Structure
Article

Spontaneous Self-Assembly of Engineered Armadillo Repeat Protein Fragments into a Folded Structure

Randall P. Watson,[1] Martin T. Christen,[1] Christina Ewald,[1,2] Fabian Bumbak,[1] Christian Reichen,[2] Maja Mihajlovic,[2] Elena Schmidt,[3] Peter Güntert,[3] Amedeo Caflisch,[2] Andreas Plückthun,[2,*] and Oliver Zerbe[1,*]
[1]Department of Chemistry, University of Zürich, 8057 Zürich, Switzerland
[2]Department of Biochemistry, University of Zürich, 8057 Zürich, Switzerland
[3]Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt am Main, 60438 Frankfurt am Main, Germany
*Correspondence: plueckthun@bioc.uzh.ch (A.P.), oliver.zerbe@chem.uzh.ch (O.Z.)
http://dx.doi.org/10.1016/j.str.2014.05.002

SUMMARY

Repeat proteins are built of modules, each of which constitutes a structural motif. We have investigated whether fragments of a designed consensus armadillo repeat protein (ArmRP) recognize each other. We examined a split ArmRP consisting of an N-capping repeat (denoted Y), three internal repeats (M), and a C-capping repeat (A). We demonstrate that the C-terminal MA fragment adopts a fold similar to the corresponding part of the entire protein. In contrast, the N-terminal YM$_2$ fragment constitutes a molten globule. The two fragments form a 1:1 YM$_2$:MA complex with a nanomolar dissociation constant essentially identical to the crystal structure of the continuous YM$_3$A protein. Molecular dynamics simulations show that the complex is structurally stable over a 1 μs timescale and reveal the importance of hydrophobic contacts across the interface. We propose that the existence of a stable complex recapitulates possible intermediates in the early evolution of these repeat proteins.

INTRODUCTION

The characteristic feature of repeat proteins is that multiple, almost identical amino acid stretches (Marcotte et al., 1999, Andrade et al., 2001) fold into tightly packed modules, which rigidly associate into stable proteins. Typically, repeat modules are short motifs of 20 to 50 amino acids. Within a repeat protein, these sequences fold into nearly identical structures, with the stacked structural modules forming an extended domain with a continuous surface. Because a module's sequence can often be varied while maintaining its overall structure, it tends to naturally undergo specific interactions and may be tailored to recognize specific targets, often with high affinity (Boersma and Plückthun, 2011). Examples of such proteins include the ankyrin repeat proteins (Sedgwick and Smerdon, 1999), HEAT repeat proteins (Andrade et al., 2001), and the armadillo repeat proteins (ArmRPs) (Hatzfeld, 1999, Xu and Kimelman, 2007, Tewari et al., 2010, Marfori et al., 2011). Many of these proteins are involved in cell signaling or transport (MacDonald et al., 2009).

ArmRPs bind peptides in an extended form; thus, it is the amino acid sequence of the peptide rather than its tertiary structure that is recognized (Huber and Weis, 2001, Xu and Kimelman, 2007). In a first approximation, two consecutive side chains of the peptide are recognized per module. Accordingly, ArmRPs make particularly attractive scaffolds for protein engineering and biotechnological applications (Boersma and Plückthun, 2011). For these reasons, Parmeggiani et al. (2008) designed repeat proteins based on consensus sequences derived from the natural ArmRPs of the β-catenin and importin-α families. In this design, the elongated hydrophobic core was optimized by atomistic molecular dynamics (MD) simulations. Moreover, special N-terminal (N-cap) and C-terminal (C-cap) repeats were developed to flank the internal repeats. Recently, initial crystal structures of such constructs have been determined, which verify the consensus design (Madhurantakam et al., 2012).

Nuclear magnetic resonance (NMR) spectroscopy may complement crystallography in many aspects of the design cycle, in particular for proteins that contain flexible parts. However, assignment of chemical shifts of repeat proteins by NMR is very challenging because of the repetitive nature of their sequence (Wetzel et al., 2010). To facilitate this process, we attempted segmental labeling (Yamazaki et al., 1998) using a split intein (Ludwig et al., 2009, Muona et al., 2010) to help deconvolute the intrinsically complex and degenerate spectra. We observed that when the repeat protein was expressed as two separate fragments with their intein ligation motifs present, the fragments showed significant affinity for each other, even though no peptide bond was formed. Removal of the split intein motifs resulted in the same observation, indicating that the interaction was not mediated by the split intein. This evidence strongly suggests the formation of a stable, noncovalent complex from the two ArmRP fragments.

Here, we present structural, biophysical, and thermodynamic data to characterize this interaction. Furthermore, we analyze this interaction with reference to the structure of the complete protein and demonstrate that the same interface contacts are indeed made. Hence, the interaction occurs in a highly similar

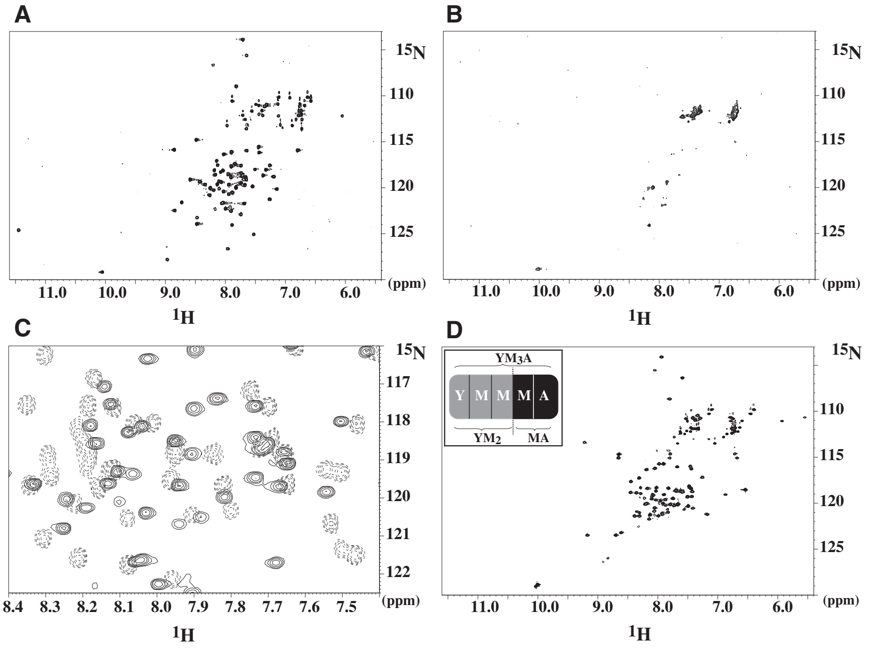

Figure 1. [^{15}N, ^{1}H]-HSQC Spectra

(A) ^{15}N-Labeled MA in the absence of YM_2.

(B). ^{15}N-Labeled YM_2 in the absence of MA.

(C) Expansion of the spectrum of ^{15}N-labeled MA in complex with unlabeled YM_2 (dotted lines) at a 1:1.2 molar ratio, superimposed with the spectrum of ^{15}N-labeled MA in the absence of YM_2 (solid lines).

(D) ^{15}N-Labeled YM_2 complexed with unlabeled MA at 1:1.2 molar ratio. A schematic drawing of the modular nature of YM_3A and the fragments is depicted in the inset in (D).

