Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

Toward a Quantitative Description of Microscopic Pathway Heterogeneity in Protein Folding

Soundhararajan Gopi,[1] Animesh Singh,[2] Swaathiratna Suresh, [3] Suvadip Paul,[2] Sayan Ranu[2] & Athi N. Naganathan*[1]*

[1]Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India.

[2]Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600036, India.

[3]Center for Biotechnology, Anna University, Chennai 600025, India.

e-mail: athi@iitm.ac.in, sayan@cse.iitm.ac.in

Wako-Saitô-Muñoz-Eaton (WSME) Model The WSME model is an Ising-like statistical-mechanical model,[1, 2] where a residue is assumed to sample only two conformational states: folded (residue with the native ϕ and ψ angles, as defined in the PDB file; denoted by *1*) and unfolded (residue in the non-native dihedral space; denoted by *0*). The binary treatment of the conformational status of residues results in an ensemble of 2^N microstates for a N-residue protein. The conformational heterogeneity in the non-native state of a residue (as $\Omega_U/\Omega_F \gg 1$, where Ω is the multiplicity) is accounted for by the fundamental parameter, the entropic cost ΔS_{conf}. The total partition function (Z) for the ensemble of 2^N microstates can be calculated as described before.[1,3,4] However, enumerating the free energies of all possible microstates is not computationally tractable and defeats a major goal of the study – to make the approach scalable.

Here, we resort to model approximations, where the number of nucleation sites is restricted to one (single sequence approximation; SSA) and two (double sequence approximation DSA).[2,4] In DSA, the two islands of folded residues are not allowed to interact. In the DSAw/L approximation, interaction between the any two native islands is allowed provided they interact in the native state.[4,5] The total partition function (Z) can be obtained by accumulating the statistical weight of microstates ($w_i = \exp(-\Delta F_i / RT)$). The free energy of a microstate (m, n), (i.e., microstate with structured residues between and including m and n) is defined as,

$$\Delta F = \Delta G_{m,n}^{stab} - T\sum_{m}^{n} \Delta S_{conf}$$

The effective stabilization free energy contribution is the sum of van der Waals interactions (E_{vdW}), electrostatic potential (E_{elec}) and solvation free energy (ΔG_{solv})[6]:

$$\Delta G_{m,n}^{stab} = E_{vdW} + E_{elec} + \Delta G_{solv}$$

The effective strength of van der Waals (*vdW*) interactions between heavy atom pairs is estimated by assigning a distance cut-off (r_{cut}) and hence,

$$E_{vdW} = \sum_{m,n} \xi_{i,j} \rho$$

where $\rho = 1$ if $r_{ij} \leq r_{cut}$ and $\rho = 0$ otherwise, and ξ is the interaction energy per native contact.

A simplified Debye-Hückel (DH) formalism is employed to quantify the electrostatic interaction energy between charged residues[6,7]:

$$E_{Elec} = \sum_{m,n} K_{Coulomb} \frac{q_i q_j}{\varepsilon_{eff} r_{ij}} \exp(-r_{ij}\kappa)$$

where K_{Coulomb} is the Coulomb constant (1389 kJ.Å.mol^{-1}), q_i and q_j are the charges on i^{th} and j^{th} atoms and r_{ij} is the distance between them, ε_{eff} is the effective dielectric constant and $1/\kappa$ is the Debye screening length, which is a function of ε_{eff}, solvent ionic-strength (I) and temperature (T).

The solvation free energy scales with the number of formed contacts with ΔC_p^{cont} being the proportionality constant[6,8]:

$$\Delta G_{solv} = x_{cont}^{m,n} \Delta C_p^{cont} [(T - T_{ref}) - T \ln(T/T_{ref})]$$

where, $x_{cont}^{m,n}$ is the number of native contacts within the microstate, ΔC_p^{cont} is temperature-independent heat capacity change per native contact and at a reference temperature (T_{ref}) of 385 K.[9,10]

Estimating experimental observables The excess heat capacity is obtained from the total partition function (Z) as,

$$C_p^{ex} = 2RT\left(\frac{\partial \ln Z}{\partial T}\right) + RT^2\left(\frac{\partial^2 \ln Z}{\partial T^2}\right)$$

The overall probability of a particular residue to be folded is calculated as,

$$\chi_i = Z^{-1}\sum\nolimits_k \exp(-\Delta F_k / RT)$$

where, ΔF_k is the free energy associated with the microstate k at temperature T, R is the gas constant and k runs over all the microstates in which the residue i is structured. The unfolding curves of the protein segments are obtained by averaging the individual residue folding probabilities of the constituent residues ($\left\langle \chi_i \right\rangle_T$).

Markov Clustering (MCL) Clusters in a network are characterized by the presence of many edges between nodes within that cluster, and fewer edges between nodes across clusters. Given these typical structures of clusters in a network, consider simulating a random walk starting from arbitrary node in the network. Since the probability of transitioning to a node within the same cluster as that of the staring node is far higher, it is very likely that the random walk would be visiting nodes within the cluster than those outside. Thus, in initial iterations of the random walk, the probability mass accumulates among nodes belonging to the cluster of the starting node. However, after large number of iterations, this effect disappears as the random walk eventually escapes outside the cluster and the flow distributes among other nodes.

MCL utilizes the correspondence explained above between random walks and its ability to identify clusters. First, the network is encoded in the form of a transition matrix M. In this matrix cell M_{ij} represents the number of times a transition has

happened from state *i* to state *j* in the transition paths of our simulated folding trajectories. Next, we add a self-loop to each node with a weight 1, i.e. $\forall i, M_{i,j} = 1$. Self-loops make the Markov chain on *M* aperiodic. Finally, we turn *M* into a row-stochastic transition matrix by dividing each value in row *i* with $\sum_{j=1}^{n} M_{ij}$, where *n* is the number of states in the network.

The MCL algorithm simulates random walks within a graph by alternation of two operators called *expansion* and *inflation*. Expansion involves taking the power of the transition matrix *M* using the normal matrix product (i.e. matrix squaring); more simply, M^t is the flow after *t* iterations of the random walk. *t* is called the expansion parameter in MCL. To avoid escaping outside the cluster MCL performs inflation, which essentially stops the random walk after *t* iterations and boosts the strong transitions in M^t further and demoting the weak transitions. Mathematically, this is achieved by taking the Hadamard power of M^t (taking powers entry-wise) with inflation factor *r*, followed by a scaling step, such that the resulting matrix is row-stochastic again, i.e. the matrix elements (on each row) correspond to probability values. More specifically,

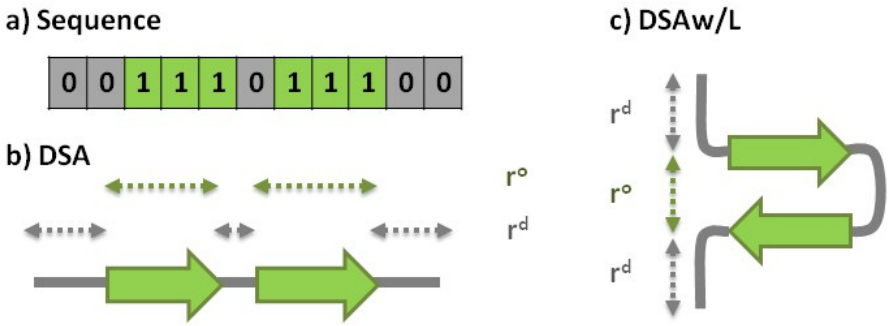
$$inflation(M_{ij}^t, r) = (M_{ij}^t)^r \Big/ \sum_{j=1}^{n} (M_{ij}^t)^r .$$

Here M_{ij}^t represents the flow from state *i* to *j* after *t* iterations. The inflation parameter *r* controls the extent to which strong transitions are strengthened and weak transitions are weakened thus influencing the granularity of clusters.

