SI Text

Ensemble Diagrams. Each of the ensembles Γ corresponds to a different sequence of locally structured segments α_k. The ensembles $\Gamma_n \in \Gamma$ are generated by labeling the α_k as $\alpha_k \longrightarrow \alpha_k^a$ to indicate when the segments are in contact. The pattern of labels corresponding to any ensemble Γ_n is defined so that (i) two segments are in contact if they have the same label, but (ii) two segments can have the same label only if contacts exist between them in the native fold. Therefore, the labels can be taken to index the nuclei in Γ_n. For example, if Γ has three folded segments, all of which are in contact in the native structure, there are five possible diagrams: The extended diagram Γ_0 (all labels different), the collapsed diagram Γ_1 (all labels identical) and diagrams for the label sequences 112, 121, and 211.

SI Fig. 5a shows the diagram corresponding to label sequence 121, which is also discussed in the text. The lines represent unfolded segments along the protein and the vertices represent the nuclei. The unfolded segments are considered to interact with the rest of the protein by excluded volume. We approximate the entropy cost to close Γ_0 into SI Fig. 5a in terms of the entropy cost to close a loop of length $l = l_1 + x + l_2$, where

$$x \; = \; \left[\frac{r'}{a}\right]^{1+\delta} \tag{1}$$

describes the distance in links bridged by the segment $\alpha_2^2(\Gamma)$ and r' is the space distance between α-carbons at either end of $\alpha_2^2(\Gamma)$. The effective entropy cost is then

$$\Delta s^{\star}_{\text{loop}} \; \simeq \; -k\,\frac{3}{2}\,\Sigma(l,x)\,\left[\ln(l) \; - \; \frac{r^2}{la^2}\right] \; + \; C \; , \tag{2}$$

where $C = 0.3251\,k$ is the entropy of the end residue (the constant term in Eq. 3 of the text), and

$$\Sigma(l,x) \; = \; 1 \; - \; \frac{x+1}{l} \tag{3}$$

(note that $x = 0$ for a single folded residue). The parameter δ modulates the entropy cost that we remove from Δs_{loop} to account for the fact that α_2^2 is confined by the local order parameter threshold.

Loops that are not broken by folded segments α_k^a are everywhere described by $\Delta s_{\text{loop}}(l, r)$. The few complex diagrams that require the approximation above are shown in SI Fig. 5b and c. Nested loops in SI Fig. 5b are described as separate loops, and in SI Fig. 5c (coupled loops) we construct a lower bound to the entropy cost by dividing the diagram into two effective loops. The rightmost diagram in Fig. 5c is adjusted by the factor $\Sigma(l_2 + l_3, y + l_2)$ to account for the entropy of the 'shared' segment l_2. The diagrams in SI Fig. 5b and c contribute minimally to the results.

Diagram Participation The probability distribution for the number of loops in a diagrams is

$$P_n(q) \;=\; \sum_{\Gamma_m}^{q} \delta(n - n_\ell(\Gamma_m))\, e^{-F(\Gamma_m)/kT} \;, \tag{4}$$

where $n_\ell(\Gamma_n)$ is the number of loops in Γ_n and the sum is over diagrams with q folded residues. $P_n(q)$ is shown in SI Fig. 6 [here $P_4(q)$ describes the fraction of diagrams of type SI Fig. 5c]. Multiple loop diagrams are suppressed somewhat when the entropy cost is adjusted to account for excluded volume. SI Fig. 7 compares the free energy profiles neglecting loops (the Muñoz−Eaton approximation), and neglecting the extension term r^2/la^2 with the result in Fig. 2a. The text model is similar to the Muñoz−Eaton model in the denatured ensemble, but discrepencies occur crossing the transition [just after the transition the functions $P_{n>0}(q)$ reach their maximum values]. This feature is typical of the proteins we have studied.