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^{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($j \rightarrow$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

$$n_1(i) = \sum_{i=1}^{N} t_1(i,j) \qquad (2a)$$

For computational efficiency, we want the list of all $n(i,k-1)$ neighbors of i, the mth member of which is jj. This can be compactly represented by

$$nl_l(m,i) = jj \quad \text{for m=1,…, } n_1(i). \qquad (2b)$$

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,

$$neib(i,j,k') = k' \tag{3a}$$

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,

$$neib(i,j,kk') = 0 \tag{3b}$$

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

$$jj = nl_k(m,i) \tag{4a}$$

Consider now the neighbors of structure jj, the m'^{th} of which is

$$ll = nl_{k'}(m', jj) \tag{4b}$$

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

$$k'' = neib(i, jj, k) + neib(jj, ll, k') \tag{5a}$$

and we set

$$neib(i, ll, k'') = k''. \tag{5b}$$

and

$$t_{k''}(i,ll) = 1 \tag{5c}$$

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

$$n_{k''}(i) = \sum_{i=1}^{N} t_{k''}(i,j) \tag{5d}$$

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 5[th] and 8[th] 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

$$f_k = \sum_{k'-1}^{k} \sum_{i=1}^{N} \sum_{j=1}^{N} t_k(i,j) \Big/ N^2 \qquad (6)$$

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, $f_{\max}$. We further wish to identify the set of all strongly connected clusters all of whose members satisfy

$$t_s(i,j) = t_s(j,i) \qquad (7)$$

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) $\mathbf{N_{in}}$, the number of nodes **j** adjacent to **i** that can reach **i** but cannot be reached from **i**, 2) $\mathbf{N_{out}}$, the number of nodes **j** adjacent to **i** that can be reached from **i** but cannot reach **i**, and 3) $\mathbf{N_{in\text{-}out}}$, the number of nodes **j** adjacent to **i** that can both reach **i** and be reached from it. We separately consider three types of relationships between two adjacent nodes **i** and **j** (**i**←**j**, **i**→**j**, and **i**↔**j**) to account for the correlation we observed between $\mathbf{N_{in}}$ + $\mathbf{N_{in\text{-}out}}$ (indegree) and $\mathbf{N_{out}}$+ $\mathbf{N_{in\text{-}out}}$ (outdegree). To generate a random graph from the reference digraph, first, we randomly select four nodes ($\mathbf{i_1}$, $\mathbf{j_1}$, $\mathbf{i_2}$ and $\mathbf{j_2}$). Then, we evaluate the following conditions: 1) color preservation, i.e., color ($\mathbf{i_1}$) = color ($\mathbf{i_2}$), and color ($\mathbf{j_1}$) = color ($\mathbf{j_2}$), 2) existence of the same type of adjacency relationship between each pair **i**, **j**, whether **i**←**j**, **i**→**j**, or **i**↔**j**, and 3) absence of adjacency between $\mathbf{i_1}$ and $\mathbf{j_2}$, and between $\mathbf{i_2}$ and $\mathbf{j_1}$. If the three conditions are fulfilled, we remove the edge/s from $\mathbf{i_1}$ to $\mathbf{j_1}$ and from $\mathbf{i_2}$ to $\mathbf{j_2}$ and draw the same type of edge/s from $\mathbf{i_1}$ to $\mathbf{j_2}$ and from $\mathbf{i_2}$ to $\mathbf{j_1}$. 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 $\mathbf{N_{in}}$,

$\mathbf{N_{out}}$ and $\mathbf{N_{in\text{-}out}}$ 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, $C\alpha_{\mathbf{i}}$ and $C\alpha_{\mathbf{j}}$ interact when the backbone fragments $C\alpha_{\mathbf{i-1}}$-$C\alpha_{\mathbf{i}}$-$C\alpha_{\mathbf{i+1}}$ and $C\alpha_{\mathbf{j-1}}$-$C\alpha_{\mathbf{j}}$-$C\alpha_{\mathbf{j+1}}$ 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 $C\alpha_{\mathbf{i-1}}$-$C\alpha_{\mathbf{i}}$-$C\alpha_{\mathbf{i+1}}$ and $C\alpha_{\mathbf{j-1}}$-$C\alpha_{\mathbf{j}}$-$C\alpha_{\mathbf{j+1}}$ fragments compatible with the pre-assigned secondary structure state of $C\alpha_{\mathbf{i}}$ and $C\alpha_{\mathbf{j}}$.
- 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_{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,

$$E_{\mathrm{HBCoop}} = -\sum_i [\Theta(i) + \delta(i)] \tag{8}$$

$$\Theta(i) = \begin{cases} 1, \text{ if } \mathrm{HB}(j,i,k) = 1 \\ 0, \text{ otherwise,} \end{cases} \tag{9}$$

where 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,

$$\delta(i) = \begin{cases} 1, \text{ if } \Theta(i) = 1 \text{ and } \Theta(i+1) = 1 \\ 0, \text{ otherwise} \end{cases} \tag{10}$$

