Supporting Materials and Methods

Go Model

In the model used in this study, a single bead centered on the C_α position represents a residue. Bond and angle potentials string together the beads to their neighbors along the protein chain. The dihedral potential encodes the secondary structure. The protein's native topology defines the network of favorable long-range tertiary interactions while all other nonbonded interactions are repulsive. The energy function for a Go model with configuration Γ is as follows:

$$H(\Gamma,\Gamma_0) = H_{\text{backbone}} + H_{\text{nonbonded}}$$

$$H_{\text{backbone}} = \sum^{\text{bonds}} K_{\text{r}}(r-r_0)^2 + \sum^{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum^{\text{dihedrals}} K_\varphi^{(n)}\left[1-\cos(n(\varphi-\varphi_0))\right]$$

$$H_{\text{nonbonded}} = \sum_{i<j-3}^{\text{native}} \varepsilon_1(i,j)\left[5\left(\frac{\sigma_{i,j}^{\text{nat}}}{\text{r}_{i,j}}\right)^{12} - 6\left(\frac{\sigma_{i,j}^{\text{nat}}}{\text{r}_{i,j}}\right)^{10}\right] + \sum_{i<j-3}^{\text{nonnative}} \varepsilon_2(i,j)\left(\frac{\sigma^{\text{non}}}{\text{r}_{i,j}}\right)^{12}.$$

The K_{r}, K_θ, and K_φ are the force constants of the bonds, angles and dihedral angles, respectively. The r, θ, and φ are the bond lengths, the angles, and the dihedral angles, with a subscript zero representing the corresponding values taken from the native configuration, Γ_0. The nonbonded contact interactions, $H_{\text{nonbonded}}$, contain Lennard-Jones 10-12 terms for the nonlocal native interactions and a short-range steric repulsive term for the nonnative pairs, corresponding to a perfectly funneled energy landscape. We chose as parameters of the energy function $K_{\text{r}} = 100\varepsilon$, $K_\theta = 20\varepsilon$, $K_\varphi^{(1)} = 1.0\varepsilon$, $K_\varphi^{(3)} = 0.5\varepsilon$, and $\varepsilon_1 = \varepsilon_2 = \varepsilon$. $\sigma_{i,j}^{\text{nat}}$ is the distance between the C_α atoms of the residues (i, j) in the native configuration and $\sigma^{\text{non}} = 4.0\,\text{Å}$ for all non-native residue pairs. The network of native contact pairs was determined using the CSU (Contacts of Structural Units)

software (1).

Reaction Coordinates

We now give a detailed description of Q_S and <*L*>. Q_S is a measure of protein folding progress that accounts for how far natively contacting residues are from their respective native distances. Q_S is similar to Q except that a Gaussian penalty is introduced if the native contact is far from its native distance. This measure is more precise than Q because it not only takes into consideration whether the contact exists but also quantifies how close the contact is to the native distance. Q_S is a summation over all native contact pairs between residues *i* and *j* defined as follows:

$$Q_S = \frac{1}{(NC-1)(NC-2)} \sum^{NC} \exp\left[-\frac{(r_{i,j}^{nat} - r_{i,j}^{comp})^2}{\sigma_{i,j}^2} \right],$$

where *NC* is the number of native contacts and $r_{i,j}^{nat}$ and $r_{i,j}^{comp}$ are the distances between residues *i* and *j* in the native and comparison structures, respectively. The well width of the Gaussian is $\sigma_{i,j} = |i-j|^{0.15}$. The possible values of Q_S ranges from 0 (native interactions are far from native distances) to 1 (exactly the same as the native distances)

One can also monitor the progress of protein folding by envisioning the protein's native topology as a network of interactions. In this approach, each residue in the protein chain is presented as a node in a network that is connected by unweighted edges that are defined by the native contacts present in the native protein structure and the backbone connectivity. The shortest path length, L_{ij}, is the minimum number of edges that connect residues *i* and *j*. For a *n*-residue protein structure with shortest path length between residues *i* and *j* given as L_{ij}, the average shortest path length over all of the residues, <*L*>, is:

$$\langle L \rangle = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} L_{ij} \ .$$

With increasing number of native contacts in a given structure, the average number of edges (distance) between residues, <*L*>, continues to decrease until a minimum corresponding to the native structure is reached.

Clustering Analysis

The protein structures belonging to the putative TSE (i.e., $P_{\text{fold}} = 0.50 \pm 0.10$) were clustered using the FITCH program from the PHYLIP package (2). The FITCH program is an algorithm that was originally designed to create phylogenic trees based on a distance measure. In our analysis, we defined the distance between any two structures (designated reference and comparison structures) in the TSE with n residues as $d = 1 - q$ where q is a normalized measure of similarity between a reference and comparison structure:

$$q = \frac{1}{(n-1)(n-2)} \sum_{i<j-1}^{n} \exp\left[- \frac{(r_{i,j}^{\text{ref}} - r_{i,j}^{\text{comp}})^2}{\sigma_{i,j}^2} \right],$$

where $r_{i,j}^{\text{ref}}$ and $r_{i,j}^{\text{comp}}$ are the C_α-C_α distance between non-neighboring pairs of residues i and j in the reference and comparison structures, respectively. The similarity measure, q, which has been used in protein structure prediction for comparison of a reference structure to a predicted structure (3), ranges from 0 (different) to 1 (similar), and, hence, d ranges from 0 (no distance) to 1 (great distance).

1. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E. & Edelman, M. (1999) *Bioinformatics* **15,** 327-332.
2. Felsenstein, J. (1989) *Cladistics* **5,** 164-166.
3. Hardin, C., Eastwood, M. P., Prentiss, M. C., Luthey-Schulten, Z. & Wolynes, P. G. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 1679-1684.