands using a cutoff distance of 8 Å and ranked by the number of ligands. Here, we consider the top five binding sites. In addition to the criterion of localization of the binding sites in the protein
structure, we impose the second criterion that a pair of proteins must be predicted to have a similar set of binding properties to a library of small molecule ligands.

The second criterion demands that there be a certain chemical similarity between molecules that bind to individual pockets. The Tanimoto coefficient (2) calculated for molecular bit strings, using SMILES or SMARTS (3), is one of the most commonly used measures in chemoinformatics to quantify the similarity between small molecules. Since binding sites detected by FINDSITE are typically associated with ligands extracted from similar sites in evolutionarily related proteins, we exploit this information to develop a more sensitive metric that is very much in the spirit of sequence profile-profile similarity measures (1). Previously, we found that the set of ligands provided by FINDSITE quite well describes the chemical aspects of binding and can be used to construct molecular fingerprint profiles for use in simple ligand-based virtual screening against a diverse compound library. As a result, the top fraction of the ranked library is significantly enriched with known binders (1). Here, we use this to construct a chemical similarity metric with respect to ligand-binding sites, referred to as a chemical correlation. The collection of ligands identified for each binding site is used to rank the KEGG compound library (Oct-07) that comprises 12,158 diverse molecules (4). Subsequently, the Pearson’s correlation coefficient (CC) is calculated using the library ranks obtained for two binding pockets. A high CC (>0.5) indicates that the pockets not only exhibit specific binding affinity toward similar ligands, but also do not bind similar ligands. A significant structure alignment and common localization of the binding pockets in conjunction with a high chemical correlation establishes a functional relationship between a pair of proteins.
Random directed graph generation

Given a colored reference digraph, our goal is to generate random directed graphs that preserve the total number of nodes and edges, and also the color and the local connectivity properties of every node. The first order local connectivity of a node $i$ is completely defined by three numbers: 1) $N_{in}$, the number of nodes $j$ adjacent to $i$ that can reach $i$ but cannot be reached from $i$, 2) $N_{out}$, the number of nodes $j$ adjacent to $i$ that can be reached from $i$ but cannot reach $i$, and 3) $N_{in-out}$, the number of nodes $j$ adjacent to $i$ that can both reach $i$ and be reached from $i$. We separately consider three types of relationships between two adjacent nodes $i$ and $j$ ($i \leftarrow j$, $i \rightarrow j$, and $i \leftrightarrow j$) to account for the correlation we observed between $N_{in} + N_{in-out}$ (indegree) and $N_{out} + N_{in-out}$ (outdegree). To generate a random graph from the reference digraph, first, we randomly select four nodes ($i_1$, $j_1$, $i_2$, and $j_2$). Then, we evaluate the following conditions: 1) color preservation, i.e., color ($i_1$) = color ($i_2$), and color ($j_1$) = color ($j_2$), 2) existence of the same type of adjacency relationship between each pair $i$, $j$, whether $i \leftarrow j$, $i \rightarrow j$, or $i \leftrightarrow j$, and 3) absence of adjacency between $i_1$ and $j_2$, and between $i_2$ and $j_1$. If the three conditions are fulfilled, we remove the edge/s from $i_1$ to $j_1$ and from $i_2$ to $j_2$ and draw the same type of edge/s from $i_1$ to $j_2$ and from $i_2$ to $j_1$. We repeat the steps of random selection of four nodes and swapping of edges until convergence of the average number of first neighbors per node that are identical to those in the original digraph. For the analyzed digraphs, the convergence occurs after approximately $n^2$ iterations, where $n$ is the number of nodes in the digraph. The properties of the resulting graph are then analyzed, and the procedure is repeated for a total of 2000 times from which the relevant statistics of the properties of the random digraphs are calculated. Since each swapping step maintains $N_{in}$,
$N_{\text{out}}$ and $N_{\text{in-out}}$ of each involved node, the nodes in the original and the randomized digraphs will have identical and equally correlated out-degree and in-degree distributions.

**TASSER Force Field**

Most of the energy potential terms in TASSER have been previously described (5-7). Here, we summarize the energy terms that are used in the folding simulations of polyvaline:

- **Generic backbone hydrogen bonding:** Two Cα-atoms, $C\alpha_i$ and $C\alpha_j$ interact when the backbone fragments $C\alpha_{i-1}-C\alpha_i-C\alpha_{i+1}$ and $C\alpha_{j-1}-C\alpha_j-C\alpha_{j+1}$ adopt geometries observed in protein structures, under the condition that a hydrogen bond forms residues $i$ and $j$.

- **A bias in the hydrogen bonding to select for geometries of the $C\alpha_{i-1}-C\alpha_i-C\alpha_{i+1}$ and $C\alpha_{j-1}-C\alpha_j-C\alpha_{j+1}$ fragments compatible with the pre-assigned secondary structure state of $C\alpha_i$ and $C\alpha_j$.

