doi:10.1016/j.jmb.2007.09.042 *J. Mol. Biol.* (2008) **375**, 837–854

JMB
Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER

Characterization and Further Stabilization of Designed Ankyrin Repeat Proteins by Combining Molecular Dynamics Simulations and Experiments

Gianluca Interlandi†, Svava K. Wetzel†, Giovanni Settanni, Andreas Plückthun* and Amedeo Caflisch*

Department of Biochemistry, University of Zürich, CH-8057 Zürich, Switzerland

Received 15 January 2007; received in revised form 11 August 2007; accepted 6 September 2007
Available online
21 September 2007

Edited by F. Schmid

Multiple molecular dynamics simulations with explicit solvent at room temperature and at 400 K were carried out to characterize designed ankyrin repeat (AR) proteins with full-consensus repeats. Using proteins with one to five repeats, the stability of the native structure was found to increase with the number of repeats. The C-terminal capping repeat, originating from the natural guanine-adenine-binding protein, was observed to denature first in almost all high-temperature simulations. Notably, a stable intermediate is found in experimental equilibrium unfolding studies of one of the simulated consensus proteins. On the basis of simulation results, this intermediate is interpreted to represent a conformation with a denatured C-terminal repeat. To validate this interpretation, constructs without C-terminal capping repeat were prepared and did not show this intermediate in equilibrium unfolding experiments. Conversely, the capping repeats were found to be essential for efficient folding in the cell and for avoiding aggregation, presumably because of their highly charged surface. To design a capping repeat conferring similar solubility properties yet even higher stability, eight point mutations adapting the C-cap to the consensus AR and adding a three-residue extension at the C-terminus were predicted *in silico* and validated experimentally. The *in vitro* full-consensus proteins were also compared with a previously published designed AR protein, E3_5, whose internal repeats show 80% identity in primary sequence. A detailed analysis of the simulations suggests that networks of salt bridges between β-hairpins, as well as additional interrepeat hydrogen bonds, contribute to the extraordinary stability of the full consensus.

Keywords: protein denaturation; protein engineering; network of salt bridges; folding pathways; ankyrin repeat proteins

Introduction

The hallmark of repeat proteins is their modular native-state architecture, which has been discovered in a variety of polypeptide families in the last decade.[1–3] The ankyrin repeat (AR) consists of 33 amino acids forming a loop, a β-turn and two antiparallel α-helices connected by a tight turn.[1] Multiple ARs are stacked in a linear array to form a rigid solenoidal native structure, which is stabilized predominantly by interactions between residues that are close in sequence (Fig. 1). Furthermore, the hydrophobic core of repeat proteins has a toroidal shape, unlike globular proteins. AR-

*Corresponding authors. E-mail addresses: plueckthun@bioc.uzh.ch; caflisch@bioc.uzh.ch.

†G.I. and S.K.W. contributed equally to this work.

Present addresses: G. Interlandi, Department of Bioengineering, University of Washington, Seattle, WA, USA; G. Settanni, MRC Center for Protein Engineering, University of Cambridge, Cambridge, UK.

Abbreviations used: AR, ankyrin repeat; DARPins, designed AR proteins; MD, molecular dynamics; PDB, Protein Data Bank; GdnHCl, guanidine hydrochloride; CD, circular dichroism; MALS, multiangle light scattering.

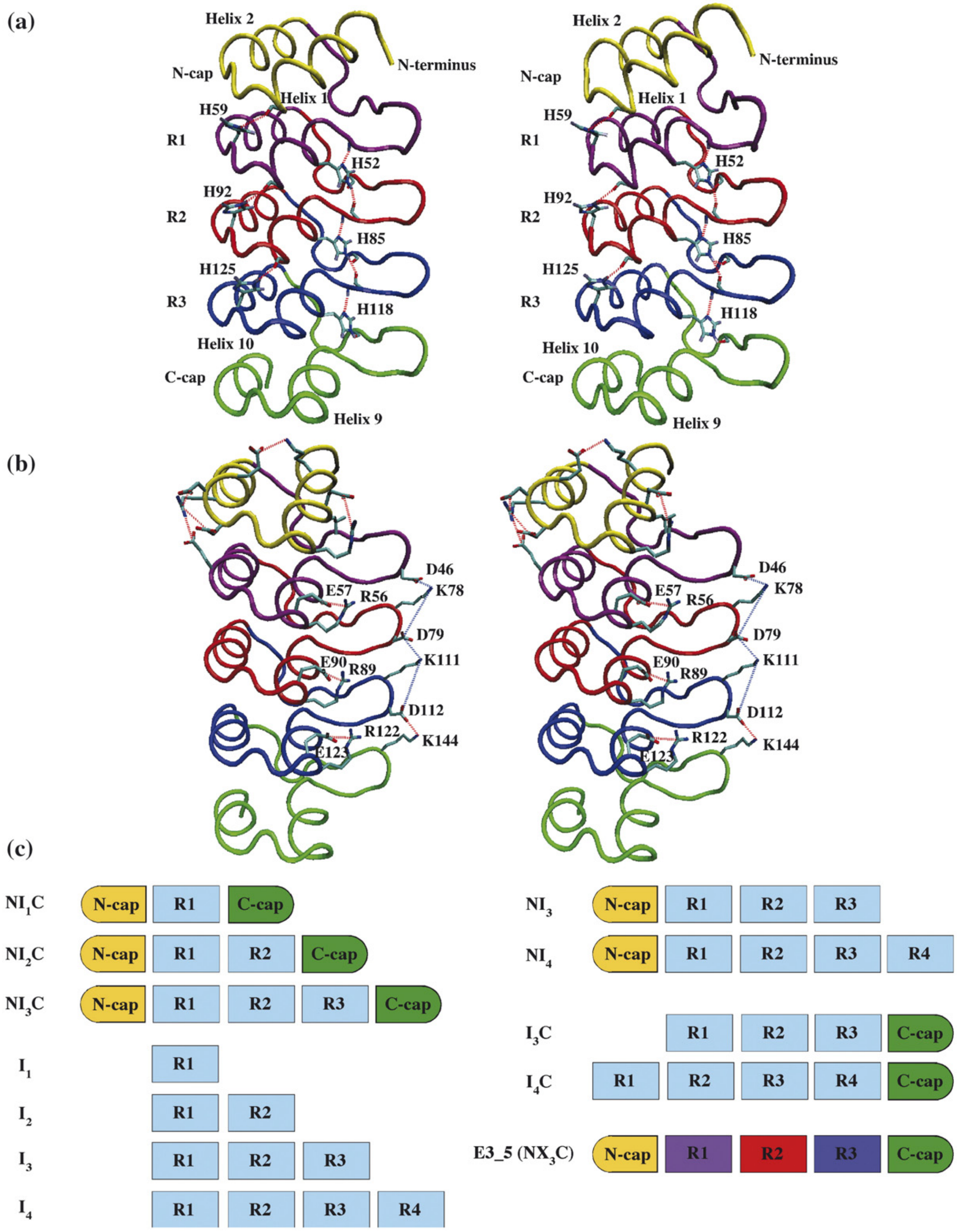

Fig. 1. (a and b) Stereo view of the DARPin NI$_3$C. The structure was modeled from the X-ray structure of E3_5,[12] which is identical in sequence in the N- and C-terminal capping repeats and is 80% identical in the internal repeats. The N-terminus is on top, and individual repeats are emphasized with different colors. Native hydrogen bonds (a) and salt bridges (b) are displayed as red dashed lines. The salt bridges involving the Lys78 and Lys111 side chains on the central loops are conserved in <50% of the simulation frames and are displayed in blue (see NI$_3$C *versus* E3_5). All side chains involved in contacts represented in (a) and (b) are shown in sticks and colored by atom type. (c) Cartoon representation of DARPins. Images (a) and (b) were prepared with the Visual Molecular Dynamics software.[40]

containing proteins are very common in nature and prevalently mediate specific protein–protein interactions.[4]

AR proteins consisting of several identical repeats have been designed and characterized biophysically.[5–7] By combining sequence and structure consensus analyses, an AR module was designed with seven randomized positions in the loop and in the first helix.[7,8] Different numbers of this module could be joined to generate combinatorial libraries of AR proteins. Because of the self-complementarity of consensus repeats, the size of the binding site can be altered simply by adding or removing repeats. To reduce the solvent exposure of hydrophobic surfaces, internal modules were flanked by N- and C-terminal "capping" repeats, which were borrowed from the guanine-adenine (GA)-binding protein[9] and slightly modified to approach the consensus and for cloning purposes.[7] Library members were then selected to function as specific binders or even enzyme inhibitors.[7,10,11] Designed AR proteins (DARPins) were shown experimentally to be thermodynamically stable, soluble and highly expressed in native form in bacteria.[7,12]

Up to now, only a few studies have been carried out to gain insight into the folding or unfolding pathways of AR proteins. The unfolding of the four-repeat tumor suppressor p16^{INK4a}, an inhibitor of a cyclin-dependent kinase, starts at the two N-terminal repeats, as suggested by mutagenesis experiments[13] and as verified by molecular dynamics (MD) simulations.[14] These repeats deviate more from the consensus sequence and may thus be intrinsically less stable. The role of topology in the energy landscape of AR proteins has recently been investigated by a simplified structure-based model.[15] The equilibrium folding behavior[16,17] and a kinetic on-pathway intermediate[18] have been experimentally characterized for the *Drosophila* Notch receptor and variants with different numbers of repeats. However, a much more detailed understanding of the folding and unfolding mechanism is essential to shed light on the stabilizing factors of AR proteins, especially in view of the increasing interest in their application in biotechnology.

Here, stabilizing interactions at room temperature and the unfolding pathways of several AR proteins and mutants thereof have been investigated by a combination of equilibrium unfolding experiments and multiple MD runs in explicit water for a total simulation time of >2 μs (Table 1). We first addressed the question of how the number of repeats affects stability. For this purpose, a full-consensus sequence with identical repeats was chosen.[19] We denote these proteins as NI$_x$C, where N and C refer to the N- and C-terminal capping repeats, respectively, I refers to the internal "full"-consensus repeat and the subscript x gives the number of identical internal-consensus repeats. Furthermore, the protein E3_5, a member of the NX$_3$C library[12] (where X denotes a library repeat module) with an 80% sequence identity to the full-consensus repeats, was chosen for MD analysis.

As mentioned above, the primary structure of flanking repeats differs from the consensus design. Hence, interrepeat interfaces involving the terminal repeats are different from interfaces between internal repeats.

To understand the role of the capping repeats in favoring solubility but potentially limiting stability, they were removed both experimentally and in MD simulations at room temperature and at high temperature. Moreover, mutations to further improve the stability of the C-terminal cap were suggested, and the MD simulations and equilibrium unfolding experiments of six NI$_1$C mutants and two NI$_3$C mutants were performed at room temperature.