All spectra were recorded on a 700 MHz spectrometer at 307 K in 50 mM sodium phosphate buffer with 150 mM NaCl, 2% glycerol, 0.02% NaN_3, and 10% D_2O (pH 7.4).

if not identical manner to that found in the native, uninterrupted protein. This finding not only has implications for future biotechnological applications of ArmRPs but may also shed light on the evolution of repeat proteins.

RESULTS

Self-Assembly of a Split Consensus ArmRP

We investigated a consensus ArmRP consisting of three identical internal repeats, flanked by N- and C-terminal capping repeats. The armadillo fold is predominantly α-helical, whereby each of the repeating modules encompasses three helices of different lengths (H1 and H2, ~10 residues; H3, ~16 residues) that are connected by short loops. H1 and H3 are oriented perpendicular to each other, H2 connects the two at an angle of ~30°, creating a triangular “spiral staircase” arrangement.

The majority of detailed NMR investigations were carried out with two fragments, an N-terminal fragment consisting of an N-terminal capping repeat (N-cap) and two internal repeats (hereafter called YM_2, the Y denoting the *y*east origin of the N-cap [Parmeggiani et al., 2008] and the M denoting the *M*D origin of the internal repeats [Alfarano et al., 2012]) and a C-terminal fragment (termed MA, A for *a*rtificial) consisting of one internal repeat and a C-terminal capping repeat (Parmeggiani et al., 2008). Amino acid sequences of the fragments are shown in Figure S1A (available online). The fragments were expressed and purified as two individual proteins from separate *E. coli* cultures as described in Experimental Procedures and Tables S1 and S2.

Heteronuclear NMR Demonstrates that the Fragments Interact Specifically

The [^{15}N, ^{1}H] heteronuclear single quantum coherence (HSQC) spectrum of the C-terminal MA fragment alone displays good signal dispersion and narrow peaks; both features indicate a well-folded protein (Figure 1A; Figure S4A). In contrast, spectra of the N-terminal YM_2 fragment alone are essentially devoid of peaks from backbone resonances, a behavior typically associated with a protein lacking well-defined tertiary structure such as a molten globule (Figure 1B) (Dyson and Wright, 2004). Upon mixing of ^{15}N-labeled MA with unlabeled YM_2, many of the MA resonances shift to new positions, indicating a change associated with the formation of a complex with YM_2 (Figure 1C; Figure S4B). In a complementary experiment, ^{15}N-labeled YM_2 was mixed with a slight excess of unlabeled MA. Interestingly, the [^{15}N, ^{1}H]-HSQC spectrum for the complexed YM_2 was now indicative of a well-behaved and folded protein (Figure 1D; Figure S5). Circular dichroism (CD) spectra (Figure S2A) of the individual fragments display features typical of α-helical proteins. Although MA, as expected from its NMR spectrum, is clearly in a predominantly helical state, surprisingly, CD data of YM_2 also indicate a high degree of helical secondary structure. Melting curves of MA followed by CD show a marked transition at ~62°C for MA, characteristic of cooperative folding (Figure S2B). CD spectra of the YM_2:MA complex showed the same profile as those of YM_3A (Figure S2A). The melting curves for YM_2, however, are essentially linear with a poorly defined transition (Figure S2B).

In addition, we have also investigated if ArmRPs can be split at other sites and reconstituted (Table S4). For that purpose, we have looked at the fragment complex YM:M_2A by NMR (see Figure S3A). Again, a well-resolved spectrum of a sample containing labeled YM and unlabeled M_2A indicates the presence of a well-folded N-terminal part, and signals are generally located close to positions observed for YM_2 (Figure S3B). Finally, we have investigated complementary fragments of VG_328, an ArmRP of the format YMRRRMA containing randomized repeats “R” (Table S4) that can bind the peptide neurotensin (Varadamsetty et al., 2012). For N-terminal fragments YM, YMR, and YMRRR, spectra corresponding to well-folded proteins were observed only in the presence of complementary C-terminal fragments (see Figures S3C and S3D). The spectrum of uncomplexed YMRRR (Figure S3E) is similar to uncomplexed YM_2 described above and indicative of a molten globule. To summarize, ArmRP can be split into complementary

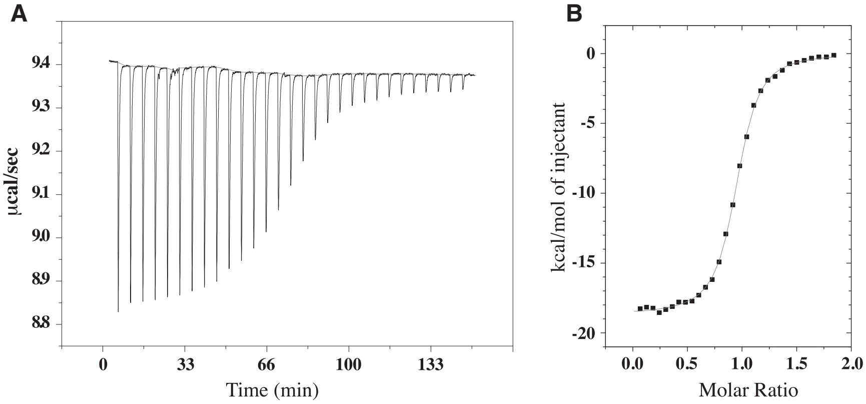

Figure 2. ITC Isotherm and Curve Fitting for the YM_2:MA Interaction

(A) ITC isotherm.

(B) Curve fitting.

fragments after each repeat, although we suspect that the exact sequence of the internal repeats will influence the stability of the complex.

In the following, we describe the complementary pair YM_2:MA in detail using solution NMR and other biophysical methods.

The YM_2 and MA Fragments Form a Complex with Nanomolar Affinity

The NMR experiments described above strongly indicate that the complementary fragments form a stable complex in solution. In order to measure the binding affinity of the two fragments for each other and to determine the thermodynamic properties of the interaction, we used isothermal titration calorimetry (ITC), with buffers and temperature identical to those used in the NMR experiments. A titration experiment in which MA was added to YM_2 yielded a K_d of ~126 ± 5 nM with corresponding ΔH of -78.2 kJ mol^{-1} and $-T\Delta S$ of 37.9 kJ mol^{-1} (Figure 2). The measured stoichiometry of 0.94:1 is indicative of a 1:1 complex, the discrepancy most likely due to a small percentage of the YM_2 being in a binding-incompetent state.

This low K_d value, characterizing a rather tight interaction, measures the overall equilibrium between the folded complex on one hand and a molten-globule N-terminal fragment and an individual C-terminal fragment in a somewhat different conformation on the other. The interaction energy between folded fragments must be very favorable, as the folding of the N-terminal fragment upon complex formation is entropically unfavorable. It is unlikely that the burial of hydrophobic surface area fully compensates for this entropy loss.

The N-Terminal Fragment Oligomerizes at Higher Concentrations

The absence of well-resolved peaks in the [^{15}N, ^{1}H]-HSQC spectrum of uncomplexed YM_2 (Figure 1B) is most likely due to poor packing of side chains and the associated conformational exchange. To exclude the possibility that the lack of peaks might be due to oligomerization, we characterized the entities by analytical size-exclusion chromatography (SEC) and multiangle light scattering (MALS) (Figure 3; Table S3). The position of the elution peak of YM_2 (MW 12.2 kDa) shows a marked correlation between concentration and oligomeric state. At 6.25 μM protein concentration, a single narrow and symmetric peak was observed, whereas multiple peaks were observed at higher concentrations, with the rightmost peak shifted and an additional broad peak at lower elution volume appearing.

