Supporting Information

Reddy et al. 10.1073/pnas.1201808109

SI Text

Methods. ***SOP-Side Chain (SOP-SC) model for GFP:*** The effective energy of a polypeptide chain in the coarse grained SOP-SC model is a sum of bonded (B) and non-bonded (NB) interactions. The non-bonded interactions are a sum of native (N) and non-native (NN) interactions. If two sites that are in contact in the native state are separated by at least two other sites along the sequence, and the distance between them is less than a cutoff distance, R_c (Table S1), in the coarse grained crystal structure then the interaction between them is classified as native. The total effective energy is given by:

$$E_{\mathrm{TOT}} = E_{\mathrm{B}} + E_{\mathrm{NB}}^{\mathrm{N}} + E_{\mathrm{NB}}^{\mathrm{NN}}. \qquad \textbf{[S1]}$$

The Finite Extensible Nonlinear Elastic (FENE) potential describes the bonding potential (E_{B}) between various bonded beads and is given by

$$E_{\mathrm{B}} = -\sum_{i=1}^{N_{\mathrm{B}}} \frac{k}{2} R_o^2 \log\left(1 - \frac{(r_i - r_{\mathrm{cry},i})^2}{R_o^2}\right) \qquad \textbf{[S2]}$$

where $N_{\mathrm{B}}(= 459)$ is the total number of bonds between the backbone-backbone beads and the backbone-side chain beads in the coarse grained model of protein.

The functional form for the non-bonded native interactions, $E_{\mathrm{NB}}^{\mathrm{N}}$, between various beads in Eq. **S1** is,

$$\begin{aligned} E_{\mathrm{NB}}^{\mathrm{N}} = &\sum_{i=1}^{N_{\mathrm{N}}^{\mathrm{bb}}} \epsilon_h^{\mathrm{bb}} \left[\left(\frac{r_{\mathrm{cry},i}}{r_i}\right)^{12} - 2\left(\frac{r_{\mathrm{cry},i}}{r_i}\right)^{6}\right] \\ &+ \sum_{i=1}^{N_{\mathrm{N}}^{\mathrm{bs}}} \epsilon_h^{\mathrm{bs}} \left[\left(\frac{r_{\mathrm{cry},i}}{r_i}\right)^{12} - 2\left(\frac{r_{\mathrm{cry},i}}{r_i}\right)^{6}\right] \\ &+ \sum_{i=1}^{N_{\mathrm{N}}^{\mathrm{ss}}} 0.5(\epsilon_i^{\mathrm{ss}} - 0.7) \left[\left(\frac{r_{\mathrm{cry},i}}{r_i}\right)^{12} - 2\left(\frac{r_{\mathrm{cry},i}}{r_i}\right)^{6}\right] \end{aligned} \qquad \textbf{[S3]}$$

where $N_{\mathrm{N}}^{\mathrm{bb}}(= 698)$, $N_{\mathrm{N}}^{\mathrm{bs}}(= 1814)$, and $N_{\mathrm{N}}^{\mathrm{ss}}(= 824)$ are the numbers of backbone-backbone, backbone-sidechain, sidechain-sidechain native interactions, respectively, r_i is the distance between the ith pair of residues, and $r_{\mathrm{cry},i}$ is the corresponding distance in the crystal structure. The strength of interaction between the pair of side chain beads i, ϵ_i^{ss}, is taken from the Betancourt-Thirumalai statistical potential (1).

The non-native interactions, $E_{\mathrm{NB}}^{\mathrm{NN}}$, in Eq. **S1** between various beads is taken to be

$$E_{\mathrm{NB}}^{\mathrm{NN}} = \sum_{i=1}^{N_{\mathrm{NN}}} \epsilon_l \left(\frac{\sigma_i}{r_i}\right)^6 + \sum_{i=1}^{3N-4} \epsilon_l \left(\frac{\sigma_{i,i+2}}{r_{i,i+2}}\right)^6 \qquad \textbf{[S4]}$$

where $N_{\mathrm{NN}}(= 100{,}860)$ is the total number of non-native interactions, σ_i is the sum of the radii of the ith pair of residues, $\sigma_{i,i+2}$ is the sum of the radii of the interaction sites i and $i+2$, and $r_{i,i+2}$ is the distance between the sites i and $i+2$. The radii for side chains of amino acids are given in Table S2. Various other interaction parameters used in the energy function are given in Table S1.

