SUPPLEMENTARY MATERIAL. Mapping the topography of a protein energy landscape.

Richard D. Hutton,[†,#] James Wilkinson,[†] Mauro Faccin,[‡] Elin M. Sivertsson,[¶] Alessandro Pelizzola,[§] Alan R. Lowe,[*,‖] Pierpaolo Bruscolini,[*,⊥] and Laura S. Itzhaki[*,¶]

Hutchison/MRC Research Centre, Hills Road, Cambridge CB2 0XZ, UK, ICTEAM Université Catholique de Lovain Euler Building 4, Avenue Lemaître B-1348 Louvain-la-Neuve Belgium, Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1PD, UK, Dipartimento di Scienza Applicata e Tecnologia, CNISM and Center for Computational Studies, Politecnico di Torino, Corso Duca degli Abruzzi 24, I–10129 Torino, Italy; INFN, Sezione di Torino, via Pietro Giuria 1, I-10125 Torino, Italy;Human Genetics Foundation, HuGeF, Via Nizza 52, I-10126 Torino, Italy, Institute for Structural and Molecular Biology & London Centre for Nanotechnology, University College London and Birkbeck College London, UK., and Departamento de Física Teórica & Instituto de Biocomputacíon y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, c/ Mariano Esquillor s/n, 50018, Zaragoza, Spain

E-mail: a.lowe@ucl.ac.uk; pier@unizar.es; lsi10@cam.ac.uk

[*]To whom correspondence should be addressed
[†]Hutchison/MRC Research Centre
[‡]Université Catholique de Lovain
[¶]University of Cambridge
[§]Politecnico di Torino
[‖]University College London and Birkbeck College London
[⊥]Universidad de Zaragoza
[#]Current address: Canterbury Scientific Ltd., 71 Whiteleigh Avenue, Christchurch 8011, New Zealand

Simulations Methods

Model

We use a modified version of the WSME model[1–4], with a suitable redefinition of the interactions. As a native-centric model[5], WSME requires the knowledge of the native state of a protein to describe its equilibrium and kinetics. Its binary variables m_k, with $k \in [1, N]$ for a N-residues protein, are related to the backbone and side chain angles, and describe the state of each residue as native, $m_k = 1$, and unfolded, $m_k = 0$.

Its effective energy can be written as

$$H = \sum_{i=1}^{N} \sum_{j=i}^{N} h_{i,j} \prod_{k=i}^{j} m_k \,, \tag{S1}$$

where

$$h_{i,j} = -\epsilon_{i,j} \Delta_{i,j} (1 - \delta_{i,j}) + T q_i \delta_{i,j} \,, \tag{S2}$$

gives an interaction energy $\varepsilon_{i,j}$ between residues i, j (if $i \neq j$), provided that the interaction also exists in the native contact map $\Delta_{i,j}$, and an entropy cost q_i for folding residue i, when $i = j$ ($\delta_{i,j}$ indicates Kronecker delta); non–native interactions are disregarded, and the completely unfolded state with all $m_i = 0$ is set to zero energy. Eq. S1 can be recast as

$$H = \sum_{i=1}^{N} \sum_{j=i}^{N} H_{i,j} \sigma_{i,j} \,, \tag{S3}$$

with $H_{i,j} = \sum_{k=i}^{j} \sum_{l=k}^{j} h_{k,l}$ represents the whole (effective) energy contribution from a native structure spanning the region (i, j) and $\sigma_{i,j} = (1 - m_{i-1}) \prod_{k=i}^{j} m_k (1 - m_{j+1})$ (here we set $m_0 = m_{N+1} = 0$). The latter expression is more suitable to present the changes we introduce in the WSME model, to describe more realistically the chemical denaturation of gankyrin. With $A^U_{i,j}$, $A^F_{i,j}$ being the solvent-accessible surface area (ASA) of the unfolded and folded regions (i, j), respectively, and $A^{U,\mathrm{m.c.}}_{i,j}$, $A^{F,\mathrm{m.c.}}_{i,j}$ the analogous quantities calculated just for

the main chain atoms, we set:

$$H_{i,j} = -\epsilon\Delta_{i,j}^{U-F} + \alpha c\Delta_{i,j}^{U-F,\mathrm{m.c.}} + Tq(j-i+1)\,, \tag{S4}$$

where $\Delta_{i,j}^{U-F} = A_{i,j}^{U} - A_{i,j}^{F}$, $\Delta_{i,j}^{U-F,\mathrm{m.c.}} = A_{i,j}^{U,\mathrm{m.c.}} - A_{i,j}^{F,\mathrm{m.c.}}$, c is the denaturant concentration, T the absolute temperature, and $\epsilon, q, \alpha > 0$. Here the first term represents the energy gain upon folding the region (i,j), and the second term represents the cost of folding it, associated with the coupling with the denaturant, in agreement with the findings in[6,7] that the interaction with the denaturant dominantly involves the main chain. Note that the $i = j$ case is included in the description above, and also that, for some choice of i, j corresponding to some short regions, the quantities $\Delta_{i,j}^{U-F}$, $\Delta_{i,j}^{U-F,\mathrm{m.c.}}$ can also be negative if the native structure is more stretched than the average unfolded one. The latter term is proportional to the entropic cost per residue, q, accounting for the fact that the unfolded state $m_k = 0$ corresponds to a much larger number of microscopic realizations than the folded one.

Calculation of the ASAs of gankyrin To calculate the folded ASAs, we refer to the NMR structure of gankyrin deposited in the Protein Data Bank (PDB code: 1TR4, first model). It is implicit in the WSME model that $H_{i,j}$ in equation S3 represents the energetic contribution of all interactions within the stretch i, j, with no contributions involving amino acids outside that region. It is important to take this into account when defining the $H_{i,j}$ in terms of the ASAs: if we consider a conformation with a structured island from i to j, with unfolded stretches separating it from other native regions, the folded ASA, $A_{i,j}^{F}$, should account from the solvent protection coming only from within the region (i, j) (according to the spirit of the WSME model, we disregard the protection coming from the other native stretches). Thus, it should not be calculated as the ASA of the region (i, j) within the whole native structure; the ASAs of the isolated region (i, j) should be calculated instead. To this end, we extract from the native structure the coordinates of the atoms in the region $(i-1, j+1)$ (unless i is the first or j the last residue), keeping only the main chain for

residues $i-1$, $j+1$, to mimic Gly-(i,j)-Gly: in this way, we seek to minimize the border effects of the isolated peptide while keeping the disturbance of the rest of the chain at its minimum value. We use ALPHASURF[8] on the coordinates above to calculate the ASAs of the isolated folded region (i,j), while the unfolded ASA, $A^U_{i,j}$, is calculated as the sum of the values obtained with Gly-X_k-Gly, with X_k the type of amino acid k ($k=i,\ldots,j$).[1] The same procedures apply to the calculation of $A^{U,\mathrm{m.c.}}_{i,j}$, $A^{F,\mathrm{m.c.}}_{i,j}$.

Mutants are described, in our approach, just by substituting the ASA of the mutated residue with that of its main chain, which roughly mimics a mutation of the residue to Glycine. The energy parameters calculated from the fits for the WT species (see below) are also used for mutants.