[^{15}N,^{1}H]-HSQC spectra of ^{15}N-labeled YM_2, measured at a concentration at which the monomer was the predominant species on SEC, confirmed that the lack of NMR peaks is not due to oligomerization. Also, the uncomplexed fragments display large deviations from the expected size, a behavior that is consistent with less compact packing (Table S3). In contrast, the YM_2:MA complex elutes at the same volume as full-length YM_3A. MALS analysis of the main peaks observed in analytical SEC (i.e., for YM_2, the peak with the largest elution volume) confirms that they represent the monomeric species of each protein (Table S3).

It is therefore likely that the YM_2 fragment is in a molten globule-like state. As a result of its poor packing, we hypothesize that a significant amount of exposed hydrophobic surface renders the fragment susceptible to limited oligomerization. In contrast, the MA fragment remains monomeric up to at least 800 μM.

Structures of the Fragments in the Complex Closely Mimic the Structure of the Covalently Linked Full-Length Armadillo Protein

Solution NMR techniques were used to determine the structures of the two fragments, both isolated and when in complex with each other. The solution structure of uncomplexed MA was calculated using 883 nuclear Overhauser effect (NOE)-derived distance restraints, of which 127 were long-range ($|i - j| \geq 5$), that is, approximately four per restrained residue (Table 1). The root-mean-square deviation (rmsd) among the 20 lowest energy conformers is 0.74 ± 0.18 Å for backbone atoms in the ordered regions, that is, residues 130 to 156 and 161 to 198 (Table 1). Herein, residue numbering refers to the full-length construct throughout the paper (cf. Figure S1). The calculated structure shows good structural similarity with the corresponding region from the crystal structure of the full-length protein (Protein Data Bank [PDB] accession number 4DBA) (Figure 4A).

Interestingly, signals from the 14 N-terminal residues of MA (corresponding to helix H1 of the internal consensus repeat) were absent in the spectra, suggestive of conformational exchange. The rmsd of the closest-to-average conformer in the NMR bundle to the crystal structure is 2.25 Å for backbone atoms and 3.09 Å for all heavy atoms in the ordered regions, that is, residues 130 to 156 and 161 to 198.

In addition, the solution structure of the MA fragment was solved in complex with unlabeled YM_2 (Figure 4B), wherein only three of the initially unassigned 14 N-terminal residues remained unassigned. The rmsd in the ordered regions (i.e., residues 119–157, 161–180, and 183–196) across the 20 lowest

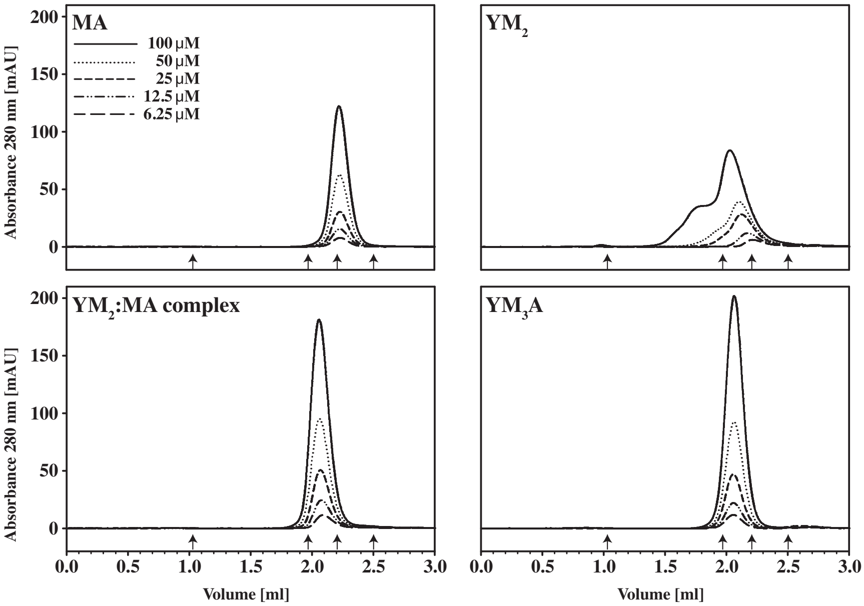

Figure 3. Analytical Size Exclusion Analysis

YM$_2$ behaves as a monomer at low concentrations and shows oligomerization at higher concentrations (top right). MA elutes as one monomeric peak at all tested concentrations (top left). The complex of YM$_2$ and MA elutes as one monomeric peak (bottom left) with the same elution volume as full-length YM$_3$A (bottom right) at all tested concentrations. The arrows indicate positions (from left to right) of exclusion volumes corresponding to the void volume and proteins with molecular weights of 44.3, 25.0, and 13.7 kDa.

energy conformers is 0.46 ± 0.06 Å for backbone atoms and 0.90 ± 0.08 Å for all heavy atoms. The structure reveals that the additionally assigned N-terminal residues of the MA fragment span a stable α helix. Overall, the structure now matches the corresponding region in the crystal structure remarkably well, the closest-to-average NMR conformer aligning to the crystal structure in the ordered regions with an rmsd of 1.09 Å and 1.39 Å for backbone and all heavy atoms, respectively (Figure 4B).

Finally, we investigated the structure of the self-assembled YM$_2$:MA complex formed in solution by the two fragments. During the assignment process, it became clear that even in the complex, a considerable proportion of the YM$_2$ fragment is in conformational exchange; specifically, the N-terminal capping repeat (Y) and the H1 helix of the first internal repeat (M^1) show excessive peak broadening. Although the construct contains two identical repeats, it was possible to obtain nearly complete backbone and side-chain assignments, with the exception of the aforementioned exchange-broadened N-terminal residues (Figures S5 and S6).

The assignment and structure calculation procedures are described in detail in the Supplemental Information. Briefly, a well-defined solution structure of the complex could be determined by augmenting the intramolecular distance restraints for YM$_2$ and MA with interfacial NOEs, the latter being identified via ^{13}C-filtered/edited NOE spectroscopy (NOESY) spectra using two complementarily ^{13}C-labeled samples (Otting and Wüthrich, 1990). In the end, the solution structure of the assembled YM$_2$:MA complex was calculated from a total of 2,195 restricting constraints (Table 2). Of these, 404 were long-range distance constraints (approximately four per restrained residue) with 77 restricting intermolecular distances, defining the tertiary structure of the fragments along the entire length of the interface (Figure 5).

The aligned structures in the refined ensemble of 20 YM$_2$:MA structures determined by NMR are shown in Figure 6A. With the notable exception of the almost entirely unrestrained residues 1 to 34 of YM$_2$, the structure of the complex is well defined (Table 2); for clarity, description of this N-terminal tail is omitted in the following discussion. Ordered regions (i.e., residues 44–54, 56–114, 119–157, 161–180, and 183–196) exhibit averaged rmsd values of 0.64 ± 0.08 Å for superimposed backbone atoms and 1.01 ± 0.08 Å for all heavy atoms (Figure S7). The NMR solution structure of the complex differs from the crystallographic structure of the full-length protein (PDB accession number 4DBA) by a backbone heavy atom rmsd of only 1.34 Å (Figure 6B).