Multicanonical molecular dynamics simulations: Thermodynamic sampling of the coarse grained model of GFP is obtained using multicanonical molecular dynamics simulations (2, 3). In multicanonical ensemble, each state of the protein is weighted by a non-Boltzmann factor $W_{\mathrm{mu}}(E)$ so that a uniform distribution of energy, $P_{\mathrm{mu}}(E)$ is obtained,

$$P_{\mathrm{mu}}(E) \propto \Omega(E) W_{\mathrm{mu}}(E) \equiv \text{constant} \qquad \textbf{[S5]}$$

where $\Omega(E)$ is the density of states. We rearrange Eq. **S5** as

$$W_{\mathrm{mu}}(E) \equiv e^{-\beta_o E_{\mathrm{mu}}(E;T_o)} = 1/\Omega(E). \qquad \textbf{[S6]}$$

$\beta_o = 1/k_{\mathrm{B}}T_o$, and T_o is an arbitrary reference temperature, k_{B} is the Boltzmann constant, and the multicanonical potential, $E_{\mathrm{mu}}(E;T_o)$, is defined as

$$E_{\mathrm{mu}}(E;T_o) = k_{\mathrm{B}}T_o \ln(\Omega(E)) = T_o S(E) \qquad \textbf{[S7]}$$

with $S(E)$ being the entropy in the microscopic ensemble. Molecular dynamics in the multicanonical ensemble follows from Eq. **S7** and is performed by solving the modified Newtons equation:

$$\begin{aligned} \dot{\vec{P}}_k &= -\frac{\partial E_{\mathrm{mu}}(E;T_o)}{\partial \vec{r_k}} - \zeta\frac{\vec{P_k}}{m_k} + \Gamma = \frac{\partial E_{\mathrm{mu}}(E;T_o)}{\partial E}\vec{f_k} - \zeta\frac{\vec{P_k}}{m_k} + \Gamma \\ &= \frac{T_o}{T(E)}\vec{f_k} - \zeta\frac{\vec{P_k}}{m_k} + \Gamma \end{aligned} \qquad \textbf{[S8]}$$

where $\vec{r_k}$, $\vec{P_k}$, $\vec{f_k}$ are the position, momentum and deterministic force respectively for the quasi-particle k, ζ is the friction coefficient, Γ is the random force with a white noise spectrum and autocorrelation function, $\langle\Gamma(t)\Gamma(t')\rangle = 2\zeta k_{\mathrm{B}}T\delta(t-t')$, and $\delta(t)$ is the Dirac delta function. The effective temperature, $T(E)$, is defined as $\frac{1}{T(E_a)} = \frac{\partial S(E)}{\partial E}|_{E=E_a}$. The equation of motion, Eq. **S8**, is integrated using Verlet algorithm (4–6) with a friction coefficient, $\zeta = 0.05\ \mathrm{m}\tau^{-1}$. The canonical probability distribution for a wide range of temperatures, T, is obtained by the reweighting technique:

$$P_B(\beta, E) \propto P_{\mathrm{mu}}(E) W_{\mathrm{mu}}^{-1}(E) e^{-\beta E}. \qquad \textbf{[S9]}$$

Because $\Omega(E)$ is unknown a priori it has to be determined numerically. The method prescribed (7) by Okamoto and Hansmann is used to iterate for $\Omega(E)$ and the procedure is described below:

1. In practice it is possible to satisfy Eq. **S5** only for an interval of E, and the interval we choose in our simulation is $E_{\mathrm{min}}(= -596$ kcal/mole$) \le E \le E_{\mathrm{max}}(= 246$ kcal/mole$)$ for the temperature range $T_{\mathrm{min}}(= 295\ \mathrm{K}) \le T \le T_{\mathrm{max}}(= 445\ \mathrm{K})$. Here $\langle E\rangle_{T=T_{\mathrm{min}}} = E_{\mathrm{min}}$ and $\langle E\rangle_{T=T_{\mathrm{max}}} = E_{\mathrm{max}}$ are determined from a Langevin dynamics simulation.
2. Initial guess for $\ln(\Omega(E))^{(0)}$ with the bin width $\delta E = 1$ kcal/mole is obtained from the Langevin dynamics simulations and weighted histogram method (8). A canonical molecular dynamics simulation at $T_o = 1200$ K is performed by integrating the equations of motion using the energy function given in Eq. **S7** as described by Eq. **S8**. The numerical derivative, $\frac{\partial E_{\mathrm{mu}}(E;T_o)}{\partial E}$, required to integrate the equation of motion is obtained by fitting $\ln(\Omega^{(0)}(E))$ in the range $(E-50)$kcal/mole $\le E \le (E+50)$kcal/mole to a 10-degree polynomial. If $E > E\,\mathrm{max}$, then $T(E) = T\,\mathrm{max}$. If $E < E_{\mathrm{min}}$,

then $T(E) = T_{\min}$ is used. The energy distribution obtained from the simulation, $H(E)$, is constructed with the bin width $\delta E = 1$ kcal/mole, and is checked for reasonable flatness. Simulation iterations are performed until a reasonably flat $H(E)$ is obtained.

3. If $H(E)$ obtained from the (i)th simulation iteration is not flat, then $(i+1)$th simulation is performed with a modified $\ln(\Omega(E))^{(i+1)}$ obtained using $\ln(\Omega(E))^{(i+1)} = \ln(\Omega(E))^{(i)} + \ln(H(E))^{(i)}$. We performed iterations until we obtained a $\Omega(E)$, which forces the protein to sample the complete energy space of interest ($E_{\min} \leq E \leq E_{\max}$). In other words, an unfolded protein folds into a native state and unfolds back into a denatured state.
4. From a long multicanonical simulation trajectory where we observe multiple protein folding-unfolding events, the average value of a physical quantity A at any temperature, $T(\beta = 1/k_B T)$, in the canonical ensemble is calculated using

$$\langle A \rangle_T = \frac{\sum_E A(E)H(E)e^{\beta_o E_{mu}(E,T_o)-\beta E}}{\sum_E H(E)e^{\beta_o E_{mu}(E,T_o)-\beta E}} = \frac{\sum_{i=1}^{N_{conf}} A(E_i)e^{[\beta_o E_{mu}(E_i,T_o)-\beta E_i]}}{\sum_{i=1}^{N_{conf}} e^{[\beta_o E_{mu}(E_i,T_o)-\beta E_i]}}. \quad \textbf{[S10]}$$

where N_{conf} is the total number of protein conformations in the trajectory.

5. From N_{sim} independent MC simulation trajectories, we calculated the average value of a physical quantity A at any temperature, T, in the canonical ensemble using the equation:

$$\langle A \rangle_T = \frac{\sum_{j=1}^{N_{sim}} \sum_{i=1}^{N_{conf}^j} A(E_{ij})e^{[\beta_o E_{mu}(E_{ij},T_o)-\beta E_{ij}]}}{\sum_{j=1}^{N_{sim}} \sum_{i=1}^{N_{conf}^j} e^{[\beta_o E_{mu}(E_{ij},T_o)-\beta E_{ij}]}}. \quad \textbf{[S11]}$$

where N_{conf}^j is the total number of protein conformations in the jth trajectory.