Choice of the parameters

We fix the parameters ϵ, q, α of Eq. S4 by fitting the fluorescence and circular dichroism (CD) experimental signals for the WT species, with those predicted by the model. To this end, we first extract a net signal from the raw experimental data, determining the native and unfolded baselines and removing them. For fluorescence, we have fitted the native baseline $x_n^{(F)} = a^{(F)}c + b^{(F)}$ from the raw signal $I^{(F)}$ at 341 nm emission wavelength, in the denaturant concentration range $c \in [0, 2.1]$, and the unfolding baseline $x_u^{(F)} = c^{(F)}c + d^{(F)}$ with $c \in [5.6, 8]$. Then, we have normalized the experimental signal on the two baselines, as

$$y^{(F)} = \frac{(I^{(F)}(c) - x_u^{(F)}(c))}{(x_n^{(F)}(c) - x_u^{(F)}(c))}, \tag{S5}$$

and we estimate errors on $y^{(F)}$ by propagating the errors on $a^{(F)}$, $b^{(F)}$, $c^{(F)}$, $d^{(F)}$ resulting from the fit,

[1] The ideal procedure would be to calculate $A^U_{i,j}$ using PROTSA[9], which gives a better estimates of unfolded ASAs. However, applying PROTSA to the N(N+1)/2 possible subregions of the protein is exceedingly time-consuming. Actually, in the computer program implementing the model, we define $\Delta^{U-F}_{i,j} = (A^U_{i,j} - A^F_{i,j})\chi$, where $\chi = (A^{U,PROTSA}_{1,N} - A^F_{1,N})/(A^U_{1,N} - A^F_{1,N})$, thus normalizing the areas according to the PROTSA value for the unfolded state of the whole protein. However, this just corresponds to rescaling the ϵ and α in Eq. S4, so we have removed it from Eq. S4 for the sake of simplicity.

For the far-UV CD ellipticity data at 222 nm, we have fitted the native baseline $x_n^{(CD)} = a^{(CD)}c + b^{(CD)}$ from the raw signal $I^{(CD)}$ in the denaturant concentration range $c \in [0, 2.1]$, and the unfolding one $x_u^{(CD)} = c^{(CD)}c + d^{(CD)}$ with $c \in [5.6, 8]$.

Then, we have normalized the experimental signal on the two baselines, as $y^{(CD)} = (I^{(CD)}(c) - x_u^{(CD)}(c))/(x_n^{(CD)}(c) - x_u^{(CD)}(c))$, following the same procedure as for fluorescence data. Table S1 reports the values of the parameters obtained in the fit.

Table S1: Values of the fitted parameters (errors in parentheses) characterizing the folded and unfolded baselines $x_{u,n}^{(F)}$, $x_{u,n}^{(CD)}$. The raw fluorescence data are expressed in arbitrary units, while ellipticity data are expressed in units of millidegrees.

	a	b	c	d
F	26.2 (1.2)	548.9 (1.5)	8.8 (0.5)	157 (3)
CD	-0.65 (0.13)	-37.10 (0.14)	0.55 (0.04)	-10.8 (0.3)

The normalized signals $y^{(F,CD)}$ serve as a reference to adjust the signal predicted with the model that we obtain as follows:

For fluorescence, we have assumed that the signal at the given wavelength (341 nm) represents the contributions from tryptophan residues, while tyrosine residues are disregarded. We further assume that the contribution to the signal of a protein conformation is proportional to the fraction of the area of the fluorescent residues protected from the solvent in that conformation. Namely we calculate the normalized predicted signal $y^{(F,theo)}$ as:

$$y^{(F,theo)} = \frac{(I^{(F,theo)}(c) - I^{(F,theo)}(8))}{(I^{(F,theo)}(0) - I^{(F,theo)}(8))} \ , \tag{S6}$$

where

$$I^{(F,theo)}(c) = \frac{1}{\sum_{i\in\mathcal{F}}((1-\mu_{i,i}) + \sum_{j<=i}\sum_{k>=i}\nu_{j,k}))} \sum_{i\in\mathcal{F}}(\phi_U^i(1-\mu_{i,i}) + \sum_{j<=i}\sum_{k>=i}\phi_{j,k}^i\nu_{j,k})) \tag{S7}$$

is the average predicted signal, at concentration c, caused by tryptophans listed in the set $\mathcal{F}$ (in the case of Gankyrin 1TR4: W46 and W74), $\mu_{i,j}$ represents the equilibrium probability

that the region between i and j is native:

$$\mu_{i,j} = \langle \prod_{k=i}^{j} m_k \rangle \,, \tag{S8}$$

and $\nu_{i,j}$ represents the equilibrium probability that the region between i and j is an isolated native region, capped by unfolded residues:

$$\nu_{i,j} = \langle \sigma_{i,j} \rangle = \langle (1 - m_{i-1}) \prod_{k=i}^{j} m_k (1 - m_{j+1}) \rangle \,. \tag{S9}$$

The protected surface fraction of residue i in a native island from j to k is calculated as:

$$\phi^i_{j,k} = 1 - \frac{A^i_{j,k} - A^i_{\min}}{A^i_{\max} - A^i_{\min}} \,, \tag{S10}$$

where $A^i_{j,k}$ is the ASA of residue i contained in the folded island (j,k),

$$A^i_{\min} = \min(A^U_{i,i}; A^i_{j,k}, \forall j <= i, k >= i) \tag{S11}$$

while $A^i_{\max}$ is defined in an analogous way. Thus, the first term in Eq. S7 represents the fraction of protection of fluorescent residue i in the unfolded states, while the second accounts for configurations in which i is contained in a native region ranging from j to k.

To model the CD signal, we have used the helices in the PDB file, and we have assumed that the normalized signal can be associated solely with the unfolding of the 15 helices: Thus, we have calculated the normalized predicted signal $y^{(CD,theo)}$ defined as:

$$y^{(CD,theo)} = \frac{(I^{(CD,theo)}(c) - I^{(CD,theo)}(8))}{(I^{(CD,theo)}(0) - I^{(CD,theo)}(8))} \,, \tag{S12}$$

where

$$I^{(CD,theo)}(c) = \frac{1}{n_{h.r.}} \sum_{i \in \mathcal{H}} \mu_{i,i} \,, \tag{S13}$$

with $\mathcal{H}$ the set of residues belonging to a helix, and $n_{h.r.}$ their total number. Eq S13 represents the average native fraction, restricted to residues that belong to helices in the native structure.

With the above definitions of the predicted normalized fluorescence and CD signals, we perform a search in the parameters space, to minimize the difference between predicted and experimental signals. First, for each experimental technique we calculate the distance between experimental and predicted signals, weighted by the experimental errors on each point:

$$d_\alpha = \frac{1}{\sqrt{\sum_{k=1}^{\mathcal{N}_\mathcal{D}^\alpha}(w_k^\alpha)^2}} \sqrt{\sum_{k=1}^{\mathcal{N}_\mathcal{D}^\alpha}(w_k^\alpha)^2 (y_k^{(\alpha,theo)} - y_k^{(\alpha,exp)})^2} \tag{S14}$$

where α =F,CD; N_D^α is the number of data at different denaturant concentration for technique α, and the weight $w_k^\alpha = 1/\rho_k^\alpha$ is the inverse of the error ρ_k^α on the k-th data from technique α. Then, to compare and sum the distance proceeding from different techniques, we introduce a "relative error" by using the average distance A_α of each experimental signal from its minimal value:

$$A_\alpha = \sum_{k=1}^{\mathcal{N}_\mathcal{D}^\alpha}(y_k^{(\alpha,exp)} - y_{min}^{(\alpha,exp)}) \ , \tag{S15}$$

which gives an estimate of the magnitude of the change in the experimental signal. Finally, the function to minimize is given by:

$$\rho = \frac{1}{2}\left(\frac{d_F}{A_F} + \frac{d_{CD}}{A_{CD}}\right) \ . \tag{S16}$$

Clearly, this choice of the "relative error" is not unique, but we just need a reasonable estimator of how good the predictions are, compared to experiments.