MCL alternates between expansion and inflation repeatedly till convergence of the transition matrix. The expansion operator is responsible for the random walk to reach different regions of the graph and the inflation operator limits flow across clusters. Convergence happens in the form of a doubly idempotent matrix, i.e. a matrix that does not change with further expansion or inflation steps. The graph associated with such a matrix consists of different connected and directed components. Each component is interpreted as a cluster, and has a star-like form, with few attractor nodes in the center and edges going from all nodes of that component to the attractor. Mathematically, attractor nodes in the converged matrix have at least one positive

transition flow in their corresponding column. Each attractor attracts all nodes with positive values in its column. Clusters are identified by grouping attractor nodes and all nodes with positive values in their columns. Multiple attractors in a cluster are possible, but they do not change the interpretation of the cluster.

Calculation of End-to-End Distances For every microstate (node) in a cluster, the ordered and disordered regions are identified based on the sequence of 1*s* and 0*s* and the node type (SSA, DSA or DSAw/L). For example, in a node of type DSA with sequence 00111011100 (see cartoon *a* below), the region spanning residues 3 to 5 and residues 7 to 9 constitute ordered regions (cartoon *b*) whereas for the same sequence of type DSAw/L the residues spanning 3 to 9 constitute ordered region as the two structured islands interact with each other (cartoon *c*).



The end-to-end distance of the ordered region (r^o) comprising of residues from position i to j is calculated from the PDB coordinates (i.e. the distance between the amide nitrogen of i^{th} residue and the amide nitrogen of the $(j+1)^{th}$ residue), whereas the end-to end distance of the disordered regions (r^d) are approximated to a linearly stretched polymer as described below,

$$r_{i,j}^{d} = \sum_{k=i}^{j} l_k$$

where the left hand side corresponds to the distance between residues i and j, l_k is the length of the k^{th} residue and is calculated as the distance between amide nitrogen of k^{th} residue and the amide nitrogen of $(k+1)^{th}$ residue. The end-to-end distance (r) of the microstate is calculated as

$$r = \sum_i r^o + \sum_j r^d$$

Here, i and j run over all the ordered and disordered regions, respectively, of the microstate of interest. The effective r of the cluster k ($\langle r \rangle_k$) is then obtained from

$$\langle r \rangle_k = \sum p_i r_i$$

where i runs over all the microstates constituting the cluster k, p_i and r_i are the probability and end-to-end distance of the microstate i, respectively. The error associated with the end-to-end distance of a cluster k is calculated as

$$\sigma(r_k) = \sqrt{\sum p_i \left(r_i - \langle r \rangle_k\right)^2}$$

The end-to-end distance based on the freely-jointed chain (FJC) model is obtained following the similar methodology except for the calculation of r^d. In FJC model, the end-to-end distance of disordered region is approximated as

$$\left\langle r_{i,j}^d \right\rangle = \sqrt{2 l_p b \,|\, i - j \,|}$$

where the left hand side corresponds to the mean distance between residues i and j, l_p is the persistence length (fixed to 4 Å) and b is the bond-length (fixed to 3.8 Å).

Visualization of Folding Networks at the Level of Microstates All nodes from the original folding trajectories are partitioned into 10 equal sized bins based on the number of structured residues (n). The nodes in each bin are sorted in descending order based on their frequency in 10000 trajectories, where the frequency is defined as the number of trajectories in which the node occurs. The top 100 nodes are selected from each bin as representative nodes to construct the network. The representative set of 1000 nodes from all 10 bins accounts for about 25-50% of the original nodes comprising the protein-folding network. The construction and visualization of the network are performed using Gephi stand-alone package. The edge weights between two nodes in that direction can be written as,

$$EdgeWeight(n_1 \rightarrow n_2) = \log_{10}(TransitionFrequency \in n_1 \rightarrow n_2 direction)$$

The edges with weight less than ~2 (0-4 depending on the protein) are removed for the clarity of visualization. We use a built-in force minimization algorithm (ForceAtlas2) to reorient the nodes. The algorithm attracts nodes with force (F_a)

proportional to the edge weight, repulses the nodes with force (F_r) proportional to the product of the degrees (degree is a measure of the total number of edges from the given node to any other node) and prevents the isolated nodes (*i.e.* nodes with very low edge weights) from drifting away with force (F_g) proportional to the degree.

$$F_a(n_1, n_2) = d(n_1, n_2)$$

$$F_r(n_1, n_2) = k_r(\deg(n_1)+1) \times (\deg(n_2)+1) \div d(n_1, n_2)$$

$$F_g(n) = k_g \times (\deg(n)+1)$$

Here, $d(n_1, n_2)$ is the distance between two nodes, which is equal to the edge weight. k_r and k_g are the proportionality constants. All other parameters are set to their default values except the edge weights that are scaled 1.3 times the original value to emphasize the weight effects. The proportionality constant for F_g was fixed to 0.08 and the network was minimized allowing overlap of nodes.

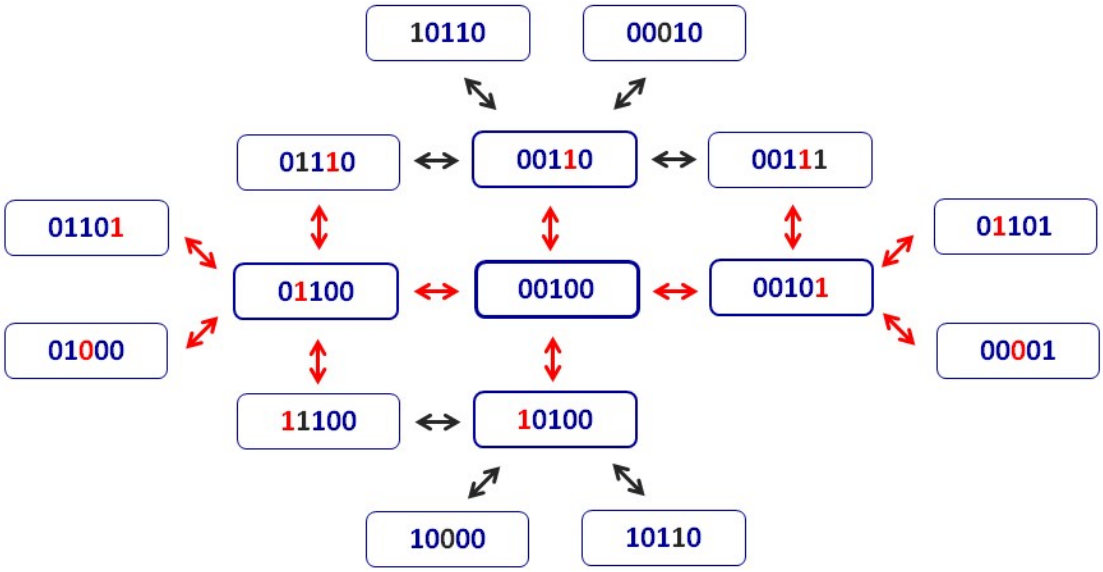


Figure S1. The interconnectivity of microstates exemplified here with a 5-residue sequence in the variant of the WSME model that allows for up to two structured islands (two stretches of *1*s). The color of the arrows indicates the residue that is flipped.

