Mapping the topography of a protein energy landscape

Richard D. Hutton,[†,#] James Wilkinson,[†] Mauro Faccin,[‡] Elin 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

Abstract

Protein energy landscapes are highly complex, yet the vast majority of states within them tend to be invisible to experimentalists. Here, using site-directed mutagenesis and exploiting the simplicity of tandem-repeat protein structures, we delineate a network of these states and the routes between them. We show that our target, gankyrin, a 226-residue 7-ankyrin-repeat protein, can access two alternative (un)folding pathways. We resolve intermediates as well as transition states, constituting a comprehensive series of snapshots that map early and late stages of the two pathways and show both to be polarized such that the repeat array progressively unravels from one end of the molecule or the other. Strikingly, we find that the protein folds via one pathway but unfolds via a different one. The origins of this behavior can be rationalized using the numerical results of a simple statistical mechanics model that allows us to visualize the equilibrium behavior as well as single-molecule folding/unfolding trajectories, thereby filling in the gaps that are not accessible to direct experimental observation. Our study highlights the complexity of repeat-protein folding arising from their symmetrical structures; at the same time, however, this structural simplicity enables us to dissect the complexity and thereby map the precise topography of the energy landscape in full breadth and remarkable detail. That we can recapitulate the key features of the folding mechanism by computational analysis of the native structure alone will help towards the ultimate goal of designed amino-acid sequences with made-to-measure folding mechanisms - the Holy Grail of protein folding.

Introduction

The folded states of proteins are in dynamic equilibrium with many partially unfolded states, leading directly to functional regulation in some cases. The result is a multitude of conformations that, together with the kinetic barriers that separate them, constitute each protein's energy surface or "landscape"[1]. However, in striking contrast to their complexity, our ability to visualise these energy landscapes has to date been very limited: Most studies have

focused on small, globular proteins, the vast majority of which fold in a two-state manner (i.e. only the native and fully denatured states are populated), and therefore the range of landscape that can be accessed by experiment is extremely narrow and confined to a single, homogeneous transition state ensemble.

Repeat proteins comprise tandem arrays of small structural motifs (20-40 residues) that pack in a roughly linear fashion to produce elongated and super-helical architectures[2–4]. They are composed of only short-range interactions, between residues within a repeat or in adjacent repeats, and in this way they contrast with globular proteins whose stabilities rely on multiple sequence-distant interactions frequently resulting in complex topologies. The folding and function of repeat proteins have been studied by both experiment and simulation[5–24], and they have been found to possess certain features that distinguish them from the more commonly studied globular proteins and that arise from the symmetry inherent in their structures and the absence of long-range interactions. In particular, the modularity of repeat proteins leads to relatively easy dissection of their biophysical properties and consequently they are highly amenable to redesign - of their thermodynamic stability, folding mechanisms and molecular recognition.[14,15,25–37].

The 226-residue gankyrin is an oncoprotein involved in multiple protein-protein interactions and a negative regulator of principal tumour suppressors p53 and pRB[38]. Gankyrin has seven repeats of the ankyrin motif, which comprises a β-turn followed by two anti-parallel α-helices and a loop (Fig. 1). Here we use site-directed mutagenesis to map out the folding energy landscape of gankyrin. We observe that the mutant chevron plots are strikingly different in shape from that of the wild type, behavior that cannot be explained by a simple folding mechanism. Instead, the results are consistent with two alternative pathways in which the ankyrin repeats fold/unfold from either the N-terminus or the C-terminus. Importantly, for both pathways we are able to characterise intermediate states in addition to transition states, and consequently we can acquire a comprehensive set of snapshots of both early and late stages of the reactions. Remarkably, the results show that the protein folds

via one pathway but unfolds via the other.

In order to understand the physical basis of the experimental findings, we perform a theoretical analysis involving a simple statistical-mechanics model, a modification of the WSME model[39–42], which we use to characterize both the equilibrium and the kinetics at the single-molecule level. We delineate the structures of the metastable equilibrium states, and we follow the folding and unfolding of single-molecule trajectories at different denaturant concentrations. We show that this very simple model using only native contacts is able to recapitulate all of the key experimental results, namely the greater stability of the N-terminal versus the C-terminal repeats, the order in which the repeats fold and unfold, pathway heterogeneity, and the difference in the pathway for folding versus unfolding. Moreover, we are able to fill in the details that are not accessible experimentally, allowing us to explain the physical basis of the experimental results. This work helps towards the ultimate goal of designed proteins in which one can dial in to the amino-acid sequence a made-to-measure folding mechanism.

Results

Equilibrium unfolding of wild-type gankyrin and mutant variants: Fluorescence and far-UV circular dichroism (CD) were used to monitor the urea-induced unfolding of gankyrin. Gankyrin has two tryptophan residues, located at positions 46 (repeat 2) and 74 (repeat 3) (Fig. 1). The refolding denaturation curve, monitored by fluorescence, is in agreement with the unfolding denaturation curve, indicating reversibility. The denaturation curve obtained at an emission wavelength of 341 nm (or indeed at other wavelengths) can be fitted to a two-state equation with a midpoint of unfolding of 4.1±0.1 M urea and an m-value of 2.6 ± 0.1 kcal mol^{-1} M^{-1}(Fig. 1); the free energy of unfolding in water is 10.8 ± 0.2 kcal mol^{-1}. The same value was obtained when the fluorescence data were plotted at other wavelengths. The denaturation curve obtained using the ellipticity at 222 nm to monitor
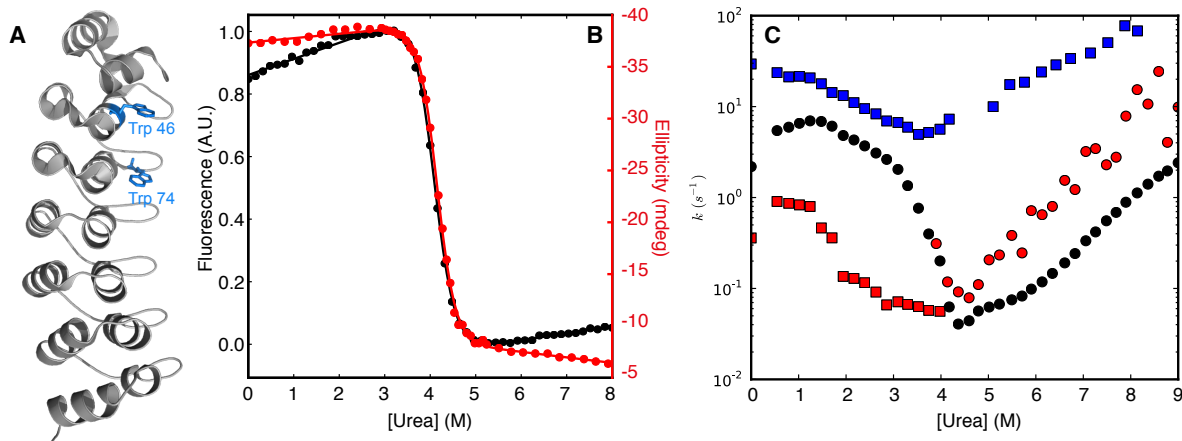
Figure 1: Equilibrium unfolding of wild-type and mutant gankyrin. (A) Schematic of the structure of gankyrin showing the location of the two tryptophan residues. (B) Denaturation curves of wild-type gankyrin monitored by fluorescence at an emission wavelength of 341 nm and by CD at 222 nm. (C) Urea dependence of the rate constants of refolding and unfolding of wild-type gankyrin, 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 (in blue) were obtained using interrupted refolding experiments.

helical structure can also be fitted to a two-state equation and it gives the same midpoint and m-value as those obtained using fluorescence, indicating that secondary and tertiary structure are lost concomitantly upon unfolding and that the two tryptophan residues are reporting on a global structural change (Fig. 1).

Nineteen conservative (non-disruptive) single-site mutations were made throughout the structure. These include mutations of alanine to glycine, and we have confirmed (by measuring the CD spectra of the mutants and showing that they overlay with that of the wild type) that neither the helical content nor structure of gankyrin is perturbed by this type of mutation. As for the wild type, the fluorescence-monitored denaturation curves of the mutants could be fitted to a two-state equation (Fig. S1 and Table S3).

The unfolding and refolding kinetics of gankyrin are multiphasic: The unfolding and refolding kinetics of gankyrin were monitored over a range of urea concentrations using stopped-flow fluorescence and stopped-flow far-UV CD. The refolding kinetics monitored by fluorescence can be fitted to the sum of three exponential phases and the unfolding kinetics

to the sum of two exponential phases (Fig. S2A,B; Fig. 1C). When the kinetics is monitored by CD, one unfolding phase and two refolding phases are observed, and the rate constants for these phases are in agreement with those of the major phases observed by fluorescence (Fig. S2C,D; Fig. S3A).

Endpoint analysis of the kinetic traces monitored by fluorescence shows that there is a small deviation of the start-point for refolding from the value predicted by linear extrapolation of the endpoint of the unfolding reaction (Fig. S3C). However, endpoint analysis of the CD data shows no deviation between initial and final signals, indicating that the burst-phase species has little native α-helical structure (Fig. S3D).

A positive rather than negative denaturant dependence of the rate constant of the major refolding phase is observed at very low urea concentrations, suggesting the formation of a misfolded state that needs to unfold in order for folding to proceed. This behavior cannot be explained by transient oligomerisation as there was no protein concentration dependence of the refolding kinetics from sub-micromolar up to 20 μM. It may instead be due to mis-docking of the repeats or of the individual helices, as observed previously[43], and for a number of other large repeat proteins including the 12-ankyrin-repeat D34 and a consensus-designed tetratricopeptide repeat proteins comprising 10 repeats.

We next measured the folding and unfolding kinetics of the mutant proteins by stopped-flow fluorescence. For all of the mutants, the refolding kinetics could be fitted to the sum of three exponential phases that have similar relative amplitudes to those of the wild type; the unfolding kinetics can be fitted to the sum of two exponential phases, again with similar relative amplitudes to those of the wild type. Chevron plots of representative mutants are shown in Fig. 2B-E together with the wild type.

Interrupted refolding experiments permit identification and characterisation of a folding intermediate: Sequential mixing (double-jump) experiments can be used to resolve the origins of kinetic heterogeneity and determine the nature of the different phases
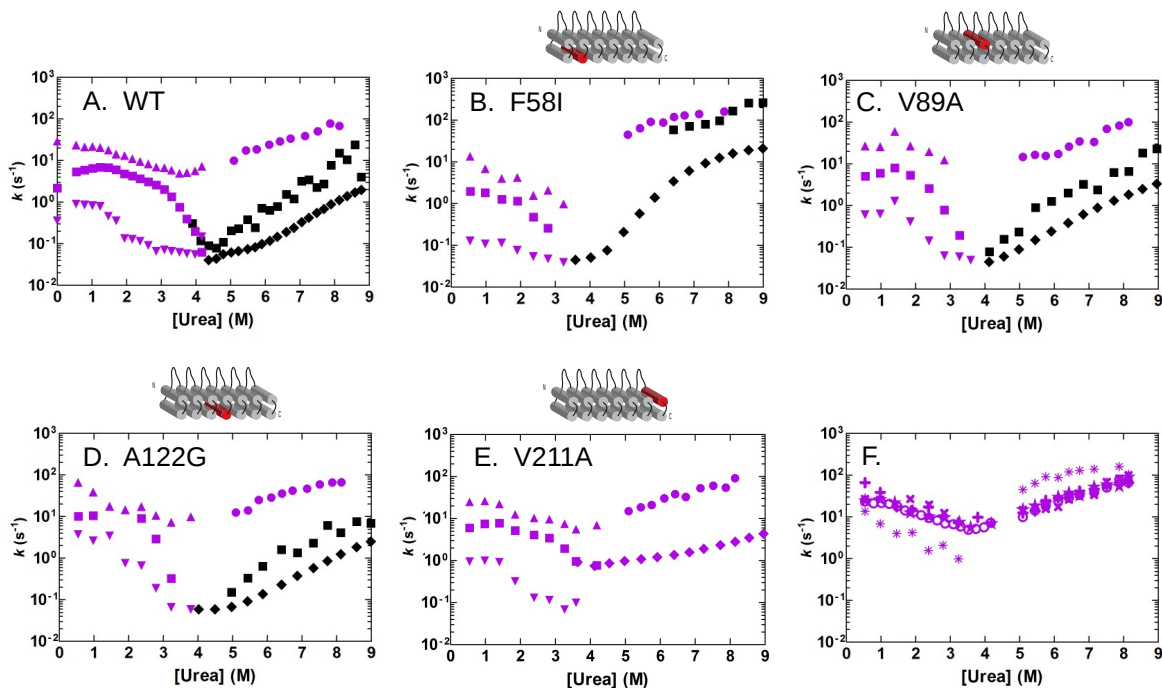
Figure 2: (B)-(E) Plots of the urea dependence of the unfolding and refolding rate constants (measured by stopped-flow fluorescence) for four representative mutants - F58I (repeat 2), V89A (repeat 3), A122G (repeat 4) and V211A (repeat 7). For comparison, wild type is shown in (A). As described in the Discussion, those phases that we can assign exclusively to Pathway A are shown in purple; the other phases are shown in black. The three refolding phases are shown in triangles, squares and inverted triangles. The fastest unfolding phase (circles) was detected by interrupted refolding experiments; the two other unfolding phases shown (squares and diamonds) were those detected in the normal single-jump set-up. (F) Comparison of the rate constants for the fastest refolding/unfolding phase for wild type and the four representative mutants. The mutant F58I that has the largest effect on this phase is shown with asterisk symbols.