Thermodynamics

The equilibrium values of all thermodynamic quantities are calculated resorting to the exact solution of the model[10,11]. In particular, we will study the behavior of the quantities Eqs. S6,

S7, S12, S13, as well as the average values $\mu_{i,i}$ from Eq. S8, representing the probability that residue i is folded at the given temperature and denaturant concentration, $\nu_{i,j}$ from Eq. S9, and the mean native fraction and standard deviation of the variable $m = \frac{1}{N}\sum_{i=1}^{N} m_i$:

$$\mu = \langle m \rangle \ , \ \ \sigma = \sqrt{\langle m^2 \rangle - \langle m \rangle^2}\,. \tag{S17}$$

Finally, to investigate the structure of the equilibrium metastable states, we calculate the contributions to $\mu_{i,i}$ proceedings from just the states with M native residues:

$$\mu_{i,i}^{M} = \frac{1}{Z_M} {\sum_{\{m_k\}}}' m_i \exp\left(-H/RT\right). \tag{S18}$$

In particular, we consider the quantity

$$A(M) = \frac{1}{S_2(M)} \sum_{i=1}^{N} (i - \frac{N+1}{2}) \mu_{i,i}^{M}\,, \tag{S19}$$

with

$$S_2(M) = \sum_{i=1}^{N} \mu_{i,i}^{M}\,. \tag{S20}$$

$A(M)$ describes the asymmetry N-C of the population at a given total number M of native residues: the (signed) distance of residue i from the middle of the protein is weighted by the probability that residue i is native within configurations of a given M, and summed. Thus, a negative $A(M)$ implies that the ensemble of configurations with fixed total number of native residues M, present a dominance of structure at the N-terminus.

Kinetics

The kinetic evolution of the model is described through a discrete–time master equation, $p_{t+1}(x) = \sum_{x'} W(x' \to x) p_t(x)$, for the probability distribution $p_t(x)$ at time t, where $x = \{m_k,\, k = 1, \ldots N\}$ denotes the state of the system. The kinetics will be studied by means

of Monte Carlo simulations: as in previous works[12,13], the transition matrix W is specified by a single residue flip Metropolis rule, which implies that a flip is accepted or rejected according to its equilibrium probability, at the temperature and denaturant concentration specified for the simulation. We study the kinetics in both folding (T=298.15 K, $c = 0$ M) and unfolding conditions (T=298.15 K, $c = 6, 7, 8$ M): in folding (or unfolding) simulations, the initial state is a completely unfolded (or folded) state, with $m_i = 0$ (resp. 1) for each i. For each simulation, we keep track of the formation/disruption of secondary structure elements until the First Passage Time in the final state. The latter is defined as the first time that the fraction of native residues of the whole protein enters the range $[\mu(c) - \sigma(c), \mu(c) + \sigma(c)]$, where μ, σ, defined in Eq. S17, are the equilibrium values, at concentration c, for the average and standard deviation of the stochastic variable m. The secondary structure elements we keep track of are defined as the regions spanning helices 1, 2, ..., 15: namely, the 120 regions $R_{a,b}$ starting at the first residue of helix a and ending at the last residue of helix b, with $a, b = 1, \ldots, 15$. In folding simulations, we keep track of the times $t_{a,b}$ as the first time such that the region $R_{a,b}$ remains fully folded for $t \geq t_{a,b}$, while in unfolding simulations, $t_{a,b}$ is the last time such that $R_{a,b}$ is fully folded for $t = t_{a,b} - 1$. In this way, for each folding/unfolding trajectory i, we can define an ordering $\{h_1^{(i)}, h_2^{(i)}, \ldots, h_n^{(i)}\}$, where $h_k^{(i)}$ is the number of the kth secondary structure element of a given kind to fold (or unfold) in trajectory i, and n is their total number. In particular, we study the order of formation of helices, pair of neighboring helices ("helix pairs"), repeats and group of repeats; Table S2 lists the definitions of these structural elements. We collect approximately 5000 trajectories for both folding and unfolding processes.

Rates and amplitudes

To extract the folding/unfolding rates and amplitudes, we collect the trajectories in groups of 200 or 400 trajectories each (depending on the total number of available runs), and average

Table S2: Definition of the secondary structure elements used to delineate the different folding/unfolding pathways, with the number n_e of elements of each kind, and the start and end residue of each structure "i". Note that the first ankyrin repeat contains three helices (the third one being a small 3_{10} helix), whereas all of the other repeats contain two helices each. The three "groups" of repeats are defined as follows: group 1 comprises repeats 1-2, group 2 repeats 3-5, and group 3 repeats 6-7: this grouping of repeats was suggested by the experimental results.

helices		helix pairs		repeats		groups	
n_e:	15	n_e:	14	n_e:	7	n_e:	3
1	9-16	1	9-28	1	9-33	1	9-62
2	19-28	2	19-33	-	-	-	-
3	31-33	3	31-50	-	-	-	-
4	43-50	4	43-62	2	43-62	-	-
5	53-62	5	53-82	-	-	-	-
6	76-82	6	76-93	3	76-93	2	76-160
7	86-93	7	86-115	-	-	-	-
8	109-115	8	109-129	4	109-129	-	-
9	119-129	9	119-148	-	-	-	-
10	142-148	10	142-160	5	142-160	-	-
11	152-160	11	152-182	-	-	-	-
12	175-182	12	175-195	6	175-195	3	175-225
13	185-195	13	185-213	-	-	-	
14	208-213	14	208-225	7	208-225	-	-
15	216-225	-	-	-	-	-	-

the signal of the native fraction $m(t) = \frac{1}{N}\sum_{i=1}^{N} m_i$ within each group; namely:

$$\bar{\mu^g}(t) = \frac{1}{N_{tr}^g} \sum_{k \in G_g} m_k(t), \tag{S21}$$

with g the group index, G_g the g-th group, and N_{tr}^g the number of trajectories belonging to group G_g. We have verified that, using a few hundreds of trajectories in each group, the resulting average signal is "clean" enough, to allow an easy determination of the rates and amplitudes. The latter are found by fitting $\bar{\mu^g}(t)$ tothe sum of two exponentials:

$$f(t) = m_\infty + a\exp(-k_1 t) + b\exp(-k_2 t), \tag{S22}$$

where a, b are the amplitudes associated to the rates k_1, k_2, and $m_\infty = m(t \to \infty) = \mu(c)$ is the asymptotic equilibrium value, calculated from Eq. S17, corresponding to denaturant concentration c (after the concentration jump). We have verified that a two-exponential fit gives significantly better results than a single-exponential one, whereas a three-exponential fit provides little further improvement.

In this way, using approximately $N_g = 10$ groups of trajectories for each value of the denaturant concentration, we get N_g independent estimates of amplitudes and rates from which we obtain an estimate of the average and error (as the standard deviation) for the amplitudes and rates.

Determination of kinetics pathways

We determine the pathways using two different and complementary approaches: clustering of trajectories with Affinity Propagation (AP) algorithm[14], and use of suitable order parameters to classify the sequence of folding/unfolding events. In order to apply AP, we need to define a distance between any two trajectories. To this end, instead of using the full description $x(t)$ of each trajectory, we consider the sequence of ordering/disordering events of the n_e (super)secondary structure elements (15 helices or 14 pairs of helices) that constitute our molecule. Consider the ith folding trajectory: the time $t^{(i)}_{k,k}$ at which element k folds is used to define a sequence of structure elements, and the position of element k in this sequence is denoted by $p^{(i)}_k$. For a given trajectory i, $\{p^{(i)}_k, k = 1 \cdots n_e\}$ is then a permutation of $\{1 \cdots n_e\}$, and the distance between any two trajectories i and $j \neq i$ is defined as:

$$d_{ij} = \sqrt{\sum_{k=1}^{n_e} \left[p^{(i)}_k - p^{(j)}_k\right]^2}. \tag{S23}$$

A similar definition is used for the unfolding trajectories.