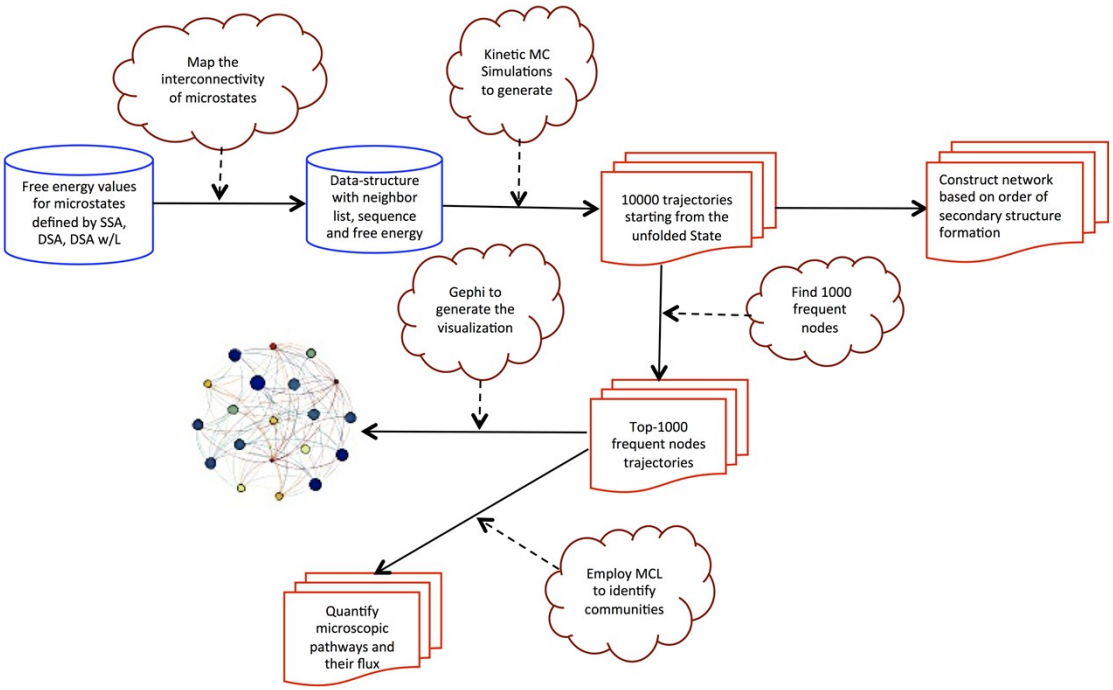


Figure S2. Flowchart highlighting the series of steps involved in the construction of the folding network from secondary-structure based approach or from Markov clustering (MCL).

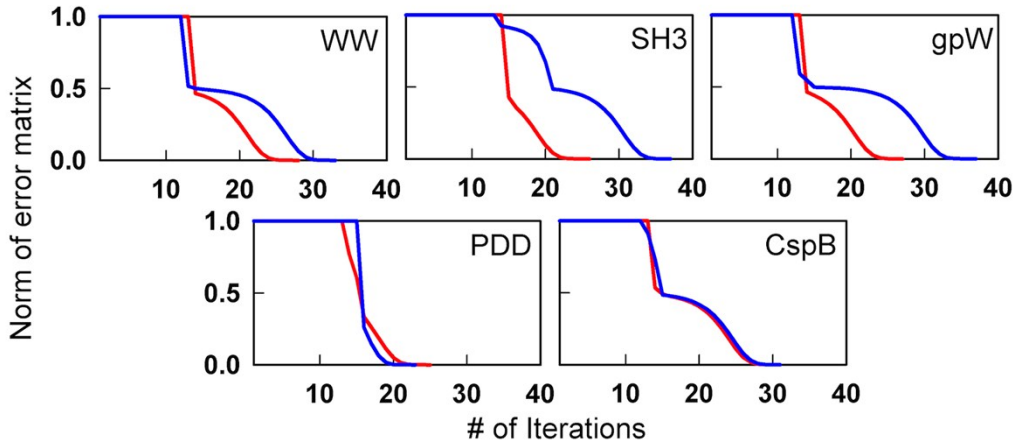


Figure S3. Plot of the norm of the error matrix as a function of the number of iterations in the Markov clustering procedure at 298 K (blue) and T_m (red). Specifically, let M* be the final converged transition matrix and let M_t be the transition matrix after *t* iterations. The ordinate plots the max-norm of the error matrix $E=M^*-M_t$ as the number of iterations, *t*, is varied. The max-norm of the error matrix E is the largest absolute value of all elements within E. As expected, the convergence is fast. Theoretically, the norm is expected to decrease quadratically.

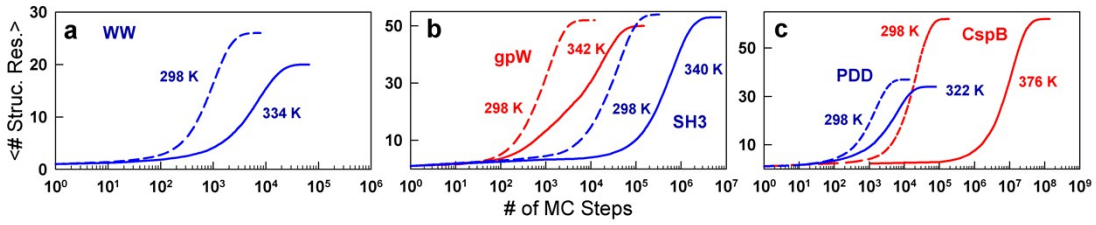


Figure S4. Calculated kinetic relaxation profiles for the 5 proteins at the indicated conditions. Note that gpW exhibits bi-exponential folding relaxation at 342 K (continuous red in panel b).

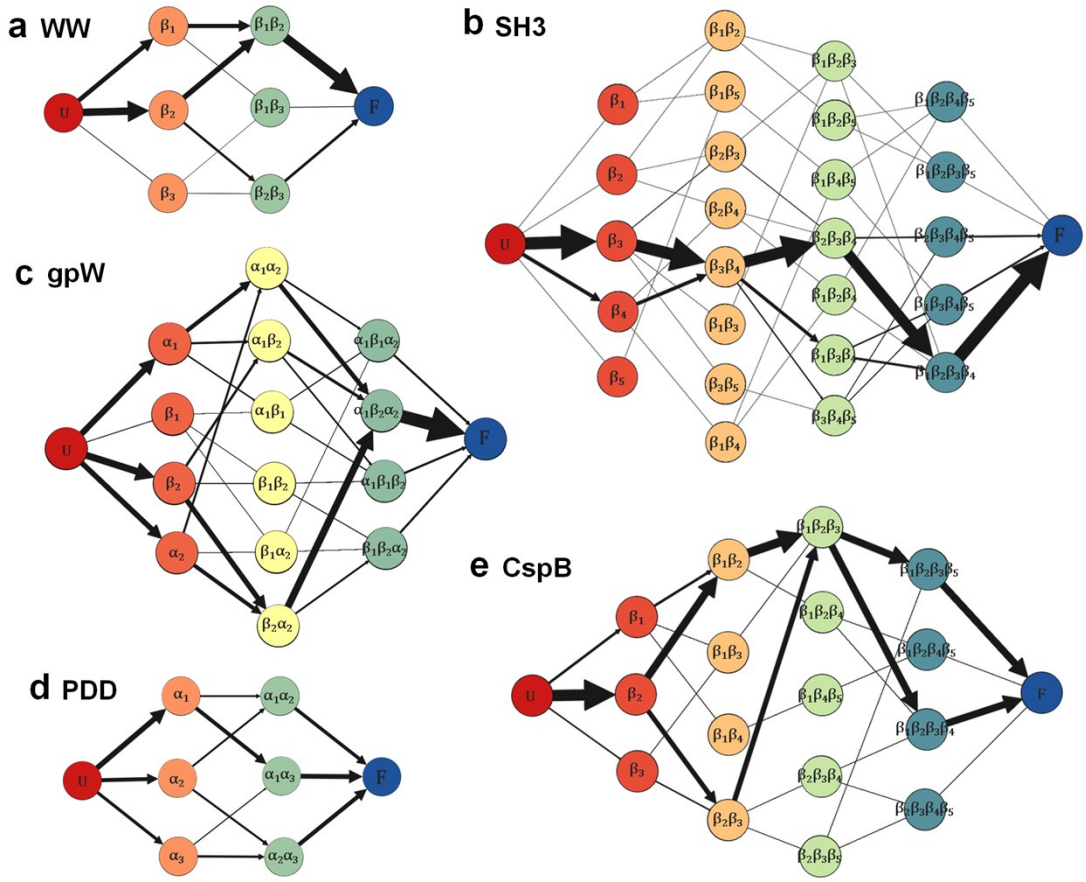


Figure S5. The secondary-structured based folding network at the midpoint conditions for WW domain (panel a), SH3 (panel b), gpW (panel c), PDD (panel d) and TmCspB (panel e). Thicker arrows represent higher flux paths.

