doi:10.1016/j.jmb.2005.09.078 J. Mol. Biol. (2005) 354, 679–692

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Energy Landscape of Modular Repeat Proteins: Topology Determines Folding Mechanism in the Ankyrin Family

Diego U. Ferreiro[1,2], Samuel S. Cho[1,2], Elizabeth A. Komives[1,2] and Peter G. Wolynes[1,2,3]*

[1]*Center for Theoretical Biological Physics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*

[2]*Department of Chemistry and Biochemistry, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*

[3]*Department of Physics University of California at San Diego, 9500 Gilman Drive La Jolla, CA 92093, USA*

**Corresponding author*

Proteins consisting of repeating amino acid motifs are abundant in all kingdoms of life, especially in higher eukaryotes. Repeat-containing proteins self-organize into elongated non-globular structures. Do the same general underlying principles that dictate the folding of globular domains apply also to these extended topologies? Using a simplified structure-based model capturing a perfectly funneled energy landscape, we surveyed the predicted mechanism of folding for ankyrin repeat containing proteins. The ankyrin family is one of the most extensively studied classes of non-globular folds. The model based only on native contacts reproduces most of the experimental observations on the folding of these proteins, including a folding mechanism that is reminiscent of a nucleation propagation growth. The confluence of simulation and experimental results suggests that the folding of non-globular proteins is accurately described by a funneled energy landscape, in which topology plays a determinant role in the folding mechanism.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: folding mechanism; energy landscape; ankyrin; non-globular protein; topology

Introduction

Energy landscape theory suggests that efficient and robust folding has been achieved by the evolution of protein molecules that satisfy the principle of minimal frustration, in which native contacts are more stable than random non-native ones.[1] Proteins therefore should have an energy landscape that resembles a partially rugged funnel.[1–3] In many cases, the energetic frustration is so small that topology becomes the key determinant of folding mechanisms. For a perfectly funneled landscape, the structural heterogeneity observed in folding transition states and the partially folded ensembles is determined by geometrical constraints reflecting the trade-offs between chain entropy and folding stabilization energy. These trade-offs can be inferred reasonably well once the native protein structure is known.[4] The basic validity of these concepts has been confirmed by comparing simulations and experiments for numerous globular folds with a variety of topological motifs.[5–7] On the other hand, a large number of proteins are built from highly homologous (although not always identical) structural blocks, with repeating amino acid motifs, which can be considered as modular units. Unlike the typical globular domains that fold to compact structural entities, these sequences are thought to fold in non-globular arrays that are formed by repeating motifs comprised of 20 to 40 amino acid residues, that stack together producing extended superhelical structures.[8] Globular proteins exhibit diverse and complex topologies that have numerous long-range interactions. In contrast, the repetitive and elongated architecture of non-globular folds is stabilized by contacts that are either within one repeat or between adjacent repeats, with no contacts between residues that are distant in sequence space, i.e. they have low contact order.

Abbreviations used: AR, ankyrin repeat; TSE, transition state ensemble.

E-mail address of the corresponding author: pwolynes@ucsd.edu

0022-2836/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.

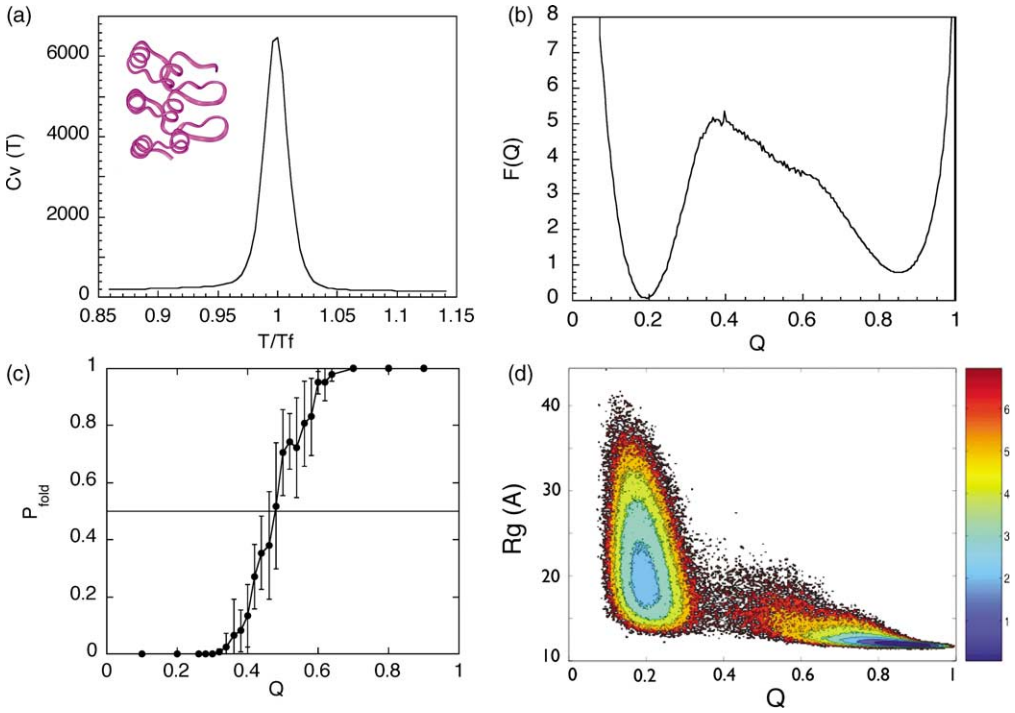

Figure 1. Folding simulations of 3ANK. The folding of 3ANK was simulated with energetically unfrustrated models. (a) The native structure is displayed in ribbon diagram, next to the heat capacity as a function of temperature derived from several constant temperature runs. The peak in the plot corresponds to the folding temperature (T_f). (b) The free energy $F(Q)$ profile as a function of the reaction coordinate Q was calculated at T_f. The thermal fluctuations around the lowest energy state account for the motions around the free energy minimum, therefore the native state ensemble has a minimum at $Q \approx 0.9$, and the unfolded ensemble at $Q \approx 0.2$. (c) The folding probability (P_{fold}) of structures with different Q was calculated. The filled circles correspond to the mean and the error bars to one standard deviation. (d) Free energy surface as a function of the radius of gyration and Q.

In addition, the near-symmetry of different repeats allows, in principle, the possibility of extensive domain swapping. This would be a sign of frustration, perhaps unavoidable if the repeats were exact copies. One of the most extensively studied class of non-globular folds is the ankyrin repeat (AR) family.[9] Here we examine whether minimally frustrated energy landscapes can reproduce the main aspects of the experimentally observed folding mechanisms.

Ankyrin repeats have a long evolutionary history. Bacterial, fungal, plant and animal genomes contain the codes for ankyrin repeat proteins that carry out a broad range of functions.[10] Members of the family have been found to act as transcription factors, cell cycle regulators, signaling proteins, cytoskeletal constituents or adaptor proteins. Ankyrin repeats are broadly distributed in subcellular localization: they may be nuclear, cytoplasmic, membrane-bound or secreted.[9] The ankyrin repeat, a 33 amino acid residue sequence motif, was first identified in the yeast cell cycle regulator Swi6/Cdc10,[11] and was eventually named for the cytoskeletal protein Ankyrin, which contains 24 copies of this repeat.[12] Since first being discovered, over 4700 ankyrin repeat proteins have been identified, some containing as few as four repeats, ranging up to one with as many as 29 continuous repeats†.[13] A unifying trait of the ankyrin repeat proteins characterized to date, is that all function in mediating specific protein–protein interactions.[14]

High-resolution structures of 13 naturally occurring AR proteins, and of three designed ones have been obtained.[15] In all cases, the AR motif adopts a highly similar fold: each repeat forms a β-hairpin, two antiparallel α-helices and a loop of variable length (see Figure 1(a)). The repeats stack against each other in a linear fashion with hydrophobic interactions predominating between the helices and hydrogen bonding between the hairpin–loop regions. This pattern results in a right-handed solenoid structure with a continuous hydrophobic core and a large solvent-exposed surface area.[14] Although they are always found in multiple adjacent copies, in principle each ankyrin repeat may constitute a modular structural element. Is the folding of the whole domain a two-state process or does folding occur through a number of populated intermediates, having some

† http://smart.embl-heidelberg.de

repeats folded and others unfolded? How does the number of repeats affect the folding mechanisms that we observe?