The AP algorithm is a message–passing algorithm that clusters data by associating an *exemplar* to each data point, and putting in the same cluster data points with the same

exemplar. The exemplar can be thought of as the best representative for the data points in a cluster. The degrees of freedom of the algorithm, defining the search space, can be represented as binary variables[15] c_{ij}, where $c_{ij} = 1$ (respectively 0) if j is (resp. is not) the exemplar of i, while $c_{ii} = 1$ (respectively 0) if i is (resp. is not) an exemplar. Each data point i must belong to exactly 1 cluster, that is it must have exactly one exemplar, unless it is an exemplar itself: hence, the constraints $\sum_j c_{ij} = 1$ must be enforced. Moreover, if j is an exemplar for some $i \neq j$ it must be an exemplar for itself, that is if $c_{ij} = 1$ it must also be $c_{jj} = 1$. The total *similarity*

$$S = \sum_{i,j} c_{ij} s_{ij}, \tag{S24}$$

is then minimized, subject to the above–mentioned constraints, using a message–passing algorithm. Here we take $s_{ij} = -d_{ij}^2$ for $i \neq j$, a common choice in AP (the larger the distance between 2 data points, the smaller their similarity). The values s_{ii}, called *preferences* in AP, are a measure of the a priori likelihood of data point i to be an exemplar. Since all our trajectories are a priori equivalent, we choose uniform preferences $s_{ii} = s$, $\forall i$, and the preference s turns out to control the number of clusters predicted by the AP algorithm: the larger s, the larger the number of clusters.

Finding a clustering is, however, not sufficient to classify folding and unfolding pathways of our molecule. The clusters obtained must be characterized physically, and to this end we have introduced some suitable order parameters. Since we are interested in the order of the folding/unfolding events, a particularly suitable order parameter is given by the Kendall rank-correlation between the order of formation/disruption of secondary structures in each trajectory, and a reference ordering $\{x_1, \ldots, x_n\}$ (indeed a permutation of $\{1, 2, \ldots, n\}$, with n the number of secondary structures of the kind under study). Upon considering the pairs (x_i, h_i) between corresponding elements in the reference set and in the sequence of secondary-structure elements of a trajectory, we say that two pairs (x_i, h_i) and (x_j, h_j), $j \neq i$, are "concordant" if $\mathrm{sgn}((x_j - x_i)(h_j - h_i)) = 1$, with sgn the sign function (that is, both $x_i > x_j$ and $h_i > h_j$ or both $x_i < x_j$ and $h_i < h_j$). They are "discordant" if

$\mathrm{sgn}((x_j - x_i)(h_j - h_i)) = -1$, and there is a "tie" if $\mathrm{sgn}((x_j - x_i)(h_j - h_i)) = 0$. The Kendall rank-correlation is then defined as:

$$\tau_B = \frac{n_c - n_d}{\sqrt{n_0(n_0 - n_2)}} \tag{S25}$$

where n_c, n_d are the number of concordant and discordant pairs, respectively; $n_0 = n(n-1)/2$, n is the number of secondary structures, $n_2 = \sum_j u_j(u_j - 1)/2$ with u_j the number of tied values in the j^{th} group of ties, in the case that some secondary structures are formed or disrupted at the same time. Since the experimental results suggest the predominance of two main pathways, progressing from the N- to the C-terminus or vice versa, we use the natural order $x_i = i$ as the reference sequence. However, the above definition Eq. S25 can be used to study the similarity of the pathways to any other order of events.

Results

Experimental

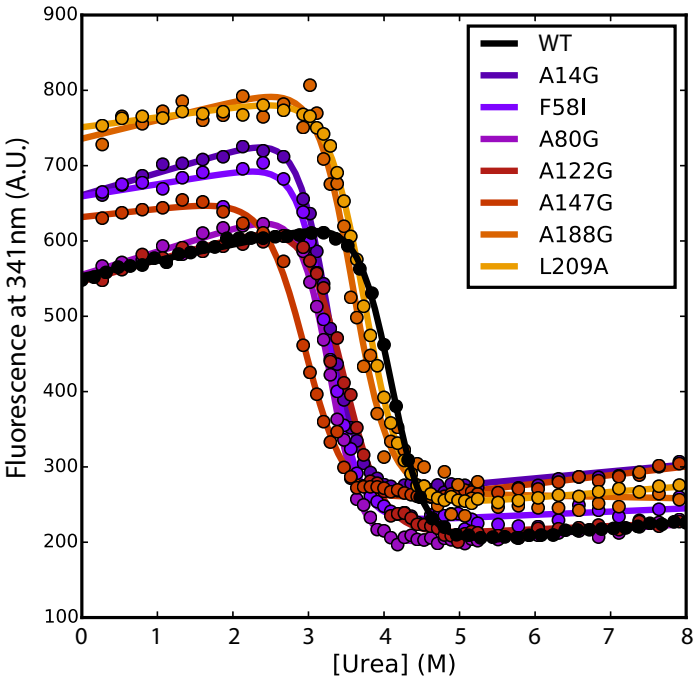


Figure S1: Denaturation curves of representative mutant proteins monitored by fluorescence. Wild-type gankyrin is shown for comparison. The data were fitted to a two-state equation.

Table S3: Equilibrium data for wild type and mutants of gankyrin. Measurements were made at 25 °C in 50 mM Tris-HCl buffer pH 8, 5 mM DTT. The protein concentration was 2 μM. The denaturation curves were fitted to a two-state equation. The standard errors are listed. Values of $\Delta\Delta G^{H_2O}_{D-N}$ were calculated using a weighted average m-value of 2.70 ± 0.10 kcal mol^{-1} M^{-1} obtained from the mutants and repeat measurements of wild type.

Protein	m-value (kcal.mol^{-1} M^{-1})	[Urea]$_{50\%}$ (M)	$\Delta\Delta G^{H_2O}_{D-N}$ (kcal mol^{-1})
Wild Type	2.63 ± 0.05	4.09 ± 0.01	-
ANK1			
V10A	3.10 ± 0.06	3.60 ± 0.01	1.31 ± 0.05
A14G	2.80 ± 0.07	3.25 ± 0.01	2.27 ± 0.08
I26A	2.74 ± 0.07	3.23 ± 0.01	2.32 ± 0.08
ANK2			
A43G	3.12 ± 0.10	3.12 ± 0.01	2.40 ± 0.09
A47G	3.09 ± 0.06	3.43 ± 0.01	1.78 ± 0.07
F58I	2.78 ± 0.10	3.25 ± 0.01	2.27 ± 0.08
ANK3			
I79V	2.65 ± 0.05	3.74 ± 0.01	0.94 ± 0.03
A80G	2.93 ± 0.08	3.29 ± 0.01	2.16 ± 0.08
A81G	2.25 ± 0.09	3.00 ± 0.01	2.93 ± 0.11
V89A	3.03 ± 0.09	3.47 ± 0.01	1.66 ± 0.06
A91G	3.12 ± 0.08	3.44 ± 0.01	1.75 ± 0.06
ANK4			
A113G	1.49 ± 0.03	3.18 ± 0.01	2.46 ± 0.06
A122G	2.36 ± 0.13	3.46 ± 0.02	1.68 ± 0.06
ANK5			
A147G	2.17 ± 0.10	2.95 ± 0.02	3.08 ± 0.07
I155V	2.54 ± 0.08	3.85 ± 0.01	0.65 ± 0.02
I157V	2.42 ± 0.13	3.85 ± 0.02	0.64 ± 0.02
ANK6			
A188G	2.47 ± 0.21	3.58 ± 0.03	1.38 ± 0.05
ANK7			
L209A	2.40 ± 0.05	3.74 ± 0.01	0.94 ± 0.03
V211A	2.39 ± 0.05	3.76 ± 0.01	0.89 ± 0.03
A188GV211A	1.79 ± 0.08	3.49 ± 0.01	1.62 ± 0.03