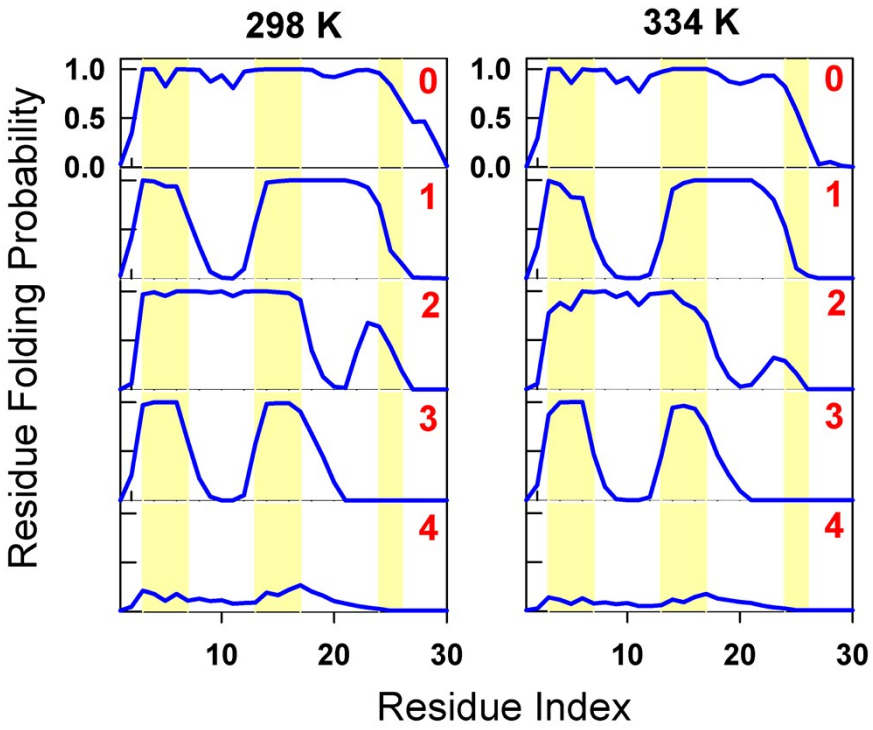


Figure S6. Structural features of clusters identified by the Markov clustering method (MCL) for the WW domain. The clusters are numbered from 0 to 4 with 0 corresponding to the folded state (sink) and 4 representing the fully unfolded state (source). The shaded regions highlight the secondary structure elements (beta strands).

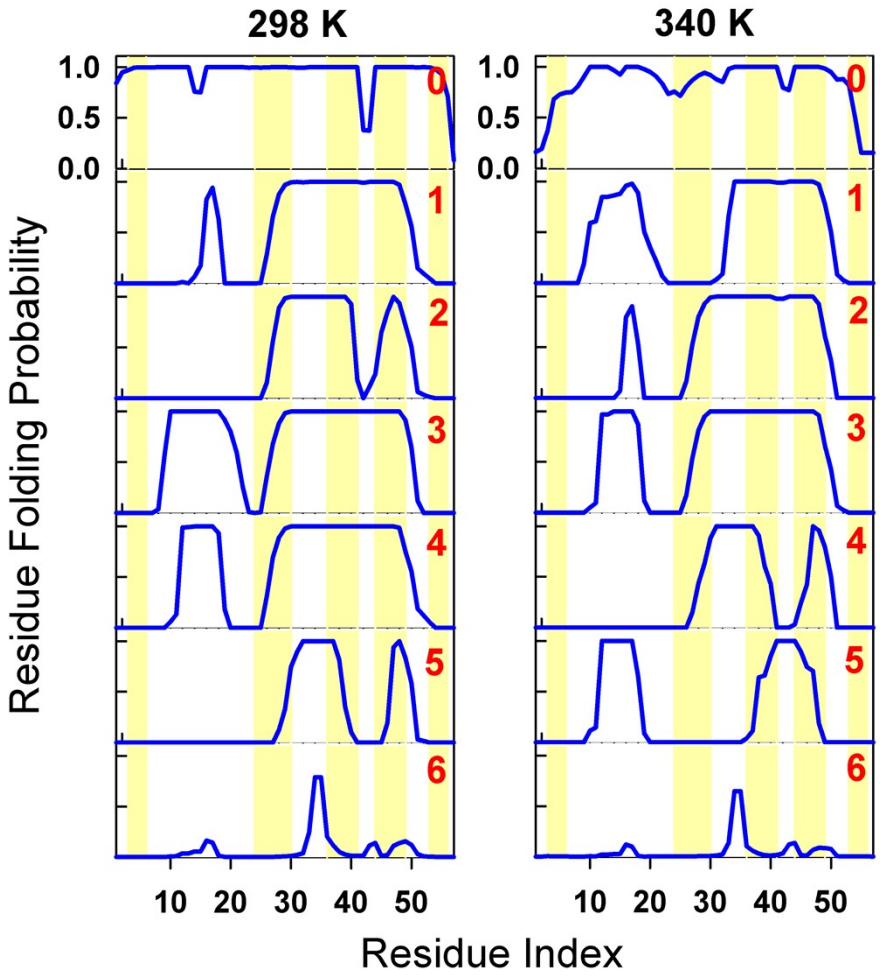


Figure S7. Same as Figure S6, but for SH3.