Early truncation studies on the cytoskeletal protein Ankyrin, suggested that four repeats represent the minimum number of repeats necessary to maintain a stably folded structure.[12] More recently, Zhang & Peng[16] reported that the two C-terminal ankyrin repeats of the tumor suppressor protein p16 may fold independently of the other repeats, although with marginal stability, suggesting that this could be the minimal folding unit of an ankyrin repeat. Nonetheless, studies of naturally occurring ankyrin repeat proteins often suggest that a thermodynamic two-state model is sufficient to describe the experimental data. Equilibrium folding experiments on p16,[17] myotrophin,[18] and Notch,[19] all indicate that the domains made of several repeats populate either the completely folded or completely unfolded states. Thus, the folding transition appears to be highly cooperative and partially folded intermediates are not detected in such equilibrium unfolding studies, indicating that some type of coupling mechanism must exist between different repeats, which enables the whole domain to fold up cooperatively *a la* globular domain. If this is the case, we expect the folding landscape of a non-globular domain to be minimally frustrated, and it is likely that their energy landscape is funneled towards native-like structures. When energetic frustration is low enough, topology becomes the key factor governing folding reactions, and it has been shown that the structures of transition states,[5,6] the existence of folding intermediates,[2,20] dimerization mechanisms,[21] and domain swapping events[22] are often well predicted in models where energetic frustration has been removed and topological information of the native state is the sole input.

Here, we address whether topology plays a determinant role in the folding of these non-globular folds, as it does in the globular case. We focus on the particular case of AR domains, but it is possible that the themes that we find are more widely applicable. We conducted a survey of the folding mechanisms of AR containing proteins, by simulating the folding processes with a Gõ-type potential,[23] which is based on native topology only. This model does not account for domain-swapping events, which are likely to occur in repeat proteins, so it must be taken as a zeroth order approximation. In order to cover a broad range of structural varieties and to explore the inter-repeat coupling mechanisms, we selected proteins containing between three and seven AR repeats. We further restricted our analysis to proteins for which a high-resolution structure is known and for which experimental evidence on the folding mechanism is available. We have found that these off-lattice simplified models with a minimally frustrated landscape capture the essential features of the AR folding mechanisms found in the laboratory, and that the cooperative nature of the inter-repeat folding can be ascribed mainly to topological factors.

Results and Discussion

The contribution of native topology to the folding landscape of several ankyrin repeat proteins was studied by simulating their folding transitions with a Gõ-type model. This potential only takes into account attractive interactions found in the native state, and thus lacks energetic frustration, including that which could allow domain swapping. In this model, each residue is represented by a bead centered at the C^{α} position, and the attractive contacts are defined by a distance cut-off between the residues' atoms. Details of the potential can be found in Model and Methods and have been described.[2,6] In all cases, annealing runs were first performed to estimate the folding temperature. To elucidate mechanistic details, constant temperature runs were then carried out near the estimated folding temperature. The folding temperature (T_f) was determined by the peak in the heat capacity change as a function of temperature (C_v plot), derived from the weighted histogram analysis (WHAM)[24] of several constant temperature runs. In general, between 50 and 100 independent runs were carried out for each protein. The global fraction of native contacts (Q) and the fractions in each repeat were chosen as order parameters to follow the folding transitions.

Three repeats: 3ANK

The smallest ankyrin repeat protein whose structure has been solved is the designed protein 3ANK. This protein consists of three identical copies of an AR bearing a consensus AR sequence.[25] Simulated annealing runs showed a single folding transition temperature (data not shown). The constant temperature simulations around the folding temperature showed that this domain displays a single ordering transition, with a sharp peak in the C_v plot (Figure 1(a)). The free energy diagram illustrates that only the unfolded and the fully folded states are well populated at equilibrium (Figure 1(b)). These ensembles can be further distinguished by their respective radii of gyration (Figure 1(d)). Thus, 3ANK can be described as having a two-state folding transition, in agreement with experimental observations.[25] The folding transition state ensemble (TSE) is somewhat broad with the highest energy barrier occurring at $Q \approx 0.4$. To further pin down the position of the TSE, we also performed a large number of folding simulations starting with structures with different Q values and computed the probability of the system folding before it unfolds (P_{fold}).[26] The results are shown in Figure 1(c). Ideally, the kinetic transition state would be defined as the collection of microstates with $P_{fold} = 0.5$. The simulations show that this

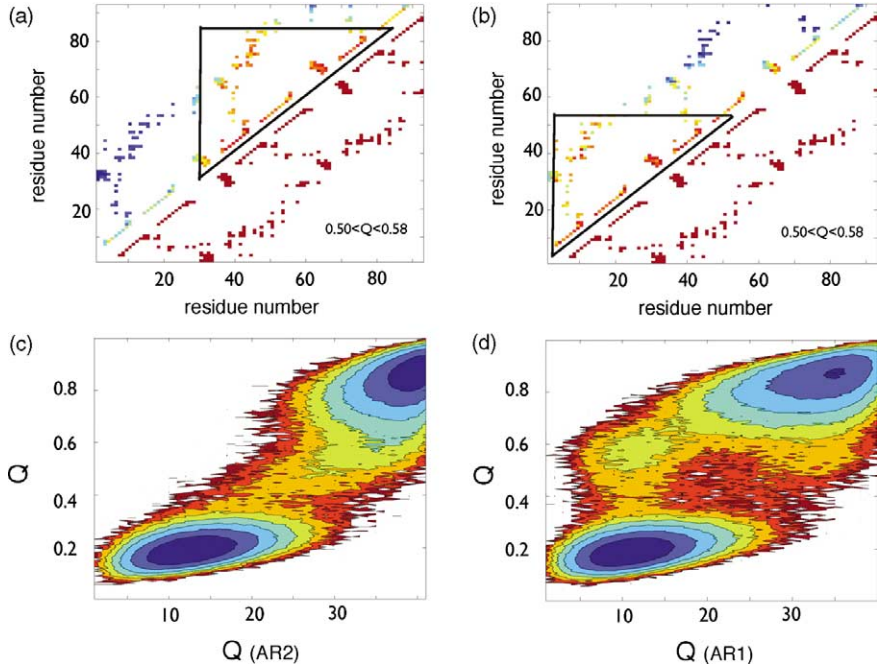

Figure 2. Folding nucleus of 3ANK. From the folding simulations of 3ANK the probability of native contact formation at the transition state ensemble is shown as a contact map. Over the diagonal, the color for each contact corresponds to values that range from 0 to 1 as quantified by the color scale. Below the diagonal the native contact map is shown for reference. Two structural ensembles were identified, and these are shown in (a) and (b). The free energy surface of the folding of p16 is plotted as a function of native contacts of the central ankyrin repeat (Q_{AR2}) *versus* the total fraction of native contacts Q in (c) and for the N-terminal ankyrin repeat (Q_{AR1}) in (d).

ensemble corresponds to structures with Q between 0.43 and 0.48.

Analysis of the structural distribution at the TSE indicates that folding nucleation occurs by the concomitant folding of one AR and the interface with an adjacent AR. Either the first AR and first helix of the second AR, or the whole second AR and first helix of the third AR pair are consistently found folded with high probability in the TSE region (Figure 2(a) and (b)). Analysis of the free energy surface with respect to the folding of the individual repeats is shown in Figure 2(c) and (d). These reveal that AR2 always contributes to the folding nucleus, and that a high-energy intermediate can be formed by the folding of AR2 and the interface with AR3, while AR1 remains unfolded. Thus, the folding of 3ANK can proceed through two distinct routes, nucleating within the interface of consecutive AR pairs. Since 3ANK is a highly symmetric protein, it is not surprising that the folding can proceed through parallel routes. Our results suggest that not every individual ankyrin repeat can act as a folding nucleus, but that an interface with a subsequent AR is also required.

