SI Text

Spatial localization of frustration measurement and databases used. In general terms, localizing energetic frustration requires the evaluation of the energy of a protein in its native state and comparison to the energies of a set of 'decoy' states. The algorithm requires as input a high resolution structure and an accurate energy function.

Energy function and frustration index definition. We chose to base our energy function on the the Associative Memory Hamiltonian optimized with water-mediated interactions (AMW). This is a natural choice given its success in predicting structure from sequence (1). The optimization scheme is also based on Energy Landscape ideas, and should be well suited in cases where sub-optimal interactions occur. We evaluate only the sequence-specific contact and burial terms of the AMW ($\mathcal{H}_{contact}$, $\mathcal{H}_{water}$, $\mathcal{H}_{burial}$) (1). These terms depend on the amino-acid identities (λ), densities (ρ) and interaction distances (r_{ij}) of all residues involved. Three types of contacts are categorized between residues: short range (distance between C_β below $6.5\mathring{A}$), long range (between 6.5 and $9.5\mathring{A}$), and water-mediated (long range and exposed to solvent). In addition, the model includes a single-residue burial term that takes into account the atomic density around each amino acid, thus accounting for solvent exposure. The contact and burial terms were previously parameterized for the interactions between the most commonly occurring 20 natural aminoacids. The total energy of a protein (E_0) in a given configuration is then computed as the sum over all the residue burial and inter-residue contact terms.

The local frustration index is a site-specific measure of the energetic fitness for a given set of residues λ_i and λ_j at residue positions i and $j > i+1$. Two definitions are presented based on alternative underlying assumptions and goals (referred to herein as "configurational" and "mutational" frustration; see main text).

Local configurational frustration index at a given site is defined as:

$$F_{ij}^c = (\mathcal{H}_{ij}^N - < \mathcal{H}_{i',j'}^U >)/\sqrt{1/N\sum_{k=1}^{n}(\mathcal{H}_{i',j'}^U - < \mathcal{H}_{i',j'}^U >)^2}, \quad [1]$$

where $\mathcal{H}_{ij}^N = \mathcal{H}_{contact}^{i,j} + \mathcal{H}_{water}^{i,j} + \mathcal{H}_{burial}^{i} + \mathcal{H}_{burial}^{j}$ is the native energy with native parameters (λ_i, λ_j, ρ_i, ρ_j, r_ij). We obtain the average and standard deviation of a set of reference energies, $\mathcal{H}_{i'j'}^U = \mathcal{H}_{contact}^{i',j'} + \mathcal{H}_{water}^{i',j'} + \mathcal{H}_{burial}^{i'} + \mathcal{H}_{burial}^{j'}$, by randomly selecting the parameters (λ_i', λ_j', ρ_i', ρ_j', r_ij') according to the native distributions.

With this definition, for a given protein sequence composition and structure, the average and standard deviation of reference energies are the same for all interacting residue pairs i, j. When $\mathcal{H}_{ij}^N = < \mathcal{H}_{i',j'}^U >$, the native energy is not discriminated from a typical energy at a random site, and $F_{ij}^c = 0$; For the present study, highly frustrated interactions are those where $F_{ij}^c < -1$. Arguments from theory suggest an interaction is minimally frustrated when $F_{ij}^c > 0.78$ (see main text).

For configurational frustration, our treatment of decoy states approximates the discrimination of the native pairwise interaction from those expected in a molten-globule state. In the native state, however, other interactions are directly influenced by pairwise mutations. To capture this affect, we define 'mutational frustration':

$$F_{ij}^m = (\mathcal{H}_{i,j}^{T,N} - < \mathcal{H}_{i',j'}^{T,U} >)/\sqrt{1/N\sum_{k=1}^{n}(\mathcal{H}_{i',j'}^{T,U} - < \mathcal{H}_{i',j'}^{T,U} >)^2}, \quad [2]$$