The YM$_2$ and MA fragments self-assemble in solution, forming a complex with a structure that is highly similar to the uninterrupted YM$_3$A protein. The conformers in the bundle exhibit an average interfacial contact surface of 827.1 ± 31.5 Å^2. This area is remarkably congruent to the contact surface of the uninterrupted protein in the crystal structure (808.2 Å^2), despite a small cavity (surface area ~59 Å^2) in the solution structure located between residues Val91 and Ala128. Closer inspection of the interaction surface reveals that it is dominated by van der Waals contacts between hydrophobic side chains (Figure 7). Few interfacial hydrogen bonds are observed in the NMR structure, namely, Ser110-O$^{\gamma}\cdots$HN$^{\delta2}$-Asn153, Ser114-O$^{\gamma}\cdots$HN$^{\delta2}$-Asn153, and Ala113-CO$\cdots$HN$^{\varepsilon2}$-Gln119. The only observed backbone-backbone interaction involves the charged termini of the two fragments, with the salt bridge Gly115-COO$^{-}\cdots$H$_3$N^{+}-Gly116 mimicking the peptide bond in the covalently bound crystal structure.

MD Simulations Confirm the Structural Stability of the Assembly

Multiple 1 μs MD simulations in explicit solvent were carried out to investigate the stability of the two-fragment assembly and the intrinsic plasticity of the N-terminal segment (for an overview, see Table S5). The results for the following three starting structures are discussed here (further MD simulations are described in the Supplemental Information): (1) the lowest energy conformer of the NMR structural bundle of the YM$_2$:MA complex refined in explicit transferable intermolecular potential three-point (TIP3P) water, (2) the crystal structure (PDB accession number 4DBA, chain A), and (3) an artificial complex derived from the entire crystal structure, in which the amide bond between Gly115 and Gly116 was "hydrolyzed" by replacing it

Table 1. NMR Constraints and Structure Statistics for Uncomplexed MA

Variable	Value
Total No. of Restricting Constraints	883
NOE constraints	1,089 (16.3 per constrained res.)
Unambiguous distances	758
Ambiguous distances	331
Restricting distances	756 (11.3 per constrained res.)
Intraresidual	260
Sequential	221
Medium range (2–4)	148
Long range ($\geq$5)	127 (3.7 per constrained res.)
Torsion angle constraints	127
Satisfaction of Experimental Constraints	
NOE distance constraints	
Violations > 0.5 Å per structure	0
Violations > 0.2 Å per structure	10.5 ± 1.7
Average violation (Å)	0.0124 ± 0.0007
Rmsd of violations (Å)	0.0447 ± 0.0014
Angular Constraints	
Number of violations > 10°	0
Number of violations > 1°	4.4 ± 1.8
Average violation (°)	0.1000 ± 0.0001
Rmsd of violations (°)	0.3976 ± 0.0662
Energies (kJ/mol)	
Total	−12,854.62 ± 191.63
Distance restraints	323.92 ± 19.58
Dihedral restraints	10.17 ± 2.22
Ramachandran Plots	
PROCHECK	
Core regions (%)	95.7
Allowed regions (%)	4.3
Generously allowed regions (%)	0.0
Disallowed regions (%)	0.0
MolProbity	
Favored (98%) regions (%)	98.5
Allowed (>99.8%) regions (%)	1.5
Disallowed regions (%)	0.0
Residue Properties	
Close contacts[a]	9
MolProbity clashscore	33.74 (*Z* score = −4.26)
Global Quality Scores	
Verify3D	0.38 (*Z* score = −1.28)
PROSA II	0.80 (*Z* score = 0.62)
PROCHECK (phi-psi)	0.05 (*Z* score = 0.51)
PROCHECK (all)	−0.17 (*Z* score = −1.01)
Idealized Geometry Rmsd (Å)[b]	
Bonds (Å)	0.018
Angles (°)	1.5
Averaged Structure Kabsch Rmsd (Å)[c]	
Backbone (N, CA, C′, O)	0.74 ± 0.18
All heavy atoms	1.16 ± 0.15

Statistics over the selected bundle of 20 NMR structures. res., residue.

[a] Within 1.6 Å for hydrogens and 2.2 Å for heavy atoms.

[b] Idealized covalent geometry based on PDB validation software.

[c] Rmsd values are for ordered regions as selected by PDBSTAT (i.e., residues 130–156 and 161–198).

with a negatively charged carboxy group and a positively charged amino group at Gly115 and Gly116, respectively. This simulation, called "split-xtal," is a reference simulation with the same two-fragment assembly as in the NMR experiments.

The simulations reveal that all repeats with the exception of the N-cap are structurally stable on a 1 μs timescale (Figure 8A; Figure S8). The time series of the rmsd of the Cα atoms (Figure 8A) and the time evolution of the secondary structure (Figure S9) indicate that the tertiary structure of repeats in the YM_2:MA complex is conserved. Importantly, control simulations starting from the crystal structure (PDB accession number 4DBA), both with and without the covalent amide bond between Gly115 and Gly116, display very similar structural stability for the whole protein (Figure 8A) and a flexibility profile similar to that found by simulations started from the NMR conformer. The N-cap (Y) is more flexible in all runs and can assume a helical structure, which is transient (Figure S9).

Moreover, the time series of the interaction energy between pairs of repeats reveals that interactions between the covalently linked M^1 and M^2 repeats are similar when compared with those between M^2 and M^3 in the YM_2:MA complex, that is, along the trajectory of the simulations starting from the NMR coordinates or the artificially split crystal conformer (Figure S8). Van der Waals interactions between repeats are very similar irrespective of the starting structure and presence of the peptide bond between Gly115 and Gly116 and, with exception of the M^3A pair, distinctly more favorable than Coulombic interactions.

The distance between the termini of the fragments remains short throughout most of the simulation (Figure 8B). The two termini form a direct salt bridge or one separated by just one H_2O molecule during most of the MD sampling. Crucially, even when separations of the two termini of up to 20 Å occur during the simulation (e.g., for NMR3 in Figure 8B), the hydrophobic contacts between the interfacial helices remain essentially unchanged, underlining the importance of hydrophobic interactions for complex stability (Movie S1). Stability of the complex in solution is given primarily through hydrophobic interactions at the core of the molecule, with complementary polar affinities only participating at the periphery of the contact area.

In conclusion, the simulation results are consistent with the NMR data and provide further evidence that the conformation of the split heterodimer is stable except for the N-cap, whose intrinsic flexibility is due to its sequence and suboptimal packing against the adjacent M repeat in both the complex and the full-length protein.

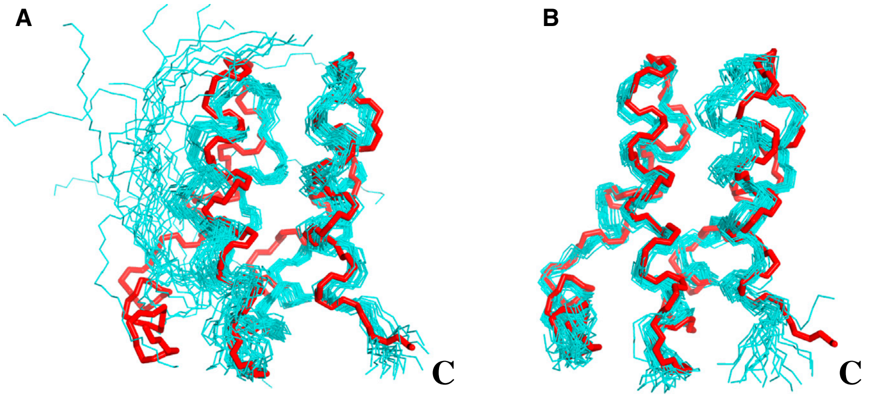

Figure 4. NMR Structures of MA

(A) Twenty lowest energy conformers of uncomplexed MA.