Four repeats: p16

p16 is a naturally occurring ankyrin repeat protein, whose primary function is related to cell cycle progression, by interacting and inhibiting Cdk6 and Cdk4.[27] It has been reported that the equilibrium folding mechanism can be spectroscopically characterized as two-state, where only fully folded and unfolded conformations are populated.[16,17] However, the kinetic folding mechanism does not conform to such a simple two-state model, and a fast formed kinetic refolding intermediate was characterized.[17]

Initial simulated annealing runs showed a single folding T_f. Results from the simulations of the folding of p16 around T_f agree remarkably well with the experimentally determined equilibrium (Figure 3). Two states are mainly populated at equilibrium at T_f: the unfolded ($Q \approx 0.2$) and the fully folded state ($Q \approx 0.9$). However, the formation of a transient intermediate ($Q \approx 0.45$) is also readily apparent (Figure 3(b)). The free energy profile positions this species as a high-energy state, located on the unfolded side of the rate-limiting step, which occurs at $Q \approx 0.6$. To further characterize the structural ensemble of the intermediate, we measured the radius of gyration and the probability distribution of the native contacts. The broad distribution of the radius of gyration at $Q \approx 0.45$ suggests that a significant portion of the protein is not collapsed (Figure 3(d)). In all cases, the kinetic intermediate is formed by the folding of two terminal ankyrin repeats along with the interface between them, with almost no native contacts being formed by the remainder of the chain (Figures 3(c)

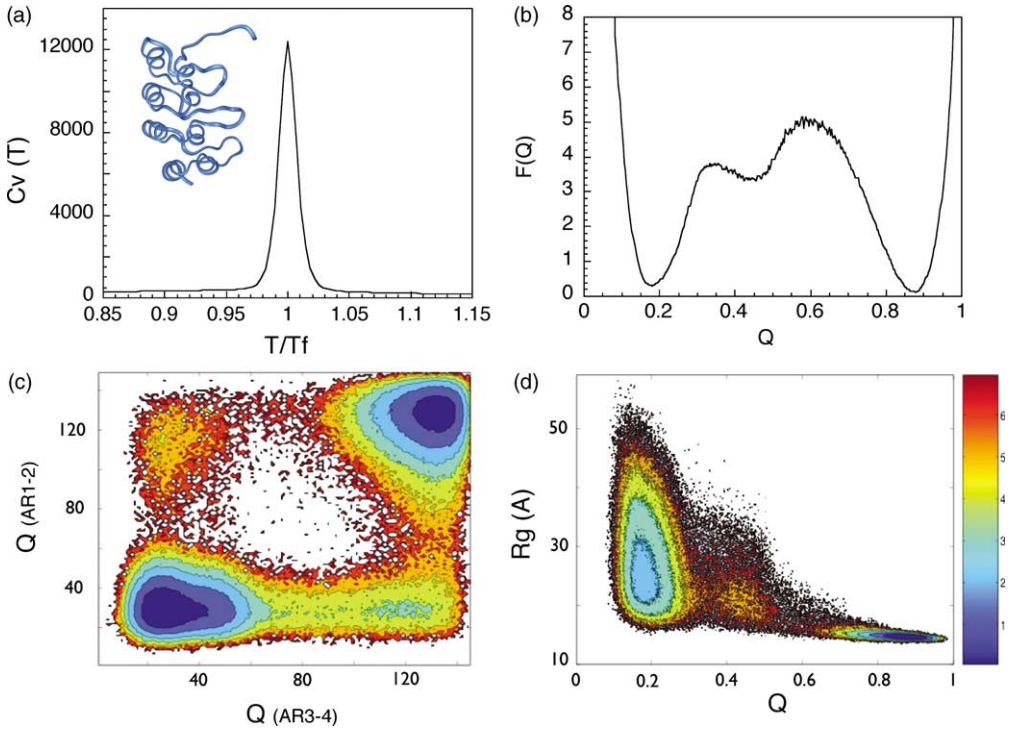

Figure 3. Folding simulations of p16. (a) The native structure is displayed in ribbon diagram, next to the heat capacity as a function of temperature. (b) The free energy $F(Q)$ profile as a function of the reaction coordinate Q was calculated at T_f. A high energy intermediate is readily apparent at $Q \approx 0.45$. (c) The free energy surface of the folding of p16 is plotted as a function of native contacts of the two N-terminal ($Q_{AR1\text{-}2}$) and the two C-terminal ($Q_{AR3\text{-}4}$) ankyrin repeats. (d) Free energy surface as a function of the radius of gyration and Q.

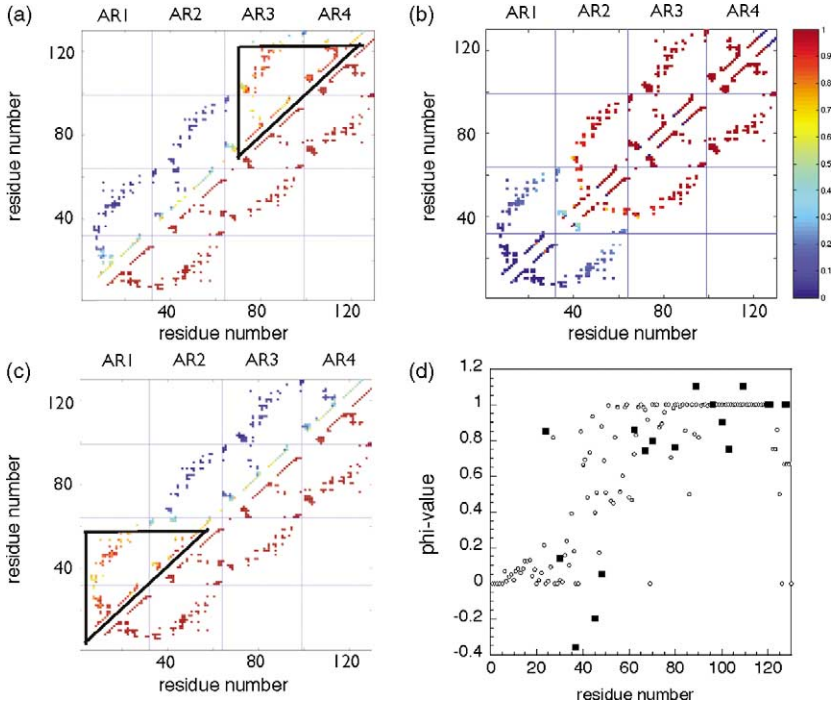

Figure 4. Kinetic intermediate and transition state ensemble of p16. From the simulations of the folding of p16, the probability of native contact formation of the kinetic intermediate was calculated. Two ensembles were identified, and shown in (a) and (b). These correspond to the ensembles identified in Figure 3(c). (c) The folding phi-values for each native contact were computed and shown as a contact map. (d) Comparison of the simulated (open circles) and experimental (filled squares)[29] phi-values as a function of sequence space.

and 4(a) and (b)). Based on the kinetic *m*-values, Itzhaki and co-workers proposed that the kinetic intermediate was 55% as compact as the native state.[17] To distinguish if the ensembles of microstates populating the kinetic intermediate correspond to parallel folding pathways or eventual kinetic traps, we took structures from each ensemble and performed 500 independent runs. We found that 5(±2)% of the runs actually reach the folded state before unfolding, regardless of the starting structure. These simulation results confirm that the intermediate basin is on the unfolded side of the rate-limiting step, and that both ensembles correspond to intermediates of parallel folding routes.

Deletion studies of p16 have shown that a fragment encompassing the two C-terminal ankyrin repeats indeed does form a stable cooperative structure, while none of the individual repeat peptides are stable by themselves.[16] On the other hand, the two N-terminal repeats could not be recombinantly expressed, and are believed to be intrinsically unstable.[16] In our simulations in ~70% of the folding events the intermediate is formed by the two C-terminal ankyrin repeats, and in the remaining ~30% of the trajectories an intermediate is formed by the two N-terminal ankyrin repeats, but never is an intermediate found with the central AR pair being folded (Figure 3(c)). Since the Gõ-type potential does not take into account any energetic frustration arising from non-native contacts, we can only attribute this behavior to topological differences in the native contacts (i.e. 10% more contacts are present in the interface AR3-4 than in the interface AR1-2).