The aim of the present study, combining simulations and experimental work, is to dissect the architecture of ARs to identify mutations that are critical for stability and to shed light on the role of the capping repeats and the relationship between stability and number of repeats. Furthermore, by an analysis of the high-temperature unfolding mechanism at the atomic level of detail, weak links may

Table 1. Simulation systems

Protein structure	Number of repeats[a]	Number of amino acids	Net charge[b] (electron units)	Box size[c] (Å)	Simulation time[d] (ns) 300 K	Simulation time[d] (ns) 400 K
NI$_1$C	3 (1)	90	−8	65.1	50, 40[e]	150, 200[e]
NI$_2$C	4 (2)	123	−10	80.6	50	100, 150[e]
NI$_3$C	5 (3)	156	−12	80.6/99.2	50	100
E3_5	5 (3)	156	−16	80.6/99.2	50	100
I$_1$	1 (1)	22[f]	−1	49.6	40	40
I$_2$	2 (2)	55[f]	−3	55.8	50	150
I$_3$	3 (3)	88[f]	−5	62.0	50	150

[a] Total number of repeats and, in parentheses, the number of noncapping repeats. In NI$_1$C, NI$_2$C and NI$_3$C, the N- and C-capping repeats flank the indicated number of identical consensus repeats. In I$_1$, I$_2$ and I$_3$, these capping repeats are missing, and only the indicated number of identical repeats is present. E3_5 is a member of the NX$_3$C library and differs at about seven positions per internal repeat, but has identical capping sequences.

[b] Total net charge of the protein at pH 7.

[c] Initial side length of the cubic box. The box adjusts its volume according to the given temperature and pressure during the simulation (see Materials and Methods). A larger water box is used to simulate NI$_3$C and E3_5 at 400 K.

[d] The total simulation time of the 300-K runs includes the initial 10 ns that were discarded during the analysis (see Materials and Methods).

[e] Two 300-K trajectories of NI$_1$C and two 400-K trajectories of NI$_1$C and NI$_2$C were run with different initial random assignments of the velocities.

[f] The 11 residues preceding the first helix of the first repeat were deleted because large displacements were observed in preliminary simulations of these proteins.

become apparent, in turn helping to understand the experimental unfolding data and the further design of ankyrins.

Results

Fluctuations and stabilizing interactions at room temperature

Comparison with crystallographic B-factors

In the loop regions of E3_5, larger fluctuation values of C_α atoms were observed along the MD trajectory at 300 K than those derived from *B*-factors (Fig. 2a). This discrepancy is probably due to intermolecular contacts between Loops 1 and 2 and two neighboring protein molecules in the crystal [Protein Data Bank (PDB) code 1MJ0[12]]. On the other hand, very low *B*-factors and fluctuations during the MD simulations characterize the helical and tight-turn regions.

Stabilizing interactions

DARPins are stable at 300 K. The values of the C_α root mean square deviation (RMSD) averaged over the interval 10–50 ns are 1.55±0.32 Å, 1.94±0.24 Å, 1.66±0.30 Å and 1.87±0.30 Å for NI_1C, NI_2C, NI_3C and E3_5, respectively. The corresponding all-atom RMSDs are 2.43±0.20 Å, 2.70±0.21 Å, 2.38±0.24 Å and 2.59±0.24 Å. Several polar interactions observed during 300-K runs contribute to the observed stability (see Materials and Methods for the definition of native contacts from the simulations). Interesting examples are the His59, His92 and His125 side chains (numbering according to PDB file 1MJ0) at the tight turn between helices 1 and 2 of each repeat (Fig. 1a), which are involved as donors in hydrogen bonds with carbonyl groups in the C-terminal turn of the first helix of the respective preceding repeat. These histidine side chains not only provide an interrepeat interaction but also contribute to the shielding of the interrepeat interface from the solvent. Additional hydrogen bonds involve the side chains of His52, His85 and His118, which are part of a conserved TPLH motif at the beginning of the first helix of each internal repeat, and the backbone carbonyl group of Ala75, Tyr81, Ala108, Tyr114 and Ala141 in the loop of the following repeat (Fig. 1a). Furthermore, each of these histidines accepts a hydrogen bond from the main-chain NH of the residue $n-3$ and thereby links two adjacent repeats. Shielding of these hydrogen bonds from the solvent is provided mainly by the side chain of the tyrosine in the loops of NI_3C (positions 48, 81 and 114).

NI_3C versus E3_5

In the 300-K simulations, there is a slightly higher number of hydrogen bonds in the full-consensus protein NI_3C than in the library member E3_5, in particular in the internal repeats (Table 2), which is consistent with the smaller fluctuations in NI_3C than in E3_5 (Fig. 2a). Furthermore, the larger number of charged residues in NI_3C compared to E3_5 (48 *versus* 36) leads to an increased number of salt bridges in NI_3C (Table 2 and Fig. 1b). In each

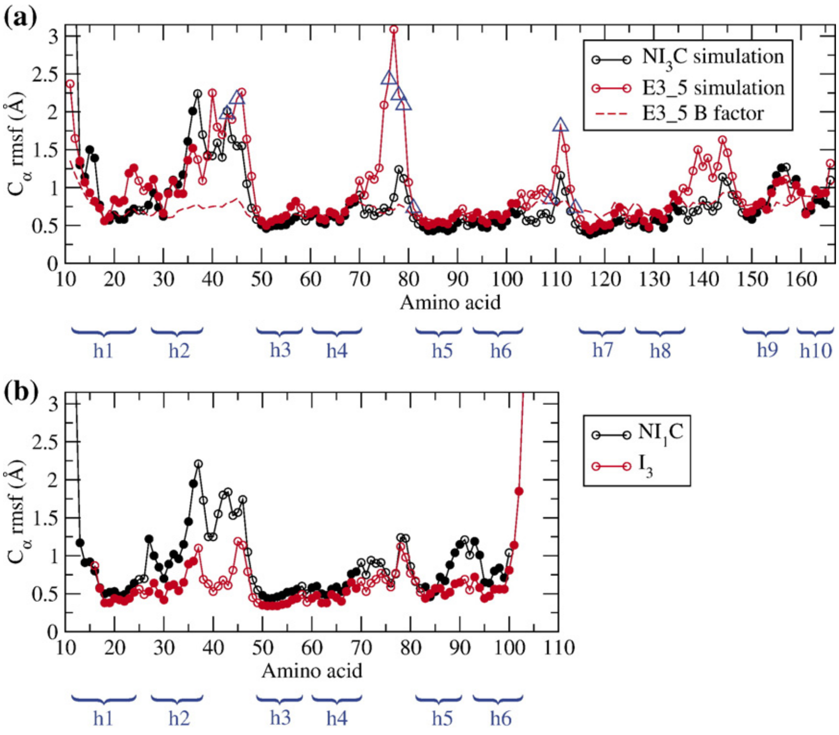

Fig. 2. C_α root mean square fluctuations (RMSFs) at 300 K. (a) NI_3C *versus* E3_5. Values of RMSF derived from crystallographic *B*-factors of the E3_5 C_α atoms (PDB file 1MJ0) were calculated using the formula $\mathrm{RMSF}_{i,\mathrm{exp}} = \sqrt{\frac{3}{8\pi^2}B_i}$, where B_i is the *B*- factor of C_α residue *i*. The interval 10–50 ns of each trajectory was used to calculate RMSF values. For each trajectory, a very similar behavior is observed for the interval 10–50 ns and for the eight 5-ns intervals (not shown). Blue triangles indicate the residues where E3_5 differs from NI_3C in primary sequence. (b) NI_1C *versus* I_3 (50-ns run). In the 40-ns run, NI_1C has a C_α RMSF sequence profile similar to that in the 50-ns trajectory. The amino acids located in the helices are represented as filled circles. Helical segments are emphasized by curled braces below the *x*-axis. h_1 to h_{10} denote helices 1–10.

Table 2. Native hydrogen bonds, salt bridges and C_α contacts

Protein structure	Total	N	R1	R2	R3	C	N-R1[a]	R1–R2	R2–R3	Rx-C[b]
Hydrogen bonds intraprotein										
X-ray[c]	147	25	26	27	25	25	3	3	6	7
E3_5	113	20	20	20	21	21	1	2	4	4
NI_3C	122	19	22	24	22	21	2	4	5	3
NI_2C	100	21	21	26		24	2	4		2
NI_1C	63	18	19			21	2			3
I_3	70		16	26	20			4	4	
I_2	37		17	17				3		
Salt bridges										
X-ray[c]	3	1	1	0	0	1	0	0	0	0
E3_5	3	2	0	0	0	0	1	0	0	0
NI_3C	9	4	1	1	1	0	1	0	0	1
NI_2C	7	2	1	1		1	1	0		1
NI_1C	5	3	1			0	1			0
I_3	3		1	1	1			0	0	
I_2	0		0	0				0		
C_α contacts										
X-ray[c]	290	44	43	43	42	39	20	19	21	19
E3_5	250	43	39	37	41	34	17	9	16	14
NI_3C	258	41	39	43	43	37	14	13	15	13
NI_2C	207	43	38	43		40	16	15		12
NI_1C	148	42	42			36	14			14
I_3	149		34	43	45			12	15	
I_2	87		34	44				9		

The values listed refer to polar interactions or C_α contacts present in >50% of the simulation time at 300 K (see Materials and Methods for distance threshold definitions). N and C denote the N- and C-caps, respectively, whereas R1 to R3 denote noncapping repeats 1 to 3, respectively (Fig. 1a).

[a] Interactions between the N-terminal cap and first repeat; the other interactions are denoted analogously.

[b] $x=3$ for E3_5 and NI_3C; $x=2$ for NI_2C; and $x=1$ for NI_1C.

[c] Crystallographic structure of E3_5 (PDB 1MJ0).

internal repeat of NI_3C, there is a salt bridge between an arginine and a glutamate, which are nearest neighbors in sequence and located at the C-terminal turn of the first helix. The Lys78 and Lys111 side chains (in the loop of repeats R2 and R3, respectively) are involved in a salt bridge with either the neighboring Asp79 (43% of the time) and Asp112 (23% of the time), respectively, or Asp46 (6% of the time) and Asp79 (39% of the time), respectively, located in the loop of the preceding repeat (Fig. 1b and Supplementary Fig. 2). Moreover, in NI_3C, the side chains of Asp112 and Lys144 are always at salt-bridge distance. These salt-bridge networks are likely to contribute to the higher thermodynamic stability[20,21] of NI_3C than E3_5 observed in the equilibrium unfolding experiments reported below.

I_1–I_3

To test the influence of the caps, structures were generated *in silico* where the capping repeats have been removed. The proteins I_2 and I_3, consisting only of two and three identical repeats, respectively, were stable at room temperature during the total simulation time of 50 ns. In contrast, I_1, consisting of one single AR, reached a C_α RMSD from the starting conformation of 6 Å already after about 30 ns. These simulation results report only on the kinetic stability of the folded state, which is the height of the activation barrier towards unfolding. Yet, they are consistent with experimental data indicating that at least two repeats are necessary for thermodynamic stability[22] (i.e., the energy difference between the folded state and the unfolded state). Interestingly, there are more hydrogen bonds in I_3 than in NI_1C (Table 2), which is likely to be one of the reasons for the smaller fluctuations in I_3 than in NI_1C (Fig. 2b).

In particular, NI_1C and all proteins with capping repeats lack two interrepeat hydrogen bonds involving the N-capping repeat, which are present between the internal repeats of NI_xC. These hydrogen bonds involve the side chain of the histidine of the TPLH motif, located at the beginning of helix 1 in the consensus design (see Stabilizing Interactions and Fig. 1a). The N-capping repeat contains, at the corresponding position (which is residue 19 according to the numbering in the PDB file 1MJ0), a leucine instead of a histidine.