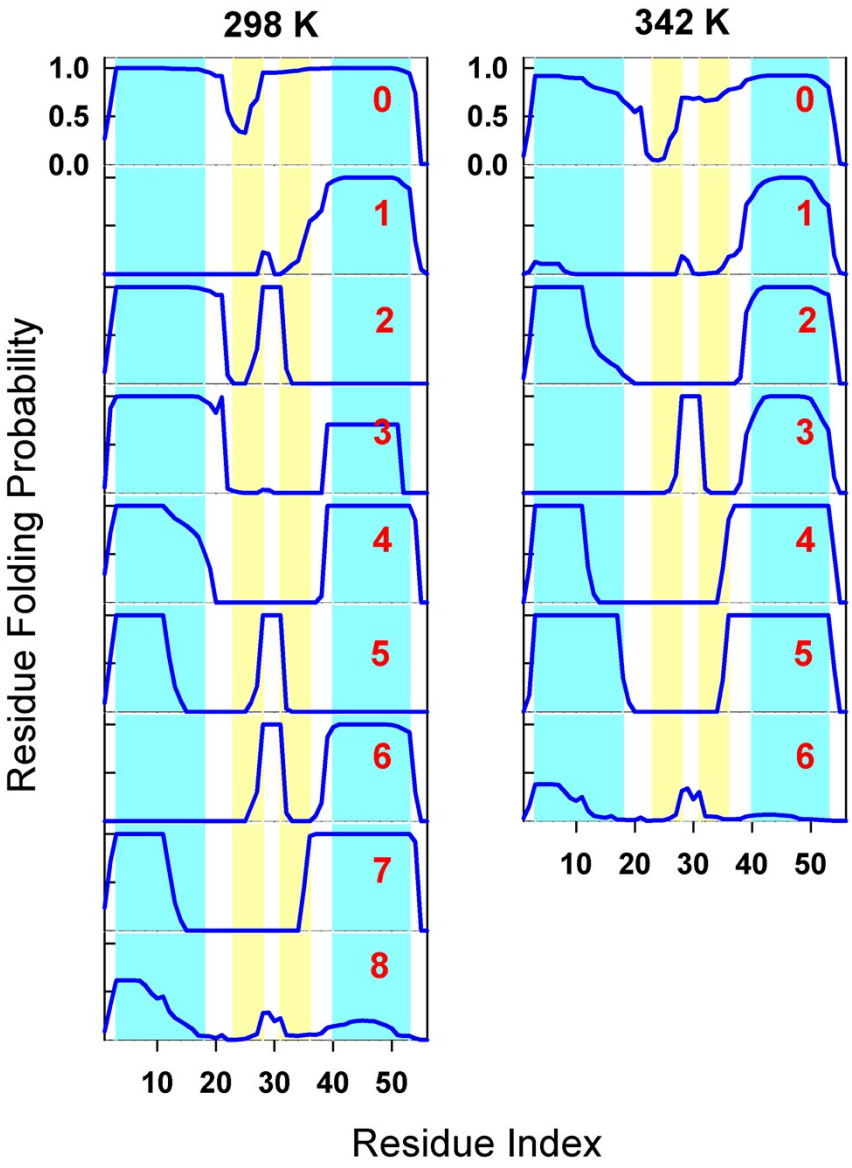


Figure S8. Same as Figure S6, but for gpW. The cyan shades represent alpha-helices.

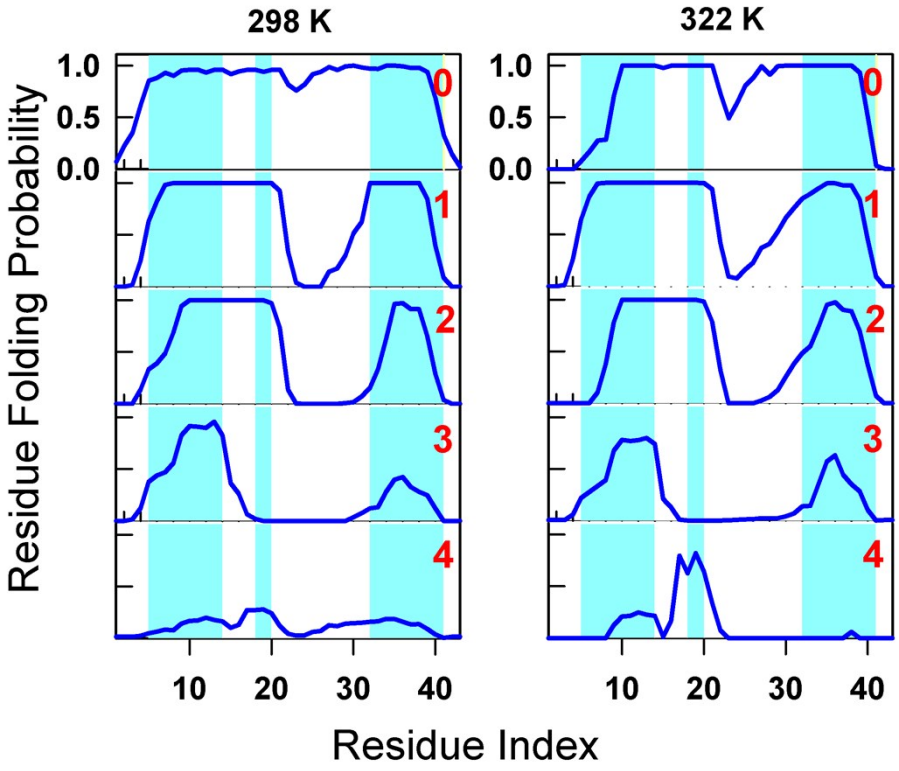


Figure S9. Same as Figure S8, but for PDD. Note that the 'folded' state (cluster 0) is more unfolded at 322 K compared to 298 K indicative of fraying of the helices. This is also evident in gpW and WW, but not obvious in SH3 or CspB.

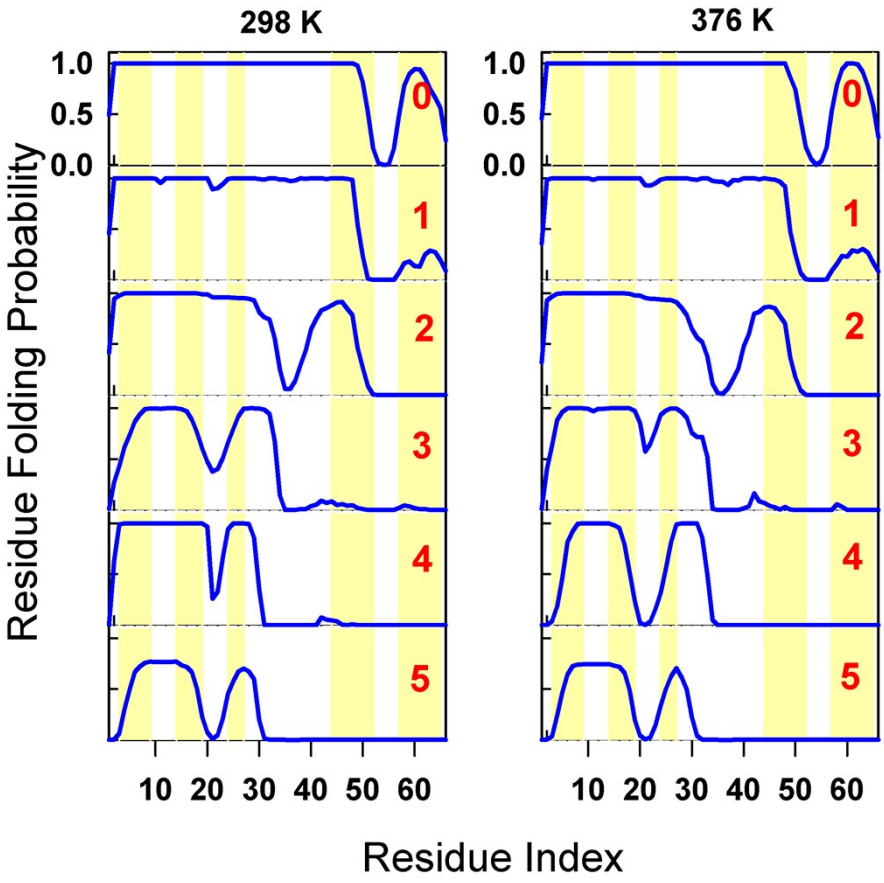


Figure S10. Same as Figure S6, but for CspB from *Thermotoga maritima*. Note that the fully unfolded state (*i.e.* a residue folding probability of ~0 across all residues) is never populated in our method and hence only 6 clusters are predicted.