One experimental tool used to characterize TSEs in the laboratory is the so-called phi-value analysis,

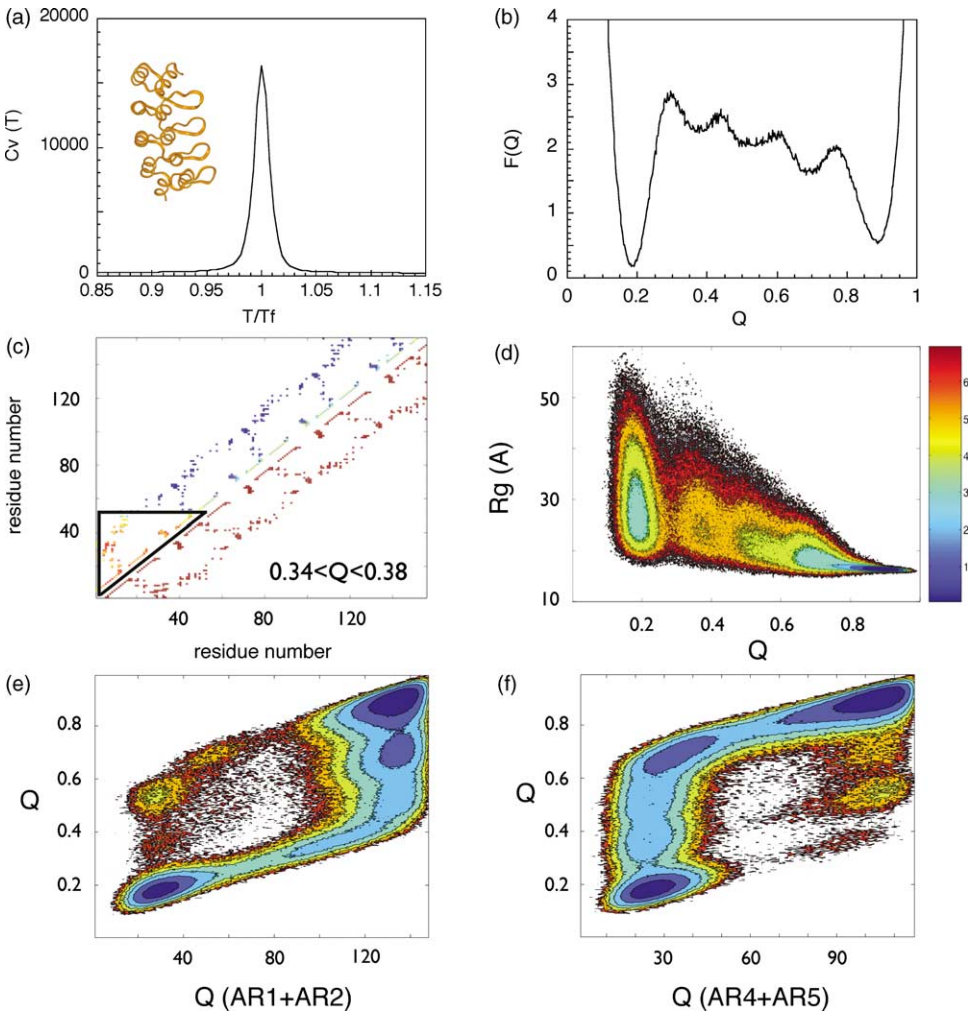

Figure 5. Folding simulations of E3_5. (a) The native structure is displayed in ribbon diagram, next to the heat capacity as a function of temperature. (b) Free energy $F(Q)$ profile as a function of Q, where three kinetic intermediates are apparent. (c) Probability of native contact formation of the ensemble at $Q \approx 0.4$. (d) Free energy surface as a function of the radius of gyration and Q. The free energy surface is plotted as a function of native contacts of the N-terminal ankyrin repeats (Q_{AR1+2}) *versus* the total fraction of native contacts Q in (e) and for the C-terminal ankyrin repeats (Q_{AR4+5}) in (f).

where the energetic effect of an amino acid substitution on the kinetic folding reaction is considered relative to the perturbation of the equilibrium caused by the substitution.[28] Since a similar analysis can be performed with the simulations by measuring the contact probability distribution at the TSE,[2] phi-value analysis is a convenient tool for validating the simulations. p16 is the only ankyrin repeat protein for which experimental phi-values have been measured.[29] The correlation with the simulated phi-values is shown in Figure 4(d). In both simulation and experiment, there is a clear asymmetry in the distribution of phi-values in sequence space, where high phi-values tend to cluster to the C-terminal half of the domain. We do not expect a strict quantitative correlation between the simulated and the experimental phi-values, since our model lacks energetic terms that are known to be relevant to the folding kinetics. The asymmetry found in the phi-value distribution is in accordance with the differential stability of the terminal ARs discussed above, and has to arise from topological factors since a simple native-contacts-based model captures the trait.

Five repeats: E3_5

The modular and versatile nature of AR proteins has encouraged the construction of combinatorial libraries where novel specific binding activities can be found.[25,30–32] The five AR protein E3_5 was selected from one of these libraries and used as a model system to study the folding of the library members.[31] The amino acid sequences of the AR repeats of E3_5 are $>80\%$ identical, conserving the consensus AR signature, and displaying variations in the common interaction sites, namely the β-fingers.[31] The members of this library show higher thermodynamic stability than do natural AR domains, and have been described as thermodynamic two-state folders.[31] Since this protein is made up of nearly identical repeats, it represents a good model for inquiring how the cooperative folding of AR proteins arises. If the folding profile is highly cooperative, no intermediates will be detected, while if each repeat folds in an independent manner, six equivalent states will be distinguished using Q as the reaction coordinate, i.e. the unfolded state, the fully folded states, and four intermediates. Several possibilities exist in between these extremes. Initial simulated annealing runs showed a single folding T_f. The presence of folding intermediates is readily apparent in the folding trajectories, but a single peak in the C_v plot reveals that all the transitions have nearly the same folding temperature (Figure 5(a)). The free energy profile shows that the most populated states are the folded and the unfolded conformations, as expected for a two-state folding protein, but three high-energy intermediates occur (Figure 5(b)). The fact that the intermediate states are "spaced" at regular Q intervals suggests that repeating modules fold sequentially leading to metastable conformations, with decreasing radius of gyration (Figure 5(d)). Analysis of the distribution of the contact probabilities at each basin shows a consistent "module" formed by the folding of the secondary structural elements of one complete AR and the first helix of the neighboring AR, together with the interfaces between them (Figure 5(c)). The rate-limiting step for this system is the nucleation of the first module at the N terminus, after which the folding proceeds downhill, gaining stabilization energy as the modules are sequentially added to the structural nucleus, essentially a nucleation-propagation mechanism. Analysis of the free energy surface with respect to the folding of the terminal repeats is shown in Figure 5(e) and (f). One major folding route is evident, ongoing from N to C termini, and a minority of the trajectories fold in the "reverse" manner. Thus, the whole protein undergoes an overall organization with slight cooperativity between folding modules, which are themselves formed in a highly cooperative fashion. This high degree of cooperativity results in the experimental observation that the thermal unfolding of E3_5 can be fitted to a two-state equilibrium model.[30] However, our simulations predict that a detailed kinetic study will reveal the presence of intermediates with increasingly consolidated structural elements.

Six repeats: IκBα

IκB proteins bind and regulate the activity and subcellular localization of the Rel/NF-κB transcription factor protein family.[33,34] All known IκB proteins contain a central ankyrin repeat domain consisting of six or seven ARs, which mediates specific interactions with NF-κB dimers, as shown in the co-crystal structure of IκBα in complex with the p50/p65 dimer.[35,36] This interaction involves contacts with several subdomains of NF-κB, each mediated by different ARs. Recent biochemical experiments show that IκBα displays a highly dynamic character when not complexed with NF-κB, benchmarked by its fast H/D exchange rates, extensive ANS binding, and its robustness to mutations at the NF-κB interaction interface.[37,38] Since IκBα might constitute a different model for AR domain folding, we performed folding simulations using the structure of IκBα from the co-crystal structure of the IκBα/NF-κB complex. In contrast to the other AR proteins described earlier, the simulated annealing runs showed more than one transition temperature and the formation of a stable structure with roughly half of the native contacts. Indeed, the WHAM analysis of several constant temperature runs shows two distinctive peaks in the C_v plot, both with equivalent energetic gains (Figure 6(a)). In order to identify the folding basin traps, free energy diagrams were computed at each folding temperature, T_f1 and T_f2 (Figures 6(b) and 7(b)). The high temperature C_v peak, T_f1, corresponds to the transitions between the