7

observed by conventional, single-jump stopped flow (e.g. [44]). Interrupted unfolding helps to resolve whether or not slow refolding phases originate in heterogeneous populations of unfolded molecules (often arising from cis/trans proline isomerisation). Interrupted refolding monitors specifically the formation of native molecules, and it can therefore be used to resolve which refolding phases corresponds to formation of the native state and which to formation of a partly folded intermediate or of a native-like intermediate that is hard to distinguish from the native state by conventional spectroscopic means. For interrupted unfolding experiments, native gankyrin was allowed to unfold in 6.25 M urea for a variable delay time and then rapidly transferred into refolding conditions (final urea concentration of 0.87 M) and the kinetics monitored. The refolding traces obtained after the different unfolding delay times were fitted globally to the sum of three exponential phases, sharing the rate constants and allowing the amplitudes to vary. The rate constants obtained for the three phases were: k_1=6$\pm$1 s^{-1}, k_2=18$\pm$1 s^{-1} and k_3=0.6$\pm$0.1 s^{-1}. These values are in agreement with the rate constants observed in the single-jump refolding experiment under the same conditions (k_1=6 s^{-1}, k_2=16 s^{-1} and k_3=0.4 s^{-1}). The accumulation of amplitude of the major refolding phase (k_1) as a function of unfolding delay time could be fitted to a single exponential phase with a rate constant of 0.20$\pm$0.03 s^{-1} which is close to the value of 0.1 s^{-1} obtained for the major unfolding phase in the single-jump experiment at 6.25 M urea (Fig. S4). The plots of the amplitudes of the two minor refolding phases as a function of unfolding delay time were very scattered because of their small amplitudes (always 5% of the total amplitude of the refolding reaction) and therefore they could not be fitted.

In the interrupted refolding experiments, denatured gankyrin in 7.6 M urea was allowed to refold in buffer at a final urea concentration of 2 M for a variable delay time. The protein was then rapidly transferred into unfolding conditions (final urea concentration of 7 M) and the kinetics monitored. After the shortest refolding delay time of 30 ms, two unfolding phases were observed with rate constants of 40$\pm$2 s^{-1} and 1$\pm$0.3 s^{-1}. After longer refolding delay times only a single unfolding phase was observed, having a rate constant of 1$\pm$0.2

s^{-1}. This phase is not the same as either of the two phases observed in the single-jump unfolding experiment at 7 M urea (0.3 s^{-1} (major) and 1.5 s^{-1} (minor)). The plot of the amplitude of the 1 s^{-1} unfolding phase versus refolding delay time could be fitted to the sum of two exponential phases with rate constants of 3.3±0.3 s^{-1} and 0.09±0.01 s^{-1} (Fig. S4); these values are in agreement with the rate constants of the two slower refolding phase in 2 M urea (3.6 s^{-1} and 0.1 s^{-1}) measured in the single-jump experiment. When refolding was performed in a manual mixing experiment and the reaction was allowed to proceed for several minutes before the protein was then unfolded, the unfolding kinetics was in good agreement with that observed in the single-jump unfolding experiment. Therefore we conclude that refolding occurs to a native-like state, which slowly converts to the native state and which has the same fluorescence as the native state and therefore this conversion is not observed in single-jump refolding experiments. The 40 s^{-1} unfolding phase was not observed in the single-jump experiment and was observed in the double-jump experiments only after a short refolding delay time, suggesting the transient accumulation of an intermediate species in the refolding reaction. The urea dependence of the rate constants for this phase was next measured. The interrupted refolding experiment was performed under the same conditions as described above, using the shortest delay time of 30 ms in order to maximally accumulate the intermediate and a range of different urea concentration in the unfolding step (shown in blue in Fig. 1C). This unfolding phase appears to correspond to the same transition state as the fastest phase recorded in the (single-jump) refolding experiments, indicating that these two phases observed for refolding and unfolding correspond to the formation and decay, respectively, of the same intermediate state.

In summary the double-jump experiments show that gankyrin folds *via* an intermediate state, I, to a native-like state and that the conversion of the latter state to the native state occurs on a slow timescale. The fastest of the three refolding phases corresponds to the formation of I. The major refolding phase correspond to the formation of the native-like state from I. Native-like species, present at very low populations at equilibrium, have been

observed previously in the literature and have in some cases been shown to differ from the native state in the isomerisation of a peptidyl-proline bond. The endpoint analysis suggests that there is in addition a burst-phase species in the refolding reaction, albeit having little α-helical structure. The inflections observed in the urea dependence of the amplitudes of the major refolding phase and the fast minor refolding phase (Fig. S3B) are also consistent with the population of multiple intermediates, as observed in other folding studies (e.g. [45]).

Discussion

Asymmetric distribution of stability between N- and C-terminal regions: There is some variability in the equilibrium m-values of the mutants (Fig. 3 and Table S3). Mutation can have two potentially opposite effects on the size of the m-value: first, the m-value may increase as the midpoint decreases, an effect that has been attributed to non-linearity in the denaturant dependence of the free energy of the unfolding[46,47]; an alternative explanation for an increase in m-value on mutation is the denatured state becoming less compact due to the disruption of residual interactions. Second, if there is an intermediate that is weakly populated and the mutation is at a site that is structured in the native state but not in the intermediate, then the mutation will destabilize only the native state and thereby increase the relative population of the intermediate relative to the native state; this behavior may be manifest in a lowering of the observed m-value compared with the wild type when the denaturation curve is fitted to a two-state model. There was no correlation between m-value and midpoint of unfolding of the mutant proteins; however, when the m-values of the mutant proteins were plotted against position in the sequence a trend could be detected (Fig. 3). The m-values were higher than that of wild type for mutants at sites in the N-terminal three repeats, and lower than that of wild type for mutants in the C-terminal four repeats. The m-values in the two regions were compared using using unequal variance student t-test, which showed that that the two populations were significantly different (P<0.05) (FIG 3C).

Figure 3: (A) Plot of m-value versus mutation. (B) Plot of midpoint of unfolding versus mutation. The red line in (A) and (B) shows the wild-type m-value and midpoint, respectively. Mutations in ankyrin repeats 1-3 are coloured gray, and repeats 4-7 are coloured white. (C) The m-values in the two regions were compared using using unequal variance student t-test, which showed that that the two populations were significantly different ($P<0.05$).

This result suggests that there is an intermediate in the equilibrium unfolding of gankyrin in which repeats 1-3 are structured and the other C-terminal repeats are at least partly unstructured. Since the m-value for wild type is of a magnitude expected for a protein of that size (e.g. the proteins p16 and myotrophin, both comprising four ankyrin repeats, have m-values of 1.7 ± 0.1 kcal mol^{-1} M^{-1} and 1.9 ± 0.1 kcal mol^{-1} M^{-1}, respectively, Notch 7-ankyrin domain has an m-value of 2.9 ± 0.1 kcal mol^{-1} M^{-1}) and consensus tetratricopeptide repeat proteins comprising two or three repeats have m-values of 1.4 ± 0.1 kcal mol^{-1} M^{-1} and 1.5 ± 0.1 kcal mol^{-1} M^{-1}, respectively, indicating that there is a reasonable correlation between m-value and protein size, as for globular proteins), then we conclude that the intermediate is only very weakly populated.

The kinetic data are not compatible with simple folding models: The chevron plots of wild type and mutants (Fig. 2) display striking complexity, with a number of features that we next attempt to fit using different kinetic models. These features are: (1) the unfolding arm of the chevron plot is not linear, and depending on the variant, we observe both upward curvature at intermediate urea concentrations (around 6 M for wild type) and downward curvature at higher urea concentrations; (2) the shape of the unfolding arm differs dramatically between different mutants and between mutants and the wild type; the mutants can be grouped into three broad categories that appear to correlate with their positions along the ankyrin repeat stack (see Fig. 4): More pronounced downward curvature in the unfolding arm relative to that of the wild type was observed for mutants in the N-terminal two repeats 1 and 2 (e.g. A14G, F58I). For mutants in the central three repeats (e.g. A80G, A122G and A147G) the unfolding arm had a similar shape to that of the wild type. For mutants in the C-terminal two repeats 6 and 7 (e.g. A188G, L209A) the unfolding arm showed little or no downward but did show upward curvature; (3) there is downward curvature in the refolding arm at low urea concentrations; (4) as well as the above features of the major phase, we have also attempted to account for the fast minor phase observed for refolding in single-jump

experiments and for unfolding in double-jump experiments (interrupted refolding) (the blue symbols in Fig. 1C). We excluded the data at the lowest urea concentrations, where we see a positive rather than a negative denaturant dependence of the refolding rate constants (we showed that the rates were not concentration dependent and therefore that this was not due to oligomerisation or aggregation); its origins are beyond the scope of this paper and will be investigated in future work.

To fit and evaluate different kinetic models we developed the software *PyFolding*[48] (Supplemental File). Looking first at the wild-type chevron plot we tested the following two simple kinetic models: a two-state model, and a three-state model in which there is a fast pre-equilibrium with a folding intermediate. Only the three-state model is able to capture the downward curvature in the refolding arm. A third model - three-state with fast phase - is able to capture the minor, fast refolding/unfolding phase in addition to the major phase.

However, when we turn to the mutants, this simple model fails dramatically to capture the details of their chevrons if we assume that the positions of the intermediate and transition states (m-values) are invariant upon mutation (i.e. the pathway does not change). Two representative mutants, A14G and L209A, one at each end of the protein, are shown in the Supplemental File, and other mutants likewise cannot be fitted by these simple models. This indicates that there are dramatic changes in m-values upon mutation, and therefore we need to consider more complex schemes. In other words, when we look at wild type on its own then a simple model is sufficient, but when we look at the whole data set of wild type and mutants in aggregate the simple models do not capture all of their features. Given that the end states are the same, we must conclude that there are multiple parallel pathways through the energy landscape.