Similarly, the C-cap contains, at the interhelical tight turn, an asparagine instead of a histidine (e.g., position 158 in NI_3C between helices 9 and 10). For this reason, the C-cap lacks the interrepeat hydrogen bond that is present between the designed consensus repeats stabilizing the macrodipole of the first helix of the preceding repeat (see Stabilizing Interactions and Fig. 1a). This observation was taken into account when suggesting the Asn92His mutation in NI_1C (see Mutations in the C-Terminal Cap). Furthermore, the terminal helix of the C-cap is three amino acids shorter than the full-consensus sequence. The missing amino acids are lysine, alanine and glycine. This causes the repeat adjacent to the C-cap in NI_3C to present a larger solvent-exposed hydrophobic surface (167± 21 $Å^2$) than the central repeat (140±19 $Å^2$) in the 300-K run. Moreover, the alanine present at the

C-terminal end of the full-consensus repeats increases the helical propensity of the C-terminal helix. This evidence was considered when suggesting a Lys-Ala-Ala extension of the C-cap terminal helix in NI_1C (see below).

Structural stability and high-temperature unfolding mechanism

Correlation between structural stability and number of repeats

The 400-K simulations with the NI_xC and I_x proteins ($x=1$, 2 and 3) allow us to analyze the influence of the number of repeats on structural stability (i.e., the kinetic stability of the native state). As mentioned above, I_1 is not stable at 300 K and fully unfolds after 15 ns at 400 K. At 400 K, >75% of the native interrepeat C_α contacts are lost at about 20 ns and 60 ns in I_2 and I_3, respectively (Fig. 3). Using the same criterion (i.e., loss of >75% of the native interrepeat C_α contacts), full unfolding is observed at about 40 ns and 160 ns in the NI_1C runs and at 90 ns in the 100-ns run of NI_2C, whereas complete unfolding is not reached during the 150-ns run of NI_2C and for NI_3C. Despite the very limited statistics that do not allow the evaluation of unfolding rates, the simulation results are consistent with the experimental observation that stability increases with the number of repeats. In fact, the experimentally measured thermodynamic stability has been shown to correlate with the number of repeats and, furthermore, to be due to slower rates of unfolding rather than faster folding rates for NI_xC ($x=1, 2, \ldots, 6$).[19] Similar results have been found for a series of tetratricopeptide repeat proteins,[23] and consistent equilibrium data have been reported for the deletion and duplication of repeats of the Notch receptor ankyrin domain.[16,17] However, this independence of folding rate from repeat number might not always be the case when consensus ARs are mixed with naturally occurring ARs. For example, the insertion of one to two consensus ARs into the five N-terminal repeats of Notch causes an increase in the folding rate.[24]

Sequence of events during unfolding at 400 K

The C-terminal repeat unfolds first for the AR proteins E3_5 and for all NI_xC molecules (Fig. 4 and Supplementary Figs. 3 and 4). The only exception is the 150-ns run of NI_2C, where the helical content and about one-third of the interrepeat contacts of the C-terminal repeat (Fig. 3 and Supplementary Figs. 3 and 4) are partially conserved with an average value of the C_α RMSD from the initial conformation of 6.9±0.5 Å during the last 10 ns. During the unfolding of the C-terminal repeat in all other molecules, the other repeats remain almost completely folded (Supplementary Fig. 4) and conserve most of their native interrepeat C_α contacts (Supplementary Fig. 4e). The C-terminal cap dislocates as a

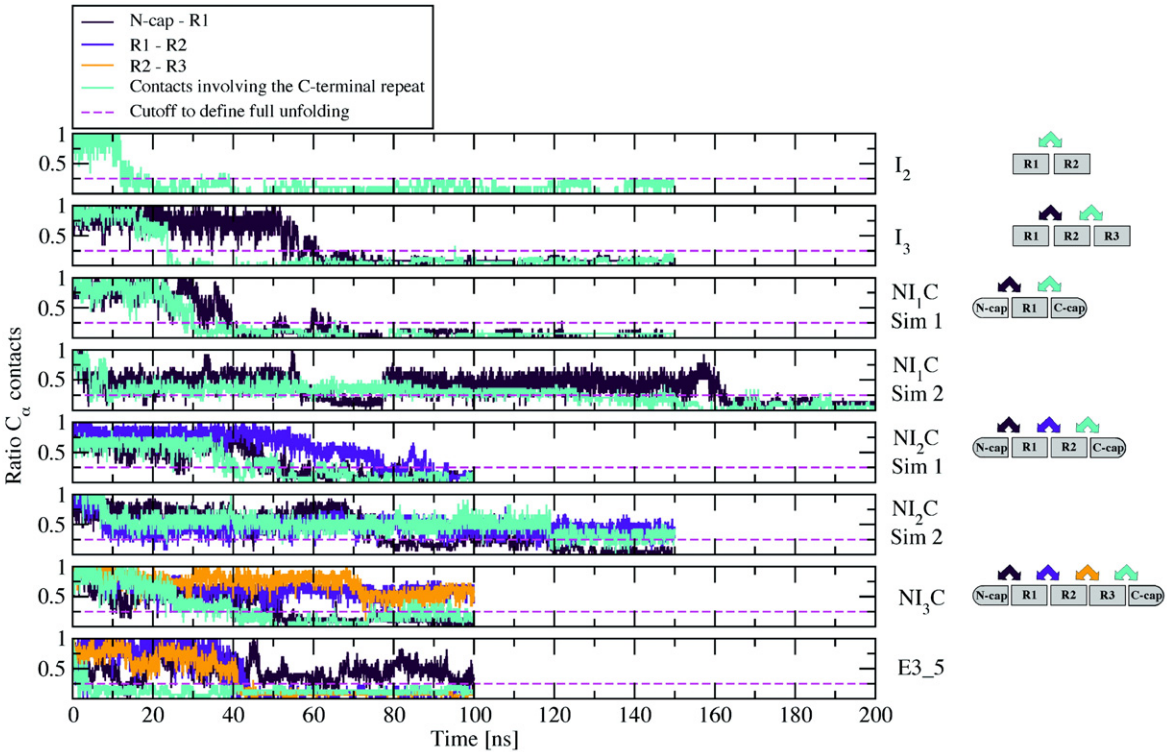

Fig. 3. Time series of the percentage of conserved native interrepeat C_α contacts at 400 K. A cartoon representation of the corresponding DARPin is presented on the right of the plot, with colored arrows indicating contacts between certain repeats. The colors correspond to the curves describing the time course of the respective interactions. For both NI_1C and NI_2C, "Sim 1" and "Sim 2" denote shorter and longer runs, respectively.

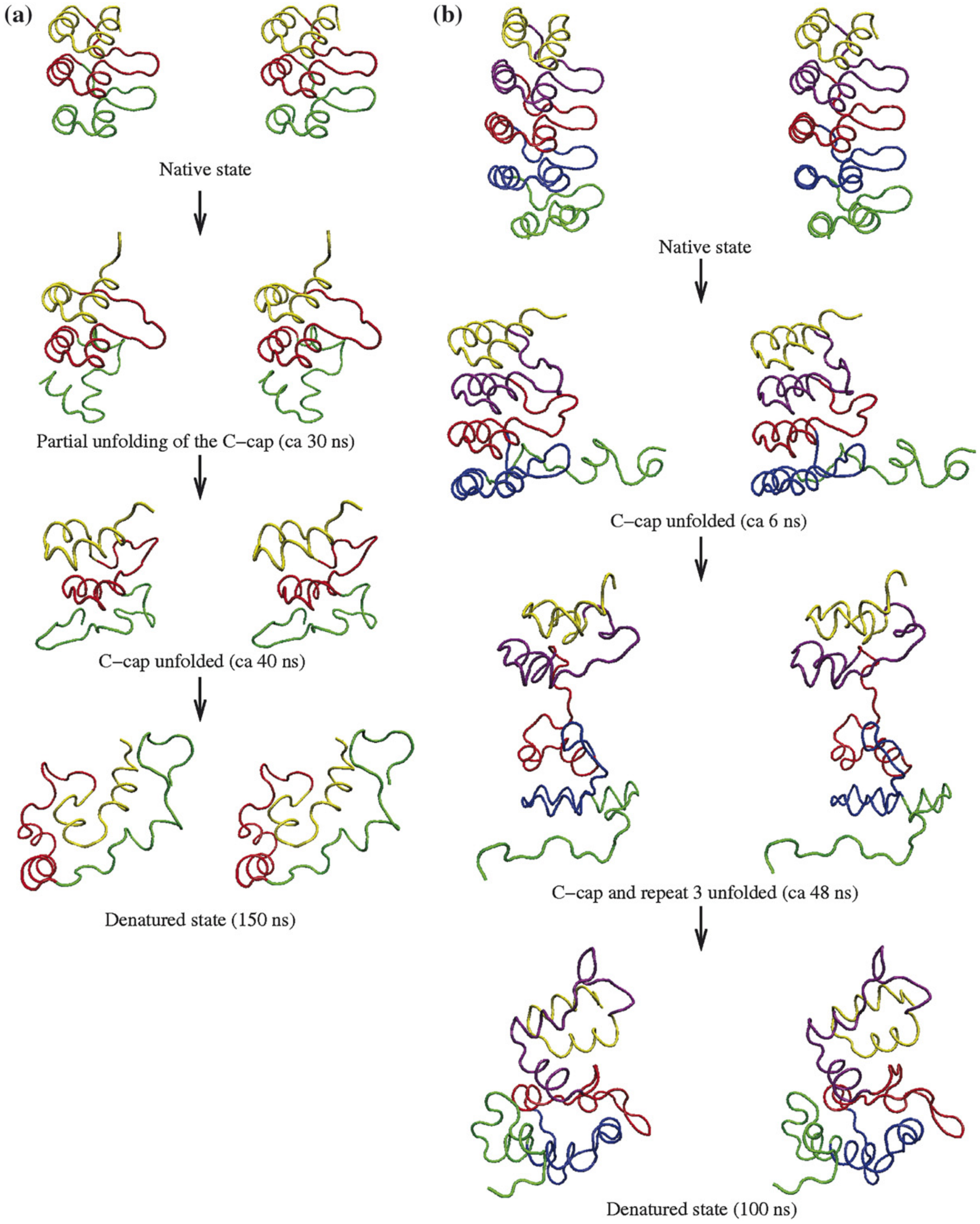
(a)
Native state
Partial unfolding of the C–cap (ca 30 ns)
C–cap unfolded (ca 40 ns)
Denatured state (150 ns)
(b)
Native state
C–cap unfolded (ca 6 ns)
C–cap and repeat 3 unfolded (ca 48 ns)
Denatured state (100 ns)

Fig. 4. Stereo view of representative conformations from the 400-K simulations of (a) NI_1C and (b) E3_5.

mostly intact unit first: The rupture of most interrepeat native C_α contacts involving the C-terminal cap (Supplementary Fig. 4e) precedes the unfolding of the helices making up the C-terminal cap (Supplementary Fig. 4c and d). The observation that the C-cap unfolds before all other repeats in the MD simulations may explain the first transition (before the main transition) in the experimentally determined guanidine hydrochloride (GdnHCl)-induced denaturation curve of NI_3C (see below).