Molecular transfer model (MTM): The energy of a protein conformation in a denaturant solution of concentration $[C]$ is assumed to be the sum of the potential energy, E, and a transfer free energy $\Delta G_{tr}([C])$ estimated using the TM model (9, 10). According to TM, the free energy of transferring a protein from water to the denaturant solution of concentration $[C]$ is equal to the sum of the transfer free energies of the backbone and sidechain beads, and is written as

$$\Delta G_{tr}([C]) = \sum_{i=1}^{2N} \delta g_{tr,i}([C])(\alpha_i/\alpha_{Gly-i-Gly}). \quad \textbf{[S12]}$$

where $\delta g_{tr,i}([C])$ is the transfer free energy of the bead i, α_i is the solvent accessible surface area (SASA) of the bead i, $\alpha_{Gly-i-Gly}$ is the solvent accessible surface area of the bead in the tripeptide $Gly - i - Gly$. The transfer energies $\delta g_{tr,i}([C])$ for backbone and sidechain chain beads are listed in Table. S3 in ref. 11. The values for $\alpha_{Gly-i-Gly}$ are listed in Table. S4 in ref. 11.

The thermodynamic properties of the protein at a nonzero denaturant concentration $[C]$ are obtained (11, 12) by reweighting the probability of the protein conformations obtained at $[C] = 0$ using the transfer energies $\Delta G_{tr}([C])$. The average value of a physical quantity A at any temperature, T and non-zero denaturant concentration $[C]$ is calculated using.

$$\langle A([C]) \rangle_T = \frac{\sum_{j=1}^{N_{sim}} \sum_{i=1}^{N_{conf}^j} A(E_{ij})e^{[\beta_o E_{mu}(E_{ij},T_o)-\beta(E_{ij}+\Delta G_{tr}^{ij}([C]))]}}{\sum_{j=1}^{N_{sim}} \sum_{i=1}^{N_{conf}^j} e^{[\beta_o E_{mu}(E_{ij},T_o)-\beta(E_{ij}+\Delta G_{tr}^{ij}([C]))]}}. \quad \textbf{[S13]}$$

where E_{ij} is the potential energy of the conformation i from the jth simulation trajectory, $A(E_{ij})$ is the physical quantity of the protein in conformation i from the jth simulation trajectory, $\Delta G_{tr}^{ij}([C]))$ is the free energy cost of transferring protein conformation i from the jth simulation trajectory from water to a denaturant solution of concentration $[C]$.

1. Betancourt M, Thirumalai D (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* 8:361–369.
2. Hansmann U, Okamoto Y, Eisenmenger F (1996) Molecular Dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble. *Chem Phys Lett* 259:321–330.
3. Mitsutake A, Sugita Y, Okamoto Y (2001) Generalized ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* 60:96–123.
4. Veitshans T, Klimov D, Thirumalai D (1996) Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding Des* 2:1–22.
5. Swope W, Andersen H, Berens P, Wilson K (1982) A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small clusters. *J Chem Phys* 76:637–649.
6. Verlet L (1967) Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys Rev* 159:98–103.
7. Okamoto Y, Hansmann U (1995) Thermodynamics of helix-coil transitions studied by multicanonical algorithms. *J Phys Chem* 99:11276–11287.
8. Kumar S, Rosenberg J, Bouzida D, Swendsen R, Kollman P (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. 1. the method. *J Comput Chem* 13:1011–1021.
9. Tanford C (1964) Isothermal unfolding of globular proteins in aqueous urea solutions. *J Am Chem Soc* 86:2050–2059.
10. Auton M, Holthauzen L, Bolen D (2007) Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc Natl Acad Sci USA* 104:15317–15322.
11. Liu Z, Reddy G, O'Brien EP, Thirumalai D (2011) Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proc Natl Acad Sci USA* 108:7787–7792.
12. O'Brien E, Ziv G, Haran G, Brooks B, Thirumalai D (2008) Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc Natl Acad Sci USA* 105:13403–13408.