where $\mathcal{H}_{i,j}^{T,N} = \mathcal{H}_{contact}^{i,j} + \sum_{k=!i,k=!j}^{N}(\mathcal{H}_{contact}^{i,k} + \mathcal{H}_{contact}^{k,j} + \mathcal{H}_{water}^{i,k} + \mathcal{H}_{water}^{k,j}) + \mathcal{H}_{burial}^{i} + \mathcal{H}_{burial}^{j}$ is the native site energy with native parameters (λ_i, λ_j, ρ_i, ρ_j, r_ij). We obtain the average and standard deviation of a set of reference energies, $\mathcal{H}_{i',j'}^{T,U} = \mathcal{H}_{contact}^{i',j'} + \mathcal{H}_{water}^{i',j'} + + \sum_{k=!i,k=!j}^{U}(\mathcal{H}_{contact}^{i',k} + \mathcal{H}_{contact}^{k,j'} + \mathcal{H}_{water}^{i',k} + \mathcal{H}_{water}^{k,j'}) + \mathcal{H}_{burial}^{i'} + \mathcal{H}_{burial}^{j'}$, by randomly selecting the amino acid identities (λ_i', λ_j') according to the protein's distribution while fixing the density and pairwise distance parameters (ρ_i, ρ_j, r_ij) to those in the native conformation. This scheme effectively evaluates every possible mutation of an amino acid pair that forms a particular contact.

An alternative scheme is also considered where mutations of only single residues are constructed. This leads to our definition of single-residue mutational frustration:

$$F_i^m = (\mathcal{H}_i^{T,N} - < \mathcal{H}_{i'}^{T,U} >)/\sqrt{1/N\sum_{k=1}^{n}(\mathcal{H}_{i'}^{T,U} - < \mathcal{H}_{i'}^{T,U} >)^2}, \quad [3]$$

where $\mathcal{H}_i^{T,N} = \sum_{k=!i}^{N}(\mathcal{H}_{contact}^{i,k} + \mathcal{H}_{water}^{i,k}) + \mathcal{H}_{burial}^{i}$ is the native site energy with native parameters (λ_i, ρ_i, r_ik). We obtain the average and standard deviation of a set of reference energies, $\mathcal{H}_{i'}^{T,U} = \sum_{k=!i}^{U}(\mathcal{H}_{contact}^{i',k} + \mathcal{H}_{water}^{i',k}) + \mathcal{H}_{burial}^{i'}$, by randomly selecting the amino acid identities (λ_i') according to the protein's distribution while fixing the density and pairwise distance parameters (ρ_i, r_ik) to those in the native conformation.

Protein monomer and complex databases. We constructed a database of high-quality monomeric proteins from the Protein Data Bank (PDB) based on a series of filtering steps. First, all PDB entries with only a single unique chain solved by X-ray diffraction were obtained. From this list, those classified by the SCOP database Classification of Proteins database (SCOP) (3) as either 'membrane', 'cell-surface', or 'small' proteins were excluded. Next, structures with relatively low-resolution (better than $3\mathring{A}$ resolution), chain breaks, or co-factors were removed. This list of high-quality structures was then filtered for redundancy at the level of 30% sequence-similarity. The remaining 314 monomeric proteins were evaluated.

To study protein-protein interfacial interactions, we used the Benchmark II (2) database based on high-resolution crystal structures from the PDB. This is a database of non-redundant multimeric protein complexes for which most of the individual monomeric crystal structures exist. Complexes are categorized based on the degree of conformational change at the interface to which the structures of the monomers alone are different in complex. We chose to study only 'Rigid-body' complexes - those in which the conformation of the individual monomers does not significantly change upon complexation.

Visualization and numerical tools. All the visual representations of the proteins were done using the program VMD (4). The contacts were drawn between the C_α atoms of each amino acid. Secondary structure assignments were based on the DSSP program (5). Pair distribution functions were calculated using Matlab, and the plots generated with Kaleidagraph.

1. Papoian GA, Ulander J, Eastwood M, Luthey-Schultes Z, Wolynes PG (2004) *Proc Natl Acad Sci USA* **101**, 3352–3357.
2. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z (2005) *PROTEINS Struct Funct Bioinformatics* **60**, 214–216.
3. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) *J Mol Biol* **247**, 536–540.
4. Humphrey W, Dalke A, Schulten K (1996) *J Mol Graphics* **14** 33–38.
5. Kabsch W, Sander C (1983) *Biopolymers* **22** 2577–2637.