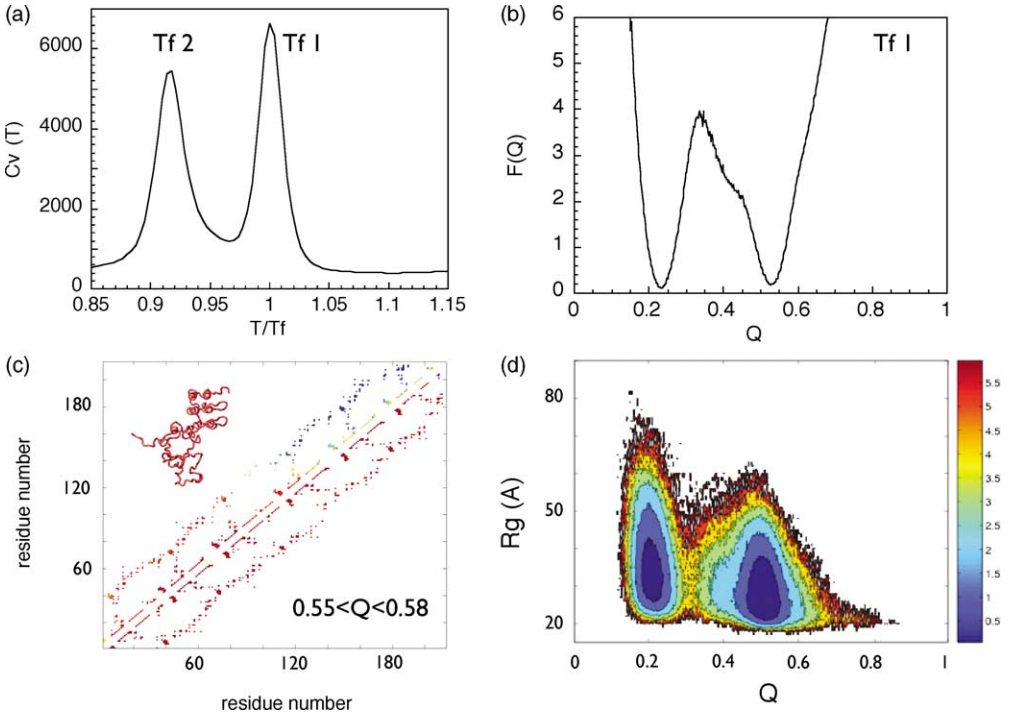

Figure 6. Folding simulations of IκBα. The heat capacity of IκBα as a function of temperature is shown in (a) which includes two peaks (T_f1 and T_f2), indicating the existence of more than one transition temperature. (b) Free energy profile at T_f1. The basin at $Q \approx 0.2$ corresponds to the unfolded state and a stable equilibrium intermediate is apparent at $Q \approx 0.55$. (c) The structural ensembles at the intermediates' free energy minimum (at $Q \approx 0.55$) are shown as contact maps, together with a representative snapshot. (d) Free energy surface at T_f1, as a function of the radius of gyration and Q.

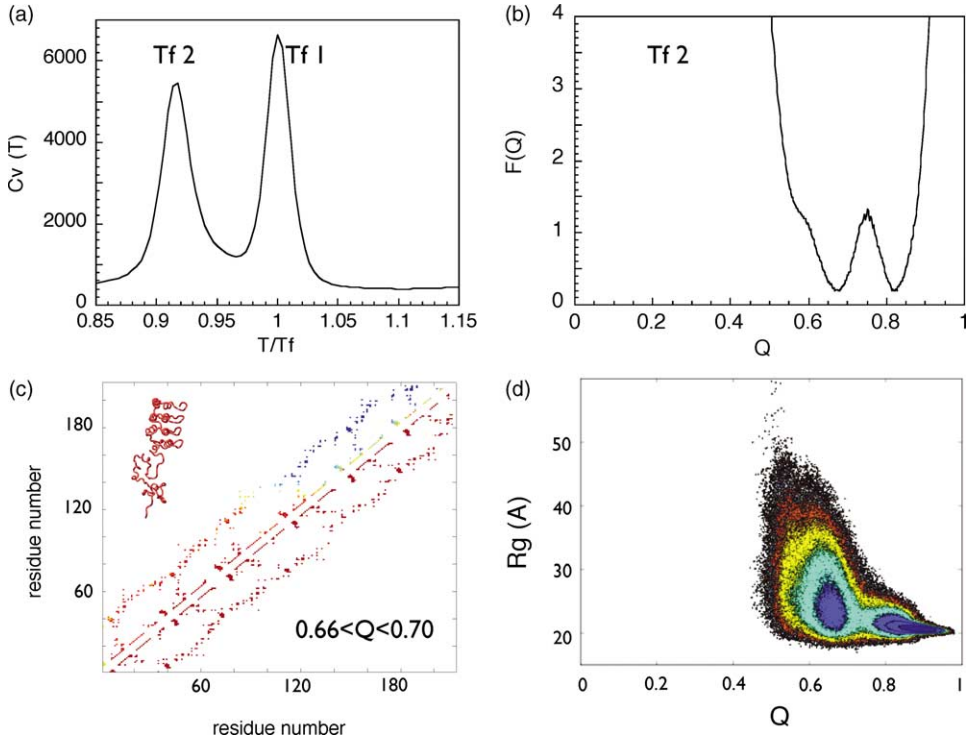

Figure 7. Folding simulations of IκBα. The heat capacity of IκBα as a function of temperature is shown in (a) which includes two peaks (T_f1 and T_f2), indicating the existence of more than one transition temperature. (b) Free energy profile at T_f2. The basin at $Q \approx 0.85$ corresponds to the folded state and a stable equilibrium intermediate is apparent at $Q \approx 0.66$. (c) The structural ensembles at the intermediates' free energy minimum (at $Q \approx 0.66$) are shown as contact maps, together with a representative snapshot. (d) Free energy surface at T_f2, as a function of the radius of gyration and Q.

unfolded basin at $Q \approx 0.2$ and a basin at $Q \approx 0.5$ contacts, that can be characterized as a two-state transition with a trough point after a high energy barrier (Figure 6(b)). The lower temperature $T_f 2$ corresponds to the transitions between basins at $Q \approx 0.7$ and $Q \approx 0.8$ with a moderate activation barrier (Figure 7(b)).

The structural make-up of these basins is shown in Figures 6(c) and 7(c). As in the case described earlier for E3_5, the folding proceeds through well-defined structural modules: an initial nucleation event takes place by the concomitant folding of the secondary structural elements of AR2 and the first helix of AR3 together with the interface between them. This is then propagated to the N-terminal AR1, forming a stable equilibrium intermediate (Figure 6(c)). In the low temperature regime, the transition corresponds to the coincident folding of the second helix of AR4, all the structural elements of AR5 and the first helix of AR6, together with the corresponding interfaces (Figure 7(c)). Thus, the IκBα protein can also be described as folding *via* a nucleation-propagation mechanism, where two separate (but not independent) nucleations are necessary to attain complete folding, each consisting of three consecutive ankyrin repeats. To date, there are no published observations on the equilibrium folding behavior of this particular protein, but we can draw some parallels with biophysical and phylogenetic data. Sequence alignments between the IκB homologs show that insertions of up to 40 amino acids can be accommodated between AR3 and AR4, which suggests that the folding of both subdomains may be uncoupled.[36] On the other hand, $H/^2H$ exchange data have revealed that the amides of AR2 and 3 exchange considerably less deuterium than do those in the AR4 to 6.[38] Our results suggest that the native-state $H/^2H$ exchange reflects the partial unfolding events in the native basin. We predict that a stable intermediate will be found for IκBα equilibrium folding mechanism, including the partial unfolding of the C-terminal ARs.