- **For sheet and coil residues, short-range backbone correlations enhance the propensity of the backbone to adopt the pre-assigned secondary structures.**

- **A centrosymmetric potential that promotes a compact globular protein conformation.**

- **Orientation-dependent, generic attractions between side-chains:** we assign a binding energy for the two Val side-chains depending on their mutual orientation and distance between their side chain centers of mass.

- **Impenetrable hard-core radii for each Cα and side-chain center of mass.**

Interestingly, we found that the original TASSER hydrogen bond scheme while capable of covering most of the PDB did not generate a library of structures that are highly connected. This effect was mainly operative for β-sheet containing proteins, which generated
non physical geometries and thereby dramatically increased the size of the conformation space so that it was not so well connected. Examination of the original hydrogen bond scheme revealed that it was far too permissive and allowed for quite twisted, non planar β strands to interact. Thus, to generate the library of compact homopolypeptide structures, we introduce a cooperative hydrogen bond term, \( E_{\text{HBCoop}} \), into the TASSER force field to promote hydrogen bond networks. We note that this hydrogen bond cooperative term especially encourages the hydrogen bond networks among β secondary structures. Only main chain hydrogen bonding is considered.

In hydrogen bond interactions, one residue can make one hydrogen bond with the other residue by playing either donor or acceptor roles or it can make two hydrogen bonds by playing both donor and acceptor roles. Then,

\[
E_{\text{HBCoop}} = - \sum_i [\Theta(i) + \delta(i)] \quad (8)
\]

\[
\Theta(i) = \begin{cases} 
1, & \text{if } \text{HB}(j,i,k) = 1 \\
0, & \text{otherwise},
\end{cases} \quad (9)
\]

where \( \text{HB}(j,i,k) \) is 1 if the \( j \)-th and \( i \)-th residues make a hydrogen bond by donor \( i \) and acceptor \( j \) residues and the \( i \)-th and \( k \)-th residues make a hydrogen bond by an acceptor \( i \)- and donor \( k \)-th residues. Moreover,

\[
\delta(i) = \begin{cases} 
1, & \text{if } \Theta(i) = 1 \text{ and } \Theta(i+1) = 1 \\
0, & \text{otherwise}
\end{cases} \quad (10)
\]

The \textit{ab initio} version of TASSER is employed to sample the conformational space of a given polyvaline sequence (8). SPICKER (9) is used to cluster the resulting set of structures with up to the top eight most populated clusters selected for subsequent analysis.
SI RESULTS

PDB300 set results

The full PDB300 set contains 5906 compact proteins between 40 and 300 residues for which all-against-all structural alignments were done. (The full list, the PDB300\textsuperscript{holo} list, and for each target protein, the set of proteins with associated TM-scores $\geq 0.40$ are found at http://cssb2.biology.gatech.edu/skolnick/files/FoldSpaceContinuity/). Here, proteins up to 300 residues are included in the analysis. Figure SI 2A shows the fraction of proteins that are no more than $k$=1,2,4,16 and 32\textsuperscript{nd} neighbors, given that first neighbors have a TM-score $\geq d$. The red line shows the asymptotic result for $f_{\text{max}}$ for the PDB200\textsuperscript{holo} set (the subset of the full PDB300 set where only proteins whose lengths range from 40 to 200 residues for which functional assignments are available; see Methods). Similarly, Figure SI 2B shows, at $d$=0.40, the relative size of the LSCC as a function of the $k$\textsuperscript{th} neighbor cutoff for the PDB300 set (black thick line) and the PDB200\textsuperscript{holo} set (red line). The thin line shows random digraph results given the same first order local connectivity for each node as in the PDB300 set. Figure SI 2C shows the length distribution of the relative abundance of proteins excluded from the LSCC at $d$=0.40. Again, the same trends are seen as in the PDB200\textsuperscript{holo} and PDB200\textsuperscript{x} sets.

These results clearly show that protein structure space is almost completely connected, with a dataset size dependence typical of a cooperative transition (10, 11). Thus, the results reported in the main text for proteins below 200 residues are in fact more general. We note that in the main text we have restricted our analysis to these smaller proteins in order to be able to compare results more directly with the polyvaline homopolypeptide library.
Length of the shortest path for different secondary structure classes

Figure SI 3 shows the distributions of the shortest path length \(k\) (neighboring order) for proteins pairs from the PDB200\(^x\) set that belong to the specified secondary classes. For example, the plot labeled \(\alpha-\alpha\) corresponds to all possible pairs of \(\alpha\) protein templates linked to \(\alpha\) protein targets, while the plot labeled \(\alpha-\beta\) corresponds to all possible pairs of \(\alpha\) protein templates linked to \(\beta\) protein targets and \(\beta\) protein templates linked to \(\alpha\) protein targets. The median value of the shortest path length for protein pairs of different secondary structure class is \(k=3\), which is identical to that corresponding to \(\alpha-\alpha\) or \(\alpha\beta-\alpha\beta\) protein pairs, and only one unit larger than that for \(\beta-\beta\) protein pairs.