After 100 ns of simulation with NI_3C, only the C-terminal cap is unfolded (Supplementary Figs. 3 and 4); thus, it is not possible to extract the complete sequence of unfolding events from the NI_3C run. In the other proteins, most of the interrepeat tertiary contacts are lost within the time scale of simulation (Fig. 3). The sequence of unfolding events is as follows (nomenclature as in Fig. 1; see Supplementary Figs. 3 and 4): C-cap/R1/N-cap in both runs of NI_1C (see also Fig. 4a), C-cap/N-cap/R1/R2 in the

100-ns run of NI_2C (as mentioned above, in the 150-ns run of NI_2C, the helical content was partially conserved) and C-cap/R2/R3/R1/N-cap in E3_5 (see also Fig. 4b). The unfolding of the central repeat R2 of E3_5 is preceded by the rupture of the hydrophobic interface between repeats R1 and R2, as indicated by the decrease of their native interrepeat C_α contacts (turquoise line in Fig. 3). The fact that the unfolding of repeat R2 of E3_5 directly follows the denaturation of the C-cap is consistent with the smaller number of native interrepeat hydrogen bonds between repeats R1 and R2 of E3_5, compared to NI_3C (Table 2), and to the relatively small number of native C_α contacts between repeats R1 and R2 of E3_5 (Table 2). Moreover, the fact that some internal repeats (e.g., repeat R2 of E3_5 and repeat R1 of NI_1C) unfold before the N-terminal cap provides evidence that the very high temperature (400 K) does not lead to artificial deformations at the protein surface in the simulations. Thus, it can be excluded that the observed early unfolding of the C-cap in the simulations is an artifact caused by the high temperature. Interestingly, the N-terminal cap seems to be more stable than the C-terminal cap, which is consistent with the fact that the N-terminal cap is more similar to the consensus repeat.

Denatured state

During the interval 155–200 ns in one of the two unfolding runs of NI_1C, helix 2 in the N-capping repeat elongates up to residue 47, with π-helical turns at its C-terminal region (Supplementary Fig. 3). The same elongation of helix 2 in the N-capping repeat takes place for NI_3C, where otherwise only the C-terminal capping repeat unfolded during a total simulation time of 100 ns (Supplementary Fig. 3). Nonnative π-helical structure is also present at residues 126–146 of E3_5 (Supplementary Fig. 3). Similarly, at the end of one of the two unfolding runs of NI_2C, residues 107–121 form a nonnative α-helical structure.

Experimental studies on the equilibrium unfolding of NI_3C and variants without capping repeats

Previously, the stability of several DARPins has been measured by both GdnHCl and thermal denaturation.[7,12] These proteins were unselected members of the DARPin library. They were all highly stable, even though some differences between individual library members can be noted. There was a general trend towards higher stability with an increasing number of repeats. However, since the individual library members of the same length covered a range of stabilities,[25] a quantitative relationship could not be established. Most of the proteins tested have previously shown highly cooperative reversible transitions that were consistent with a two-state equilibrium system.

The full-consensus proteins are even more stable, and the details of the dependence of folding and unfolding rates on the number of repeats in the protein are reported in the accompanying manuscript.[19] By the design of the consensus sequence, the previously variable library positions were now chosen according to the most frequent residues, which turn out to be charged or polar; a possible reason for the stability is that additional favorable electrostatic interactions are formed (see NI_3C *versus* E3_5).

When the full-consensus protein NI_3C is compared to E3_5, a member of the NX_3C library, two differences in GdnHCl equilibrium denaturation experiments are apparent (Fig. 5). First, the main transition is shifted to higher GdnHCl (by 0.8 M). Second, the denaturation is no more fully cooperative and is not consistent with a two-state equilibrium system. Instead, there is a small ”pre-transition” visible at 3.7 M GdnHCl, before the main transition, which occurs at about 5.6 M GdnHCl. The two transitions can be well described by a sequential three-state model, and the calculated ΔG (19.7 kcal/mol) is about double that obtained for E3_5 (11.2 kcal/mol; Fig. 5a).

A possible explanation is that the higher stability of the central domains in NI_3C uncouples the unfolding of one or both of the capping repeats, which may therefore unfold already at a lower denaturant concentration and give rise to an equilibrium intermediate with one or both of the caps detached from the central repeats. At this intermediate state, a portion of the circular dichroism (CD) signal is lost. This explanation is also supported by the unfolding behavior of all NI_xC proteins and E3_5 in the MD simulations where the C-cap unfolds prior to the other repeats.

To test this hypothesis, additional proteins, which were devoid of one or both of the capping repeats, were constructed. We denote them NI_3 or NI_4, to indicate that they have only the N-cap and three or four full-consensus repeats, respectively, and I_3C and I_4C, to indicate that they carry only the C-cap and the number of consensus repeats indicated by the number. Finally, we also created the molecules without any caps, which consist only of the consensus repeats and are named I_3 and I_4 (see Materials and Methods and Supplementary Fig. 1 for the definition of the respective sequences).

It is immediately apparent that the molecules lacking both N- and C-terminal caps show significant amounts of insoluble protein upon expression in *Escherichia coli* (Fig. 6), in contrast to E3_5 and NI_3C, which are completely soluble. The molecules lacking only one of the capping repeats could be purified from the soluble fraction and were further analyzed using multiangle light scattering (MALS). A portion of NI_3 and NI_4 precipitated after elution from the Ni–NTA column, while I_3C and I_4C remained soluble. MALS analysis showed that the proteins I_3C and I_4C form soluble aggregates, however (data not shown). In contrast, the soluble portion of the proteins NI_3 and NI_4, lacking the C-terminal cap, remains mainly monomeric. However, these proteins do remain aggregation-prone, as they aggregate at intermediate concentrations of

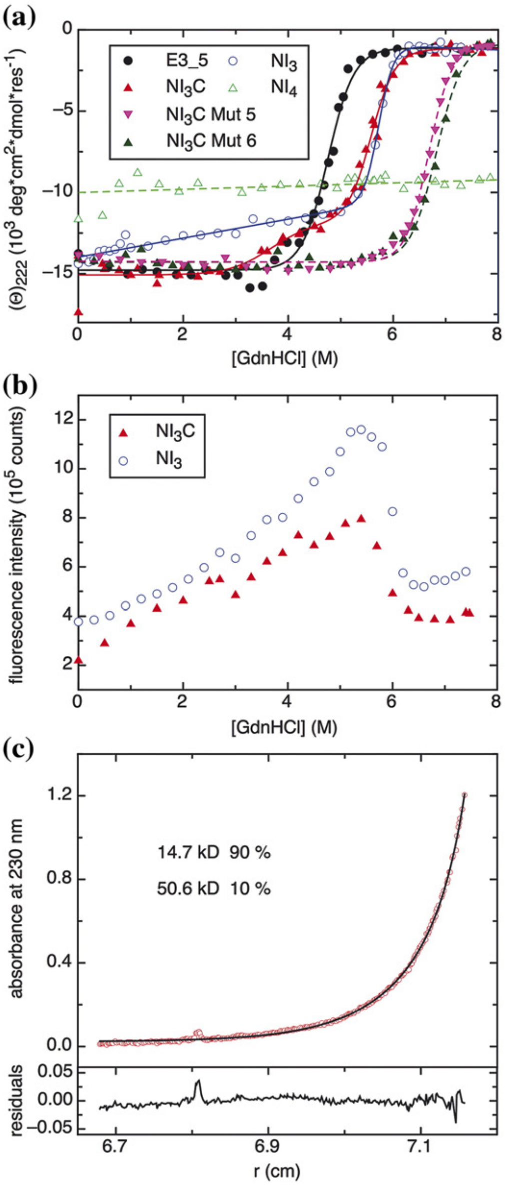

Fig. 5. GdnHCl-induced equilibrium unfolding of ankyrin proteins E3_5 (■), NI$_3$C (▲), NI$_3$ (○), NI$_4$ (△), NI$_3$C Mut 5 (▼) and NI$_3$C Mut 6 (▲) at 20 °C followed by (a) CD spectroscopy and (b) tyrosine fluorescence (see Materials and Methods). The lines represent the least-squares fit to the two-state model (for E3_5, NI$_3$) with a midpoint of denaturation D_m=4.8 M (ΔG=11.2±0.8 kcal/mol) for E3_5 and D_m=5.7 M (ΔG=23.6±2.4 kcal/mol) for NI$_3$, or a sequential three-state model[31] with D_m1=3.7 M and D_m2=5.6 M (ΔG=19.7±4.6 kcal/mol) for NI$_3$C. The protein concentrations were 10 μM (E3_5, NI$_3$C, NI$_3$C Mut 5, NI$_3$C Mut 6), 5 μM (NI$_3$) and 7 μM (NI$_4$). The midpoints are calculated from the fit according to $D_{mx} = \Delta G^\circ_x/m_x$, where x refers to the first or second transition. (c) The oligomerization state of NI$_3$ was analyzed by analytical ultracentrifugation. NI$_3$ at 7 μM in 50 mM phosphate, 150 mM NaCl and 2 M GdnHCl was analyzed by sedimentation equilibrium at 35,000 rpm and 20 °C. A global species analysis fit yielded 90% monomeric (14.7 kDa) protein and only 10% higher-molecular-weight species (50.6 kDa).

GdnHCl with increasing protein concentration, and this becomes detectable above a protein concentration of about 7 μM (data not shown). To test our interpretation of the pretransition in NI$_3$C as being due to the unfolding of the C-cap, we measured the equilibrium unfolding of NI$_3$ at protein concentrations of 5 μM and 15 μM. The transition point for the curve at 15 μM was shifted to a higher GdnHCl concentration (data not shown), while the main transition of NI$_3$ at 5 μM was superimposable with that of NI$_3$C (Fig. 5a). In order to test whether NI$_3$ is really monomeric at a GdnHCl concentration below its transition, sedimentation equilibrium experiments with two different concentrations of NI$_3$ and NI$_3$C at 2 M GdnHCl were performed in an analytical ultracentrifuge. While at 21 μM NI$_3$ forms significant proportions of higher-molecular-weight species, at 7 μM, the sample of NI$_3$ consists of 90% monomeric protein (Fig. 5c), as does the sample of NI$_3$C at a protein concentration of 10 μM (data not shown). The unfolding transition of NI$_3$ at 5 μM monitored by CD and fluorescence was therefore assigned to that of monomeric protein.

The CD equilibrium unfolding of NI$_3$ is cooperative and has a similar transition midpoint as NI$_3$C, which contains the C-terminal cap (Fig. 5a). Importantly, no pretransition at 3.7 M GdnHCl is detected. The absence of this pretransition in NI$_3$ is indeed consistent with the C-cap being denatured in the equilibrium intermediate of NI$_3$C. The main transition (and the only transition present for NI$_3$) is thus interpreted as the cooperative denaturation of the consensus repeats including the N-cap, but not of the C-cap. If the C-cap of NI$_3$C is selectively denatured in the intermediate at 3.7 M GdnHCl, a molecule identical with NI$_3$ with a denatured appendage would be obtained, which would be expected to denature under similar conditions as NI$_3$. Indeed, the equilibrium unfolding curves of NI$_3$ and NI$_3$C measured by fluorescence reveal similar transition points at around 5.6 M GdnHCl. This supports the assumption that both monitor the unfolding of monomeric protein. Nevertheless, the nature of the

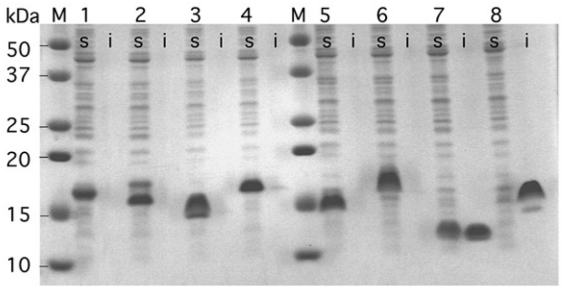

Fig. 6. Expression of ankyrin proteins E3_5 (lane 1, 17.7 kDa) and NI$_3$C (lane 2, 17.9 kDa), and the cap-lacking constructs NI$_3$ (lane 3, 14.7 kDa), NI$_4$ (lane 4, 18.8 kDa), I$_3$C (lane 5, 14.1 kDa), I$_4$C (lane 6, 17.6 kDa), I$_3$ (lane 7, 10.8 kDa) and I$_4$ (lane 8, 14.3 kDa). At OD$_{600}$=0.7, *E. coli* cultures were induced with 0.5 M IPTG and grown for 4 h at 37 °C. After cell lysis using a French press, the proteins in the soluble and insoluble fractions, s and i, were separately analyzed.