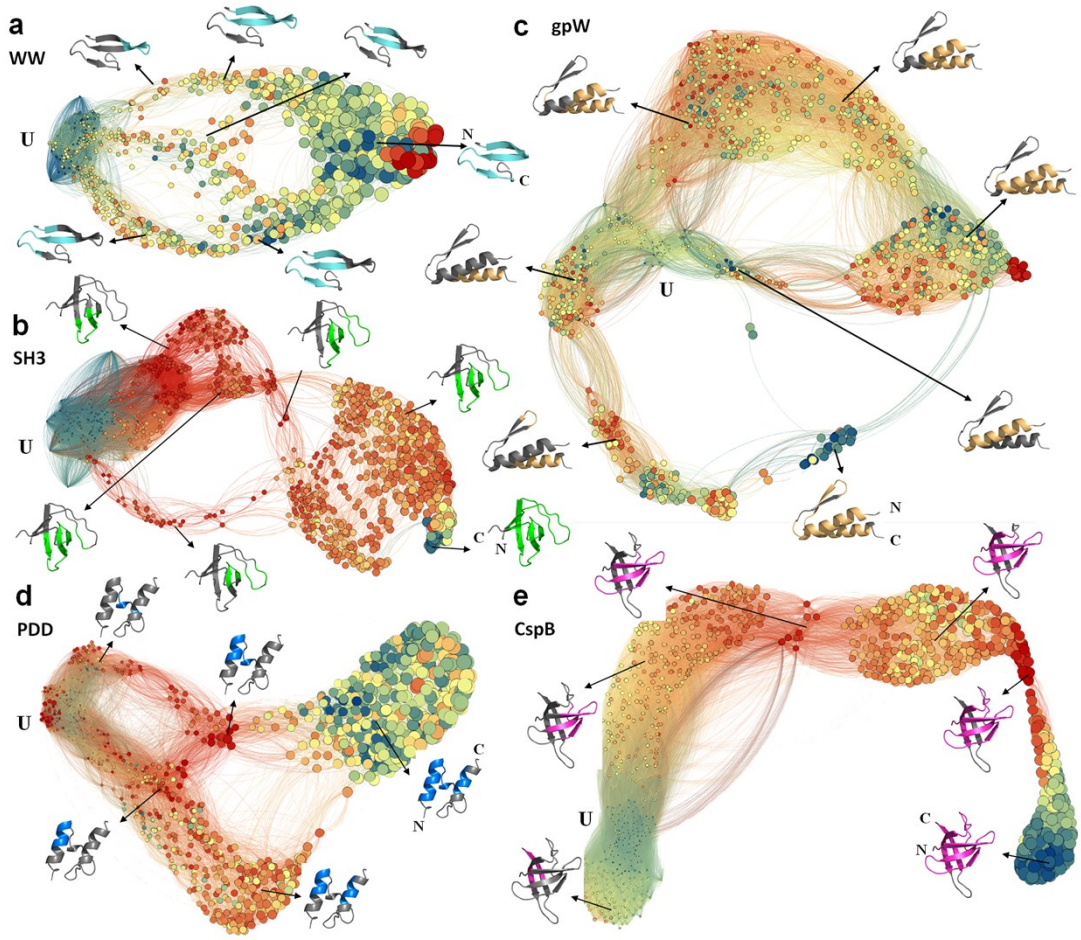


Figure S11. Panels a-e provide a simple visualization of the complex folding network of WW (panel a), SH3 (panel b), gpW (panel c), PDD (panel d) and CspB (panel e) at their respective melting temperatures. U represents the fully unfolded state. The size of the nodes is proportional to the number of structured residues and the color highlights microstate free energy in the spectral scale (blue to red represents the linearly increasing free energy value in kJ mol^{-1}). The edges are colored based on the color of the source node. The graph is organized such that the most unfolded and folded microstates are close to the extreme left and right, respectively. Structural features of some representative paths are displayed as cartoons; the gray colored regions represent unfolded residues while colored regions represent folded residues. Note the folded and unfolded ensembles have similar free energy for WW, SH3 and CspB (collection of bluish microstates to both the left and right of the graphs) as expected for midpoint conditions. The gpW and PDD folding networks are more complex because a large number of microstates have similar free energies at the midpoint. This can also be observed from the flat free-energy profiles of gpW and PDD at midpoint conditions in Figure 1 of the main text.

Supporting Table T1. Details on Proteins, WSME Model Parameters and Simulation Conditions.

Protein	**WW**		**SH3**		**gpW**		**PDD**		**CspB****	
PDB	1I5H		1SHG		1HYW		2PDD		1CSP/1G6P	
Molecular Weight (Da)	5720		6686		6976		4642		7365	
Secondary Structure	β		B		α/β		A		β	
Heavy-Atom Interaction Cut-off (Å)	5		6		5		5		6	
Number of Nearest Neighbors Excluded	0		1		0		0		1	
ξ (J mol^{-1} per contact)*	-146.0		-53.0		-135.5		-112.5		-70.3	
ΔC_p^{cont} (J mol^{-1} K^{-1} per contact)	-0.35		-0.30		-0.35		0		-0.25	
$\Delta S_{conf,o}$ (J mol^{-1} K^{-1} per residue)	-26.0		-14.2		-21.0		-18.5		-14.5	
Ionic Strength (mM)	43		10		43		43		100/160	
Total Number of Microstates	59823		821251		738189		250906		1493450	
MC Simulation Temperatures (K)	298	334	298	340	298	342	298	322	298	376
<Trajectory Length> (MC steps)	1170	5864	46232	732736	1644	16444	1220	2374	28901	11730518
<TP Length> (MC Steps)	277	175	1520	1914	880	880	220	104	9461	9451

*The C-C vdW interaction energies at 5 and 6 Å from the Amber force field are, -128.2 and -46.1 J mol^{-1}, respectively. The model estimates are of the same order as these atomic-level parameters.

**The experimental heat capacity profile of only *Bacillus subtilis* CspB (mesophile; PDB id: 1CSP) is available. Therefore, the model was employed to first fit the BsCspB DSC curve and the heat capacity profile of hyperthermophilic CspB was predicted employing identical parameters but with the input structure of *Thermotoga maritima* CspB (PDB id: 1G6P). The TmCspB heat capacity profile was predicted at an ionic strength of 160 mM mimicking experimental single molecule conditions.

Supporting Table T2. Cut-off Values Used to Delineate the Transition Paths on the Reaction Coordinate (n, number of structured residues).

Protein	@ 298K		@ T_m	
	Unfolded (n_u)	**Folded (n_f)**	**Unfolded (n_u)**	**Folded (n_f)**
WW	8	24	8	20
SH3	14	50	14	50
gpW	14	46	19	46
PDD	11	25	10	20
CspB	16	55	16	55

Supporting Table T3. Mutants that are off the experimental trend in Figure 3 of the main text.

Protein	Residue Index	Mutation
WW	10	S10G
	11	G11A
	12	R12A/R12H
	26	E26A/E23Q
SH3	24	D24A
	36	W36A
	38	K38A
PDD	1	V1A
	2	I2A/I2G
	3	A3G
	35	I35V
	37	A37G
	39	L39A
CspB	10	N10A
	15	F15A
	17	F17A
	19	E19A

Supporting Table T4. Top 10 Folding Paths in the Secondary Structure Based Network.