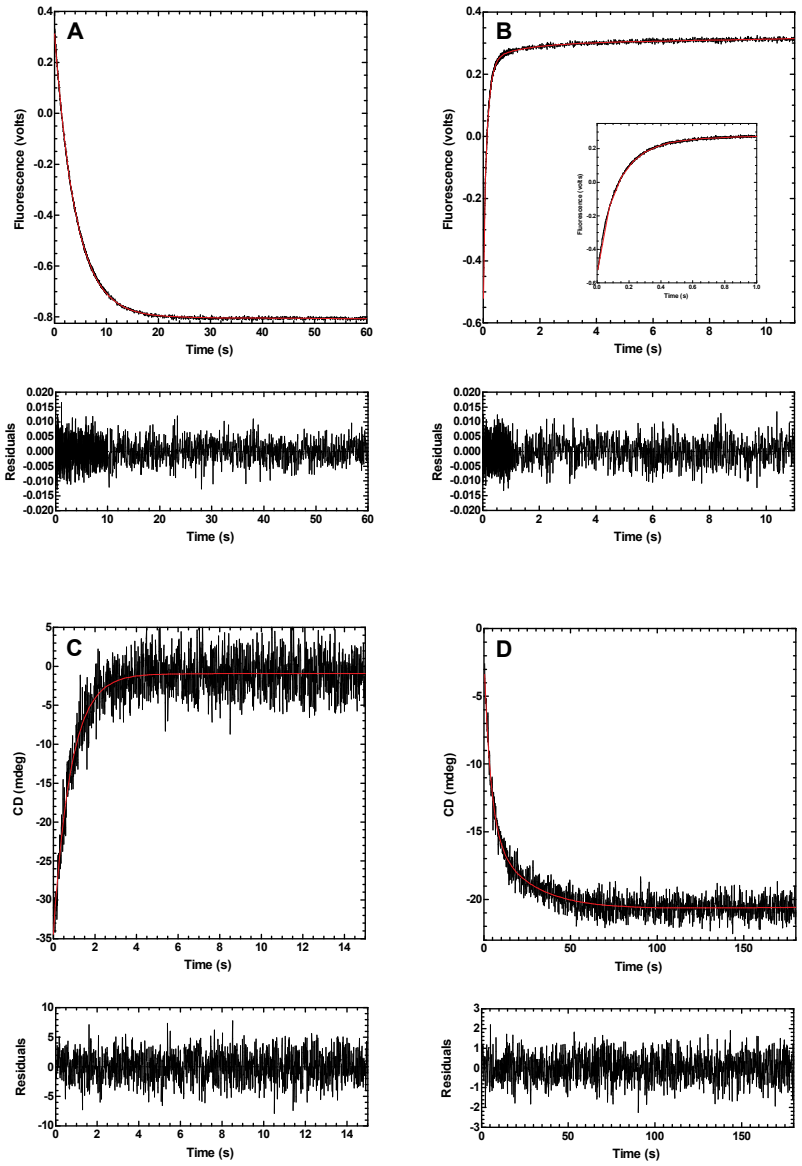


Figure S2: Representative kinetic traces (with fitting residuals shown below) for wild-type gankyrin. (A) Unfolding at 6.8 M urea, fitted to the sum of two exponentials, and (B) refolding at 1.2 M urea, fitted to the sum of three exponential phases, monitored by stopped-flow fluorescence; (C) unfolding at 8.6 M urea, fitted to a single exponential phase, and (D) refolding at 3.2 M urea, fitted to the sum of two exponential phases, monitored by stopped-flow CD.

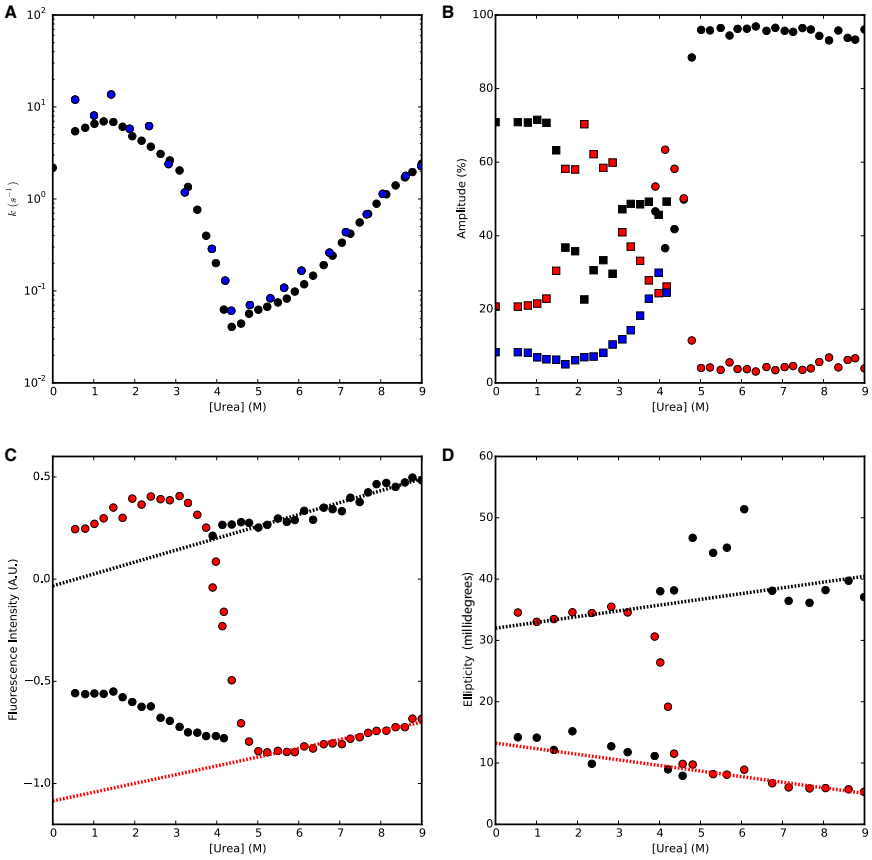


Figure S3: Unfolding and refolding kinetics of wild-type gankyrin. (A) The rate constants of the major refolding and unfolding phases observed by stopped-flow CD (blue) are shown overlaid on those observed by fluorescence. (B) Urea dependence of the amplitudes of refolding and unfolding monitored by stopped-flow fluorescence. The major phases are shown in black, and the minor phases in red and blue. Note that the data for the fastest unfolding phase (shown in blue) were obtained using interrupted refolding experiments. (C) and (D) Endpoint analysis monitored by fluorescence and CD. Starting points (fluorescence/ellipticity at t=0 s as estimated from the fit of the kinetic traces) are in black, endpoints (fluorescence/ellipticity at t=∞ as estimated from the fit of the kinetic traces) are in red. The urea dependence of the endpoints recapitulates the fluorescence-monitored equilibrium denaturation curve (see Fig. 1B in the main text). The black dashed lines show the extrapolation of the starting points of the unfolding kinetics to low urea concentrations. The red dashed lines show the extrapolation of the endpoints of the unfolding traces to low urea concentrations.

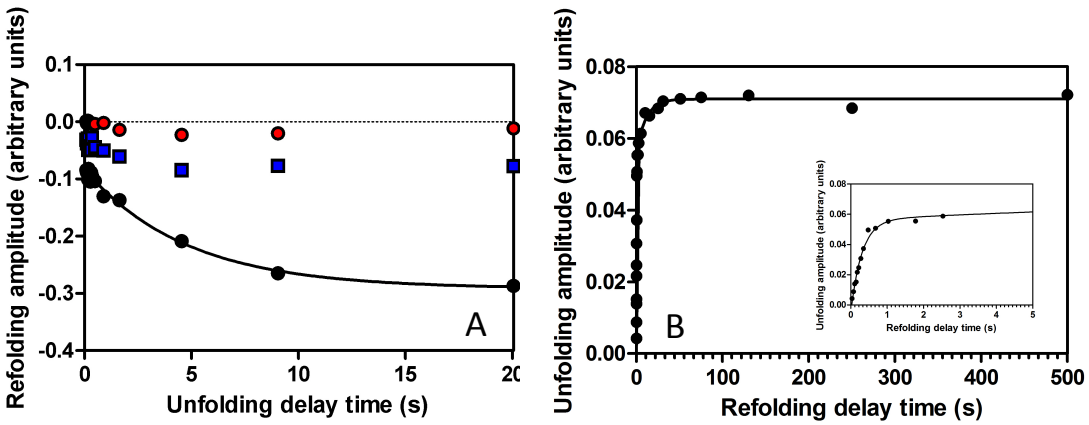


Figure S4: Interrupted unfolding (A) and interrupted refolding (B) experiments for wild-type gankyrin measured by stopped-flow fluorescence in sequential mixing mode. (A) Plot of the amplitudes of the three refolding phases as a function of unfolding delay time. The colours used are the same as those in Figure 3A. The data for the major refolding phase (black symbols) are fitted to a single exponential phase. The minor, fast refolding phase is shown in blue symbols and the minor, slowest refolding phase is shown in red symbols. (B) Plot of the amplitude of the major unfolding phase as a function of refolding delay time. The data are fitted to the sum of two exponential phases; inset is a close-up of the data at short refolding delay times.