Because of the complexity of the entire chevron, we focus next for simplicity on the unfolding arm; we need a model involving paralel pathways that is able to capture the dramatic changes in shape upon mutation that we showed above cannot be captured with a simple, single pathway model[13,15,49,50]. The downward curvature in the unfolding arm that is clear

for mutants such as A14G at high denaturant concentrations can be captured with a sequential barriers model, in which there is a switch upon increasing denaturant concentration between two sequential transition states separated by a high energy, metastable intermediate[51,52]. We therefore use a scheme comprising two alternative pathways, A and B, wherein for pathway A we assume a linear relationship between unfolding rate and urea concentration and for pathway B we assume a sequential barriers model (as drawn schematically in Fig. 10, described later). We define the microscopic rate constants $k_{N \to I'}, k_{I' \to N}$, and $k_{I' \to D}$ as the rate constants for transitions between the Native (N), metastable intermediate (I') and Denatured (D) states, according to the scheme:

$$D \underset{k_{I \to D}}{\overset{k_{D \to I}}{\rightleftharpoons}} I' \underset{k_{N \to I}}{\overset{k_{I \to N}}{\rightleftharpoons}} N \tag{1}$$
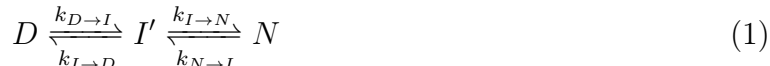
(I' indicates that this metastable unfolding intermediate is different from the intermediate, I', that is transiently populated under refolding conditions). For Pathway B, under unfolding conditions where either of the two sequential transition states (TS1 and TS2) are always rate limiting, we define the unfolding rates as $k_u^{TS1} = \frac{k_{N \to I'}}{k_{I' \to N}} \cdot k_{I' \to D}$ and $k_u^{TS2} = k_{N \to I'}$. As in ref[49] we assume that the intermediate species is always metastable by setting the values of a $k_{I' \to D} = 1 \times 10^4$ s^{-1} and $m_{I' \to D} = 0$ M^{-1}. The experimental data for wild type and mutants were fitted to the sum of the rates of unfolding through pathways A and B, by globally sharing m-values but allowing all rate constants to vary freely (constrained to be positive values and that ΔG_{D-N} for the two pathways were the same). Initial parameters for the fitting were chosen by free fitting of the wild-type data. Fractional flux through each pathway is calculated as: $\rho_A = \frac{k_u^A}{k_u^A + k_u^B}$ and $\rho_B = 1 - \rho_A$.

The fit of the unfolding data to this parallel pathways model is shown for wild type and a set of representative mutants in Fig. 4 together with the hypothetical unfolding arms for each pathway and the flux through each pathway as a function of urea concentration. The β-Tanford value for the transition state of pathway A is 0.89, and 0.26 and 0.91 for TS1 and
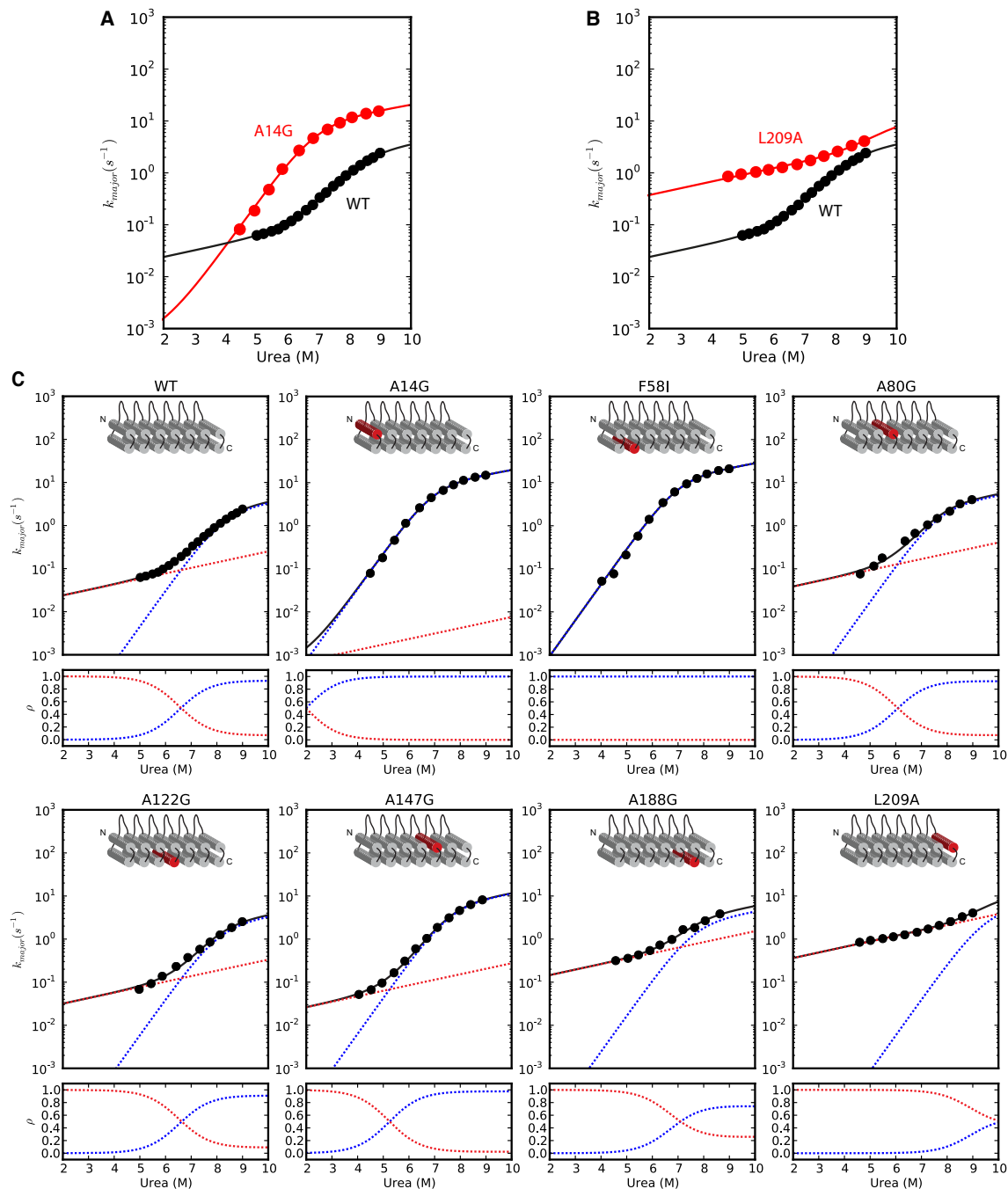
Figure 4: Top shows the urea dependence of the unfolding rate constants for (A) an N-terminal mutant and (B) a C-terminal mutant with wild-type gankyrin for comparison, highlighting their strikingly different shapes. (C) The unfolding rate constants for wild-type gankyrin and representative mutants (one in each repeat) fitted to a parallel pathways model. The fit of the unfolding data to the parallel pathways model is shown in black. In red is the hypothetical unfolding arm corresponding to pathway A and in blue is the hypothetical unfolding arm corresponding to pathway B. Below the main plots are the plots of the fractional fluxes through the pathway A (dashed red line) and pathway B (dashed blue line) pathways. All data shown are those obtained by stopped-flow fluorescence.

TS2, respectively, of pathway B.

Φ–value analysis maps out the transition state structures for the two alternative pathways: As described above, the qualitative picture that we can obtain by visual inspection of the unfolding arms, shows that the mutants differentially affect (i.e. destabilise) the two unfolding pathways depending on their location along the repeat array. The energetic effects of the mutations on the two pathways can be quantified with Φ–values, which allow us to infer the structures of the transition states along these pathways. We calculate the Φ–values by using the rate constants obtained from the fits of the unfolding data to the parallel pathways model (Fig. 4 and Table 1). As stated earlier, the data for all mutants and the wild type were fitted globally, sharing the m-values but allowing all rate constants to vary freely. A Φ–value of 1 indicates that the mutation destabilises the transition state by as much as the native state, from which we can infer that the site of mutation is as highly structured in the transition state as in the native state. Conversely, a Φ–value of 0 indicates that the mutation does not destabilise the transition state, from which we can infer that the site of mutation is as unstructured in the transition state as in the denatured state. For mutations in the N-terminal two repeats, we can see from the parallel pathways fit that pathway A is almost completely depopulated over the whole urea range in which unfolding was measured, and therefore the rate constants cannot be accurately defined. However, this behavior does tell us, qualitatively at least, what the Φ–values are: if the mutations destabilize the transition state for pathway A to such an extent that they shift the flux almost exclusively to pathway B, then this part of the protein must be highly structured in the transition state for pathway A, i.e. the Φ–values are high. The converse is true for mutations in the C-terminal two repeats. When we look across all the mutants the structural map that we obtain from the Φ–values is as follows: along the sequential barriers pathway (pathway B), TS1 is highly polarized with the N-terminal moiety unstructured and the C-terminal moiety highly structured. TS2 has higher Φ–values than does TS1, indicating

that structure is lost progressively along this unfolding pathway. The transition state along pathway A shows the opposite pattern of polarization, with the N-terminal moiety being the more highly structured; only the C-terminal two repeats have Φ–values that are not 1. The central repeats (3, 4 and 5) are fully or highly structured both in the transition state of pathway A and in TS2 of pathway B and they are partly structured in TS1 of pathway B. We show in the next sections that, like for unfolding pathway B, the structure progressively unravels along unfolding pathway A also: we show that we can delineate an intermediate in which repeats 1-3 are highly structured and the transition state for its folding/unfolding has repeats 1-2 structured.

Lastly, for pathway A the Φ–values for residues in repeat 7 are less than zero, and for pathway B the Φ–values for residues in repeats 1 and 2 are less than zero. One interpretation of this type of non-classical Φ–values is that there are non-native interactions in the transition state: the transition state is stabilized by the mutation whereas the native state is destabilized, giving rise to faster than expected unfolding rates. Alternatively, negative Φ–values can arise if the mutation has too small an effect on the equilibrium stability, but that is not the case here (in all cases where negative Φ–values were observed the change in free energy of unfolding was greater than 0.9 kcal mol^{-1}).

Interrupted refolding of the mutants shows that the kinetic folding intermediate has structured N-terminal ankyrin repeats: A trend can be observed in the effects of the mutations on the fastest refolding phase, which corresponds to folding to the intermediate state I (see the representative mutants in Fig. 2). Mutations in repeats 1 and 2 result in a decrease in the rate constant, whereas mutations in all of the other repeats have rate constants that are similar to those of the wild type. This observation suggests that only repeats 1 and 2 are structured in the transition state for the formation of the intermediate. The effects of the mutations on the rate constants of this phase are relatively small compared with the effects of the mutations on the stability of the native state, suggesting that the

Table 1: Kinetic parameters derived from the fit of the wild-type and mutant data to the parallel pathways model. All experiments were carried out in 50 mM Tris-HCl buffer pH 8.0, 5 mM DTT at 25°C and a protein concentration of 2 μM. The data were fitted globally, sharing the m-values, which were as follows: for pathway A $m_u = 0.29 \pm 0.02$; for pathway B $m_{I' \to D} = 0.3 \pm 0.8$, $m_{I' \to N} = 1.3 \pm 0.8$, $m_{N \to I'} = 0.24 \pm 0.01$. *For these mutants pathway A is not sufficiently populated for an accurate determination of the kinetic parameters, but this nevertheless indicates that the Φ–values are high. #For these mutants pathway B is not sufficiently populated for an accurate determination of the kinetic parameters, indicating that these Φ–values are high. ‡The Φ–value has been calculated for A188G in the context of V211A.