The *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^{holo} list, and for each target protein, the set of proteins with associated TM-scores ≥ 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 32nd neighbors, given that first neighbors have a TM-score ≥d. The red line shows the asymptotic result for f_{max} for the PDB200^{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^{th} neighbor cutoff for the PDB300 set (black thick line) and the PDB200^{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^{holo} and PDB200^{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 α-α corresponds to all possible pairs of α protein templates linked to α protein targets, while the plot labeled α-β corresponds to all possible pairs of α protein templates linked to β protein targets and β protein templates linked to α 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 α-α or αβ-αβ protein pairs, and only one unit larger than that for β-β protein pairs.

Sequence alignments corresponding to the proteins structurally aligned in Figure 1

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

```
1gnyA GNVVIEVDMANGWRGNASGSTSHSGITYSADGVTFAALGDGVGAVFDI--
1ekrA ----------------------------------------------GE

1gnyA --ARPT-----TLEDAVIAMVVNVS--AE-FK---AS---EAN--LQ--I
1ekrA AHMVDVSAKAETVREARAEAFVTMRSETLAMIIDGRHHKGDVFATARIAG

1gnyA F--AQ-LKE-----DWSKG-EWDCLAGSSELTA-DTDLTLTCTIDEDDDK
1ekrA IQAAKRTWDLIPLCHPLMLSKVEVNL---QAEPEHNRVRIETLCRL--TG

1gnyA FNQTAR-DVQ--V--GIQ--AKG-TPAG--T--ITIKSVTI-TLAQEA--
1ekrA -KTG--VEMEALTAASVAALTI-YDMCKAVQKDMVIGPVRLLAKSSGDFK
```

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

```
1ekrA GEAHMVDVSAKAETVREARAEAFV-------T-MRSETLAMIIDGRHHKG
101m  ------------------------MVLSEGEWQLVLHVWAKVEADV---A

1ekrA DVFATARIAGIQ-AAKRTWDLIPLCHPLMLSKVE----------------
101m  GHGQDILIRLFKSHPET------------LEKFDRVKHLKTEAEMKASED

1ekrA ---V---NL--QA--EPEHNRV----RI--ET--LCRLTGKTG-VEMEAL
101m  LKKHGVTVLTALGAILKKKG--HHEAELKPLAQSHATK-----HK--IPI

1ekrA TAASVAALTIYDMCK---A-VQK-D-M--VI---GP-VR--L-LAKSSGD
101m  KYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKELGY

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

REFERENCES

1. Brylinski M & Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* 105(1):129-134.
2. Tanimoto TT (1958) An elementary mathematical theory of classification and prediction. in *IBM Internal Report*).
3. Anonymous (2007) Daylight Theory Manual (Daylight Chemical Information Systems, Inc., Aliso Viejo, CA), 4.9.
4. Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27-30.
5. Zhang Y, Kolinski A, & Skolnick J (2003) TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J* 85(2):1145-1164.
6. Zhang Y, Arakaki AK, & Skolnick J (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* 61 Suppl 7:91-98.
7. Zhang Y, Devries ME, & Skolnick J (2006) Structure Modeling of All Identified G Protein-Coupled Receptors in the Human Genome. *PLoS Comput Biol* 2(2):e13.
8. Borreguero JM & Skolnick J (2007) Benchmarking of TASSER in the ab initio limit. *Proteins* 68(1):48-56.
9. Zhang Y & Skolnick J (2004) SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* 25(6):865-871.
10. Poland D & Scheraga HA (1966) Phase transitions in one dimension and the helix-coil transition in polyamino acids. *J Chem Phys* 45(5):1456-1463.
11. Poland D, Vournakis JN, & Scheraga HA (1966) Cooperative interactions in single-strand oligomers of adenylic acid. *Biopolymers* 4(2):223-235.