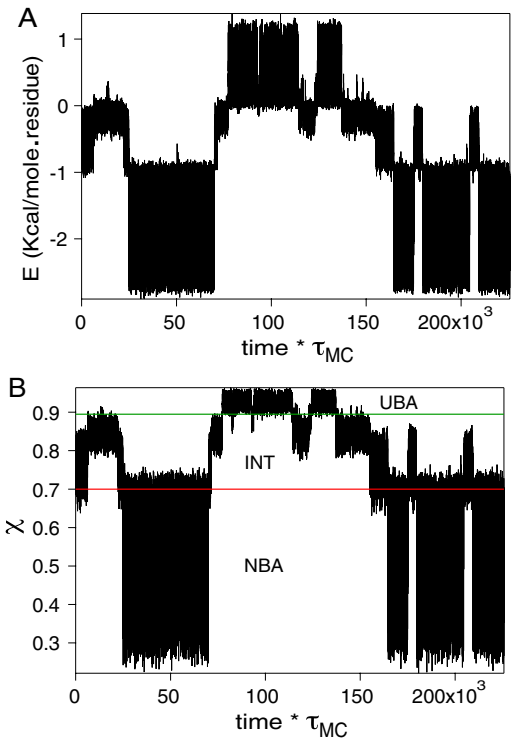


Fig. S1. (*A*) A typical Multicanonical simulation trajectory demonstrating that GFP reversibly folds, thus effectively sampling the relevant energy space. (*B*) A plot of the structural overlap function, χ, (defined in the main text) as a function of t for a Multicanonical simulation trajectory. The red and green lines are drawn at $\chi = 0.7$ and 0.895 respectively. The conformations with $0 < \chi \leq 0.7$ are folded, an intermediate state has $0.7 < \chi \leq 0.895$ and conformations with $\chi > 0.895$ are unfolded.

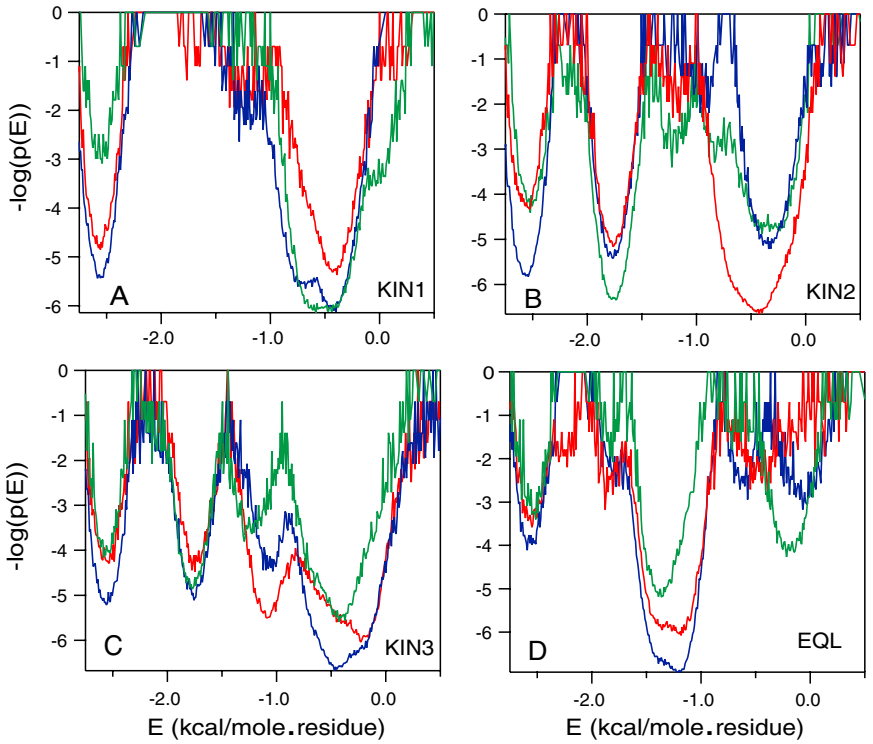


Fig. S2. Criteria used to identify the four distinct folding pathways for GFP. For each trajectory, logarithm of energy distribution, $\log(P(E))$ is plotted as a function of E. The trajectories fall into four distinct pathways as shown in panels *A*, *B*, *C* and *D*. For each pathway three different trajectories are shown in red, green and blue respectively. The energy in trajectories in A hop between two values one corresponding to the folded state and the other to the unfolded state. When plotted as $\log P(E)$ we see that there are two distinct basins. In contrast, multiple basins are found in *B*, *C*, and *D*.