Protein	Pathway A k_u (s^{-1})	Pathway B k_u^{TS1} (s^{-1})	k_u^{TS2} (s^{-1})	A Φ_u	B Φ_u^{TS1}	Φ_u^{TS2}
WT	0.013 ± 0.003	3.1e-07 ± 2.3e-07	0.3 ± 0.06			
ANK 1						
V10A	0.004 ± 0.0055	5.1e-06 ± 3.7e-06	1.5 ± 0.23	-	-0.27	0.28
A14G*	0.0004 ± 0.0065	1.6e-05 ± 1.1e-05	1.7 ± 0.26	-	-0.027	0.54
I26A*	5.3e-05 ± 0.0054	1.9e-05 ± 1.4e-05	1.8 ± 0.27	-	-0.051	0.55
ANK 2						
A43G*	9.3e-08 ± 0.0076	2.7e-05 ± 1.9e-05	2.5 ± 0.37	-	-0.1	0.48
A47G*	0.005 ± 0.0076	1.1e-05 ± 7.6e-06	1.1 ± 0.16	-	-0.17	0.58
F58I*	9.6e-08 ± 0.0066	2.0e-05 ± 1.4e-05	2.5 ± 0.38	-	-0.089	0.45
ANK 3						
I79V	0.0095 ± 0.0049	2.6e-06 ± 1.9e-06	1 ± 0.17	1.2	-0.33	0.22
A80G	0.022 ± 0.0052	1.2e-06 ± 8.6e-07	0.45 ± 0.076	0.87	0.63	0.89
A81G	0.027 ± 0.0052	1.7e-06 ± 1.2e-06	0.42 ± 0.067	0.86	0.66	0.94
V89A	0.019 ± 0.0045	1.2e-06 ± 8.5e-07	0.39 ± 0.059	0.88	0.52	0.91
A91G	0.019 ± 0.0065	4.0e-06 ± 2.8e-06	0.74 ± 0.12	0.88	0.14	0.7
ANK 4						
A113G	0.025 ± 0.0054	7.0e-07 ± 5.2e-07	0.32 ± 0.059	0.85	0.8	0.98
A122G	0.018 ± 0.0043	4.0e-07 ± 3e-07	0.3 ± 0.058	0.9	0.91	1
ANK 5						
A147G	0.015 ± 0.0044	3.0e-06 ± 2.1e-06	1 ± 0.17	0.98	0.57	0.77
I155V	0.012 ± 0.0035	3.2e-07 ± 6.8e-07	0.33 ± 0.66	1.1	0.97	0.92
I157V	0.014 ± 0.0036	3.6e-07 ± 2.6e-07	0.32 ± 0.067	0.95	0.87	0.94
ANK 6						
A188G	0.081 ± 0.012	1.1e-06 ± 8.4e-07	0.39 ± 0.072	0.23	0.44	0.89
A188G/ V211A#‡	0.71 ± 0.078	5.8e-05 ± 6.1e-05	0.67 ± 0.11	0.061	-	-
ANK 7						
L209A#	0.2 ± 0.028	5.9e-08 ± 6.7e-08	0.44 ± 0.4	-0.71	-	-
V211A#	0.22 ± 0.033	5.2e-08 ± 6.8e-08	1.5 ± 1.7	-0.87	-	-

interactions are only partly formed. Folding from the burst-phase intermediate detected by end-point analysis would also explain the small effects of the mutations on the rate of formation of I. The intermediate I resembles the intermediate state detected at equilibrium: the low m-values of mutants in repeats 4-7 indicate an equilibrium intermediate which is structured in repeats 1-3 (Fig. 3).

In order to investigate further the structure of the intermediate and of the transition state for its unfolding, the interrupted refolding experiment (in which the denatured protein is allowed to refold for a very short period of time (30 ms) in order to populate the folding intermediate and then the unfolding of this intermediate is initiated by mixing a range of different urea concentration) was next performed for four representative mutants. The urea dependence of this unfolding phase is shown in Fig. 2F together with the fastest refolding phase which corresponds to the formation of the intermediate. For three of the four mutations (located in repeats 3-7) both the rate of folding to the intermediate and the rate of unfolding from the intermediate are the same as the wild-type values, which indicates that the site of mutation is unstructured in the transition state between the unfolded state and the intermediate. In contrast, for the other mutation, located in repeat 2 (F58I), folding was slower and unfolding faster than wild type, indicating that this site is partly structured in the transition state. These results show, therefore, that only the N-terminal part of the protein has some (weak) structure in the transition state for unfolding of I.

A different pathway dominates for folding versus unfolding: Mutations at opposite ends of the protein have strikingly different effects not only on the major refolding phase/unfolding phase but also on the minor phases (see mutants F58I and V211A in Fig. 2B and E respectively). The behavior of the mutants allow us to put together a complete picture of gankyrin's energy landscape over the whole range of reaction conditions. As discussed above, F58I has a pronounced effect on the folding and unfolding phases that correspond, respectively, to formation and decay of the intermediate; the mutation also slows

down the slower, major refolding phase. These observations suggest that under native conditions, gankyrin folds via the N-polarised pathway (i.e. the pathway in which the N-terminal repeats are structured in the intermediate/transition state and the C-terminal repeats are unstructured - abbreviated subsequently to N-path). This picture is consistent with the fitting of the unfolding data of wild type and mutants to the parallel pathways model (Fig. 4), which indicates that the N-path is favored under mildly denaturing conditions and there is a switch to the C-path under strongly denaturing conditions. For destabilizing mutations in the C-terminal repeats, the C-path is selectively destabilized and the route is shifted to the N-path throughout virtually the whole unfolding urea range; the converse shift of route is seen for destabilizing N-terminal mutations. The minor unfolding phase (in the single-jump unfolding experiments) shows mostly small perturbations upon mutation, but larger effects are observed for mutations at the termini such as F58I and V211A; the phase is greatly speeded up for F58I and is absent for V211A; these observations suggest that this phase is associated with a C-type path. The slowest, minor refolding phase also shows relatively small perturbations upon mutation with the exception of N-terminal mutations, such as F58I, which slow it down significantly and therefore point to an N-type path. In summary, the kinetics are multiphasic, and although it is not possible for us to quantitatively model all of these phases, they are nevertheless consistent with a picture in which gankyrin folds along a different route from that along which it unfolds, i.e. what folds first does not necessarily unfold last (discussed further in the Summary). We note that in the interrupted refolding experiments folding at 2 M urea occurs via pathway A, and so when highly denaturing conditions (7 M urea) are applied after a very short refolding time, the starting state is now the N-terminally structured intermediate and the unfolding reaction then proceeds along pathway A, rather than predominantly along pathway B as when starting from the native state in the single-jump experiments. So, unlike single-jump unfolding, which proceeds along pathway B predominantly, we are watching unfolding along pathway A in the interrupted refolding experiments performed with short delay times.

Simulations reproduce the equilibrium experiments: We next performed a theoretical analysis using a simple statistical mechanics model, in order to provide an independent and microscopic insight, at the residue level, of the equilibrium and kinetics of the folding process of gankyrin. As explained in the Methods, we fix the three model parameters (related to the interactions between residues) on the equilibrium signals of the wild type, and then use the model to predict other equilibrium and kinetic quantities. Fig. 5A shows the model fluorescence and CD signals obtained after fitting the parameters of the model (see SI for details). The predicted signals for the two probes overlap, supporting a two-state model; indeed, a two-state fit yields an unfolding midpoint of 4.09 M urea, an m-value of 2.64 kcal mol^{-1} M^{-1} and a free energy of unfolding in water of 10.78 kcal mol^{-1}, in agreement with the experimentally determined values. The model allows a direct inspection of the structured
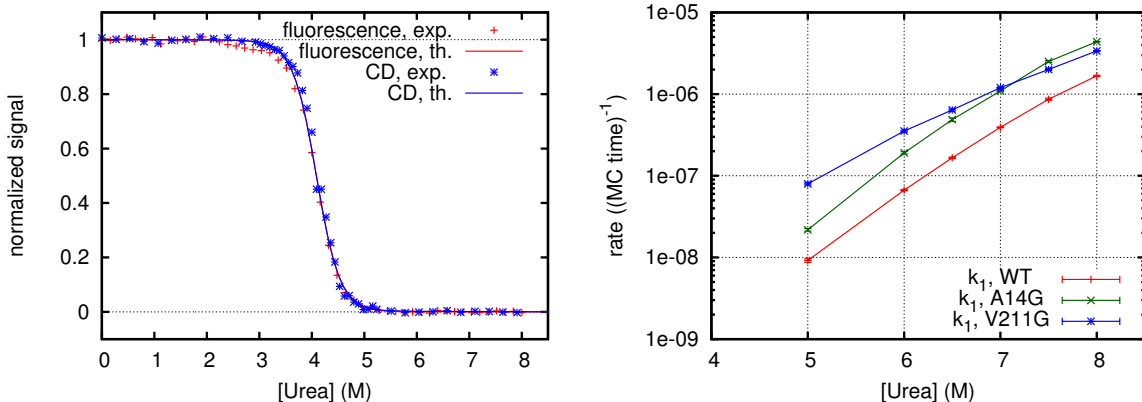

Figure 5: *Left*: Fluorescence (F) and circular dichroism (CD) signals, as predicted with the model, together with the corresponding normalized experimental signals (see "Methods" section in SI), as a function of urea concentration c. *Right*: Rates of the slower unfolding phase (in units of inverse simulation time), for the WT and two mutants, one in the N- and the other in the C-terminal part of the protein. Estimated errors are smaller than the size of the plot symbols.

regions: Figure S6 shows the populations $\nu_{i,j}$ of isolated native regions between residues i and j, revealing a slight dominance of N-terminal structures starting at the first residue of the protein and spanning a length that is dependent on the denaturant concentration. Such slightly asymmetrical distribution of the structure is also apparent, in a more quantitative way, in Fig. S7. "Internal" structures, involving repeats between 2 and 6, always appear to

21

be less stable than the corresponding N-terminal structures (i.e. those ending at the same place but starting at the first repeat).

Simulations support pathway heterogeneity: Fig. 5B shows the predicted rates of the major unfolding phase, obtained from the analysis of the fraction of native residues,[53] for the wild type and two representative mutants, A14G and V211G[54]. The urea dependence of these rates are in qualitative agreement with the experimental data and are able to recapitulate the key experimental observations, namely that upon mutation at the N-terminus (A14G) the unfolding arm become slightly steeper and the downward curvature is more pronounced, whereas for V211G the opposite is observed. Although the changes are much less pronounced than those observed experimentally, the picture is qualitatively the same. Thus, we see that the theoretical analysis is good enough to describe, at the qualitative level, the general aspects of the experimental relaxations even if the energy landscape of the model does not reproduce the experimental one at the level of detail required for quantitative predictions. We next analyse individual trajectories in order to shed light on the microscopic aspects that are not directly accessible to experiment.

Simulations of single-molecule relaxations at different concentrations show great heterogeneity, related to the stochastic nature of the microscopic dynamics. However, inspection of order parameters based on coarse-grained structural information allows us to see the fundamental trends in the kinetics. Fig. 6 reports the behavior of the Kendall rank correlations τ_B calculated from the order of folding/unfolding of either helices, pairs of neighboring helices (referred to as "helix pairs"[55]), repeats or groups of repeats, as defined in Table S2 of S.I., in strongly unfolding ($c = 8$, i.e. 8 M urea) and refolding ($c = 0$) conditions. As explained in the Methods and S.I., such correlations are based on the last time at which the super-secondary structure element is recorded as folded/unfolded during each single-molecule relaxation. Therefore, they are independent quantities providing complementary information.[56] At $c = 8$ we observe that the individual helices are the least committed to

a precise ordering: τ_B(helices) shows fluctuation around two partially polarized values at approximately -0.5 and 0.5, due to the fact that isolated helices appear to be quite stable, even when the protein can be considered essentially unfolded. Thus, supersecondary structures that include at least a pair of neighboring helices provide better information about the progression of the unfolding reaction. Fig. 6 shows that the different levels of coarse-
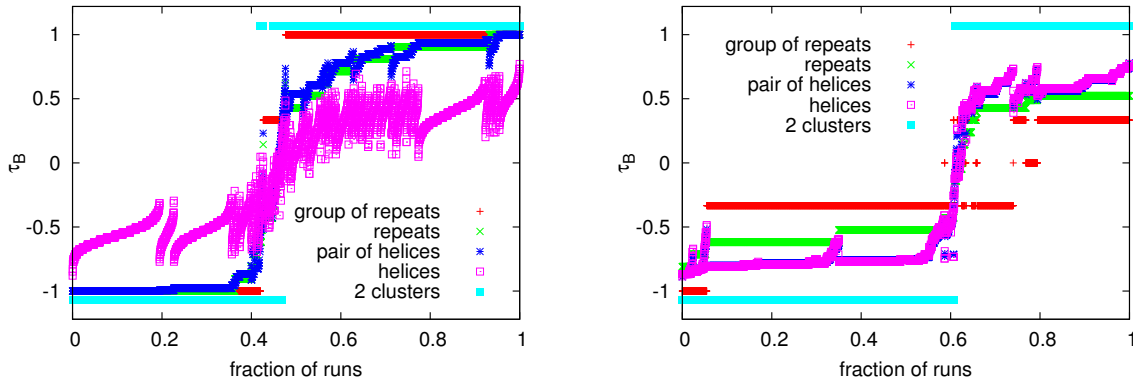