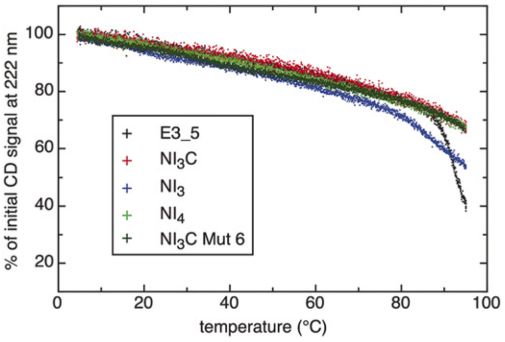

Fig. 7. Thermal melting of E3_5 (black dots), NI_3C (red dots), NI_3 (blue dots), NI_4 (green dots) and NI_3C Mut 6 (dark green dots) from 5 °C to 95 °C was followed by CD spectroscopy (see Materials and Methods). The heating gradient was 0.5 °C/min, and melting was only partially reversible (i.e., only 70% of the signal was regained upon cooling).

rather steep slope and of a possible kink in the pretransition baselines remains unclear (Fig. 5b).

The fact that there appears to be only a single cooperative transition in NI_3 in equilibrium unfolding measurements does not, of course, preclude that the kinetic process of unfolding is stepwise, with a preferred or random order of repeats unfolding. Indeed, possible unfolding pathways of DARPins emerge from the analysis of the MD simulations.

NI_4, which also has no C-cap but one consensus repeat more (and can be thought of as NI_3C, with the C-cap being "replaced" by another consensus repeat), shows no transition with GdnHCl, indicating an improved stability over NI_3C (Fig. 5a). This indicates that the consensus repeat provides much more stability than the C-capping repeat, albeit at the price of reduced solubility, especially under the conditions of folding in the cell. The natural evolution of the C-capping repeat must therefore have been governed predominantly by solubility and resistance to aggregation as the driving force. Nevertheless, these findings do lead to the question of whether a C-cap that combines high solubility and still shows further improved stability compared to the natural C-cap can be designed.

In thermal denaturation, all proteins (NI_3C, NI_3 and NI_4) are very stable and cannot be melted by heating up to 95 °C (Fig. 7), while E3_5, an unselected library member of the NX_3C library, begins denaturation at about 90 °C.

Mutations in the C-terminal cap

The flanking repeats of DARPins are necessary to provide solubility, particularly to allow folding in the cell (see above). However, they have been derived from the naturally occurring GA-binding protein, and their amino acid sequence differs from that of the designed consensus. For this reason, the interface between the capping repeats and their neighboring repeats is not as optimized as the interface between internal repeats. Figure 2b shows that, at 300 K, the C_α atoms of I_3 (three full-consensus repeats) fluctuate less than the C_α atoms of NI_1C, in particular in the external capping repeats. This is consistent with a better packing of the hydrophobic interface in I_3 than in NI_1C. Hence, to improve stability further, the internal surface of the C-terminal capping repeat should be engineered using the designed consensus sequence as a guide,[7] while the solvent-exposed residues should remain unmodified, as they are necessary to avoid aggregation.

Six multiple point mutants of NI_1C are listed in Table 3, and the side chains involved are shown in Fig. 8a. They are denoted NI_1C Mut 1 to NI_1C Mut 6. To validate our suggestion *in silico*, two additional 300-K runs were performed for each of the six mutants of NI_1C. The point mutants were inspired by bringing the C-terminal repeat closer to the consensus. Ala83Pro introduces a proline present in the consensus repeat, being part of the conserved Thr-Pro-Leu-His motif that is missing in the C-terminal repeat. The mutations Ile86Leu, Ser87Ala, Leu95Ile and $Ile^{98}Val$ might improve the packing of the interrepeat hydrophobic interface. Moreover, Ser87Ala could potentially increase the helical propensity. To facilitate the formation of a salt bridge observed in the internal repeats, the mutations Asp89Arg and Asn90Glu were also tested. In addition, Asn92 was changed to a histidine to favor a hydrogen bond with the carbonyl oxygen of residue 56 in the C-terminal turn of the first helix of R1 (see Stabilizing Interactions). Finally, the second helix of the C-terminal repeat was extended, since it is shorter by

Table 3. Mutated residues in the C-terminal cap of NI_1C

	Hydrophobic					Hydrophilic			α-Helix elongation		
Mutant	A83P	I86L	S87A	L95I	I98V	D89R	N90E	N92H	101K	102A	103A
Mut 1	X	X	X								
Mut 2				X	X						
Mut 3	X	X	X	X	X						
Mut 4									X	X	X
Mut 5	X	X	X	X	X				X	X	X
Mut 6	X	X	X	X	X	X	X	X	X	X	X

The consensus sequence[7] was used to suggest the type of side chain for each substitution. Additionally, the three-residue segment 101K–102A–103A was used to elongate the C-terminal helix. Two 300-K simulations of 40 ns and 50 ns with different initial assignments of velocities were performed for each mutant.

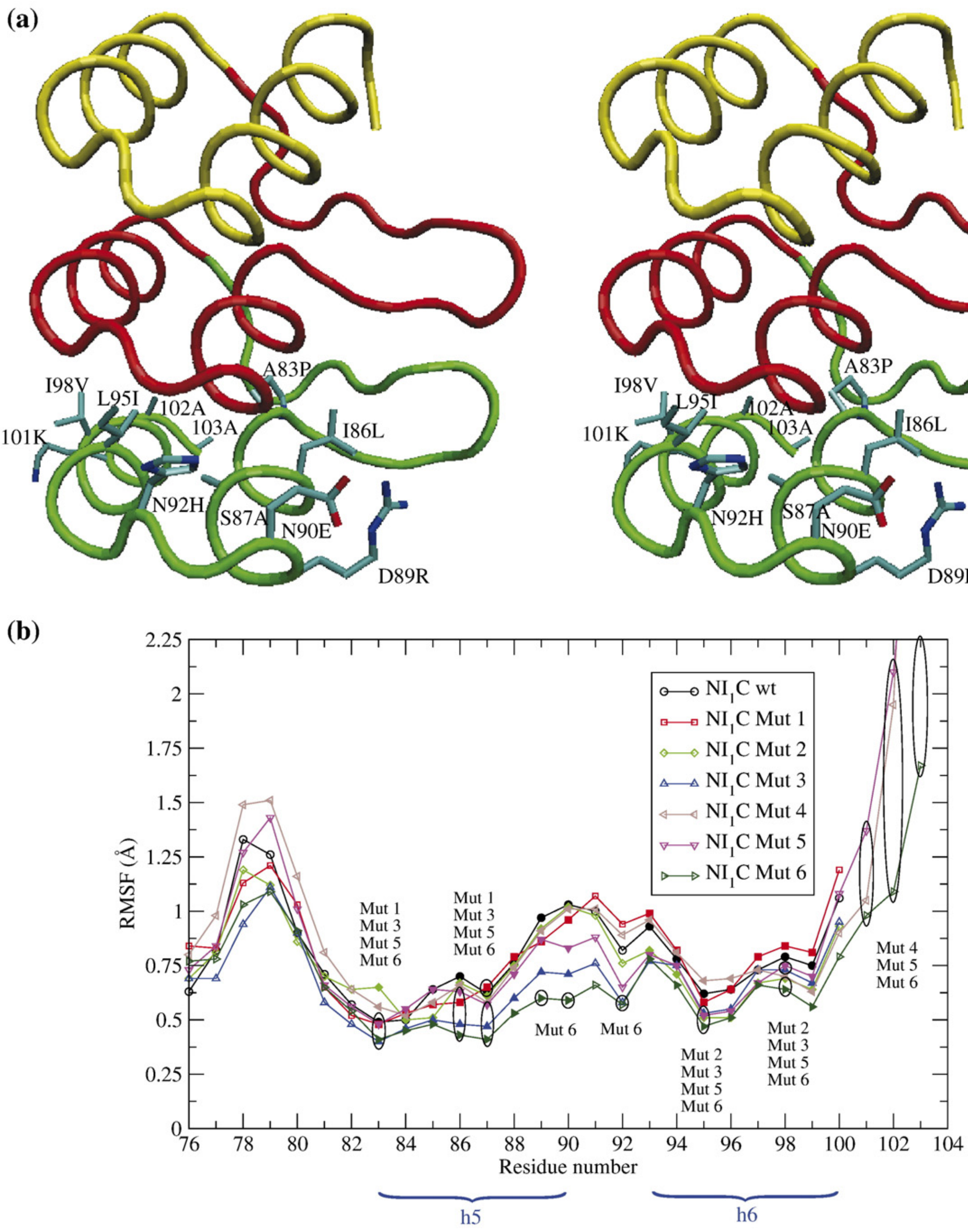

Fig. 8. (a) Stereo view of the NI_1C mutant with the eight point mutations and the three-residue C-cap extension 101K–102A–103A. (b) C_α RMSF of the C-terminal cap of wild-type NI_1C and the six mutant proteins, containing multiple substitutions (Table 3), at 300 K. The sites of mutation are enclosed by ellipses, and the names of the NI_1C mutant proteins that contain a substitution at the position indicated are given above or below the ellipse. The residues in helices are emphasized by filled symbols. The RMSF of the residue at position 103 is not displayed for Mut 4 and Mut 5. Their values are 4.54 Å and 3.44 Å, respectively. The last 40 ns of the 50-ns runs at 300 K were used to calculate the fluctuations. Quantitatively similar results are obtained using the 40-ns runs. A particular mutation is considered to increase the stability of the folded structure of NI_1C if the C_α RMSF at the site of the mutation is lower than that of the wild-type.

three residues than the consensus second helix of each repeat. The residues Lys-Ala-Ala were chosen for this extension because they increase the helical propensity.

Two main conclusions can be drawn from the 12 MD runs that represent an aggregated simulation time of 0.54 μs. First, the 11-point mutant of NI_1C (Mut 6) shows the lowest fluctuations (Fig. 8b) and the smallest number of different clusters of conformations (determined using the leader-clustering algorithm;[26,27] data not shown). This indicates that Mut 6 explores a more confined conformational space with respect to the other mutants and the wild-type. Second, individual hydrophobic side-chain replacements do not contribute significantly to the structural stability of the folded structure (Table 4). On the other hand, the mutations Asp89Arg and Asn90Glu together seem to be favorable, as they allow the Arg89–Glu90 salt bridge to form. In fact, the Arg89–Glu90 salt bridge is present in 84% of all frames sampled during the intervals 10–40 ns and 10–50 ns of the two 300-K simulations with Mut 6.