(B) Twenty lowest energy conformers of MA in complex with unlabeled YM_2.

In both cases, the conformer bundles are superimposed with the corresponding region from the crystal structure (PDB accession number 4DBA, red) of the entire protein, YM_3A. The location of the C terminus of MA is indicated.

DISCUSSION

Large proteins fold into individual domains, which are defined as (essentially) autonomous folding units. Formation of native-like contacts in these units occurs in a synchronized manner, resulting in cooperative folding behavior (Fersht, 2000). Conversely, truncated forms of most globular proteins do not allow the formation of all required interactions. Indeed, proteins that are truncated within domain borders are usually insoluble as a consequence of the hydrophobic core being exposed to the solvent, resulting in severe aggregation and possibly precipitation (Thirumalai et al., 2003). Therefore, most globular proteins cannot be reconstituted from fragments. We note that this phenomenon is not limited to globular proteins but was also observed by us in repeat proteins, such as designed ankyrin repeat proteins, where fragments missing one or both capping repeats show a high tendency to aggregate (Interlandi et al., 2008).

In stark contrast to these previous observations, we demonstrate in this work using solution NMR methods that a consensus-designed ArmRP, when split into two fragments, is indeed capable of regaining the structure of the parent protein through the formation of a noncovalent complex. Crucially, the C-terminal MA fragment is structured to a large degree. In this way, we postulate that it serves as a template onto which the N-terminal YM_2 fragment can attach in a coupled folding-binding event with remarkably high affinity and in a structurally well-defined manner, characteristic of a very specific interaction. Furthermore, biophysical analysis reveals that the complex remains monomeric and assembles in a defined 1:1 manner that is highly similar to the covalently linked full-length protein.

A number of other proteins exist that can be reconstituted from complementary fragments. One famous example is the entire class of split inteins that when mixed form a splicing-competent protein (Wu et al., 1998). Other well-known proteins that can be reconstituted from complementary fragments are ribonuclease A (Richards, 1958) and ubiquitin (Johnsson and Varshavsky, 1994). However, in all these examples, the site of the split cannot be easily shifted to a remote location. Whether the behavior in the ArmRP described in this paper is a generic feature of repeat proteins remains to be investigated in the future.

So what makes the ArmRPs so special that their fragments remain in solution and are able to reconstitute the entire protein when mixed together? Obviously, the individual fragments must remain soluble, and at least one of the two fragments should exist in a nonaggregated state. Moreover, complex formation via coupled folding-binding is also facilitated if both fragments assume a structure that is not too different from the native state, that is, one in which the helical secondary structure elements are formed and not too many long-range contacts are disrupted within the fragments. Moreover, in globular proteins, contacts are routinely formed between residues far apart in sequence, the extent of which is quantified by the contact order (Makarov et al., 2002). In contrast, repeat proteins intrinsically possess low contact order, as contacts can be formed only between residues of neighboring repeats (Cortajarena and Regan, 2012). In a repeat protein fragment, all intrafragment contacts therefore remain present, and only the contacts to one neighboring repeat are lost; the latter interactions may be easily reestablished during complexation.

What about the stability of the individual fragments? Surprisingly, the program AGADIR, which estimates propensities for helix formation on the basis of amino acid sequence (Muñoz and Serrano, 1994), predicts a helical content of only 1.2% for the entire YM_3A protein, even though it is almost completely helically folded. This may be explained by the fact that the particular triangular spiral staircase arrangement of helices in ArmRPs results in a large number of tertiary contacts, contributions that are not taken into account by the AGADIR software. Despite the presence of many such tertiary contacts, individual repeats remain essentially unstable. Indeed, most of their stability appears to arise solely from their interactions with neighboring repeats. The Ising model, which is commonly used to describe the energetics of repeat protein folding, allows differentiation between the intra- and interrepeat contributions to the global free energy of folding (Zimm and Bragg, 1959, Mello and Barrick, 2004, Kajander et al., 2005, Wetzel et al., 2008). For most repeat proteins, the interresidue coupling energy is much more favorable than the contributions from within individual repeats. Accordingly, the stability of repeat proteins generally increases linearly with increasing number of repeats, to the extent that consensus ankyrin repeat proteins eventually become so stable that they can no longer be unfolded thermally (Wetzel et al., 2010). Conversely, this means that a single repeat is unlikely to fold on its own, but two or three connected repeats may constitute a stable unit, although with still limited stability.

For the reasons presented here, we believe that repeat proteins are inherently more suitable than globular proteins to enable reconstitution of the full-length protein from two such fragments. It is apparent that the formation of sufficiently stable

Table 2. NMR Constraints and Structure Statistics of the YM_2:MA Complex

Variable	Value
Total No. of Restricting Constraints	2,195
NOE constraints	2,634 (15.8 per constrained res.)
Unambiguous distances	1,937
Ambiguous distances	697
Restricting distances	1,916 (11.5 per constrained res.)
Intraresidual	508
Sequential	506
Medium range (2–4)	498
Long range ($\geq$5)	404 (3.6 per constrained res.)
Interchain ($YM_2 \leftrightarrow$ MA)	77 (3.7 per constrained res.)
Torsion angle constraints	279
Satisfaction of Experimental Constraints	
NOE distance constraints	
Violations > 0.5 Å per structure	0
Violations > 0.2 Å per structure	9.9 ± 2.5
Average violation (Å)	0.0085 ± 0.0005
Rmsd of violations (Å)	0.0323 ± 0.0011
Angular constraints	
Number of violations > 10°	0
Number of violations > 1°	5.9 ± 1.9
Average violation (°)	0.0900 ± 0.0308
Rmsd of violations (°)	0.2947 ± 0.0602
Energies (kJ/mol)	
Total	−30,415.88 ± 334.76
Distance restraints	428.02 ± 28.33
Dihedral restraints	6.40 ± 2.51
Ramachandran Plots	
PROCHECK	
Core regions (%)	95.8
Allowed regions (%)	3.4
Generously allowed regions (%)	0.4
Disallowed regions (%)	0.3
MolProbity	
Favored (98%) regions (%)	95.3
Allowed (>99.8%) regions (%)	3.9
Disallowed regions (%)	0.8
Residue Properties	
Close contacts[a]	32
MolProbity clashscore	25.50 (*Z* score = −2.85)
Global Quality Scores	
Verify3D	0.31 (*Z* score = −2.41)
PROSA II	0.96 (*Z* score = 1.28)
PROCHECK (phi-psi)	0.01 (*Z* score = 0.35)
PROCHECK (all)	−0.22 (*Z* score = −1.30)
Idealized Geometry Rmsd (Å)[b]	
Bonds (Å)	0.016
Angles (°)	1.3
Averaged Structure Kabsch Rmsd (Å)[c]	
Backbone (N, CA, C′, O)	0.64 ± 0.08
All heavy atoms	1.01 ± 0.08

Statistics over the selected bundle of 20 NMR structures. res., residue.

[a] Within 1.6 Å for hydrogens and 2.2 Å for heavy atoms.

[b] Idealized covalent geometry based on PDB validation software.