Figure 6: *Left*: Values of τ_B at $c = 8$, for different choices of (super)secondary structures, together with the 2-cluster picture from Affinity Propagation. x coordinates correspond to fraction of trajectories; the latter are ordered according to increasing values of τ_B(groups), with the ties resolved according to increasing values of τ_B(repeats), then τ_B(helix pairs), then τ_B(helices). $\tau_B = -1$ implies that the structure is gradually lost from the C-terminus towards the N-terminus (in the "leftward" direction), whereas $\tau_B = 1$ implies the opposite ("rightward") unfolding direction, from the N- to the C-terminus. Accordingly, intermediate negative (respectively: positive) values of τ_B imply structure "polarization" at the N-terminus (respectively: C-terminus), with the overall order just partially respected. *Right*: values of τ_B at $c = 0$, for the same choices of (super)secondary structures as before, together with the 2-cluster view from Affinity Propagation. For the sake of simplicity, here we plot $-\tau_B$, so as to keep the same interpretation of negative (respectively, positive) values corresponding to polarization at the N-terminus (respectively, C-terminus) For rendering reasons, trajectories are sorted according to the values of τ_B(repeats), then of τ_B(groups), τ_B(helix pairs) and finally of τ_B(helices).

graining, from the helix pairs to the groups of repeats, agree with each other and point to the same overall order of unfolding, indicating that the results are robust and independent of the level of coarse-graining and of the precise definition of the supersecondary structures (Table S2). This behavior is especially clear for the N-polarized trajectories: in these we find that τ_B=-1 for helix pairs, repeats and groups, which implies that unfolding takes place

in a perfectly sequential fashion, from the most C-terminal pair of helices towards the most N-terminal one. The C-polarized trajectories show greater heterogeneity, which is reflected in the gradual increase in the order parameters of the helix pairs and repeats. The presence of a sharp shift in the order parameter at a value of the fraction of runs, x, around $x = 0.45$, from a value close to -1 to a value greater than 0.5, suggests that indeed there are two well-distinguished classes of trajectories that we can map, according to their N- or C-terminal structure polarization, onto pathway A and pathway B, respectively. The clustering by Affinity Propagation (AP), which is independent of the choice of any reference order, supports this view: First, the analysis of the number of clusters as a function of the "preference" parameter shows a major plateau at $n = 2$, indicating that there are two main classes of trajectories (see Figure S8). Second, the clustering by AP agrees with the values of the τ_B(groups) in Fig. 6: most trajectories "naturally" cluster into two main classes characterized by the unfolding of the groups in the order 123 and 321, with only small fractions following the order 132 or 312 (see Table S6 in SI). Thus, simulations clearly and robustly indicate that single-molecules trajectories cluster into two pathways, which agree with those emerging from the experiments.

The plots of the order parameters at $c = 6$, $c = 7$ are similar to that at observed $c = 8$ (Fig. S9), with a progressive increase in the fraction of N-polarized trajectories relative to the C-polarized ones upon lowering c (see also Table S6), that again agrees with the experimental findings. The situation is different for folding trajectories, at $c = 0$: Fig. 6 shows that none of the τ_B is as polarized as in the unfolding trajectories: a perfect folding order ($|\tau_B| = 1$) from N to C or vice versa is never observed at the level of repeats, helix pairs or individual helices. The information coming from helices and helix pairs is essentially the same, indicating that secondary and tertiary structures are formed concomitantly. For each trajectory, τ_B decreases in magnitude with increasing levels of structural coarse-graining, suggesting that the proposed coarse-graining and/or reference order used to calculate the τ_B are not the best suited for the folding kinetics. However, the dominance of the N-polarized

trajectories, and the sharp transition from negative to positive τ_B values, nevertheless points to the existence of two main pathways. Indeed, the analysis of the number of clusters as a function of the "preferences" supports the two-cluster view (see Figure S8), and the partitioning into two classes is in agreement with the sign of the τ_B's: we can see in Fig. 6 that all the N-polarized trajectories (60% of the total) belong to one cluster, and the C-polarized trajectories belong to the other.

Pre-transition fluctuations of the native state reflect the weakness of the C-terminus relative to the N-terminus, as also observed at equilibrium: Inspection of individual trajectories also yields an explanation for the apparent paradox of a dominant C-terminal-structured unfolding pathway despite the equilibrium results pointing to an N-terminal prevalence in the distribution of the structure. Fig. 7 shows the average evolution of each residue during unfolding calculated over all trajectories belonging to the same pathway, as well as one example of single-molecule relaxation from each pathway. Naively, one would expect that the progressive unfolding from C- to N-terminus along pathway A would produce, in the average plot, a roughly triangular shaped structured (i.e. yellow) region below a "hypotenuse" going from the top left to the bottom right of the plot; likewise, the opposite should hold for pathway B, unraveling from N to C, with a yellow right-angled triangle positioned above the bottom-left/top-right diagonal of the plot. Instead, however, the averages reveal that C-terminal residues spend a considerable fraction of their time unfolded, even in the case of pathway B. This counter-intuitive behavior can be rationalized by looking at the single-molecule runs, which reveal that the structure at both ends (but especially the C-terminus) frays and reforms several times before the unfolding reaction proceeds to completion along either pathway.

Identifying rate-determining energy barriers from the simulations: The weakness of the C-terminus is consistent with the equilibrium finding of greater stability of the N-terminal structures; moreover, it suggests that the apparent paradox observed between
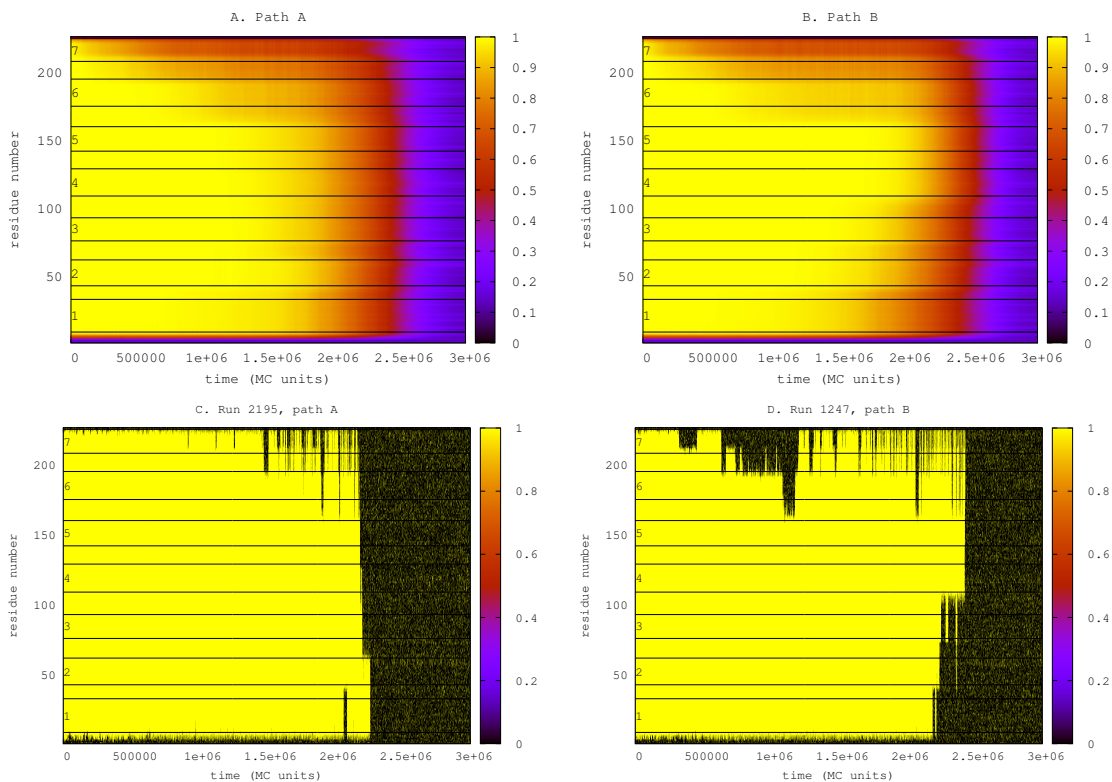
25

Figure 7: Average unfolding at $c = 8$, along pathway A (τ_B(helix pairs)< 0, *Panel A*) and pathway B (τ_B(helix pairs)> 0, *Panel B*). The average is performed on all the trajectories belonging to the same path, at equal time (on the x-axis). At any time t, light/dark colors correspond to residues with a high/low probability of being folded t. *Panel C,D*: Examples of single molecule relaxations along pathway A or B at $c = 8$. Color coding as before, but now each residue can just be folded (yellow) or unfolded (black).

kinetics and equilibrium (i.e. unfolding from the N-terminus via pathway B despite the weakness of the C-terminus) must be related to the nature of the rate-limiting unfolding barriers for the two pathways. Therefore, we next sought to locate these two rate-limiting barriers in the simulations. From the analysis of individual trajectories it is not easy to locate transition states and intermediates, since they emerge as ensemble properties. However, the analysis of the average lifetimes or "dwell times" of each helix-pair[57] (Fig. 8), calculated for groups of trajectories that share the same unfolding sequence, can give us an idea of which are the slowest steps. The upper plots in Fig. 9 tell us that the longest dwell time always corresponds to the unfolding of the first structural element; all initial fluctuations from the native state are hidden here, since the reported time corresponds to the last one when the helix-pair was structured. Here we see that such fluctuations do not involve an intermediate characterized by a part of the protein being unfolded: the structure frays at the ends and goes back to the native state several times before unfolding starts. The lower plots indicate that unfolding along pathway A is more abrupt and rapid than along pathway B and that the second-longest dwell time corresponds to unfolding the last repeat (repeat 1 or 2). Along pathway A, the disruption of repeat 4 and of the interface between repeats 2 and 3 (corresponding to the regions m_2 and m_1 respectively, in Fig. 9) is also a slow step. We note that the former roughly correspond to the experimentally observed transient intermediate "I" along pathway A. Moreover, a negligible time is spent at repeat 5, corresponding to crossing TS_A. The long-lived structures along pathway B are the interface between repeats 1 and 2, and repeat 4 together with its interface with repeat 3, while candidate transition states are located at repeat 2 (consistent with the TS_2 inferred from experiments) and the interface between repeats 4 and 5.

We note that, due to the sequential nature of folding/unfolding along the two pathways, all relevant steps in each pathway should be characterized by just one structured region, which elongates or shrinks at its ends, with a very marginal role for coalescence of two previously formed non-adiacent regions. Importantly, this property allows us to read out
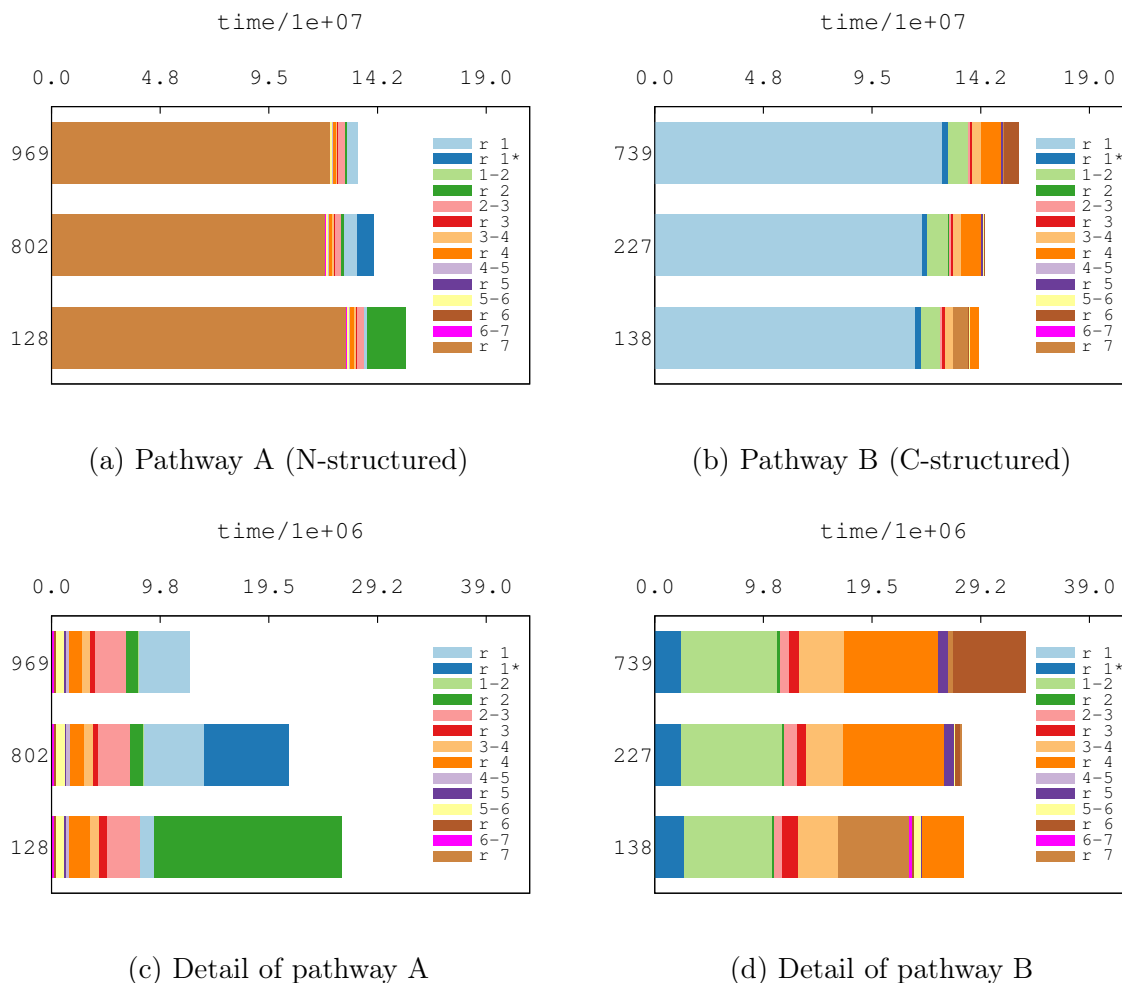
(a) Pathway A (N-structured)