To test these suggestions, six mutants of NI_1C (see Supplementary Data), which contain the six multiple point mutations, were constructed (Table 3) in the C-terminal capping repeat. Expression in *E. coli* led to completely soluble proteins for all six mutants, and MALS analysis showed that all of the purified proteins are monomeric (data not shown). The stabilities of the six mutants were compared to the wild-type protein NI_1C using thermal denaturation experiments (Fig. 9a). While Mut 2 shows a T_m= 57 °C that is slightly below the transition midpoint of NI_1C wild-type (i.e., T_m=60 °C), the other mutants show increased T_m values in the following order: Mut 4 with T_m=64 °C, Mut 1 and Mut 3 with T_m=68 °C, and Mut 5 and Mut 6 with 77 °C.

These results validate the hypotheses that replacing hydrophobic residues in the interrepeat interface with those present in the consensus sequence and elongation of the C-terminal helix can further improve the stability of DARPins. Indeed, Mut 1 and Mut 3, which present mutations of hydrophobic residues but no helix elongation, have an increased melting point of 8 degrees. On the other hand, Mut 4, which differs from the wild-type only by the elongation of the helix, shows a 4 degrees increase in stability. When mutations of hydrophobic residues and elongation of the C-terminal helix are combined, as in Mut 5 and Mut 6, an even larger increase in the melting point (17 degrees in total) is observed. However, mutations of hydrophobic residues in the C-terminal helix (i.e., L95I and I98V) cause a slight destabilization (the T_m of Mut 2 is 3 degrees less than for the wild-type), while these mutations have essentially no effect if they are combined with mutations in the first helix (Mut 3 and Mut 1 behave almost identically). Furthermore, the hypothesis of increased stability by additional electrostatic interactions (i.e., a salt bridge between the side chains at positions 89 and 90, and an interrepeat hydrogen bond involving a histidine at position 92) could not be confirmed. In fact, Mut 5 shows the same transition midpoint as Mut 6, although Mut 6 differs from Mut 5 by the additional mutations at positions 89, 90 and 92.

In summary, the stabilizing effect of the 11 mutations present in Mut 6 seems to be largely caused by six of them: the three present in Mut 1, which all bring the C-cap closer to the consensus (A83P, introducing a conserved Pro; I86L and S87A, increasing the helical propensity), as well as those present in Mut 4 (the extension of the second helix by Lys-Ala-Ala).

Mut 5 and Mut 6 were chosen for GdnHCl equilibrium unfolding measured by CD (Fig. 9b). The transitions of both mutants have the same midpoint, and they are cooperative, consistent with a two-state model. The calculated ΔG value (7.8 kcal/mol) is more than double that obtained for NI_1C wild-type (3.7 kcal/mol) and is 85% of the value for the four-repeat molecule NI_2C containing the wild-type C-cap.[19]

Because of the cooperative nature of NI_1C folding, this small protein serves well to quantify the effects of the stabilizing cap mutations. However, we also wished to test their contributions in the context of the larger NI_3C, to test their effect on the pretransition and the main transition. Therefore, in addition to introducing the C-cap mutations in NI_1C, they were also introduced in NI_3C. NI_3C Mut 5 and NI_3C Mut 6 are fully soluble, as are all the other mutants. However, NI_3C Mut 5 and NI_3C Mut 6 are a mixture of monomer and dimer (with about 15% dimer) at

Table 4. Side-chain contributions to the energy difference between Mut 6 and wild-type

	Side chain/system[a]		Side chain/protein[b]	
Mutation	Total	van der Waals	Total	van der Waals
Hydrophobic				
A83P	−3±2 (−5±1)	−4±1 (−4±1)	−6±2 (−4±1)	−8±1 (−4±1)
I86L	−1±1 (−13±1)	−1±1 (−12±1)	−3±1 (−10±1)	−2±1 (−10±1)
S87A	11±1 (−16±3)	1±1 (−6±1)	11±1 (−16±2)	0±1 (−5±1)
L95I	−1±1 (−13±1)	−1±1 (−12±1)	−1±1 (−12±1)	−1±1 (−11±1)
I98V	2±1 (−11±2)	−1±1 (−11±1)	−0±1 (−7±2)	1±1 (−8±1)
Polar				
D89R	58±40 (−156±46)	−11±2 (2±3)	−34±20 (−17±23)	−2±2 (−3±1)
N90E	−147±37 (−30±6)	6±3 (−7±2)	−52±24 (−9±4)	1±2 (−6±1)
N92H	−6±7 (−28±6)	−3±2 (−8±2)	−8±3 (−10±3)	−2±1 (−7±1)

All values are expressed in kilocalories per mole. A negative value indicates that the mutation is favorable. Values outside the parentheses are the energy difference between Mut 6 and wild-type, while wild-type energy is given inside the parentheses.

Only values of Mut 6 are shown (averages over the intervals 10–40 ns and 10–50 ns of the two 300-K simulations), as this mutant contains all substitutions. The results obtained with the other mutants for the same substitutions are not shown because their values fall within the standard deviation of those reported here for Mut 6.

[a] Total and van der Waals energies are calculated for a mutated side chain and the rest of the system (i.e., protein, ions and water), excluding the backbone of the mutated side chain.

[b] Total and van der Waals energies are calculated for a mutated side chain and the rest of the protein, excluding the backbone of the mutated side chain.

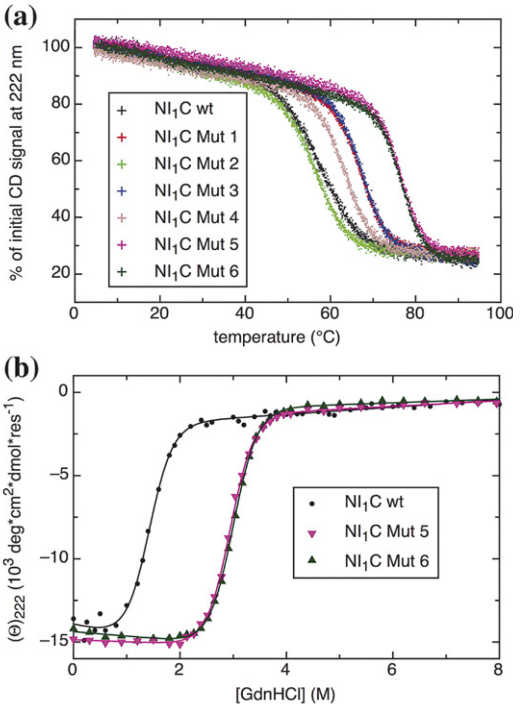

Fig. 9. Thermal and denaturant-induced unfolding of NI_1C mutants and wild-type followed by CD spectroscopy. (a) Thermal melting of NI_1C wild-type, NI_1C Mut 1, NI_1C Mut 2, NI_1C Mut 3, NI_1C Mut 4, NI_1C Mut 5 and NI_1C Mut 6 from 5 °C to 95 °C, using a heating gradient of 0.5 °C/min. Melting was only partially reversible. (b) GdnHCl-induced equilibrium unfolding of wild-type NI_1C, NI_1C Mut 5 and NI_1C Mut 6. The lines represent the least-squares fit to the two-state model with a midpoint of denaturation $D_m=1.4$ ($\Delta G=3.7\pm0.3$ kcal/mol) for wild-type, $D_m=2.94$ M ($\Delta G=7.8\pm0.2$ kcal/mol) for Mut 5 and $D_m=3.0$ M ($\Delta G=7.8\pm0.1$ kcal/mol) for Mut 6. The protein concentration was 10 μM in each case.

75 μM on MALS analysis. Under the assumption that the (presumably) even lower amount of dimer does not significantly influence the outcome of the experiment at the lower protein concentration of 10 μM, GdnHCl-induced equilibrium unfolding was measured by CD (Fig. 5a). In contrast to the unfolding curve of NI_3C wild-type, NI_3C Mut 5 and Mut 6 exhibit only a single transition at 6.7 M and at about 6.8 M GdnHCl, respectively. The increase in the D_m value is similar to the increase between NI_1C wild-type and NI_1C Mut 6. In thermal denaturation, it does not denature below 100 °C, but even wild-type NI_3C does not (Fig. 7).

Importantly, the pretransition in GdnHCl induced unfolding has disappeared in both NI_3C Mut 5 and Mut 6, fully consistent with our interpretation that this transition was due to the selective unfolding of the wild-type C-cap (Fig. 5). The more stable C-cap therefore "couples" to the rest of the protein, such that the main transition is moved to higher values. In summary, the weak link of the original C-cap (derived from the GA-binding protein) has been strengthened by our design.

Discussion

Multiple MD simulations with explicit solvent were carried out to examine the stability and unfolding behavior of DARPins. These simulations totaled >2 μs and included the 300-K runs and high-temperature unfolding simulations. They involved designed repeat proteins of different lengths, comprising either "full"-consensus repeats or, in the case of the library member E3_5, consensus repeats differing in positions that had previously been randomized, as they are part of the potential binding interface.[7,10] In addition, variants without the capping repeats were examined, as were point mutants in the C-terminal capping repeat. To validate the simulations, experiments comparing the solubility, expression properties and equilibrium denaturation behavior of variants with and without capping repeats, as well as mutated capping repeats, were conducted.

Three main conclusions emerge from these studies. First, native-state stability appears to increase with the number of repeats, as can be seen by the time points of denaturation of the central repeats in the 400-K simulation (Fig. 3). Only the C-capping repeat denatures earlier (see below) and needs to be considered separately. This increase of stability with the number of repeats is consistent with the trend that has been observed in equilibrium denaturation experiments using either GdnHCl denaturation[12] or heat denaturation[7] with several members of the DARPin libraries. Similar results have also been reported from shortened variants of the ankyrin domains of the *Drosophila* Notch receptor or by insertion of consensus repeats into Notch.[16,24] Furthermore, the comparison of the previously described consensus sequences with two to four repeats also showed the same trends.[5,6]

As in all these previous experiments, either the sequences of the repeats were not identical or the proteins were not fully soluble; full-consensus proteins with caps identical with the ones simulated here have been constructed; and both equilibrium and kinetic stability were measured in GdnHCl and by temperature-induced unfolding, which are reported in detail in the accompanying manuscript.[19] The increase of stability with the increasing number of repeats was found with these identical (full-consensus) repeats as well.

Similar results were also found in studies of other repeat proteins. As an example, in a series of tetratrico peptide repeat proteins,[23] again a stability increase with the number of repeats was observed.

In the equilibrium unfolding experiments of our DARPins, we found either a single cooperative transition involving all repeats (as in the case of E3_5; Fig. 5 and a previous study[12]) or two transitions, with the first transition being interpreted as involving the C-terminal cap (see below) and with the main transition involving all other repeats (Fig. 5). In contrast, in force-induced unfolding experiments measured by atomic force microscopy, a sequential unfolding was observed.[28] This difference in unfolding behavior is not surprising, as force-induced unfolding imposes a particular direction on the unfolding trajectory. Furthermore, the force-induced unfolding is a kinetic experiment, while the solution experiments described here have been equilibrium experiments. A stepwise unfolding is observed in the high-temperature unfolding simulations described here. This is not at variance with experimental results. In the kinetic unfolding of a NX_1C library member protein, no intermediate was detected at 5 °C, but differential scanning calorimetry experiments revealed a deviation from a two-state model at higher temperatures.[29] Also, the interpretation of the equilibrium and kinetic unfolding data of the full consensus proteins[19] are consistent with folding models different from 2-state.