Seven repeats: Notch

Bearing seven ARs, the cytoplasmic domain of *Drosophila melanogaster* Notch receptor is the largest AR protein for which extensive experimental information on the folding mechanism is available.[19,39,40] Initially, equilibrium spectroscopic analysis suggested a two-state folding behavior, showing a strong cooperativity between its ARs,[19] but a detailed mutagenesis approach aimed at destabilizing each individual AR showed that the overall cooperative behavior may become uncoupled.[39] In particular, AR1 is intrinsically unstable, while the folding of AR2-5 may be thermodynamically uncoupled from the folding of AR6-7, as a result of a destabilization of AR6 by an alanine to glycine substitution. In the latter case, the folding transitions are no longer consistent with a two-state mechanism.[39]

Initial simulated annealing runs of the Notch AR domain showed a single folding T_f. A single peak is evidenced in the C_v plot (Figure 8(a)), although a stable folding intermediate at $Q \approx 0.4$ is evidenced (Figure 8(b)). The free energy diagram shows that the folding temperature actually corresponds to two major transitions, one between the unfolded basin at $Q \approx 0.2$ and an intermediate basin at $Q \approx 0.3$ and the other transition occurring from $Q \approx 0.4$ to $Q \approx 0.8$. At T_f, the basin of the intermediate state at $Q \approx 0.3$–0.4 is as stable as the native state, a situation that can mask the formation of an intermediate if spectroscopic signals are to be recorded. The population of the intermediate basin was identified by computing the probability of contact formation in this Q range, and can be ascribed to two different situations (Figure 8(c) and (d)). The intermediate forms either by the coincident folding of AR5 and AR6 and the interface between them (Figure 8(c)), or by the folding of AR2 and the first helix of AR3 and their corresponding interface (Figure 8(d)). These parallel folding routes are evidenced in Figure 8(e). Since AR1 and AR7 are found unfolded at T_f (not shown), AR2 and AR6 constitute *de facto* terminal repeats. As in the case described earlier for p16, the folding intermediate of Notch is formed by the folding of AR repeats at either end of the protein, but never by the central AR pair (Figure 8(f)). Recent experimental evidence also suggests that the folding of Notch AR domain is preferentially nucleated at AR2 or AR5.[40] Our simulations indicate that topological differences captured by Gō-models are sufficient to account for the overall folding behavior.

Conclusions

Unlike typical globular proteins, where distant long-range interactions are of paramount importance, the folding of repeat-containing proteins must rely on more local interactions. To answer whether the same underlying principles that make the foundations of our understanding of globular proteins could be extended to non-globular repeat-containing proteins, we undertook a folding simulation survey of ankyrin repeat domain folding mechanisms using models based on native topology alone. Using simplified minimally frustrated models, we were able to reproduce several aspects of the experimentally determined folding mechanisms, and make testable predictions for the unknown ones. This observation indicates that non-globular proteins also display funneled energy landscapes, further extending the application of these now well-established concepts[1,41,42] that had been previously restricted to globular proteins.

The present study points out that the topological minimal folding module of ankyrin repeats consists of one complete AR and the first helix of the neighboring AR, folding together in a highly cooperative fashion (Figure 9). This module

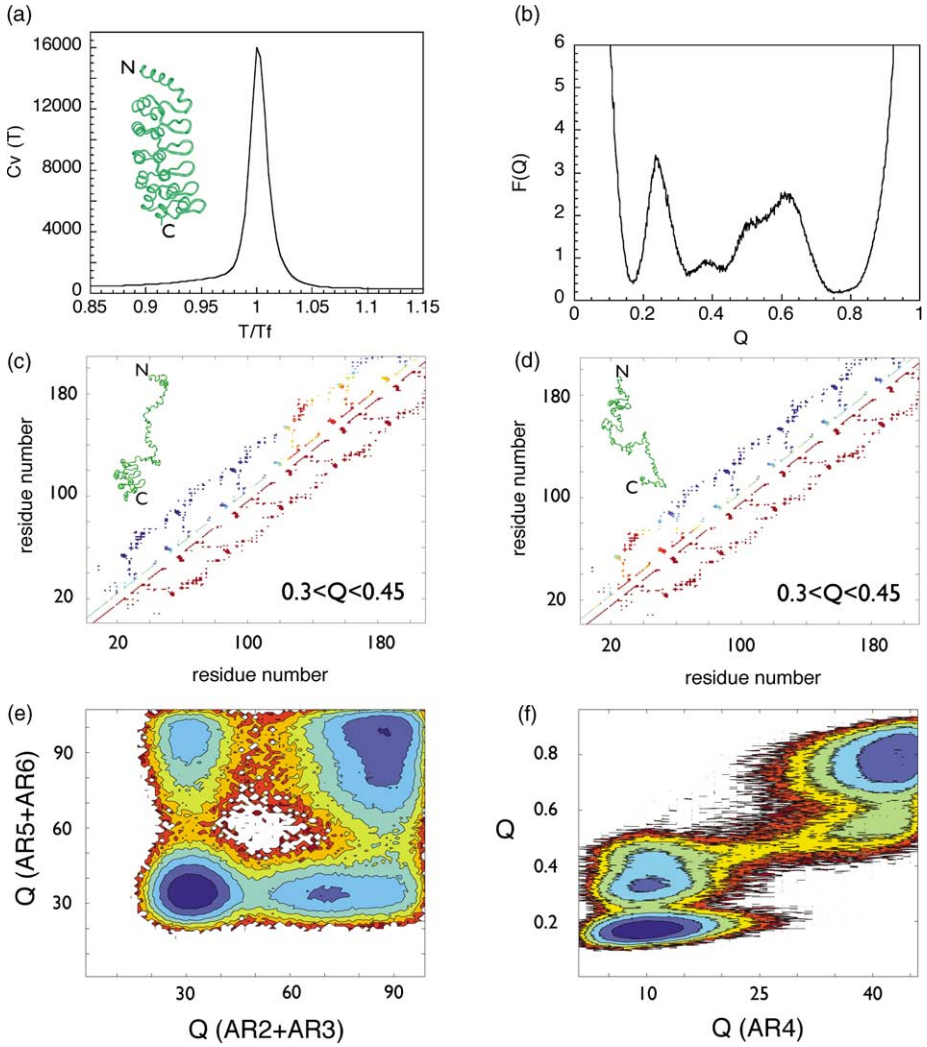

Figure 8. Folding simulations of Notch. (a) Heat capacity as a function of temperature, together with a ribbon representation of the native structure. (b) Free energy profile as a function of Q. A stable intermediate is apparent at $Q \approx 0.3$–0.4. (c) and (d) The structural population of the intermediate's basin, are shown as contact maps. Two different structural ensembles contribute to this intermediate, either folding from the C-terminal AR (c) or the N-terminal AR (d). (e) The free energy surface is plotted as a function of native contacts of the ankyrin repeats 2 and 3 (Q_{AR2+3}) and ankyrin repeats 5 and 6 (Q_{AR5+6}). (f) The free energy surface is plotted as a function of native contacts of the central ankyrin repeat (Q_{AR4}) *versus* the total fraction of native contacts Q.

includes the folding of the β-hairpin region between the AR, and is further stabilized by packing interactions at the interface of the helical motifs. It is possible that for a single ankyrin repeat, the entropic cost of folding is larger than the energetic contribution of intra-repeat interactions, a situation that may be overcome by the stabilization energy gained by the inter-repeat interactions. In this respect, our results suggest that the folding of repeat proteins is preferentially initiated at terminal repeats, for which the entropic cost of folding would probably be less than for internal repeats.

Given the modular architecture of ankyrin repeat domains, it is somehow surprising that the equilibrium folding experiments can be described by two-state folding transitions.[43] We show here that this can be the result of two different situations: on the one hand, the energetic coupling (either of enthalpic or entropic basis) between the repeats might be high relative to the stability of the individual repeats, so that a strong cooperative folding results for the entire domain, a situation likely to arise when the number of repeats is low. On the other hand, as the number of ARs increases, the folding of the structural modules decouples as a result of an imbalance of the enthalpic gain and the entropic folding cost. This situation leads to the formation of stable intermediates that require further nucleation to attain folding. However, since these intermediates are composed of topological modules that may display similar stabilities, the spectroscopic signals can still conform to overall two-state equilibrium folding models.

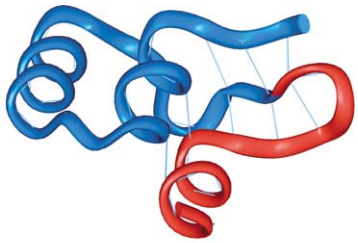

Figure 9. Topological folding nucleus of ankyrin repeats. The hypothetical minimal folding module of ankyrin repeats consists of one complete AR and the first helix of the N-terminal neighboring AR. The sequence of concensus AR[8] is shown below, and the structure is colored accordingly.