(b) Pathway B (C-structured)

(c) Detail of pathway A

(d) Detail of pathway B

Figure 8: Upper plots: Dwell times during unfolding trajectories at c=8 M; Lower plots: close-ups of the upper plots, reporting only times $\Delta t_i = t_i - t_1$ after the first element unfolds. We group together the trajectories that share the same sequence of unfolding events, and average the unfolding times within such group. The number of trajectories sharing the same sequence of unfolding events is reported at the left of the graphics; only the three largest groups of trajectories are reported for each class. In each histogram the length of the bar is proportional to the average lifetime of that helix-pair during unfolding, after the preceding one has unfolded; the color identify the helix pair. For clarity, we have indicated as "r i" the helix pair corresponding to the repeat "i", and "i-(i+1)" the helix pair corresponding to the last helix in repeat i and the first in (i+1). "r 1*" stands for the pair helix-2,helix-3, where helix 3 is the short 3-10 helix of repeat 1.

the main features of the folding/unfolding pathways from the equilibrium $\nu_{i,j}$ maps, despite the fact that the relaxation kinetics is intrinsically a non-equilibrium process whereas the maps are equilibrium averages and would describe the kinetics only if, during relaxation, the motion along the pathway was much slower than the time needed for the rest of the protein

to equilibrate. On the contrary, the analysis of the interaction energies does not provide any significant clues as to the kinetic behavior (see Figs. S5 and S9).

Explanation for the difference in dominant unfolding versus folding pathway: The top panels in Fig. 9 show the equilibrium $\nu_{i,j}$ maps under strongly native and strongly denaturing conditions. In the light of the results about the sequential unraveling of single-run trajectories, it is natural to roughly identify pathway A with the vertical line joining the native islands (5,5) and (5,225), and pathway B with the horizontal line between (5,225) and (225,225). The bottom panels show such one-dimensional cuts, in the same conditions. The counter-intuitive dominance of unfolding pathway B (preserving structure at the C-terminus, and therefore corresponding to moving horizontally away from the native state in Fig. 9, top-right panel) can be understood by observing that at $c = 8M$ there is a "forbidden" horizontal region corresponding to native islands ending at around residue 150 with TS_A representing a major barrier along pathway A. This barrier is located quite far from the native state, and all the regions in between, corresponding to structures starting at repeat 1 and ending at repeat 5 and (even more) at 6 or 7, are easily accessible from the native state by fraying the C-terminus. This explains why many single-molecule runs repeatedly follow the path A away from the native state and back again, fraying and reforming the C-terminus, before moving to pathway B, according to the behavior reported in Fig. 8. Note also that the analysis of dwell times for unfolding along pathway A is in agreement with the results for ν_{ij}.

Along pathway B there are three main barriers, β_1, β_2 and β_3 (in order of increasing structure): their positions suggest the identity of β_2 with the experimental TS_1. The experiments indicate that TS_2 involves the disruption of repeat 1 and 2; according the equilibrium $\nu_{i,j}$ of the simulations the conformations with repeat 1 and part of repeat 2 unfolded are local probability minima within a high-energy region. A close look at the equilibrium $\nu_{i,j}$ reveals that the highest-energy point along pathway B is β_3 (see also the bottom-right panel) with
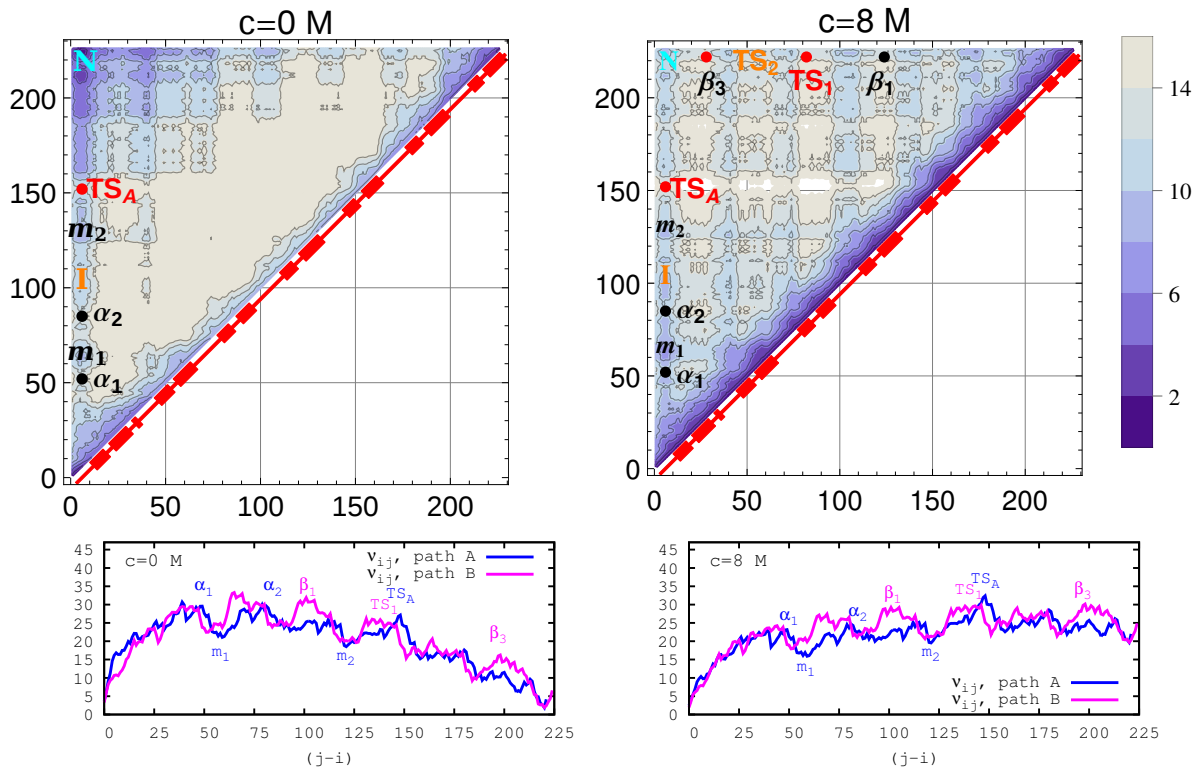
29

Figure 9: *Top:* Equilibrium populations of the native islands $\nu_{i,j}$ under strongly native ($c = 0$M) and strongly denaturing ($c = 8$M) conditions, with a scheme of the secondary structure along the diagonal. The values of $z = -\log_{10}(\nu_{i,j})$ in the range 0-16 are reported; darker colors correspond to most likely regions. Along pathway A the barrier α_3 coincides with the experimental findings for the main barrier TS_A, and the partially populated region m_2 is compatible with the experimental intermediate I. Two "barriers" α_1 and α_2 separated by a partially populated spot m_1 occupy the region where the experimental barrier, between the intermediate and unfolded state, was identified. Along pathway B, the barrier β_2 fits the experimentally determined position for TS_1, while the position of barrier β_3 appears to be slightly shifted with respect to the experimentally determined position for TS_2 (however, the latter are consistent with the findings for the dwell times). An additional barrier β_1, consistent with the model kinetics, appears in the equilibrium map but is not detected in experiments. *Bottom:* Graphics of $z = -\log(\nu_{i,j})$ along pathway A (the straight line $(5, j)$ in the top panels) and pathway B (the straight line $(i, 225)$), under folding ($c = 0$ M) and unfolding ($c = 8$ M) conditions. To make it easier to compare the two pathways, we use an x-axis corresponding to the values of $(j - i)$ are reported, so that for example x=220 corresponds in both cases to the native minimum.

β_1 and β_2 being close in energy to it. Importantly, the conformation corresponding to β_3 is more likely than that of TS_A along pathway A, i.e. the rate-limiting step on pathway A is lower in energy than that on pathway B. Thus, despite the fact that most of the structures along pathway A are lower in energy than those along pathway B, the highest rate-limiting barrier is that along pathway A (see also Fig. S10, in SI). The fact that the rate-limiting barrier β_3 along pathway B is structurally close to the native state means that all "easy" (pre-transition) fluctuations are small; bigger fluctuations at the N-terminus would involve crossing the highest barrier, and once the protein has passed over this and committed to pathway B, then crossing back is unlikely. In folding conditions, pathway B presents bigger barriers, and is thus disfavored, as can be seen from the bottom-left panel. TS_A still appears as one of the main barriers on pathway A[58]. The intermediate detected in experiments does not exactly coincide with minimum m_2 but would correspond to the broad plateau between α_2 and m_2.