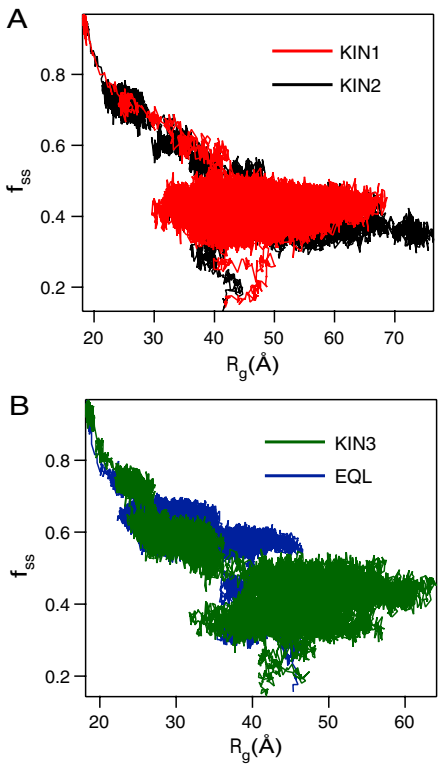
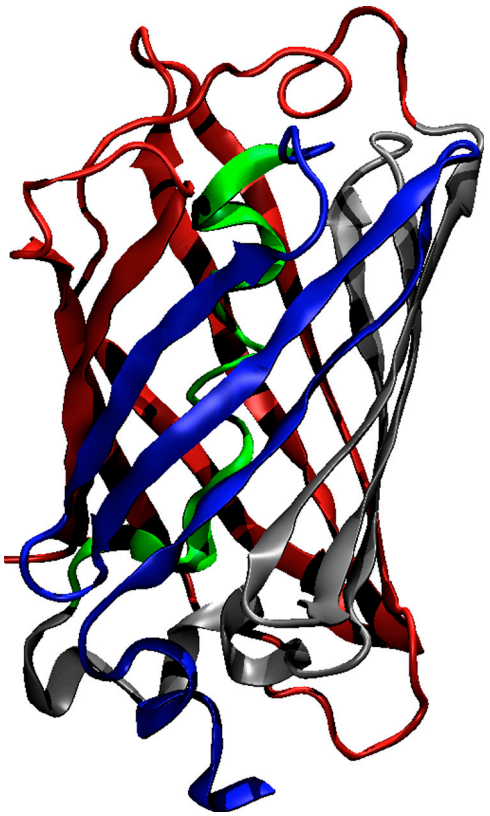


Fig. S3. (*A*) Plots of f_{ss} as a function of R_g sampled during the process of folding in pathways 1 and 2. (*B*) Same as (*A*) except these plots are for KIN3 and EQL pathways.



Movie S1. Folding trajectory of KIN1 pathway. β -sheets with local contacts shown in blue, silver and red form first on decreasing the temperature. These substructures coalesce to form the folded structure.
Movie S1 (MPG)

Table S1. Parameters for the SOP-Side Chain model used in Eqs. S1–S4

Parameters	Protein
R_o	2.0
k	20 kcal/(mol.Å^2)
R_c	8 Å
ϵ_h^{bb}	0.45 kcal/mole
ϵ_h^{bs}	0.45 kcal/mole
ϵ_l	1.0 kcal/mole
σ^{bb}	3.8 Å

Table S2. Sidechain and backbone radii of amino acids based on partial molar volumes

Residue	Radius (Å)
Gly	0
Ala	2.52
Val	2.93
Leu	3.09
Ile	3.09
Met	3.09
Phe	3.18
Pro	2.78
Ser	2.59
Thr	2.81
Asn	2.84
Gln	3.01
Tyr	3.23
Trp	3.39
Asp	2.79
Glu	2.96
*Hse**	3.04
Hsd	3.04
Lys	3.18
Arg	3.28
Cys	2.74
backbone	2.25

**Hse*-Neutral histidine, proton on NE2 atom.
Hsd-Neutral histidine, proton on ND1 atom.