On the basis of perfectly funneled landscapes simulations, we cannot rule out that domain swapped intermediates may arise. Yet, these must be clearly disfavored, since most of the experimentally observed intermediates are well accounted for by the present models that do not allow swapping. Also, preliminary results from our group (S.S.C. & P.G.W., unpublished) indicate that when domain swapping is permitted as in a "symmetrized Gõ-model",[22] very little intramolecular swapping is observed, at least in the case of a four ankyrin repeat model.

Our study shows that although the individual ankyrin repeat domains have overall similar topological features, subtle differences in the number and position of the contacts between the repeats can lead to the occurrence of dramatically different overall mechanisms. These behaviors are well captured by off-lattice Gõ models. Since AR proteins are highly symmetric, it is not trivial to determine directly from the contact maps how the different topological factors influence the outcome of the whole system of contacts. A thoughtful perturbation analysis will be required to gain insight into these conundrums.

Finally, we suggest that the balance between the folding and the coupling among the repeats may be of functional significance for repeat-containing proteins. The modification of the folding state of one repeat (by covalent modifications and/or binding to other macromolecules) can facilitate the folding of contiguous repeats, propagating the changes to distant sites. In turn, if other structural elements are present in between the repeats (something not unusual in the ankyrin repeat case) the cooperative folding of the repeats may become uncoupled. These mechanisms would provide the means to differentially transmit allosteric effects between distant sites along a non-globular array.

Model and Methods

Simulated proteins

The high-resolution structures of AR proteins studied were taken from the Protein Data Bank (PDB),[15] with the corresponding PDB code in parentheses: 3ANK (1N0Q),[25] p16 (1D9S),[44] E3_5 (1MJ0),[30] IκBα (1NFI),[35] Notch (1OT8).[45]

Simulation algorithms

All the simulations were performed with a C^{α}-based Gõ model,[2] that takes into account only interactions present in the native structure and therefore does not include energetic frustration. An interaction between a pair of residues (i, j) exists if the distance between the C^{α} atoms of the residues is smaller than 8 Å or the distance between any side-chain heavy atoms in the two residues is smaller than 4 Å. Native contacts between pairs of residues (i,j) with $|i-j|<4$ were discarded from the native contact list because contiguous residues already interact through the bond, angle and dihedral terms.

Each residue is represented by a single bead centered in its C^{α} position, and adjacent beads are strung together into a polymer chain by means of a potential encoding bond length and angle constraints. The secondary structure is encoded in the dihedral angle potential and the non-bonded (native contact) potential. The interaction energy U for a given protein conformation Γ is given by:

$$\begin{aligned} U(\Gamma,\Gamma_0) = & \sum_{\text{bonds}}^{N-1} K_b(b_i-b_{0i})^2 + \sum_{\text{angles}}^{N-2} K_\theta(\theta_i-\theta_{0i})^2 \\ & + \sum_{\text{dihedrals}}^{N-3} \{K_\Phi^{(1)}[1-\cos(1\times(\Phi_i-\Phi_{0i}))] \\ & + K_\Phi^{(3)}[1-\cos(3\times(\Phi_i-\Phi_{0i}))]\} \\ & + \sum_{\text{native contacts}|i-j|>3} \left\{\varepsilon\left[5\left(\frac{r_{0ij}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{0ij}}{r_{ij}}\right)^{10}\right]\right\} \\ & + \sum_{\text{non-native contacts},|i-j|>3} \left(\frac{C}{r_{ij}}\right)^{12} \end{aligned} \quad (1)$$

In this equation, b_i, θ_i, and ϕ_i stand for the ith virtual bond length between ith and $(i+1)$th residue, the virtual bond angle between $(i-1)$th and ith bonds, and the virtual dihedral angle around the ith bond, respectively. The parameters b_{oi}, θ_{oi}, and ϕ_{oi} stand for the corresponding variables in the native structure. In the framework of the model, all native contacts are represented by a 10–12 Lennard-Jones potential without any discrimination between the various chemical types of interactions. The r_{ij} and r_{0ij} are the C^{α}–C^{α} distance between the contacting residues i and j in conformation Γ and Γ_0 (the PDB structure), respectively. In the summation over non-native contacts, C (=4.0 Å) parameterizes the excluded volume repulsion between residue pairs that do not belong to the given native contact set. Here, all the temperatures and energies are reported in units of ε. For other parameters, we use similar values to those used in several other folding studies,[2,5] namely, $K_b=100.0$, $K_\theta=20.0$, $K_\phi^{(3)}=0.5$, $K_\phi^{(3)}=0.5$, $\varepsilon=1.0$.

We used molecular dynamics for simulating the dynamics of these protein models. We employed the

simulation package AMBER at constant temperature, except for the initial simulated annealing runs. For each model, between 100 and 200 constant temperature simulations were made and combined using the WHAM algorithm[24] to generate a heat capacity profile as a function of temperature and a free energy $F(Q)$ as a function of the reaction coordinate Q.[2] For the calculations of the probability of contact formation, a contact was considered to be formed if the distance between the C^{α} atoms is shorter than γ times their native distance r_{0ij}. It has been shown that the results do not strongly depend on the choice made for the cut-off distance γ.[6] Here we used $\gamma = 1.25$.

The validity of Q as a reaction coordinate in describing the finer details of the folding mechanism has recently been questioned. It has been argued that Q, as a collective reaction coordinate, sometimes fails to correctly identify the transition state that separates the unfolded and folded states.[26] The reaction coordinate P_{fold} measures the probability of a given conformation to fold and the transition state, by definition, corresponds to P_{fold} equal to 0.5. The method has been posited as the ideal measure of the distance of a given conformation to the putative transition state ensemble for a protein exhibiting two-state behavior.[26] However, its calculation is computationally very costly, requiring numerous folding simulations for each tested conformation, and thus must be limited to a small set of conformations. It also has no simple direct structural interpretation. The use of P_{fold} as a reaction coordinate is further complicated by the presence of intermediates in folding landscapes. When there are intermediates, P_{fold} is not a local function of conformation. The P_{fold} of any configuration will depend in a complex way on the relative free energies of the native, denatured, intermediate state(s), and transition state ensembles between them. As such, variations in the intermediates' free energy levels will be reflected in P_{fold}, even though the TSE energetics and the unfolded and folded basins remain unaffected. For these reasons, we chose Q as a reaction coordinate to analyze the transitions.

To evaluate the validity of Q as a collective reaction coordinate in our study, we calculated P_{fold} for 3ANK, the only system whose P_{fold} will not be affected by kinetic intermediates. The other proteins all have distinct folding intermediates, making P_{fold} an inappropriate reaction coordinate for the reasons stated above. For the calculations of P_{fold}, ten structures at each Q value were randomly chosen from the trajectories and 100 independent runs were performed for each one. The P_{fold} of each structure was calculated, and the mean and standard deviation are reported for each Q value. It should be noted that the Gõ-model does not consider non-native interactions, and that the contact terms are strictly pairwise additive, assumptions that may have strong affects on the folding barriers. The close agreement between Q and P_{fold} for these models may be related to these approximations, which may not hold in more complicated scenarios.

To probe the nature of the transition state ensembles (TSE), we computed the Φ_{ij} values for a native contact pair between i and j from the probability of formation P_{ij}:

$$\Phi_{ij} = \frac{\Delta\Delta F^{\text{TS–U}}}{\Delta\Delta F^{\text{F–U}}} \approx \frac{P_{ij}^{\text{TS}} - p_{ij}^{\text{U}}}{P_{ij}^{\text{F}} - p_{ij}^{\text{U}}}$$

where $\Delta\Delta F$ is the free energy difference between the wild-type and the mutant protein, P_{ij} is the probability of formation of contacts between i and j, and the superscripts F, U and TS correspond to folded, unfolded and transition state ensembles, respectively. Since all non-bonded contacts in the Gõ model have the same energetics, the Φ_i value of residue i can be calculated from the contact values, Φ_{ij}, by averaging all the Φ_{ij} values that are assigned to residue i.

The protein structures belonging to the putative TSE were clustered using the Fitch program from the Phylip package.[46] This is an algorithm that was originally designed to create phylogenic trees based on a distance measure, and it has been adapted to measure similarity between a reference and a comparison structure as described by Hardin *et al*.[47]

Acknowledgements

We thank Koby Levy for his helpful suggestions and deep insights, and the Center for Theoretical Biological Physics for the computational resources. S.S.C. is supported by a University of California San Diego Molecular Biophysics training grant. This work was funded by NIH grant RO1GM044557-16. D.U.F. is a fellow of the Jane Coffin Childs Fund for Medical Research.