Sequence alignments corresponding to the proteins structurally aligned in Figure 1

Alignment between 1gnyA (153 residues) and 1ekrA (143 residues)

1gnyA  GNVVIEVMANGWRGNASGSTSHSGITYSADGVTFAALGHDGVGAVFDI--
1ekrA  -------------------------------------------GE

1gnyA  --ARPT-----TLEDAVIAMVNVV--AE--FK---AS---EAN--LQ--I
1ekrA  AHMVVSAKAETVREARAEAPVTMRSETALMIDGRHKGDFVATAIAG

1gnyA  F---AQ--LKE------DWGK--EDCLAGSSELTA--DDTDLTLCTIDEDDDK
1ekrA  IQAAKRTWDLPLCHPLMKEVNL--QAEPEHNRVRIETLCLR--TG

1gnyA  FNQTAR--DVQ--V--GQ--AKG--TPAG--T--ITIKSVTI--TLQQSA--
1ekrA  -KIG--VEMEALTAASVAALTIT--YDMCKAYQKMDVFGVRLAKSGDFK

Alignment between 1ekrA (143 residues) and 101m (154 residues)

1ekrA  GEAHMVVSAKAETVREARAEAFV-------T--MRSETLMIDIHGK
101m  ---------------------------------------------------GE

1ekrA  DVFATARAGQ--AAKRTWDLPLCHPLMKEV-----------------
101m  GHQDILRLFKSHPET------------------LEKFDVYKHLKETAMKASED

1ekrA  ---V---NL--QA--EPEHNRV------RI--ET--LCLRTGRTG--VEMEAL
101m  LKKHGVTVLALGAILKKKG--HHEAEKPLAQSHATK-------HK--IFI

1ekrA  TAASVAALTITYDMCK---A--VQK--D--M--VI---GP--VR--L-LAKSGDG
101m  KYLEFISEAIIHVLHRSHPGNGDAQQGAMNKALELFRKDIAAKYKELG

1ekrA  FK--
101m  Q--G
REFERENCES


Fig. S1. Graphical representation of structural relationships between proteins. For TM-score cutoffs $d = 0.40$ (A–C) and $d = 0.60$ (D–F), we show (i) the matrix of TM-scores between template and target proteins (A and D), where the TM-scores greater or equal to the corresponding TM-score cutoff are highlighted in red; (ii) the graph derived from the TM-score matrix, representing the structural relationships between template and target proteins (B and E), with the LSCC of the graph contained inside the blue curve; and (iii) the matrix representing the neighboring order $k$ between each template and target protein (C and F). For the sake of simplicity, this figure does not show the distinction between small proteins (up to 200 residues) or large proteins (200–300 residues, acting as structural bridges), which can be represented as 2 possible colors associated with each vertex in the graph (see “Random Directed Digraphs” in Methods in the main text).
Fig. S2. (A) Mean fraction of proteins in the PDB300 set (black lines with diamonds) or in the PDB200holo set (red line) that are no more than \(k\)th neighbors (\(f_k\)) and whose first neighbors have a TM-score \(\geq d\). (B) Relative size of the LSCC in the PDB300 set (black lines) or in the PDB200holo set (red line) as a function of the \(k\)th neighbor cutoff, at \(d = 0.40\). The thick lines corresponds to the values observed in the original PDB300 set (black line with diamonds) or PDB200holo set (red line), whereas the thin line indicates the median values obtained from 2,000 randomly generated digraphs with the same number of nodes, edges, and first-order connectivity per node as in the original PDB300 set (error bars indicate minimum and maximum values from the 2,000 random graphs). (C) Length distribution of proteins not belonging to the LSCC, at \(d = 0.40\), relative to all proteins in the PDB300 set. The relative abundance is calculated as the fraction of the total number of proteins excluded from the LSCC that fall in a given interval of protein length divided by the fraction of the total number of proteins members of the set falling in that same interval of protein length.
Fig. S3. Distribution of the length of the shortest path $k$, linking all proteins pairs in the PDB200x set, discriminated by secondary structure class. The statistics represented in each box-and-whisker plot are the first percentile (filled circle, Bottom), fifth percentile ($\circ$, Bottom), 10th percentile (whisker, Bottom), 25th percentile (box, Bottom), median (thick blue line), average (thin red line), 75th percentile (box, Top), 90th percentile (whisker, Top), 95th percentile ($\circ$, Top), and 99th percentile (filled circle, Top).

Other Supporting Information Files

SI Appendix