[c] Rmsd values are for ordered regions as selected by PDBSTAT (i.e., residues 44–54, 56–114, 119–157, 161–180, and 183–196).

and soluble fragments imposes restraints on the sequence and architecture of the underlying modules. The sequence restraints that have been proposed for the successful design of well-folded proteins also apply in this case (Koga et al., 2012). These restraints rely on two opposing categories of properties: some that help the polypeptide chain adopt a particular fold and some that explicitly prevent misfolding, a situation that was previously described as “negative design” (Thirumalai et al., 2003). ArmRPs may be especially useful for uncovering such properties because they possess an unusually dense network of tertiary contacts, a major fraction of which remains intact even for residues at the fragment interface. ArmRP folding topology is simple enough; that is, they have no β sheet structure and very short loops, both features that can be the source of the stabilizing interactions of aggregates. Moreover, they contain a sufficiently large number of surface charges and do not easily convert into β-type structure, which helps the fragments remain soluble. This becomes apparent from the behavior of the N-terminal fragment, which in its uncomplexed form exists as a molten globule that nonetheless retains a high content of helical secondary structure.

Last, an important feature of the system under study is the mostly structured C-terminal fragment, which presents a stable, soluble protein, even though it consists of only two repeat modules. The absence of amide signals from the H1 helix in the M module suggests that even that stretch is not entirely unstructured but interconverts between different, mostly helical, conformers. MA thus serves as a template onto which YM_2 docks with high specificity and affinity. Simple packing of the interface helices against each other results in formation of a stable protein complex.

We believe that this result has important ramifications when considering the way in which these repeat proteins may have evolved in nature. Indeed, it is generally assumed that repeat proteins have arisen by gene duplication of the repeats (Haigis et al., 2002, Lee and Blaber, 2011) and that a gain of function of a longer protein drove the selection. Nonetheless, in present-day genes of natural ArmRPs, exon-intron boundaries do not correspond well to structural protein repeats, suggesting that this modular gene duplication must have occurred in prebiotic gene evolution. Indeed, the noncovalent assembly of repeats is a plausible intermediate during prebiotic protein evolution (Söding and Lupas, 2003), when particular exons may have proliferated because of their versatile assembly properties. The ArmRPs may thus recapitulate an early form of this exon.

With the designed proteins used here, which are based on the consensus sequence and may thus be close to primordial

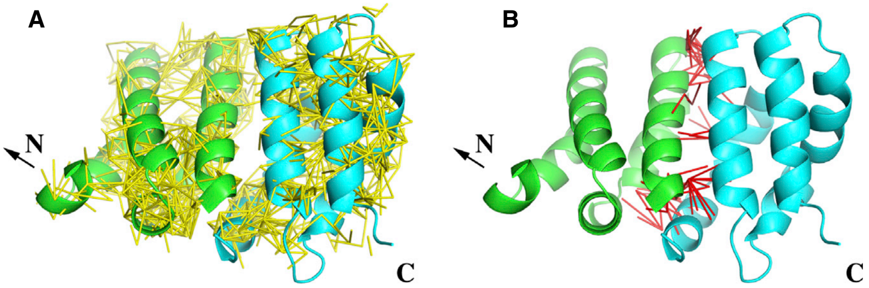

Figure 5. Distance Constraints Mapped on the Closest-to-Average NMR Conformers of the Armadillo Complex

The complexed YM_2 (green) and MA (cyan) fragments are visualized as cartoons; residues 1 to 34 are omitted for clarity. Upper-limit distance constraints are shown for intramolecular restraints (A) (yellow) and interfacial restraints (B) (red).

armadillo sequences, we can directly see their noncovalent assembly, which would be extremely difficult to achieve for most other protein fragments. In the case of consensus ArmRPs, the solubility of the fragments is high enough that assembly can be directly shown. Although many globular proteins have been split into fragments that can reassemble (Shekhawat and Ghosh, 2011), the ArmRP assembly may have lead to a rapid evolution of functional repeats that enabled their widespread use in the binding of extended peptides of different sequence, possibly including protein fragments as an early evolutionary intermediate.

By swapping large segments between repeat proteins—no longer requiring that the introns be placed at structural protein boundaries—the diversity of the pools is rapidly increased at the genetic level. Such a swap is much less likely to occur in a globular protein, because globular proteins frequently possess long-range contacts that are idiosyncratic. Their formation is important for protein stability, and it is highly unlikely that these contacts will be retained when swapping segments.

EXPERIMENTAL PROCEDURES

Cloning, Expression, and Purification

The expression vector pLIC_CR is a variant of pZE12-Luc (Lutz and Bujard, 1997) containing lacI^q, a T5/lacO promoter followed by an N-terminal MRGSH$_6$-tag, a recombinant tobacco etch virus (rTEV) recognition site and a *Sac*B gene as additional selection marker. All pLIC constructs (YM_2, M_2A, MA, and YM_3A) were amplified from pPANK-$Y_{II}M_3A_{II}$ (Madhurantakam et al., 2012) (Table S1). Ligase-independent cloning (Aslanidis and de Jong, 1990) and selection (Gay et al., 1983) were performed as previously described. YM was expressed from pLIC_RW_Trp_3C_YM as insoluble, 3C protease cleavable Trp-leader fusion (Miozzari and Yanofsky, 1978). YMR and YMRRR were expressed from a pPANK-based vector (Parmeggiani et al., 2008) with N-terminal MRGSH$_6$-tag.

Proteins were expressed in *E. coli* M15 (pREP4) in Luria broth medium for unlabeled proteins and in M9 minimal medium supplemented with ^{15}N-NH_4Cl or ^{13}C-glucose as previously described (Wetzel et al., 2010) with an induction optical density at 600 nm of 0.5. Cell pellets were resuspended in TBS_{500} (50 mM Tris·HCl pH 8.0, 500 mM NaCl, 5% glycerol) and lysed by sonication as previously described (Wetzel et al., 2010). The His-tag was cleaved with rTEV protease (molar ratio enzyme/protein 1:30 for MA and M_2A 1:5 for YM_2 during dialysis into PBS_{150} [50 mM sodium phosphate, 150 mM NaCl, 2% glycerol]) at room temperature (pH 7.4). Cleaved His-tag and uncleaved educt were removed by reverse nickel immobilized metal affinity chromatography (Ni-IMAC). YM was purified from inclusion bodies, which were washed with TBS_{500} plus 0.1% Triton X-100 twice; YM was solubilized in TBS_{500} plus 8 M urea and purified by Ni-IMAC. The His-tag-Trp-leader sequence was cleaved by 3C protease at room temperature drop-wise during refolding into TBS_{500} plus 5% glycerol. Cleaved His-tag-Trp-leader and uncleaved educt were removed by reverse Ni-IMAC, and YM was concentrated, dialyzed into TBS_{500} plus 2% glycerol, and complexed with M_2A.

All rTEV-digested fragments as well as the YM_2:MA and YM:M_2A complexes were further purified by preparative SEC (S75 16/60 HiLoad; GE Healthcare). For NMR experiments requiring complete complexation of isotopically labeled protein, an excess (1.5 equivalent) of unlabeled protein was used.

SEC and MALS

Analytical SEC was carried out in PBS_{150} containing 2% glycerol at pH 7.4 (S200 5/150 GL; Pharmacia). Samples of the same preparation used for analytical SEC, CD, and ITC measurements (see the following discussion) were analyzed using MALS as described previously (Varadamsetty et al., 2012). Minor deviations of the molecular weight determined by MALS from the expected calculated weight can be explained by the calibration to globular protein standards of the MALS analysis.