The sequence of unfolding events observed in the high-temperature simulations here is not at variance with Go-type simulations of the DARPin E3_5.[15] That study suggests that folding of E3_5 starts with the formation of the N-cap and propagates sequentially through neighboring repeats to the C-cap. However, in the unfolding simulation presented here, R3 unfolds after R2. This discrepancy could be due to the limited statistics (only one run with E3_5), the fact that unfolding might follow a slightly different pathway than folding or the presence of multiple folding pathways that are not detected by the simplified model used in that study.[15] Interestingly, that same study suggests that folding of the Notch receptor starts at the second or the sixth AR.[15] This disagrees with recent mutagenesis experiments showing that folding of the Notch receptor begins with the formation of repeats 3–5.[30] There, it is pointed out by the authors that the discrepancy between simulations and experiments of the Notch receptor could be due to the coarse-grained model used in the former.[30]

The second main conclusion derived from the present studies is that the full-consensus proteins show higher stability than even the most favorable library members. For example, E3_5, a member of the NX_3C library, can be compared with the full-consensus protein NI_3C. It should be noted that the stability of E3_5 is already very high, with a ΔG value of 11.2 ± 0.8 kcal/mol (determined by GdnHCl-induced equilibrium denaturation) and a melting temperature of >85 °C (Figs. 5 and 7). Nevertheless, further stability improvement can be observed in the full-consensus structure. Interestingly, the NI_3C molecule no longer shows a fully cooperative transition in GdnHCl-induced unfolding but a first transition where about 20% of the helical CD signal is lost. On the basis of the sequence of events observed in the MD simulations of unfolding, the first transition was interpreted as the loss of the C-cap structure (see below). Measurements with a protein lacking the C-cap, NI_3, supported this interpretation. The unfolding transition of NI_3 is cooperative, with a ΔG value of 23.6 ± 2.4 kcal/mol, and occurs at the same GdnHCl concentration as the second transition ($I \rightleftharpoons U$) for NI_3C (Fig. 5a). Although NI_3 aggregates in GdnHCl at higher protein concentrations, it was mainly monomeric at 7 μM, as shown by analytical ultracentrifugation (Fig. 5c). A ΔG value of 19.7 ± 4.6 kcal/mol was calculated from the experimental equilibrium unfolding data for NI_3C using a sequential three-state model.[31] The presence of a third species in equilibrium unfolding has been observed as well for the natural ankyrin p19;[32] however, in this protein, several repeats strongly deviate from the consensus sequence and might constitute a "weak link." Furthermore, this protein is substantially less stable ($D_{\mathrm{m,urea}}=2.9$ M). GdnHCl-induced equilibrium unfolding experiments with NI_4 showed no transition, indicating an even higher stability for a protein with four full-consensus repeats. However, the unfolding study with this protein is difficult, as it is very prone to aggregation in GdnHCl, as also observed with NI_3 in GdnHCl.

In designing this "full-consensus" sequence, those residues that mediate binding to target proteins in DARPins[7,10] were replaced by the most frequent residues,[19] and thereby a number of charged residues were newly introduced. These are involved in additional salt bridges (e.g., between β-hairpins), and they are likely to contribute to the unusual stability.

The third main conclusion is that the C-cap used here is the limiting part for the stability of the whole-consensus repeat protein. Note, however, that this becomes only experimentally noticeable in the most extremely stable molecules. In the majority of library members, which still have ΔG values of 10–20 kcal/mol measured in equilibrium denaturation experiments and melting temperatures of between 70 and 90 °C[7,12] and are thus already at the upper edge of natural proteins, a single cooperative transition characterizing obviously very stable molecules is found. This indicates that the engineering of the C-cap will be of importance only for applications under the most extreme conditions and might push the already highly stable DARPins even further.

As mentioned above, the C-cap denatured first in almost all MD runs of the designed consensus ARs. A possible reason is that the C-cap originates from the natural GA-binding protein and has not been under particular evolutionary pressure. Thus, its amino acid sequence significantly differs from the full-consensus design and is characterized by a shorter helix. These differences possibly lead to a low structural stability of the shorter helix, to a poor packing of the hydrophobic core at the interface to the preceding repeat and to the lack of one interrepeat hydrogen bond and one intrarepeat salt bridge. The last two electrostatic interactions are indeed present among consensus repeats. These observations were taken into account to suggest the design of an even more stable C-terminal capping repeat (see below).

The C-capping repeat plays a very important role: In its absence, the expression of DARPins in *E. coli* leads to a significant amount of insoluble aggregated protein. The C-capping repeat prevents formation of insoluble aggregates and the N-capping repeat prevents soluble aggregates, while the constructs I_3 and I_4 (missing both capping repeats) are expressed almost entirely in inclusion bodies (Fig. 6). These observations explain the evolution of the capping repeats to secure the cellular folding and function of AR proteins and also demonstrate the importance of the capping repeats for the practical utilization of DARPins in biotechnology.

These findings are also consistent with the report on another design study of full-consensus AR proteins[5] with a slightly different sequence and without any caps, which were only soluble at acidic pH. The introduction of positive charges in the C-cap then allowed the protein to be soluble at neutral pH, but it still had to be produced from inclusion bodies made in *E. coli* with subsequent refolding.[6] However, the gain in solubility was accompanied by a significant loss in stability at pH 4. Furthermore, the stability could not be measured at pH 7 because the protein was not soluble.

The C-cap has thus been identified as being absolutely necessary to provide a highly charged surface to the protein to allow it to fold to the native state in a bacterial expression system, but at the same time to become a liability if one wants to drive the stability of these proteins to even more extreme values. The combined simulation and experimental results led to the question on whether it might be possible to design an equally soluble C-cap that nevertheless was of a similar stability as the internal consensus repeats. Eight different point mutations were considered, as was an extension of three amino acids to the last helix. The variant containing all mutations, as well as the C-terminal extension (NI_1C Mut 6), showed significantly smaller fluctuations than the wild-type in room-temperature MD simulations (Fig. 8b).

Testing the mutations in equilibrium unfolding experiments largely confirmed the suggested design. All the six mutants are equally soluble as the wild-type protein. Both NI_1C Mut 5 and NI_1C Mut 6 show a remarkable increase in stability, as indicated by a melting point that lies 17 degrees higher and by a >2-fold increase in ΔG value when compared to the wild-type NI_1C. These results confirmed the importance of a better packing of the hydrophobic core. However, the introduction of additional electrostatic interactions does not further increase the stability, as indicated by the similar curves of Mut 5 and Mut 6 (Figs. 5a and 9). When the eight point mutations and the three-residue helical extension are introduced into NI_3C (NI_3C Mut 5 and Mut 6), we also observe a large increase in stability compared to NI_3C wild-type, but more importantly, the pretransition at 3.7 M GdnHCl is absent (Fig. 5a). This experiment is further proof that the equilibrium intermediate in NI_3C wild-type corresponds to a state wherein the less stable wild-type C-terminal capping repeat is selectively denatured.

The current study has a number of direct implications for the understanding of repeat proteins and the design of further improved libraries. It helps to rationalize the minimal number of ARs found in natural proteins, as the critical interactions between repeats are important for stabilizing the repeat domain. It also helps to understand the vital importance of the capping repeats and shows that if extreme stabilities are needed, the C-cap of GA-binding protein can become limiting. However, with the improved design of the C-cap, DARPins of even more extreme stability can be designed.

Materials and Methods

Sequences and initial conformations

Table 1 lists the systems that were simulated in the present study. Simulations of E3_5 were started from its X-ray structure[12] (PDB code 1MJ0). The initial conformations of NI_1C, NI_2C and NI_3C were modeled from the structure of E3_5. The mutated side chains were constructed with a library of rotamers using the program Insight II (Accelrys, Inc.). The experimental structure of NI_3C has meanwhile been determined and is described in the accompanying paper.[33]

The first and last residues of each repeat are defined here differently from Refs. 7, 12 and 34 for topological reasons. In the present work, each internal repeat includes six amino acids preceding the β-hairpin at the tip of the loops, while in Refs. 7, 12 and 34, that β-hairpin was used as the start (Supplementary Fig. 1). The present position Ala1 of each repeat would correspond to Ala28 of the previous repeat in Refs. 7, 12 and 34. In this way, intraloop contacts (such as hydrogen bonds and salt bridges) being counted as interrepeat contacts is avoided. This definition is also used to calculate repeatwise RMSDs from the initial conformation.

For the "full-consensus" AR proteins NI_xC, the variable positions in Ref. 7 were fixed.[19] In the notation of this manuscript, the primary structure of the full-consensus internal repeats of NI_xC is A_1DVNA**K**D **KD**G_{10}-**Y**TPLHLAA**RE**$_{20}$ GHLEIVEVLL$_{30}$K**A**G_{33}, where the newly defined residues that differ in E3_5 and other members of the library are in boldface (cf. Supplementary Fig. 1). For the discussion of E3_5 and NI_3C, the residue number refers to the numbering scheme according to PDB file 1MJ0. The constructs missing the N-cap (I_1, I_2, I_3, I_4, I_3C, and I_4C) start in front of the first helix of the internal-consensus repeat (sequence TPLHL, position 12 of the numbering scheme shown above, corresponding to position 49 in PDB file 1MJ0; see also Supplementary Fig. 1). The constructs missing the C-cap (I_1, I_2, I_3, I_4, NI_3 and NI_4) end with the second helix of the internal repeat (sequence LLKAG, position 33 of the numbering scheme shown above, corresponding to position 136 in the PDB file of E3_5; PDB code 1MJ0). This molecule, E3_5, has the same length as an NI_3C molecule (see also Supplementary Fig. 1).

For the simulations, the C-terminal cap of the NI_1C mutants was modeled by homology by superimposing the central repeat of NI_1C to the C-terminal cap to generate the coordinates of the mutated side chains and the three-amino-acid extension as well.

Simulations

The MD simulations were performed with the program NAMD2[35] using the CHARMM all-hydrogen force field (PARAM22)[36] and the TIP3P model of water. To effectively compare simulations with experimental results (e.g., a pH of 7.4 in the CD experiments; see below), side chains of aspartates and glutamates were negatively charged, those of lysines and arginines were positively charged and histidines were considered neutral. Initial conformations were minimized in vacuo by performing 100 steps of steepest descent and subsequently 500 steps of conjugate gradient minimization with CHARMM.[37] The proteins were then inserted into a cubic water box of different side lengths, depending on the number of amino acids. In the case of NI_3C and E3_5, a larger box was used for the 400-K than for the 300-K simulations. The minimal distance between the protein and the boundary of the box was 12 Å. The different box sizes and durations of the simulations are summarized in Table 1. Furthermore, for each of the six mutants of NI_1C, two simulations at 300 K (50 ns and 40 ns) were performed using a box with the same dimensions as the one used for NI_1C. Chloride and sodium ions were added to neutralize the system and approximate a salt concentration of 150 mM. The water molecules overlapping with the protein or the ions were removed if the distance between the water oxygen and any atom of the protein or any ion was smaller than 3.1 Å. The number of water molecules ranged from 3906 to 31,443, and the total number of atoms ranged between 12,087 and 96,878. To avoid finite-size effects, periodic boundary conditions were applied. Different initial random velocities were assigned whenever more than one simulation was performed with the same protein. Electrostatic interactions were calculated within a cutoff of 12 Å, while long-range electrostatic effects were taken into account by the Particle Mesh Ewald summation method.[38] Van der Waals interactions were treated with the use of a switch function starting at 10 Å and turning off at 12 Å. The temperature was kept constant by using the Berendsen thermostat[39] with a relaxation time of 0.2 ps, while the pressure was held constant at 1 atm by applying a pressure piston.