Summary We show here that the 7-ankyrin repeat protein gankyrin folds and unfolds via two alternative pathways (Fig. 10). This behavior is manifest in the wild-type protein by upward curvature in the unfolding arm of the chevron plot and by dramatic changes in the shape of the unfolding arms upon mutation (e.g. compare F58I and L209A). Single-site mutants shift the relative flux through the pathways, in some cases resulting in flux almost exclusively through a single pathway across the whole urea range (again, compare F58I and L209A in Fig. 4). For these mutants we are then able to see that one pathway (pathway B) has a broad rate-limiting energy barrier (F58I) characterised by downward curvature in the unfolding arm, whereas the other pathway (pathway A) has a rather narrow rate-limiting energy barrier characterised by a linear unfolding arm (L209A). It is striking that two seemingly very different types of transition state structure, one apparently sensitive to solvent perturbations and the other not, can be present in one protein. However, they are not fundamentally different; rather, they are both manifestations of the same underlying
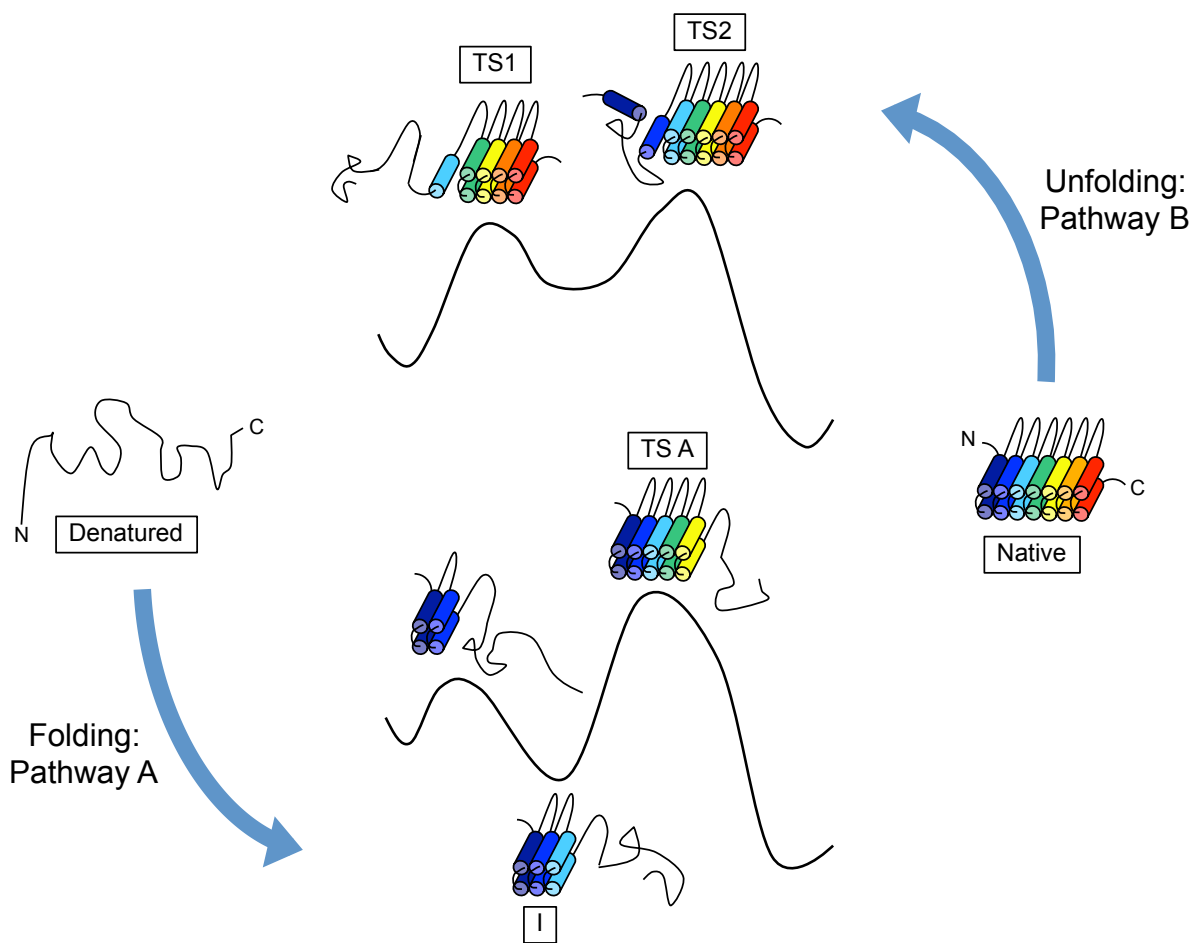
Figure 10: Schematic showing the structures of the transition states and the intermediate for pathways A and B, as mapped by Φ–value analysis. The curved arrows represent the finding that Pathway A dominates the kinetics under folding conditions whereas Pathway B dominates under strongly unfolding conditions.

mechanism of progressive (stepwise) unfolding of the repeats. As illustrated in Fig. 10, the intermediate in pathway B is a high-energy "lake" lying between a pair of transition states that switch in being rate-limiting depending on the denaturant concentration; by contrast, the intermediate state for pathway A (detected in the unfolding direction by double jump) is of sufficiently low energy that it is transiently populated in the refolding reaction and weakly populated at equilibrium also. Thus, both pathways have rough free-energy profiles but they contain different fine structures that result in the dramatically different shapes of their chevron plots[59]. The different stabilities of the two intermediates must reflect differences in the degrees of coupling between repeats along the repeat array: repeats 1-3 are sufficiently stable and decoupled from the other repeats that they constitute an independently folded domain.

Despite the large number of folding studies in the literature, parallel pathways have rarely been observed, most likely because for many small proteins there is only one low energy route to the native state. Examples include proteins that have structural symmetry, such as the GCN4 coiled-coil, protein G and tandem repeat proteins[13,15,23,28,50,60,61]; early examples were barnase, S6 and lysozyme, the last possessing two subdomains either of which can fold first[62–65]. However, even if there are parallel routes, they will not be detected in chevron plots if there is little difference in the compactness of the rate-limiting barriers for the two pathways and if measurements are not made over a sufficiently large denaturant range.

Φ–value analysis allows us to map out the structure of the intermediate and transition states for both pathways, thereby providing a comprehensive picture of the energy landscape (Fig. 10). Repeats 1-5 are fully structured in the transition state of pathway A, whereas repeats 3-7 are fully structured in TS2 of pathway B and partly structured in TS1 of pathway B. We observe an intermediate species that accumulates in the refolding reaction but not in the unfolding reaction, and its unfolding can therefore be detected only by using the double-jump method. The intermediate has structure in repeats 1-3; we therefore propose that it is

an intermediate on pathway A. The kinetic intermediate resembles the equilibrium unfolding intermediate identified by variations in the m-values. Thus, for both pathways the structure unravels progressively from one or other end of the molecule. We reiterate that we can only coarsely define the structural boundaries of these states.

A striking finding is that gankyrin folds along a different route from that along which it unfolds; i.e. what folds first does not unfold last. The folding of the N-terminal repeats from the denatured state to a discrete intermediate has a relatively low associated energy barrier, and therefore in the folding direction the N-path is favored even in the case of destabilising mutants such as F58I; this domain is also the most stable at equilibrium. However, when it comes to the unfolding reaction at high denaturant, it is the N-terminal region that dissociates the more easily from the native repeat stack. The simulations provide an explanation for this apparently contradictory behavior: although it is indeed easier to undergo pre-transition fluctuations from the C-terminus than the N-terminus, the rate-limiting barrier itself - the highest energy point - is in fact higher along pathway A (rupturing of repeat 5 from repeat 6) than that along pathway B (rupturing of repeat 2 from repeat 1).

We find that simulations based purely on the native contacts are able to reproduce the key features of the energy landscape obtained by experiment. They predict the dominance of N-polarized structure in equilibrium intermediates, and the existence of two unfolding pathways involving sequential unraveling from either end of the protein chain, that agree with the characteristics of experimental pathways A and B. Moreover, in the unfolding reaction the flux shifts from pathway B to pathway A as denaturant concentration is reduced, again in agreement with the experimental findings. Finally, the simulations shed light on key features of the energy landscape of gankyrin that are not accessible to experiment, and these details yield insights into the physical basis of the experimental results. In particular, as discussed above, they provide a framework within which to reconcile the seemingly contradictory behavior of the protein under equilibrium/refolding conditions (i.e. high stability of the N-terminal part of the polypeptide chain) when compared with kinetic unfolding conditions

(i.e. dominance of the unfolding pathway involving unravelling from the N-terminus). The detailed map of gankyrin obtained embodies all the fundamental features of protein energy landscapes, and it also demonstrates, in a striking way, how the fine structure that is an inherent characteristic of folding free energy profiles manifests itself in the experimental data. We are now primed to take the next step, namely to program the folding mechanism (including features such as order of structure formation, number of pathways accessed, and shape of the energy barrier) into the amino-acid sequence - the Holy Grail of the protein folding field.

Materials and Methods

Site-directed mutagenesis, protein expression and purification

The *E. coli* expression plasmid for gankyrin was a generous gift of Dr. A. Wilkinson, University of York, UK. Site-directed mutagenesis was performed using the Quikchange kit (Stratagene). Protein expression and purification was performed as described in[66]. Purity was determined by SDS-PAGE and mass spectrometry. All of the experiments were performed in 50 mM Tris-HCl buffer pH 8.0, 5 mM DTT (or DTE for CD experiments) and at 25 °C, unless stated otherwise.

Equilibrium denaturation

Urea solutions were prepared by mixing the appropriate volumes of a solution of buffer and 10 M urea in buffer dispensed using a Hamilton MicroLab 500 series. Protein stock was then added to a final concentration of 2 μM to each urea concentration and the samples equilibrated at 25 °C for 2 hours before measurement. For CD, the protein concentration was 20 μM. Fluorescence was measured using a Perkin Elmer luminescence spectrometer LS-55 with a 1 cm pathlength cuvette. The excitation wavelength was 280 nm and the excitation and emission slit-widths were 5 nm. Wavelength scans between 300 nm and 370

nm were performed for each sample at a rate of 1 nm s^{-1}. CD was measured using an Aviv 202 CD spectrometer with a 3 mm pathlength cuvette.

Kinetic experiments

An Applied Photophysics SX.18MV instrument was used to perform stopped-flow fluorescence experiments. The excitation wavelength was 280 nm and emission was recorded above 320 nm with the use of a cut-off filter. Unfolding was initiated by 1:10 mixing of protein in buffer and a urea solution. The data from at least 6 traces were averaged at each denaturant concentration. Refolding was initiated by 1:10 mixing of protein in urea and buffer containing low concentrations of urea. The concentration of the protein after mixing was 2 μM. Several traces were collected at each urea concentration and averaged. Stopped-flow CD was performed using an Applied Photophysics Π*180 instrument, monitoring ellipticity at 222 nm. The experiments were performed as for fluorescence except that the final protein concentration was 20 μM. For double-jump experiments the final protein concentration was 1 μM, achieved by mixing a solution of protein at 36 μM in a 1:5 ratio with either refolding or unfolding solution followed by a second mixing step in a 1:5 ratio with either unfolding or refolding solution. At least six traces at each delay time were averaged. Data were fitted using the program Kaleidagraph (Synergy Software).

Theoretical modeling and simulations

We use a modified version of the native-centric WSME model[39–42], with a suitable redefinition of the interactions to describe more realistically the chemical denaturation of gankyrin. The model has been used to study the kinetics and thermodynamics of several proteins, upon thermal[41,42,67–80], chemical[6,81,82] or mechanical denaturation[81–88]. The binary variables of the model, m_k, with $k \in [1, N]$ for a N-residues protein, describe the state of each residue as native, $m_k = 1$, and unfolded, $m_k = 0$.

Its effective energy can be written as $H = \sum_{i=1}^{N} \sum_{j=i}^{N} H_{i,j} \sigma_{i,j}$, where we set $\sigma_{i,j} =$

$(1-m_{i-1})\prod_{k=i}^{j}m_k(1-m_{j+1})$ (with $m_0 = m_{N+1} = 0$) and $H_{i,j}$ represents the whole (effective) energy contribution, including interaction energy, solvation free-energy, as well as side-chain entropy, from a native structure spanning the region (i, j). The quantities $H_{i,j}$ are written in term of the change of accessible surface area upon folding the isolated peptide corresponding to the region i, j; see S.I. for details. We fix the parameters of the model by fitting the fluorescence and circular dichroism (CD) experimental signals, after baseline removal, with those predicted by the model, for the WT species. The same parameters are then used also for mutants: see S.I. for details.

The equilibrium values of all thermodynamic quantities are calculated resorting to the exact solution of the model[89,90]. In particular, we focus on the equilibrium probability that the region between i and j is found as an isolated native region, flanked by unfolded residues:

$$\nu_{i,j} = \langle \sigma_{i,j} \rangle . \tag{2}$$

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 $x = \{m_k, k = 1, \dots N\}$ denotes the state of the system. As in previous works[68,69], we use Metropolis Monte Carlo simulations where a single residue flip is accepted or rejected according to its equilibrium probablity. We monitor relaxation in both folding (T=298.15 K, $c = 0$, completely unfolded initial state) and unfolding conditions (T=298.15 K, $c = 6, 7, 8$, completely folded initial state). For each trajectory, we keep track of the evolution of the fraction of native residues $m(t) = N^{-1}\sum_i m_i$ and of the last formation/disruption time $t_{a,b}$ of super-secondary structure elements $R_{a,b}$ until the First Passage Time in the final state. Such elements are defined as the regions spanning 1, 2, ..., 15 helices; see S.I. for details. In particular, for each folding/unfolding trajectory, we monitor the order of formation of helices, pairs of neighboring helices, repeats and groups of repeats. Since we are interested in the order of the folding/unfolding events, a particulary 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 $\{x_1, \ldots, x_n\}$ (we use the natural order $x_i = i$). A complementary approach for the identification of pathways, independent of a predefined reference order, is provided by the Affinity Propagation (AP) algorithm[91], which groups together trajectories according to their distance; the latter is defined as $d_{ij} = \sqrt{\sum_k \left[p_k^{(i)} - p_k^{(j)} \right]^2}$, with $p_k^{(i)}$ the position of the structure k in trajectory i. According to a tunable parameter (see S.I.) representing the "selfishness" of each trajectory, AP produce a variable number of clusters, each with one representative exemplar; each trajectory belongs to just one cluster. Rates and amplitudes of folding/unfolding are estimated from the evolution of the fraction of native residues as described in S.I.