WW at 298 K		WW at 334 K	
Path	**Flux (%)**	**Path**	**Flux (%)**
$\beta_2 \rightarrow \beta_1 \rightarrow \beta_3$	43.0	$\beta_2 \rightarrow \beta_1 \rightarrow \beta_3$	46.9
$\beta_1 \rightarrow \beta_2 \rightarrow \beta_3$	27.6	$\beta_1 \rightarrow \beta_2 \rightarrow \beta_3$	27.7
$\beta_2 \rightarrow \beta_3 \rightarrow \beta_1$	20.8	$\beta_2 \rightarrow \beta_3 \rightarrow \beta_1$	20.6
$\beta_3 \rightarrow \beta_2 \rightarrow \beta_1$	6.0	$\beta_1 \rightarrow \beta_3 \rightarrow \beta_2$	2.2
$\beta_1 \rightarrow \beta_3 \rightarrow \beta_2$	1.6	$\beta_3 \rightarrow \beta_2 \rightarrow \beta_1$	1.9
$\beta_3 \rightarrow \beta_1 \rightarrow \beta_2$	1.0	$\beta_3 \rightarrow \beta_1 \rightarrow \beta_2$	0.7

SH3 at 298 K		SH3 at 340 K	
Path	**Flux (%)**	**Path**	**Flux (%)**
$\beta_3 \rightarrow \beta_4 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \beta_5$	54.2	$\beta_3 \rightarrow \beta_4 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \beta_5$	52.2
$\beta_4 \rightarrow \beta_3 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \beta_5$	12.7	$\beta_4 \rightarrow \beta_3 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \beta_5$	14.0
$\beta_3 \rightarrow \beta_4 \rightarrow \beta_1 \rightarrow \beta_2 \rightarrow \beta_5$	8.4	$\beta_3 \rightarrow \beta_4 \rightarrow \beta_1 \rightarrow \beta_2 \rightarrow \beta_5$	9.5
$\beta_3 \rightarrow \beta_4 \rightarrow \beta_2 \rightarrow \beta_5 \rightarrow \beta_1$	6.2	$\beta_3 \rightarrow \beta_4 \rightarrow \beta_2 \rightarrow \beta_5 \rightarrow \beta_1$	5.5
$\beta_3 \rightarrow \beta_4 \rightarrow \beta_1 \rightarrow \beta_5 \rightarrow \beta_2$	3.9	$\beta_3 \rightarrow \beta_4 \rightarrow \beta_1 \rightarrow \beta_5 \rightarrow \beta_2$	4.8
$\beta_3 \rightarrow \beta_4 \rightarrow \beta_5 \rightarrow \beta_2 \rightarrow \beta_1$	3.6	$\beta_4 \rightarrow \beta_3 \rightarrow \beta_1 \rightarrow \beta_2 \rightarrow \beta_5$	2.6
$\beta_3 \rightarrow \beta_4 \rightarrow \beta_5 \rightarrow \beta_1 \rightarrow \beta_2$	2.7	$\beta_3 \rightarrow \beta_4 \rightarrow \beta_5 \rightarrow \beta_1 \rightarrow \beta_2$	2.5
$\beta_4 \rightarrow \beta_3 \rightarrow \beta_1 \rightarrow \beta_2 \rightarrow \beta_5$	2.6	$\beta_3 \rightarrow \beta_4 \rightarrow \beta_5 \rightarrow \beta_2 \rightarrow \beta_1$	2.0
$\beta_4 \rightarrow \beta_3 \rightarrow \beta_2 \rightarrow \beta_5 \rightarrow \beta_1$	1.5	$\beta_3 \rightarrow \beta_2 \rightarrow \beta_4 \rightarrow \beta_1 \rightarrow \beta_5$	1.8
$\beta_3 \rightarrow \beta_2 \rightarrow \beta_4 \rightarrow \beta_1 \rightarrow \beta_5$	1.2	$\beta_4 \rightarrow \beta_3 \rightarrow \beta_1 \rightarrow \beta_5 \rightarrow \beta_2$	1.4

gpW at 298 K		gpW at 342 K	
Path	**Flux (%)**	**Path**	**Flux (%)**
$\alpha_2 \rightarrow \beta_2 \rightarrow \alpha_1 \rightarrow \beta_1$	19.3	$\beta_2 \rightarrow \alpha_2 \rightarrow \alpha_1 \rightarrow \beta_1$	23.3
$\alpha_1 \rightarrow \alpha_2 \rightarrow \beta_2 \rightarrow \beta_1$	16.3	$\alpha_2 \rightarrow \beta_2 \rightarrow \alpha_1 \rightarrow \beta_1$	16.7
$\alpha_2 \rightarrow \alpha_1 \rightarrow \beta_2 \rightarrow \beta_1$	14.3	$\alpha_1 \rightarrow \alpha_2 \rightarrow \beta_2 \rightarrow \beta_1$	16.0
$\beta_2 \rightarrow \alpha_2 \rightarrow \alpha_1 \rightarrow \beta_1$	9.5	$\beta_2 \rightarrow \alpha_1 \rightarrow \alpha_2 \rightarrow \beta_1$	7.7
$\alpha_1 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \alpha_2$	8.8	$\alpha_2 \rightarrow \alpha_1 \rightarrow \beta_2 \rightarrow \beta_1$	7.2
$\alpha_2 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \alpha_1$	5.5	$\alpha_1 \rightarrow \beta_2 \rightarrow \alpha_2 \rightarrow \beta_1$	4.9
$\alpha_1 \rightarrow \beta_1 \rightarrow \beta_2 \rightarrow \alpha_2$	5.4	$\alpha_1 \rightarrow \alpha_2 \rightarrow \beta_1 \rightarrow \beta_2$	4.7
$\alpha_1 \rightarrow \beta_2 \rightarrow \alpha_2 \rightarrow \beta_1$	5.2	$\beta_2 \rightarrow \alpha_2 \rightarrow \beta_1 \rightarrow \alpha_1$	4.0
$\alpha_1 \rightarrow \alpha_2 \rightarrow \beta_1 \rightarrow \beta_2$	3.7	$\alpha_1 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \alpha_2$	2.8
$\alpha_2 \rightarrow \alpha_1 \rightarrow \beta_1 \rightarrow \beta_2$	2.7	$\alpha_2 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \alpha_1$	2.4

Supporting Table T4 (Contd.). Top 10 Folding Paths in the Secondary Structure Based Network.

PDD at 298K		PDD at 322K	
Path	**Flux (%)**	**Path**	**Flux (%)**
$\alpha_1 \rightarrow \alpha_3 \rightarrow \alpha_2$	21.49	$\alpha_1 \rightarrow \alpha_3 \rightarrow \alpha_2$	31.19
$\alpha_2 \rightarrow \alpha_3 \rightarrow \alpha_1$	19.98	$\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1$	17.42
$\alpha_2 \rightarrow \alpha_1 \rightarrow \alpha_3$	18.70	$\alpha_2 \rightarrow \alpha_3 \rightarrow \alpha_1$	16.34
$\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$	17.56	$\alpha_2 \rightarrow \alpha_1 \rightarrow \alpha_3$	15.00
$\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1$	12.46	$\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$	12.65
$\alpha_3 \rightarrow \alpha_1 \rightarrow \alpha_2$	9.81	$\alpha_3 \rightarrow \alpha_1 \rightarrow \alpha_2$	7.40