CD and ITC

CD, ITC, analytical size exclusion, and MALS analysis was carried out with 6.25 to 100 μM protein in PBS_{150} (pH 7.4) and 2% glycerol. CD measurements were carried out on a Jasco J-810 spectropolarimeter with a cylindrical cuvette (path length 0.5 mm). Data were recorded at 20°C from 190 to 250 nm (data pitch 1 nm, scan speed 20 nm/min, response time 4 s, bandwidth 1 nm). The CD signal was corrected by buffer subtraction and converted to mean residue ellipticity.

The K_d of the YM_2:MA complex assembly was determined on a VP-ITC (MicroCal) instrument at 32°C (Figure 2). Samples of YM_2 and MA were each dialyzed twice (12 h) against PBS_{150} (pH 7.4) at 4°C. YM_2 (in the cell) was diluted to 6.7 μM with dialysis buffer, and 69.7 μM MA was added in 32 10 μl steps during titration (300 s interval, cell volume 1.47 mL). Data integration and fitting were carried out using Origin Software.

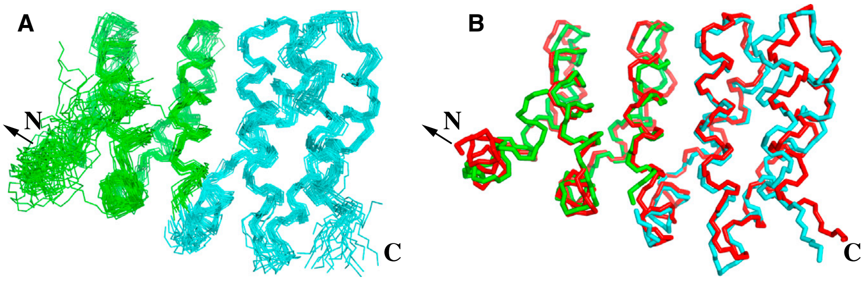

Figure 6. 3D Structures of the YM_2:MA Complex

The polypeptide backbone is shown in stick representation. NMR structures are colored by fragment: YM_2 (green) and MA (cyan). Note that the two fragments are not covalently linked. Locations of termini are indicated by letters; the unrestrained residues 1 to 34 are omitted for clarity.

(A) Superposition of 20 NMR structures.

(B) Closest-to-average NMR structure superimposed onto the crystal structure of full-length YM_3A (PDB accession number 4DBA, red).

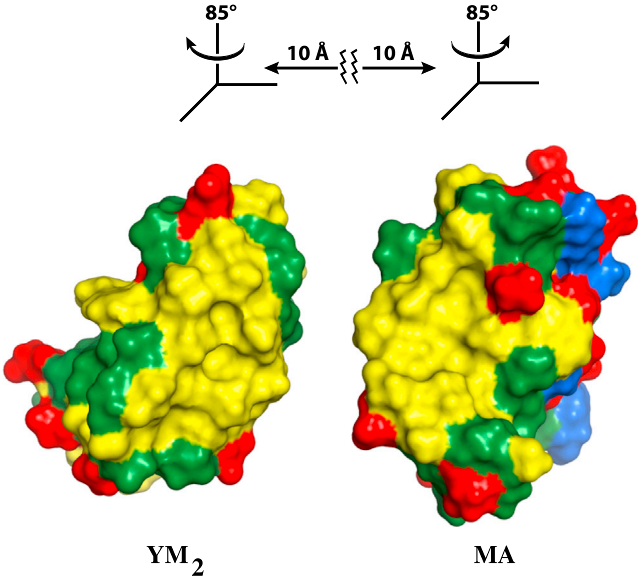

Figure 7. Interface between the YM_2 and MA Fragments in an "Open Book" View

The Connolly contact surface is colored by amino acid charge: nonpolar side chains are depicted in yellow, acidic in red, basic in blue, and polar uncharged in green. For this visualization, the closest-to-average NMR structure of the YM_2:MA complex was separated at the interface and the two fragments splayed as indicated.

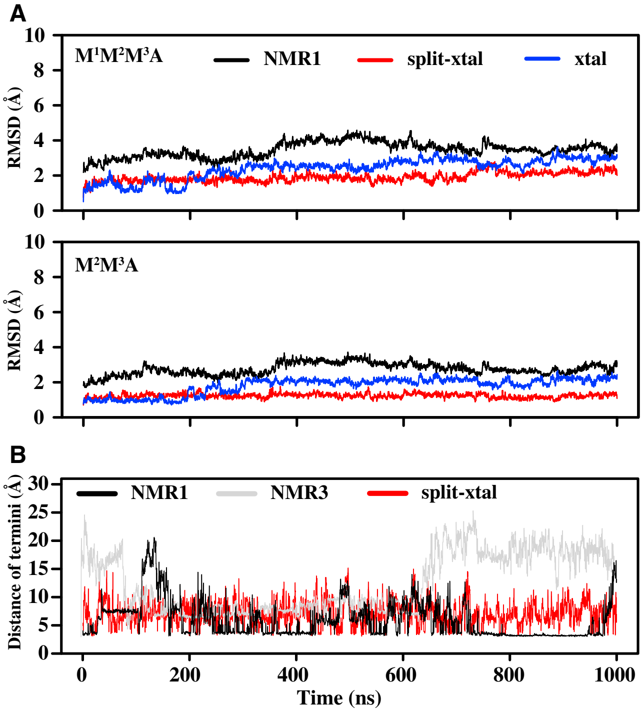

Figure 8. MD Simulations

(A) Structural stability of the YM_2:MA complex analyzed by MD simulations through comparison with the full-length protein. The time series of the rmsd from the crystal structure (PDB accession number 4DBA) were calculated for the Cα atoms of repeats $M^1M^2M^3A$ (top) and M^2M^3A (middle) to the crystal structure. Note that only in the xtal run is M^2 covalently linked to M^3.

(B) Time series of the distance between the carboxyl C atom of Gly115 and the amino N atom of Gly116 termini along the MD simulations, which started from the lowest energy NMR structure of runs NMR1 and NMR3 and the split-xtal runs.

NMR Spectroscopy, Assignments, and Structure Calculation

All NMR samples were prepared in PBS_{150} buffer (pH 7.4) with 10% D_2O, 1 mM tetramethylsilylpropanate, 0.02% NaN_3, and 2% glycerol and recorded at 34°C on Bruker Avance 600 and 700 MHz spectrometers.

Spectra for the uncomplexed C-terminal fragment were collected at a concentration of 0.75 mM labeled MA; 1.2 equivalents of unlabeled YM_2 were added to determine the structure of complexed MA. In an analogous fashion, spectra of labeled YM_2 were initially measured at 1.0 mM with 1.5 equivalents of unlabeled MA added to determine the structure of the N-terminal fragment in the complex. Resonance assignments for the complex structure (see the following discussion) were performed pairwise on the correspondingly labeled fragments in presence of the unlabeled partner.

Resonances were assigned from triple-resonance spectra. Spectra were processed using the software TOPSPIN 2.1 and analyzed in CARA (Keller, 2004) and CCPN Analysis (Vranken et al., 2005).