Before production runs, harmonic constraints were applied to the positions of all the atoms of the protein to equilibrate the system at 300 K or 400 K during a time length of 0.2 ns. After this equilibration phase, the harmonic constraints were released. For the runs at 300 K, the first 10 ns of unconstrained simulation time were also considered part of the equilibration and were thus not used for the analysis. For the six mutants of NI_1C, the equilibration was elongated by 2 ns without restraints on the mutated amino acids and its two neighbors. The dynamics was integrated with a time step of 2 fs. The covalent bonds involving hydrogens were rigidly constrained by means of the SHAKE algorithm with a tolerance of 10^{-8}. Snapshots were saved every 2 ps for trajectory analysis.

Determination of native contacts

The conformations sampled at room temperature were used to determine native hydrogen bonds, salt bridges and C_α contacts. To define a hydrogen bond, a H···O distance cutoff of 2.7 Å and a D–H···O angle cutoff of 120° were used, where a donor D could either be an oxygen or a nitrogen. An interaction was defined as a salt bridge if the N_ζ of Lys or the C_ζ of Arg was closer than 4 Å or 5 Å, respectively, from either the C_γ of Asp or the C_δ of Glu. All histidines were assumed to be neutral. A C_α contact involves two C_α atoms with a distance smaller than 6.5 Å and not adjacent in sequence (i.e., residue pairs i,j, with $j>i+2$). Only those hydrogen bonds, salt bridges and C_α contacts present in at least half of the simulation frames at 300 K were selected as native contacts (Table 2). They were used to compare the conformational flexibility of different proteins at room temperature and to monitor the changes in secondary and tertiary structures during unfolding.

Design and synthesis of DNA-encoding AR proteins, protein expression and purification

The process of the sequence design of the full-consensus ARs is described in the accompanying manuscript.[19] The cloning, expression and purification of DARPins have been performed as described elsewhere.[7] The construction of the C-cap mutants is described in the Supplementary Data.

CD spectroscopy

All CD experiments were performed in 50 mM sodium phosphate buffer (pH 7.4) and 150 mM NaCl, using 5–10 μM protein purified by immobilized metal ion affinity chromatography as described.[7,10] To measure the denaturant-induced equilibrium unfolding curves, the samples were equilibrated at 20 °C overnight at the corresponding GdnHCl concentrations. The CD signal at 222 nm was recorded on a Jasco J-715 instrument (Jasco, Japan) equipped with a computer-controlled water bath, using a cylindrical quartz cell of 1 mm path length. CD data were collected at 222 nm and 20 °C every 5 s with a bandwidth of 2 nm and a response time of 4 s, averaged over 2 min. A baseline correction was made with the buffer. The CD signal was converted to mean residue ellipticity (Θ_{MRE}) using the concentration of the sample determined spectrophotometrically at 280 nm. Thermal unfolding was recorded by continuous heating from 5 °C to 95 °C with a temperature gradient of 0.5 °C/min. CD data were collected at 222 nm every 5 s with a bandwidth of 2 nm and a response time of 4 s. Reversibility was determined from the recovery of ellipticity after cooling.

Fluorescence spectroscopy

Tyrosine fluorescence was excited at 274 nm, and emission spectra were recorded from 290 to 350 nm using a PTI Alpha Scan spectrofluorimeter (Photon Technologies, Inc.). Slid widths of 5 nm were used for both excitation and emission. Samples were prepared as for the CD measurements. After buffer correction, the intensity of the emission maximum at 304 nm or 303 nm, respectively, was plotted against the denaturant concentration.

Analytical ultracentrifugation

Sedimentation equilibrium experiments were performed with a Beckman XL-A centrifuge with a NA-50 Ti rotor at 20 °C using optical absorbance detection. NI_3 protein at two concentrations (0.1 mg/ml and 0.3 mg/ml) was measured in 2 M GdnHCl, 50 mM phosphate buffer and 150 mM NaCl (pH 7.4). Protein and buffer samples were placed in cells fitted with double-sector centerpieces and quartz windows. Sedimentation equilibrium was approached at a rotor speed of 35,000 rpm. The cells were scanned at 230 nm, and 40 scans were collected. The scans were analyzed using the software SEDPHAT‡.

‡ http://www.analyticalultracentrifugation.com/sedphat/ (by P. Schuck, National Institutes of Health, Bethesda, MD).

Solvent density was calculated from the weight of the salts, and the partial specific volume of the protein was determined from the amino acid sequence of the protein using the software UltraScan§, not taking into account the influence of dissolved GdnHCl on the partial specific volume.

Acknowledgements

We thank Dr. H.K. Binz for preparing the homology model of NI_3C, and Dr. C. Bodenreider (Universität Basel) for help with data fitting. We are grateful to Dr. C. Briand for valuable help with analytical ultracentrifugation, and Dr. P. Forrer for helpful discussions about library design and biophysical properties of AR proteins. We thank A. Widmer (Novartis Pharma, Basel, Switzerland) for providing the molecular modeling program WITNOTP, which was used for visual analysis of the trajectories. The simulations were performed on the Matterhorn Beowulf cluster at the Computing Center of the University of Zürich. We thank C. Bolliger and Dr. A. Godknecht for setting up the cluster, and the Canton of Zürich for generous hardware support. This work was supported by a Swiss National Science Foundation grant to A.C. and by a National Center of Competence in Research in Structural Biology grant to A.P.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.09.042

References

1. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131.
2. Kobe, B. & Kajava, A. V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* **25**, 509–515.
3. Groves, M. R. & Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383–389.
4. Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins: Struct. Funct. Bioinf.* **17**, 363–374.
5. Mosavi, L. K., Minor, D. L. & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA*, **99**, 16029–16034.
6. Mosavi, L. K. & Peng, Z. Y. (2003). Structure-based substitutions for increased solubility of a designed protein. *Protein Eng.* **10**, 739–745.
7. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, 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.
8. Forrer, P., Stumpp, M. T., Binz, H. K. & Plückthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett.* **539**, 2–6.
9. Batchelor, A. H., Piper, D. E., de la Brousse, F. C. & McKnight, S. L. (1998). The structure of GABP alpha/beta: an ETS domain ankyrin repeat heterodimer bound to DNA. *Science*, **279**, 1037–1041.
10. 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. *Nat. Biotechnol.* **22**, 575–582.
11. Amstutz, P., Binz, H. K., Parizek, P., Stumpp, M. T., Kohl, A., Grütter, M. G. *et al.* (2005). Intracellular kinase inhibitors selected from combinatorial libraries of designed ankyrin repeat proteins. *J. Biol. Chem.* **280**, 24715–24722.
12. Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Grütter, M. G. & Plückthun, A. (2003). Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc. Natl Acad. Sci. USA*, **100**, 1700–1705.
13. Tang, K. S., Fersht, A. R. & Itzhaki, L. S. (2003). Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure*, **11**, 67–73.
14. Interlandi, G., Settanni, G. & Caflisch, A. (2006). Unfolding transition state and intermediates of the tumor suppressor p16^{ink4a} investigated by molecular dynamics simulations. *Proteins: Struct. Funct. Bioinf.* **64**, 178–192.
15. Ferreiro, D. U., Cho, S. S., Komives, E. A. & Wolynes, P. G. (2005). The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *J. Mol. Biol.* **354**, 679–692.
16. Mello, C. C. & Barrick, D. (2004). An experimentally determined protein folding energy landscape. *Proc. Natl Acad. Sci. USA*, **101**, 14102–14107.
17. Tripp, K. W. & Barrick, D. (2004). The tolerance of a modular protein to duplication and deletion of internal repeats. *J. Mol. Biol.* **344**, 169–178.
18. Mello, C. C., Bradley, C. M., Tripp, K. W. & Barrick, D. (2005). Experimental characterization of the folding kinetics of the notch ankyrin domain. *J. Mol. Biol.* **352**, 266–281.
19. Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K & Plückthun, A. (2007). Folding and unfolding mechanism of highly stable full consensus ankyrin repeat proteins. *J. Mol. Biol.* In press. doi:10.1016/j.jmb.11.046.
20. Karshikoff, A. & Ladenstein, R. (2001). Ion pairs and the thermotolerance of proteins from hyperthermophiles: a 'traffic rule' for hot roads. *Trends Biochem. Sci.* **26**, 550–556.
21. Berezovsky, I. N. & Shakhnovich, E. I. (2005). Physics and evolution of thermophilic adaptation. *Proc. Natl Acad. Sci. USA*, **102**, 12742–12747.
22. Zhang, B. & Peng, Z.-Y. (2000). A minimum folding unit in the ankyrin repeat protein p16(ink4). *J. Mol. Biol.* **299**, 1121–1132.
23. Main, E. R. G., Stott, K., Jackson, S. E. & Regan, L. (2005). Local and long-range stability in tandemly

§ http://www.ultrascan.uthscsa.edu/ (by B. Demeler, University of Texas Health Science Center, San Antonio, TX).

arrayed tetratricopeptide repeats. *Proc. Natl Acad. Sci. USA*, **102**, 5721–5726.

24. Tripp, K. W. & Barrick, D. (2007). Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. *J. Mol. Biol.* **26**, 1187–1200.
25. Binz, H. K., Kohl, A., Plückthun, A. & Grütter, M. G. (2006). Crystal structure of a consensus-designed ankyrin repeat protein: implications for stability. *Proteins: Struct. Funct. Bioinf.* **65**, 280–284.
26. Hartigan, J. A. (1975). *Clustering Algorithms.* Wiley, New York.
27. Settanni, G., Rao, F. & Caflisch, A. (2005). Φ-Value analysis by molecular dynamics simulations of reversible folding. *Proc. Natl Acad. Sci. USA*, **102**, 628–633.
28. Li, L. W., Wetzel, S., Plückthun, A. & Fernandez, J. M. (2006). Stepwise unfolding of ankyrin repeats in a single protein revealed by atomic force microscopy. *Biophys. J.* **90**, L30–L32.
29. Devi, V. S., Binz, H. K., Stumpp, M. T., Plückthun, A., Bosshard, H. R. & Jelesarov, I. (2004). Folding of a designed simple ankyrin repeat protein. *Protein Sci.* **13**, 2864–2870.
30. Bradley, C. M. & Barrick, D. (2006). The notch ankyrin domain folds via a discrete, centralized pathway. *Structure*, **14**, 1303–1312.
31. Barrick, D. & Baldwin, R. L. (1993). Three-state analysis of sperm whale apomyoglobin folding. *Biochemistry*, **32**, 3790–3796.
32. Zeeb, M., Rosner, H., Zeslawski, W., Canet, D., Holak, T. A. & Balbach, J. (2002). Protein folding and stability of human cdk inhibitor p19(INK4d). *J. Mol. Biol.* **315**, 447–457.
33. Merz, T., Wetzel, S. K., Firbank, S., Plückthun, A., Grütter, M. G. & Mittl, P. R. E. (2007). Stabilizing ionic interactions in a full consensus ankyrin repeat protein. *J. Mol. Biol.* In press. doi:10.1016/j.jmb.11.047.
34. 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.
35. Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N. *et al.* (1999). NAMD2: greater scalability for parallel molecular dynamics. *J. Comp. Phys.* **151**, 283–312.
36. MacKerell, A. D. E. A., Jr (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.
37. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
38. Darden, T., York, D. & Pedersen, L. (1993). Particle Mesh Ewald—an $N{\cdot}\log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092.
39. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A. & Haak, J. R. (1984). Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.
40. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graphics*, **14**, 33–38.