Simulations

Equilibrium

The fit of the parameters, according to the procedure explained in Methods, produced the values reported in Table S4:

Table S4: Parameter values resulting from the fit. The last column represents relative error Eq. S16. Units for ε, q and α are kJ/(mol Å^2), kJ/(K mol) and kJ/(mol Å^2 [D]) respectively, where [D] is the denaturant concentration in M; ρ is dimensionless.

ε	q	α	ρ
0.08173	0.01521	3.108×10^{-3}	3.33×10^{-2}

Figure S5 reports the resulting maps for the effective contact energies $h_{i,j}$ between residues i and j, and the energies of whole regions $H_{i,j} = \sum_{k=i}^{j}\sum_{l=k}^{j} h_{k,l}$ Eq.S4. The analysis of the residue-residue contacts does not give any significant clue on the characteristics of the equilibrium and kinetics.

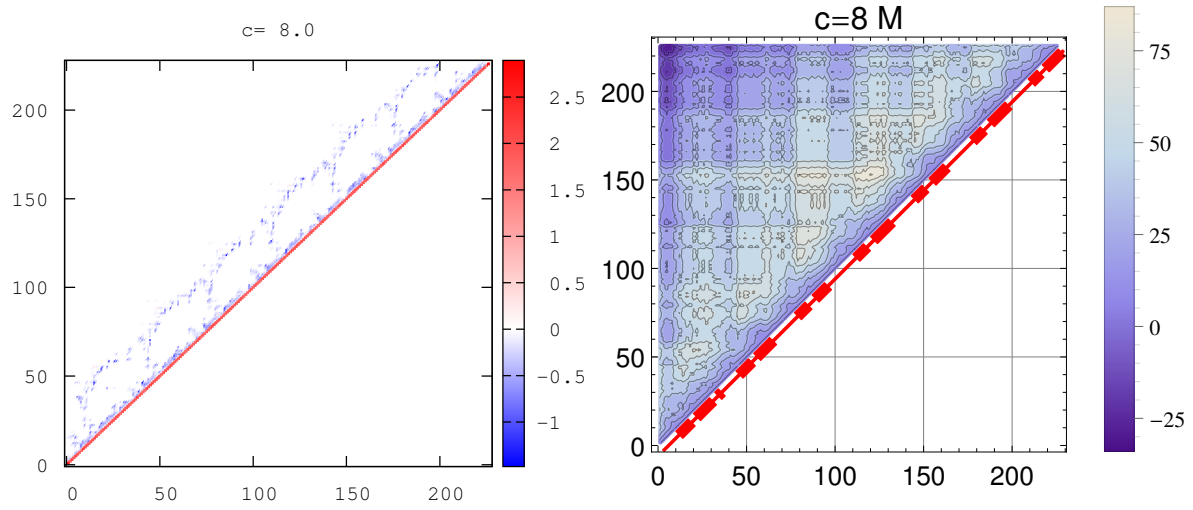
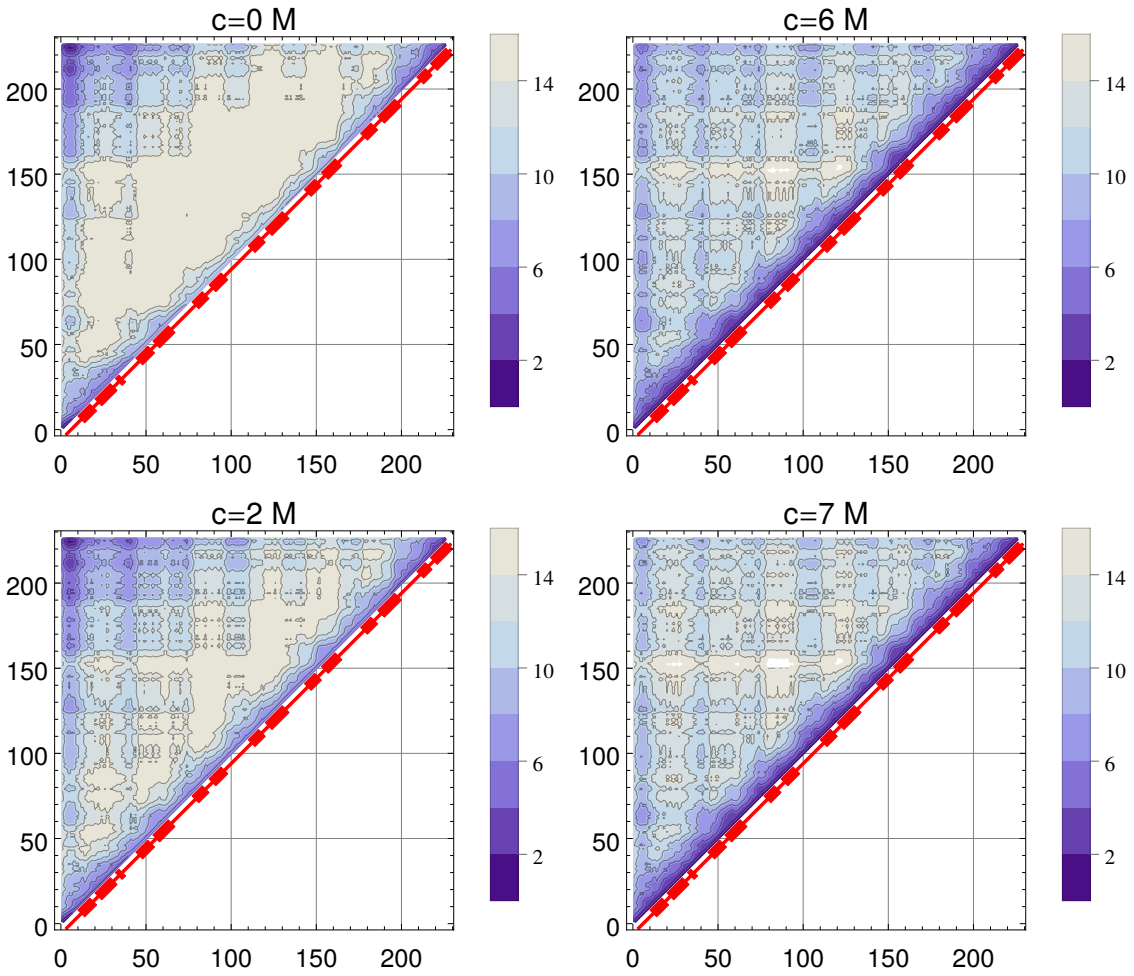


Figure S5: Residue-residue effective interaction energies h_{ij} (*Left*), and cumulative effective energy of all the contacts from region (ij), H_{ij} of Eq.S4 (*Right*). The energies are expressed in kJ/mol.