CspB at 298K		CspB at 376K	
Path	**Flux (%)**	**Path**	**Flux (%)**
$\beta_2 \rightarrow \beta_3 \rightarrow \beta_1 \rightarrow \beta_4 \rightarrow \beta_5$	32.86	$\beta_2 \rightarrow \beta_1 \rightarrow \beta_3 \rightarrow \beta_5 \rightarrow \beta_4$	26.34
$\beta_2 \rightarrow \beta_1 \rightarrow \beta_3 \rightarrow \beta_4 \rightarrow \beta_5$	19.62	$\beta_2 \rightarrow \beta_1 \rightarrow \beta_3 \rightarrow \beta_4 \rightarrow \beta_5$	26.27
$\beta_2 \rightarrow \beta_3 \rightarrow \beta_1 \rightarrow \beta_5 \rightarrow \beta_4$	18.05	$\beta_2 \rightarrow \beta_3 \rightarrow \beta_1 \rightarrow \beta_5 \rightarrow \beta_4$	15.42
$\beta_2 \rightarrow \beta_1 \rightarrow \beta_3 \rightarrow \beta_5 \rightarrow \beta_4$	10.77	$\beta_2 \rightarrow \beta_3 \rightarrow \beta_1 \rightarrow \beta_4 \rightarrow \beta_5$	15.11
$\beta_1 \rightarrow \beta_2 \rightarrow \beta_3 \rightarrow \beta_4 \rightarrow \beta_5$	5.89	$\beta_1 \rightarrow \beta_2 \rightarrow \beta_3 \rightarrow \beta_5 \rightarrow \beta_4$	6.22
$\beta_3 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \beta_4 \rightarrow \beta_5$	5.70	$\beta_1 \rightarrow \beta_2 \rightarrow \beta_3 \rightarrow \beta_4 \rightarrow \beta_5$	6.02
$\beta_1 \rightarrow \beta_2 \rightarrow \beta_3 \rightarrow \beta_5 \rightarrow \beta_4$	3.19	$\beta_3 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \beta_5 \rightarrow \beta_4$	2.05
$\beta_3 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \beta_5 \rightarrow \beta_4$	3.13	$\beta_3 \rightarrow \beta_2 \rightarrow \beta_1 \rightarrow \beta_4 \rightarrow \beta_5$	1.87
$\beta_1 \rightarrow \beta_3 \rightarrow \beta_2 \rightarrow \beta_4 \rightarrow \beta_5$	0.20	$\beta_2 \rightarrow \beta_1 \rightarrow \beta_4 \rightarrow \beta_3 \rightarrow \beta_5$	0.25
$\beta_3 \rightarrow \beta_1 \rightarrow \beta_2 \rightarrow \beta_4 \rightarrow \beta_5$	0.15	$\beta_2 \rightarrow \beta_3 \rightarrow \beta_5 \rightarrow \beta_1 \rightarrow \beta_4$	0.10

Supporting Table T5. Top 10 Folding Paths Through the MCL Method.*

WW at 298 K		WW at 334 K	
Path	**Flux (%)**	**Path**	**Flux (%)**
4→1→0	39.8	4→2→0	28.9
4→2→0	23.2	4→1→0	26.1
4→3→1→0	14.4	4→3→1→0	16.0
4→3→2→0	10.3	4→3→2→0	13.3
4→1→3→2→0	2.5	4→2→3→1→0	4.0
4→0	1.9	4→1→3→2→0	3.9
4→2→3→1→0	1.9	4→1→3→0	1.4
4→1→3→0	1.5	4→2→3→0	1.4
4→2→3→0	1.5	4→3→0	1.4
4→3→0	1.5	4→0	1.3

SH3 at 298 K		SH3 at 340 K	
Path	**Flux (%)**	**Path**	**Flux (%)**
6→0	13.0	6→1→0	11.8
6→5→2→1→0	10.9	6→1→2→3→0	7.1
6→5→2→1→4→3→0	10.9	6→2→3→0	6.0
6→2→1→0	7.3	6→2→1→0	5.5
6→2→1→4→3→0	7.3	6→4→2→1→0	5.4
6→1→0	5.9	6→4→2→3→0	5.4
6→1→4→3→0	5.9	6→1→3→0	4.0
6→1→4→0	5.4	6→2→1→3→0	4.0
6→2→1→4→0	5.4	6→4→2→1→3→0	4.0
6→5→2→1→4→0	5.4	6→5→1→0	3.6

gpW at 298 K		gpW at 342 K	
Path	**Flux (%)**	**Path**	**Flux (%)**
8→0	21.2	6→0	16.5
8→3→0	10.1	6→1→0	8.0
8→1→0	9.7	6→1→3→0	5.1
8→4→0	4.9	6→1→2→0	5.0
8→6→0	4.1	6→2→0	4.4
8→3→2→0	3.8	6→3→0	3.8
8→5→2→0	3.6	6→3→1→0	2.9
8→2→0	2.9	6→3→1→2→0	2.9
8→1→6→0	2.5	6→1→2→4→0	2.5
8→4→7→0	2.4	6→2→4→0	2.5

* Note that the structural features of the clusters differ between temperatures. See Figures S6-S8.

Supporting Table T5 (Contd.). Top 10 Folding Paths Through the MCL Method.*

PDD at 298K		PDD at 322K	
Path	**Flux (%)**	**Path**	**Flux (%)**
4→2→1→0	16.9	4→3→2→0	16.4
4→3→2→1→0	16.3	4→3→1→2→0	16.4
4→0	12.6	4→3→1→0	11.8
4→2→0	10.3	4→3→2→1→0	11.8
4→3→2→0	10.3	4→2→0	10.8
4→3→1→0	9.3	4→2→1→0	10.8
4→3→1→2→0	9.3	4→3→0	5.3
4→3→0	5.4	4→2→3→0	3.6
4→1→0	1.9	4→2→3→1→0	3.6
4→1→2→0	1.9	4→2→1→3→0	2.0

CspB at 298K		CspB at 376K	
Path	**Flux (%)**	**Path**	**Flux (%)**
5→3→2→1→0	56.0	5→4→3→2→1→0	41.9
5→4→3→2→1→0	12.3	5→3→2→1→0	38.8
5→4→2→1→0	7.4	5→2→1→0	11.6
5→3→4→2→1→0	7.4	5→4→1→0	1.1
5→2→1→0	7.0	5→3→4→1→0	1.1
5→2→0	1.4	5→2→3→4→1→0	1.1
5→4→2→0	1.4	5→3→1→0	0.8
5→3→2→0	1.4	5→4→3→1→0	0.8
5→4→3→2→0	1.4	5→2→3→1→0	0.8
5→3→4→2→0	1.4	5→1→0	0.3

* Note that the structural features of the clusters differ between temperatures. See Figures S9-S10.

Supporting Movies 1-10. Each of the 10 movies present a visual representation of the top 5 folding paths (starting from the fully unfolded state) generated from the MC simulation and as identified from the secondary-structure based network analysis at 298 K and at the respective midpoint temperatures (T_m). The unfolded segments are represented in gray while the fully folded structure is color coded from blue (N-terminal) to red (C-terminal).

Supporting References

1. H. Wako and N. Saito, *J. Phys. Soc. Japan*, 1978, 44, 1939.
2. V. Muñoz and W. A. Eaton, *Proc. Natl. Acad. Sci. U.S.A.*, 1999, 96, 11311.
3. P. Bruscolini and A. Pelizzola, *Phys. Rev. Lett.*, 2002, 88, 258101.
4. E. R. Henry and W. A. Eaton, *Chem. Phys.*, 2004, 307, 163.
5. J. Kubelka, E. R. Henry, T. Cellmer, J. Hofrichter and W. A. Eaton, *Proc. Natl. Acad. Sci. USA*, 2008, 105, 18655.
6. A. N. Naganathan, *J. Chem. Theory Comput.*, 2012, 8, 4646.
7. A. N. Naganathan, *J. Phys. Chem. B*, 2013, 117, 4956.
8. P. Bruscolini and A. N. Naganathan, *J. Am. Chem. Soc.*, 2011, 133, 5372.
9. M. M. Garcia-Mira, M. Sadqi, N. Fischer, J. M. Sanchez-Ruiz and V. Muñoz, *Science*, 2002, 298, 2191.
10. A. D. Robertson and K. P. Murphy, *Chem. Rev.*, 1997, 97, 1251.