Acknowledgement

Supporting Information Available

A detailed presentation of the theory is given. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins* **1995**, *21*, 167–95.

(2) Barrick, D.; Ferreiro, D. U.; Komives, E. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 27–34.

(3) Javadi, Y.; Itzhaki, L. S. *Curr Opin Struct Biol* **2013**, *23*, 622–31.

(4) Main, E. R. G.; Lowe, A. R.; Mochrie, S. G. J.; Jackson, S. E.; Regan, L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 464–471.

(5) Mello, C. C.; Barrick, D. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 14102–14107.

(6) Faccin, M.; Bruscolini, P.; Pelizzola, A. *J Chem Phys* **2011**, *134*, 075102.

(7) Tang, K. S.; Fersht, A. R.; Itzhaki, L. S. *Structure* **2003**, *11*, 67–73.

(8) Wetzel, S. K.; Settanni, G.; Kenig, M.; Binz, H. K.; Plückthun, A. *J. Mol. Biol.* **2008**, *376*, 241–257.

(9) Hagai, T.; Azia, A.; Trizac, E.; Levy, Y. *Biophys J* **2012**, *103*, 1555–65.

(10) Ferreiro, D. U.; Walczak, A. M.; Komives, E. A.; Wolynes, P. G. *PLoS Comput Biol* **2008**, *4*, e1000070.

(11) Ferreiro, D. U.; Cho, S. S.; Komives, E. A.; Wolynes, P. G. *J Mol Biol* **2005**, *354*, 679–692.

(12) Löw, C.; Weininger, U.; Neumann, P.; Klepsch, M.; Lilie, H.; Stubbs, M. T.; Balbach, J. *Proc Natl Acad Sci USA* **2008**, *105*, 3779–84.

(13) Werbeck, N. D.; Rowling, P. J. E.; Chellamuthu, V. R.; Itzhaki, L. S. *Proc Natl Acad Sci USA* **2008**, *105*, 9982–7.

(14) Werbeck, N. D.; Itzhaki, L. S. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7863–7868.

(15) Lowe, A. R.; Itzhaki, L. S. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 2679–2684.

(16) Junker, M.; Schuster, C. C.; McDonnell, A. V.; Sorg, K. A.; Finn, M. C.; Berger, B.; Clark, P. L. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 4918–23.

(17) Kloss, E.; Barrick, D. *Protein Sci.* **2009**, *18*, 1948–60.

(18) Javadi, Y.; Main, E. R. G. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 17383–8.

(19) Braselmann, E.; Clark, P. L. *J Phys Chem Lett* **2012**, *3*, 1063–1071.

(20) Liu, C.; Gaspar, J. A.; Wong, H. J.; Meiering, E. M. *Protein Sci.* **2002**, *11*, 669–79.

(21) Capraro, D. T.; Roy, M.; Onuchic, J. N.; Jennings, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14844–8.

(22) Longo, L. M.; Lee, J.; Tenorio, C. A.; Blaber, M. *Structure* **2013**, *21*, 2042–50.

(23) Tsytlonok, M.; Craig, P. O.; Sivertsson, E.; Serquera, D.; Perrett, S.; Best, R. B.; Wolynes, P. G.; Itzhaki, L. S. *Structure* **2013**, *21*, 1954–65.

(24) Sivanandan, S.; Naganathan, A. N. *PLoS Comput. Biol.* **2013**, *9*, e1003403.

(25) Kohl, A.; Binz, H. K.; Forrer, P.; Stumpp, M. T.; Pluckthun, A.; Grutter, M. G. *Proceedings of the National Academy of Sciences* **2003**, *100*, 1700–1705.

(26) Ferreiro, D. U.; Cervantes, C. F.; Truhlar, S. M. E.; Cho, S. S.; Wolynes, P. G.; Komives, E. A. *J Mol Biol* **2007**, *365*, 1201–16.

(27) Bergqvist, S.; Alverdi, V.; Mengel, B.; Hoffmann, A.; Ghosh, G.; Komives, E. A. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19328–19333.

(28) Tripp, K. W.; Barrick, D. *J Am Chem Soc* **2008**, *130*, 5681–5688.

(29) Main, E. R. G.; Xiong, Y.; Cocco, M. J.; D'Andrea, L.; Regan, L. *Structure* **2003**, *11*, 497–508.

(30) Cortajarena, A. L.; Yi, F.; Regan, L. *ACS Chem. Biol.* **2008**, *3*, 161–166.

(31) Cortajarena, A. L.; Mochrie, S. G. J.; Regan, L. *Protein Sci.* **2011**, *20*, 1042–7.

(32) Tamaskovic, R.; Simon, M.; Stefan, N.; Schwill, M.; Plückthun, A. *Meth. Enzymol.* **2012**, *503*, 101–34.

(33) Yadid, I.; Tawfik, D. S. *J. Mol. Biol.* **2007**, *365*, 10–7.

(34) Nikkhah, M.; Jawad-Alami, Z.; Demydchuk, M.; Ribbons, D.; Paoli, M. *Biomol. Eng.* **2006**, *23*, 185–94.

(35) Lee, J.; Blaber, M. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 126–30.

(36) Broom, A.; Doxey, A. C.; Lobsanov, Y. D.; Berthin, L. G.; Rose, D. R.; Howell, P. L.; McConkey, B. J.; Meiering, E. M. *Structure* **2012**, *20*, 161–71.

(37) Aksel, T.; Barrick, D. *Biophys. J.* **2014**, *107*, 220–232.

(38) Lozano, G.; Zambetti, G. P. *Cancer Cell* **2005**, *8*, 3–4.

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

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

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

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

(43) Capaldi, A. P.; Shastry, M. C.; Kleanthous, C.; Roder, H.; Radford, S. E. *Nature Structural Biology* **2001**, *8*, 68–72.

(44) Kiefhaber, T. *Proceedings of the National Academy of Sciences* **1995**, *92*, 9029–9033.

(45) Khorasanizadeh, S.; Peters, I. D.; Roder, H. *Nature Structural Biology* **1996**, *3*, 193–205.

(46) Matouschek, A.; Fersht, A. R. *Proc Natl Acad Sci USA* **1993**, *90*, 7814–8.

(47) Matouschek, A.; Matthews, J. M.; Johnson, C. M.; Fersht, A. R. *Protein Eng* **1994**, *7*, 1089–95.

(48) https://github.com/quantumjot/PyFolding.

(49) Sánchez, I. E.; Kiefhaber, T. *J Mol Biol* **2003**, *327*, 867–884.

(50) Wright, C. F.; Lindorff-Larsen, K.; Randles, L. G.; Clarke, J. *Nat Struct Biol* **2003**, *10*, 658–62.

(51) Jonsson, T.; Waldburger, C. D.; Sauer, R. T. *Biochemistry* **1996**, *35*, 4795–802.

(52) Sánchez, I. E.; Kiefhaber, T. *J Mol Biol* **2003**, *325*, 367–376.

(53) To be precise, the signal of the native fraction is more similar to the CD signal rather than to the fluorescence signal. However, the experiments report agreement, at least for the major phase, between the two.

(54) Although the experimental mutation is V211A, our method of mimicking the mutations involves substituting the ASA of the mutated residue with that of its main chain; this roughly corresponds to a mutation to Gly, so we refer to the in silico mutation as V211G.

(55) Note that the helix pairs do not simply correspond to the ankyrin repeats. Ankyrin repeat 1 encompasses helix pairs 1 and 2 (as there are three helices in repeat 1, the third one being a small 3_{10} helix), while repeats i where i=2,...,7 correspond to helix pairs $2i$. The other helix pairs (i.e. helix pairs 3, 5, 7, ..., 15) therefore provide information about inter-repeat contacts.

(56) The fact that, for example, a repeat is last reported as folded at a certain time does not inform us about the time or sequence of the unfolding of the constituent helices; and

conversely, the unfolding of the constituent elements can take place much later than that of the supersecondary structure of which they are a member.

(57) We use the latter, instead of the individual helices or repeats, because they provide the most detailed information with the least amount of random noise from fluctuations; see Fig.6.

(58) especially considering that barrier m_1 is accessible from more regions than those accounted in the bottom panel, and thus a lower barrier than depicted in the 1-d projection: see the darker (more likely) "triangle" close to it, in the top left panel.

(59) Oliveberg, M.; Wolynes, P. G. *Q Rev Biophys* **2005**, *38*, 245–88.

(60) Moran, L. B.; Schneider, J. P.; Kentsis, A.; Reddy, G. A.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 10699–704.

(61) Nauli, S.; Kuhlman, B.; Baker, D. *Nat. Struct. Biol.* **2001**, *8*, 602–5.

(62) Radford, S. E.; Dobson, C. M.; Evans, P. A. *Nature* **1992**, *358*, 302–7.

(63) Matthews, J. M.; Fersht, A. R. *Biochemistry* **1995**, *34*, 6805–14.

(64) Bieri, O.; Wildegger, G.; Bachmann, A.; Wagner, C.; Kiefhaber, T. *Biochemistry* **1999**, *38*, 12460–70.

(65) Otzen, D. E.; Oliveberg, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11746–51.

(66) Krzywda, S.; Brzozowski, A. M.; Higashitsuji, H.; Fujita, J.; Welchman, R.; Dawson, S.; Mayer, R. J.; Wilkinson, A. J. *J. Biol. Chem.* **2004**, *279*, 1541–5.

(67) noz, V. M. *Curr Opin Struct Biol* **2001**, *11*, 212–6.

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

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

(70) Abe, H.; Wako, H. *Phys. Rev. E* **2006**, *74*, 011913.

(71) Bruscolini, P.; Pelizzola, A.; Zamparo, M. *J. Chem. Phys.* **2007**, *126*, 215103.

(72) Bruscolini, P.; Pelizzola, A.; Zamparo, M. *Phys. Rev. Lett.* **2007**, *99*, 038103.

(73) Itoh, K.; Sasai, M. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13865–13870.

(74) Zamparo, M.; Pelizzola, A. *J. Chem. Phys.* **2009**, *131*, 035101.

(75) Abe, H.; Wako, H. *Physica A* **2009**, *388*, 3442–3454.

(76) Bruscolini, P.; Naganathan, A. N. *J. Am. Chem. Soc.* **2011**, *133*, 5372–9.

(77) Caraglio, M.; Pelizzola, A. *Phys Biol* **2012**, *9*, 016006.

(78) Naganathan, A. N. *Journal of Chemical Theory and Computation* **2012**, *8*, 4646–4656.

(79) Henry, E. R.; Best, R. B.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 17880–5.

(80) Naganathan, A. N. *J Phys Chem B* **2013**, *117*, 4956–64.

(81) Aioanei, D.; Tessari, I.; Bubacco, L.; Samorì, B.; Brucale, M. *Proteins* **2011**, *79*, 2214–23.

(82) Aioanei, D.; Brucale, M.; Tessari, I.; Bubacco, L.; Samorì, B. *Biophys. J.* **2012**, *102*, 342–50.

(83) Caraglio, M.; Imparato, A.; Pelizzola, A. *J Chem Phys* **2010**, *133*, 065101.

(84) Caraglio, M.; Imparato, A.; Pelizzola, A. *Phys Rev E Stat Nonlin Soft Matter Phys* **2011**, *84*, 021918.

(85) Imparato, A.; Pelizzola, A.; Zamparo, M. *J. Chem. Phys.* **2007**, *127*, 145105.

(86) Imparato, A.; Pelizzola, A.; Zamparo, M. *Phys. Rev. Lett.* **2007**, *98*, 148102.

(87) Imparato, A.; Pelizzola, A. *Phys. Rev. Lett.* **2008**, *100*, 158104.

(88) Imparato, A.; Pelizzola, A.; Zamparo, M. *Phys. Rev. Lett.* **2009**, *103*, 188102.

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

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

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

Graphical TOC Entry