More detailed information, providing some hints about the position, and not just the

length, of the structured islands in the intermediate states, comes from the analysis of the probability of the isolated native islands, Figure S6:



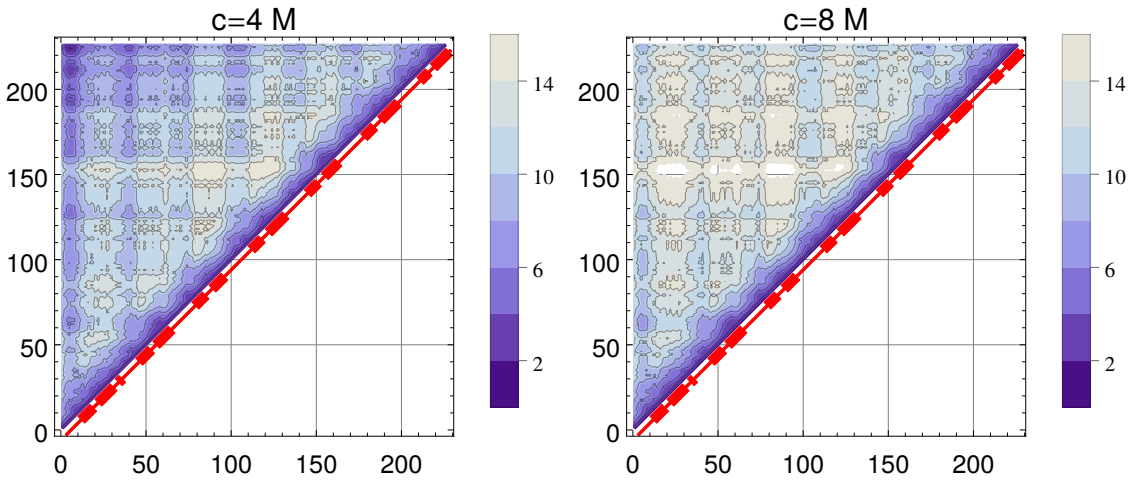


Figure S6: Populations of the "capped" native islands from residue i (x-axis) to j (y-axis), $\nu_{i,j}$ at different denaturant concentrations, together with a scheme of the helices of the secondary structure, along the diagonal. The values of $z = -\log_{10}(\nu_{i,j})$ in the range 0-16 are reported, so that darker colors correspond to most likely regions. At $c = 0$ the most likely structure (at the upper left corner, of coordinates (roughly) (3,223)) encompasses all helices, and corresponds to the fully native state, this is followed by the region (3,223) encompassing all the helices but the last one. Shorter isolated structures, both at the N-terminus and the C-terminus, are highly unlikely. However, at higher denaturant concentrations, the vertical arrangement of darker regions at the left sides of the plots unveils the existence of several isolated structures, marginally stable, encompassing the first two, four, of five repeats. For $c = 4$, C-terminal structures, roughly complementary to and less likely than the N-terminal ones, can be seen. Note that short regions of around 10 residues (i.e. points close to the diagonal: (i,i+10)) can be structured also in strongly denaturant condition, which prompts for isolated helices to be partially structured in the unfolded state.

The analysis of the asymmetry of the structure, Eq. S19, helps to quantify the prevalence of N-terminal structure. Note that the asymmetry profiles at varying M are almost the same at all concentrations c, but the weight S_2 of the ensemble of configurations with a given M deeply changes with c.

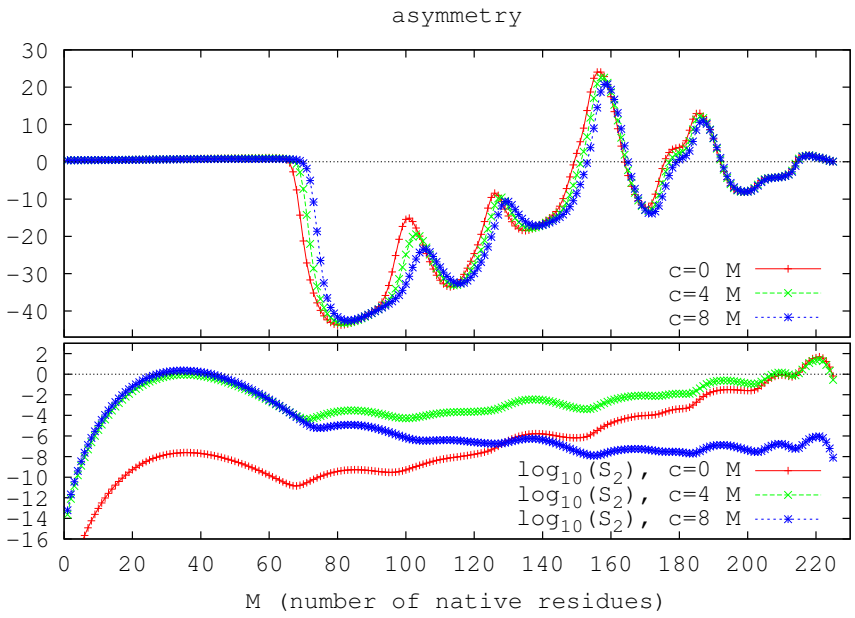


Figure S7: *Top* Plot of the asymmetry $A(M)$ as a function of the number of native residues M, at different denaturant concentrations. $A(M)$ describes the asymmetry N-C of the population at a given total number M of native residues, see Eq.S19. A negative $A(M)$ implies that the ensemble of configurations with fixed total number of native residues M, presents a dominance of structure at the N-terminus. We can see that, for most values of M, a generic configuration with M native residue will be more structured at the first, N-terminal half of the protein than at the second one. *Bottom* The weight S_2 Eq. S20 of the total fraction of native residues at a given M. Notice that while the asymmetry of the top panel is more or less the same at all denaturant concentrations, S_2 reveals that the probability of the configurations with a fixed number of native residues strongly depends on the concentration.

Kinetics

Figure S8 reports the number of clusters as a function of the preference s in folding and unfolding conditions.

For the folding case, a wide plateau, corresponding to 2 clusters, is clearly observed for $s \in [-500000, -30000]$; this suggests that the corresponding clustering is very robust. The

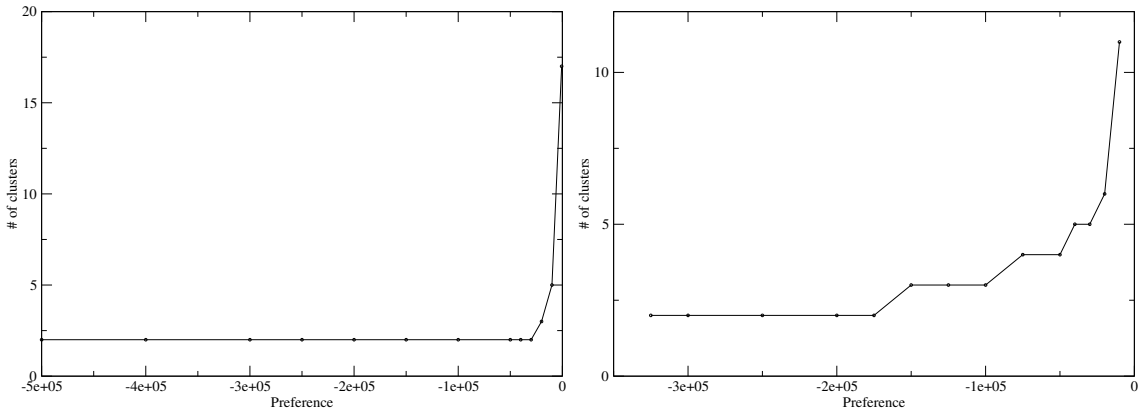


Figure S8: Number of clusters as a function of preference s for folding ($c = 0$M, *left panel*) and unfolding ($c = 8$M, *right panel*) conditions.

2-cluster solution reported here is obtained by setting $s = -50000$. In the unfolding case, upon decreasing the value of the preferences, the 2-cluster solution is preceded by smaller plateaus at 3-4 clusters, reflecting the higher heterogeneity of pathways. In this work we set $s = -225000$ to study the 2-clusters case at $c = 8$. Table S5 reports the representative sequence and number of members of each cluster.

Table S5: Representative trajectories of the different clusters, at $c = 0$M and $c = 8$M, together with their classification as pathways A or B (as defined by the experiments), and the number of trajectories that form the clusters.

c	pathway	sequence	n. of trajectories
0	A	2 3 4 5 6 7 8 9 10 11 12 1 13 14 15	2592
0	B	12 13 11 10 9 8 7 14 6 5 4 3 2 15 1	1665
8	A	13 9 15 10 11 12 7 8 6 14 5 4 1 2 3	2270
8	B	2 1 4 5 7 6 15 9 8 14 10 13 11 3 12	2710

Figure S9 reports the behavior of the Kendall τ_B order parameters for the concentrations $c = 6$, $c = 7$. Note the progressive shift from pathway A to pathway B (the fraction of trajectories with positive values of τ_B increases) as denaturant concentration increases (see also Fig. 7 in the main text).