References

1. Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524–7528.
2. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and "*en-route*" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953.
3. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441–450.
4. Baker, D. (2000). A surprising simplicity to protein folding. *Nature*, **405**, 39–42.
5. Koga, N. & Takada, S. (2001). Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J. Mol. Biol.* **313**, 171–180.
6. Clementi, C., Jennings, P. A. & Onuchic, J. N. (2000). How native-state topology affects the folding of dihydrofolate reductase and interleukin-1beta. *Proc. Natl Acad. Sci. USA*, **97**, 5871–5876.
7. Clementi, C., Garcia, A. E. & Onuchic, J. N. (2003). Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. *J. Mol. Biol.* **326**, 933–954.
8. Main, E. R., Jackson, S. E. & Regan, L. (2003). The folding and design of repeat proteins: reaching a consensus. *Curr. Opin. Struct. Biol.* **13**, 482–489.

9. Sedgwick, S. G. & Smerdon, S. J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem. Sci.* **24**, 311–316.
10. Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins: Struct. Funct. Genet.* **17**, 363–374.
11. Breeden, L. & Nasmyth, K. (1987). Similarity between cell-cycle genes of budding yeast and fission yeast and the Notch gene of Drosophila. *Nature*, **329**, 651–654.
12. Michaely, P. & Bennett, V. (1993). The membrane-binding domain of ankyrin contains four independently folded subdomains, each comprised of six ankyrin repeats. *J. Biol. Chem.* **268**, 22703–22709.
13. Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J. *et al*. (2004). SMART 4.0: towards genomic data integration. *Nucl. Acids Res.* **32**, D142–D144.
14. Groves, M. R. & Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383–389.
15. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
16. Zhang, B. & Peng, Z. (2000). A minimum folding unit in the ankyrin repeat protein p16(INK4). *J. Mol. Biol.* **299**, 1121–1132.
17. Tang, K. S., Guralnick, B. J., Wang, W. K., Fersht, A. R. & Itzhaki, L. S. (1999). Stability and folding of the tumour suppressor protein p16. *J. Mol. Biol.* **285**, 1869–1886.
18. Mosavi, L. K., Williams, S. & Peng Zy, Z. Y. (2002). Equilibrium folding and stability of myotrophin: a model ankyrin repeat protein. *J. Mol. Biol.* **320**, 165–170.
19. Zweifel, M. E. & Barrick, D. (2001). Studies of the ankyrin repeats of the *Drosophila melanogaster* Notch receptor. 2. Solution stability and cooperativity of unfolding. *Biochemistry*, **40**, 14357–14367.
20. Levy, Y., Caflisch, A., Onuchic, J. N. & Wolynes, P. G. (2004). The folding and dimerization of HIV-1 protease: evidence for a stable monomer from simulations. *J. Mol. Biol.* **340**, 67–79.
21. Levy, Y., Wolynes, P. G. & Onuchic, J. N. (2004). Protein topology determines binding mechanism. *Proc. Natl Acad. Sci. USA*, **101**, 511–516.
22. Cho, S. S., Levy, Y., Onuchic, J. N. & Wolynes, P. G. (2005). Overcoming residual frustration in domain-swapping: the roles of disulfide bonds in dimerization and aggregation. *Phys. Biol.* **2**, S44–S55.
23. Go, N. & Taketomi, H. (1978). Respective roles of short- and long-range interactions in protein folding. *Proc. Natl Acad. Sci. USA*, **75**, 559–563.
24. Swendsen, R. H. (1993). Modern methods of analyzing Monte Carlo computer-simulations. *Phys. A*, **194**, 53–62.
25. Mosavi, L. K., Minor, D. L., Jr & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA*, **99**, 16029–16034.
26. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. S. (1998). On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334–350.
27. Serrano, M., Hannon, G. J. & Beach, D. (1993). A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature*, **366**, 704–707.
28. Fersht, A. R. (1999). *Structure and Mechanism in Protein Science*, Freeman, New York.
29. Tang, K. S., Fersht, A. R. & Itzhaki, L. S. (2003). Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure (Camb)*, **11**, 67–73.
30. Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Pluckthun, A. & Grutter, M. G. (2003). Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc. Natl Acad. Sci. USA*, **100**, 1700–1705.
31. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Pluckthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489–503.
32. Binz, H. K., Amstutz, P., Kohl, A., Stumpp, M. T., Briand, C., Forrer, P. *et al*. (2004). High-affinity binders selected from designed ankyrin repeat protein libraries. *Nature Biotechnol.* **22**, 575–582.
33. Baeuerle, P. A. & Baltimore, D. (1988). I kappa B: a specific inhibitor of the NF-kappa B transcription factor. *Science*, **242**, 540–546.
34. Baldwin, A. S., Jr (1996). NF-kappa B and I kappa B proteins: new discoveries and insights. *Annu. Rev. Immunol.* **14**, 649–683.
35. Jacobs, M. D. & Harrison, S. C. (1998). Structure of an IkappaBalpha/NF-kappaB complex. *Cell*, **95**, 749–758.
36. Huxford, T., Huang, D. B., Malek, S. & Ghosh, G. (1998). The crystal structure of the IkappaBalpha/NF-kappaB complex reveals mechanisms of NF-kappaB inactivation. *Cell*, **95**, 759–770.
37. Huxford, T., Mishler, D., Phelps, C. B., Huang, D. B., Sengchanthalangsy, L. L., Reeves, R. *et al*. (2002). Solvent exposed non-contacting amino acids play a critical role in NF-kappaB/Ikappa-Balpha complex formation. *J. Mol. Biol.* **324**, 587–597.
38. Croy, C. H., Bergqvist, S., Huxford, T., Ghosh, G. & Komives, E. A. (2004). Biophysical characterization of the free IkappaBalpha ankyrin repeat domain in solution. *Protein Sci.* **13**, 1767–1777.
39. Bradley, C. M. & Barrick, D. (2002). Limits of cooperativity in a structurally modular protein: response of the Notch ankyrin domain to analogous alanine substitutions in each repeat. *J. Mol. Biol.* **324**, 373–386.
40. Mello, C. C. & Barrick, D. (2004). An experimentally determined protein folding energy landscape. *Proc. Natl Acad. Sci. USA*, **101**, 14102–14107.
41. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct. Funct. Genet.* **21**, 167–195.
42. Plotkin, S. S. & Onuchic, J. N. (2002). Understanding protein folding with energy landscape theory. Part I: basic concepts. *Quart. Rev. Biophys.* **35**, 111–167.
43. Mosavi, L. K., Cammett, T. J., Desrosiers, D. C. & Peng, Z. Y. (2004). The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci.* **13**, 1435–1448.

44. Yuan, C., Li, J., Selby, T. L., Byeon, I. J. & Tsai, M. D. (1999). Tumor suppressor INK4: comparisons of conformational properties between p16(INK4A) and p18(INK4C). *J. Mol. Biol.* **294**, 201–211.
45. Zweifel, M. E., Leahy, D. J., Hughson, F. M. & Barrick, D. (2003). Structure and stability of the ankyrin domain of the Drosophila Notch receptor. *Protein Sci.* **12**, 2622–2632.
46. Felsenstein, J. (1989). PHYLIP—Phylogeny inference package. *Cladistics*, **5**, 164–166.
47. Hardin, C., Eastwood, M. P., Prentiss, M. C., Luthey-Schulten, Z. & Wolynes, P. G. (2003). Associative memory Hamiltonians for structure prediction without homology: alpha/beta proteins. *Proc. Natl Acad. Sci. USA*, **100**, 1679–1684.

Edited by J. Thornton

(Received 21 December 2004; received in revised form 13 September 2005; accepted 27 September 2005)

Available online 13 October 2005

Note added in proof: While this article was under revision, a detailed analysis of the folding kinetics of Notch ankyrin repeat domain was published (Mello, C. M. *et al. J. Mol. Biol.* (2005). **352**, 266–281). As our model predicted, a refolding intermediate was characterized, and the rate-limiting step was attributed to the formation of this intermediate.