Table S6 describes in more detail, at the level of groups of repeats, this shift between the two main pathways, which qualitatively agrees with the changes in the flux between pathways A and B emerging from the experiments (even if the predicted fraction of flux

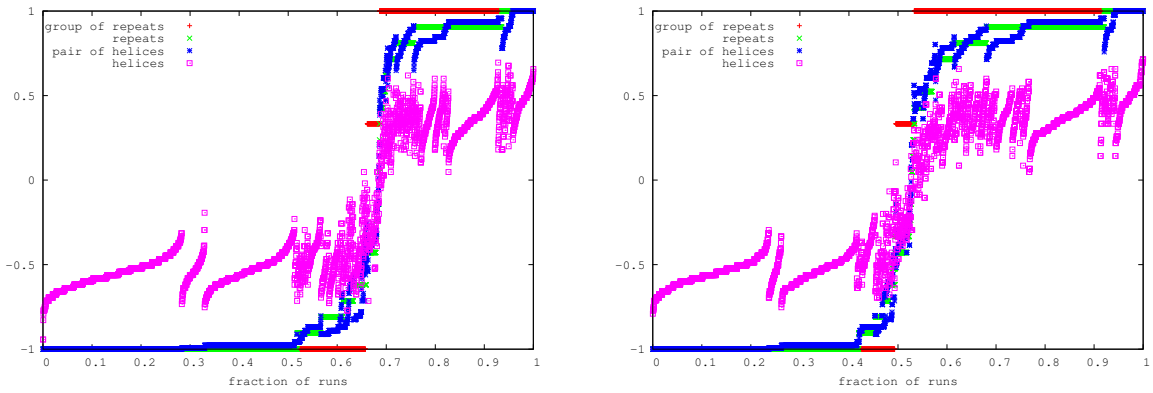


Figure S9: Values of the Kendall τ_B calculated at $c = 6$ (*left panel*) and $c = 7$M (*right panel*), for different choices of (super)secondary structures. x coordinates correspond to trajectories; the latter have been ordered according to increasing values of τ_B(groups), with the ties resolved according to increasing values of τ_B(repeats), and then τ_B(helix-pairs) and finally τ_B(helices). A value $\tau_B = -1$ implies that the structure is sequentially lost from the C-terminus towards the N-terminus ("leftward" ordering), while $\tau_B = 1$ implies the opposite ("rightward") unfolding direction, from the N- to the C-terminus. Intermediate values imply that an overall order is only partially respected.

along each pathway does not reproduce exactly the experimental values).

Table S6: Percentage of trajectories following the different unfolding pathways, using the groups-of-repeats description, at different urea concentrations c. The value of τ_B(groups) is reported along the string specifying the unfolding sequence of events; The orders 213 and 231 are not reported, as we did not find any trajectories where the central group of repeats unfolds before the others. The last column reports the total number of trajectories considered for each urea concentration.

c	123	132	312	321	n
τ_B	1	1/3	-1/3	-1	
6.0	31.8	1.8	0.3	66.2	4329
7.0	47.8	3.5	0.3	48.3	4797
8.0	52.3	5.0	0.6	42.0	4980

Figure S10 reveals the profiles of the effective energies $H_{i,j}$ of Eq. S4 and of the native island population $\nu_{i,j}$ of Eq.S9 along pathway A and B. It is clear that the $\nu_{i,j}$, properly accounting for the configuration entropies, are more useful than the $H_{i,j}$ to understand the structure and relevance of the minima and maxima along each pathway, since the $H_{i,j}$ do not allow to estimate which are the most relevant barriers, even though they reflect the same pattern of minima and maxima as the $\nu_{i,j}$.

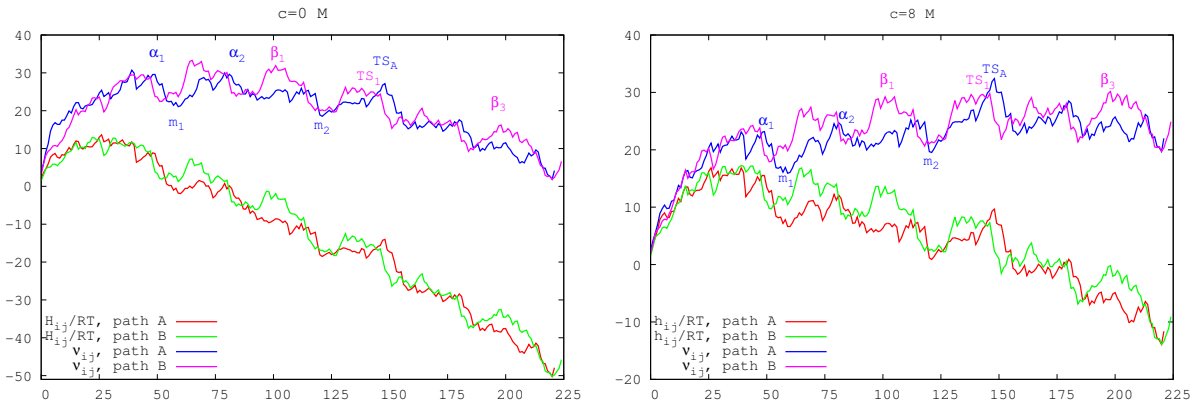


Figure S10: Energy and probability of native regions at low (*left panel*) and high (*right panel*) denaturant concentrations c, along the two pathways. Bottom lines: graphics of $H_{i,j}/RT_0$ along pathway A (the straight line $(5, j)$ in Fig. S5, right panel) and B (the straight line $(i, 225)$). Top lines: Graphics of $z = -\log(\nu_{i,j})$ along pathway A and B (see also Figure 10, right panel, in the main text). In order to make the comparison between the two pathways easier, in the x-axis, the values of $(j - i)$ are reported, so that x=220 corresponds in both cases to the native minimum. Labels refer to the naming of the minima and barriers reported in Fig. 10 in the main text.

References

(1) Wako, H.; Saito, N. *J. Phys. Soc. Jpn.* **1978**, *44*, 1931–1938.

(2) Wako, H.; Saito, N. *J. Phys. Soc. Jpn.* **1978**, *44*, 1939–1945.

(3) Muñoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5872–9.

(4) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11311–6.

(5) Ueda, Y.; Taketomi, H.; Gō;, N. *Biopolymers* **1978**, *17*, 1531–1548.

(6) Street, T. O.; Bolen, D. W.; Rose, G. D. *Proceedings of the National Academy of Sciences* **2006**, *103*, 13997–14002.

(7) Auton, M.; Holthauzen, L. M. F.; Bolen, D. W. *Proceedings of the National Academy of Sciences* **2007**, *104*, 15317–15322.

(8) Edelsbrunner, H.; Koehl, P. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2203–8.

(9) Estrada, J.; Bernadó, P.; Blackledge, M.; Sancho, J. *BMC Bioinformatics* **2009**, *10*, 104.

(10) Bruscolini, P.; Pelizzola, A. *Phys. Rev. Lett.* **2002**, *88*, 258101.

(11) Pelizzola, A. *J. Stat. Mech.* **2005**, P11010.

(12) Zamparo, M.; Pelizzola, A. *J. Stat. Mech.* **2006**, P12009.

(13) Zamparo, M.; Pelizzola, A. *Phys. Rev. Lett.* **2006**, *97*, 068106.

(14) Frey, B. J.; Dueck, D. *Science* **2007**, *315*, 972–6.

(15) Givoni, I. E.; Frey, B. J. *Neural Comput* **2009**, *21*, 1589–600.