The N-terminal YM_2 fragment was difficult to assign because the protein contains two repeats with identical amino acid sequence. Although the amide moieties could be linked unambiguously from the same set of triple-resonance spectra, side-chain assignments were more challenging because of the increased overlap in the [^{13}C,^{1}H]-HSQC. In the first step, the FLYA module of CYANA was used for automatic side-chain assignments (Schmidt and Güntert, 2012). However, likely because of the degeneracy of the amino acid sequence, the automatic procedure yielded only a few correct assignments. The most valuable data for side-chain assignment were provided by the HN(CO)CCCH experiment, which helped connect amide moieties to side-chain spin systems; the 4D HCCH-TOCSY experiment allowed the recognition of entire side chains from either a single ^{13}C-α entry or a methyl group, and four-dimensional NOESY-HSQC aided greatly in the disambiguation of Leu side-chain resonances. After several assignment rounds, 95.2% of all expected backbone amide and 90.9% of all proton resonances could be annotated for residues 30 to 105. Final chemical shift assignments, obtained after manual verification of the peak lists for individual spectra, were used for the automatic peak annotations of NOESY peak lists. The obtained assignments for HN, Hα, Cα, Cβ, CO, and N chemical shifts were also used to predict backbone ϕ/ψ dihedral angle restraints using TALOS-N (Shen and Bax, 2013). For annotated [^{15}N,^{1}H]-HSQC spectra of MA and YM_2, see Figures S4 and S5.

Distance restraints for the uncomplexed and complexed MA as well as for the complexed YM_2 fragments were obtained from both ^{13}C- and ^{15}N-resolved 3D HSQC-NOESY spectra (τ_{mix} = 75 ms) (Zhang et al., 1994). To solve the structure of the YM_2:MA complex, we recorded two sets of ^{13}C,^{15}N-filtered,^{13}C-edited as well as ^{13}C,^{15}N-filtered,^{15}N-edited NOESY spectra on samples in which only one of the two partners was doubly labeled. Careful comparison with the standard NOESY experiments of the separate fragments allowed the identification of intermolecular cross-peaks. Interfacial H-H distances < 5 Å were extracted from the crystallographic structure of the corresponding single-chain armadillo construct (PDB accession number 4DBA) and used to generate a synthetic NOE peak list to help guide the assignment process, which resulted in 66 interfacial distance restraints. These distance restraints were artificially loosened by 1 Å and added to the structure calculation in order to partially constrain the complex while allowing for local rearrangement. Using all these restraints plus the TALOS-N derived dihedral restraints for YM_2 and MA, a structural ensemble was calculated for the entire complex.

In the final refinement, the conformational ensembles for both structures were subjected to refinement in explicit TIP3P water using the parallhdg5.3 parameters implemented in the nmr_waterrefine extension (Linge et al., 2003, Nabuurs et al., 2004) to XPLOR-NIH. For statistics of assignments and further details of structure calculations and verifications, see Tables 1 and 2, as well as Figures S6 and S7.

MD Simulations

We have performed five MD simulations starting from the coordinates of the NMR structure of the YM_2:MA complex and two starting from the crystal structure (for an overview, see Table S5).

MD simulations were carried out using GROMACS version 4.5.5 (Van der Spoel et al., 2005; Hess et al., 2008), with the CHARMM36 force field (Best et al., 2012). All ionizable residues were modeled in their standard state at pH ~7.4; that is, Asp and Glu side chains and C termini were negatively charged, Arg and Lys side chains and N termini were positively charged, and the His side chains were neutral. Each of the protein structures (NMR1–3, 4DBA-NMR, 4DB6-NMR, xtal, and split-xtal) was individually solvated in a dodecahedral box of TIP3P (Jørgensen et al., 1983) water molecules, with the box edge at a distance of at least 1.2 nm from the protein surface. Ions (Na^+ and Cl^-) were added to neutralize the total charge of the system at a concentration of 150 mM. After energy minimization, a 0.1 ns equilibration at constant molecular number, volume, and 310 K temperature, with positional restraints on protein, was performed. The pressure was equilibrated in a 0.9 ns position-restrained simulation at constant molecular number, pressure, and temperature (NPT). The 1 ns equilibration was followed by unrestrained NPT simulations (i.e., productive runs) at 310 K (the temperature of NMR data acquisition) and a length of 1 μs each. The temperature and pressure (1 bar) of the system were controlled with the modified Berendsen (velocity-rescaling) (Bussi et al., 2007) and the Berendsen (Berendsen et al., 1984) algorithms, respectively. To avoid finite-size effects, periodic boundary conditions were applied in all three dimensions. Coulomb and van der Waals interactions, as well as the short-range neighbor list, were cut off at 1.0 nm, whereas the particle mesh Ewald summation method (Darden et al., 1993) was used for the calculation of long-range electrostatics. The Lincs algorithm was used to constrain covalent bonds to their equilibrium lengths, allowing a time step of 2 fs.

ACCESSION NUMBERS

The coordinates of uncomplexed MA and of the complex formed by YM_2 and MA have been deposited in the PDB database under accession numbers 2RU5 and 2RU4, respectively. Chemical shifts and experimental restraints were deposited in the Biological Magnetic Resonance Bank database under accession numbers 11548 and 11544 for MA and the YM_2:MA complex, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, ten figures, five tables, and one movie and can be found with this article online at http://dx.doi.org/10.1016/j.str.2014.05.002.

ACKNOWLEDGMENTS

We acknowledge financial support by the SINERGIA program of the Swiss National Science Foundation (grant no. 122686). We are indebted to Dr. I. Jeselarov for help with the ITC measurements and to the NMR team at the EU-NMR center in Warsaw for measuring and processing the four-dimensional spectra. MD simulations were carried out on the Schrödinger cluster of the University of Zurich.

Received: March 19, 2014
Revised: April 30, 2014
Accepted: May 2, 2014
Published: June 12, 2014

REFERENCES

Alfarano, P., Varadamsetty, G., Ewald, C., Parmeggiani, F., Pellarin, R., Zerbe, O., Plückthun, A., and Caflisch, A. (2012). Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy. Protein Sci. *21*, 1298–1314.

Andrade, M.A., Perez-Iratxeta, C., and Ponting, C.P. (2001). Protein repeats: structures, functions, and evolution. J. Struct. Biol. *134*, 117–131.

Aslanidis, C., and de Jong, P.J. (1990). Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res. *18*, 6069–6074.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., and Haak, J.R. (1984). Molecular dynamics with coupling to an external bath. J. Chem. Phys. *81*, 3684–3690.

Best, R.B., Zhu, X., Shim, J., Lopes, P.E., Mittal, J., Feig, M., and Mackerell, A.D.J., Jr. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ(1) and χ(2) dihedral angles. J. Chem. Theory Comput. *8*, 3257–3273.

Boersma, Y.L., and Plückthun, A. (2011). DARPins and other repeat protein scaffolds: advances in engineering and applications. Curr. Opin. Biotechnol. *22*, 849–857.

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. J. Chem. Phys. *126*, 014101.

Cortajarena, A.L., and Regan, L. (2012). Comprehensive Biophysics. (New York: Elsevier), pp. 267–289.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. J. Chem. Phys. *98*, 10089–10092.

Dyson, H.J., and Wright, P.E. (2004). Unfolded proteins and protein folding studied by NMR. Chem. Rev. *104*, 3607–3622.

Fersht, A. (2000). Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. (New York: Freeman).

Gay, P., Le Coq, D., Steinmetz, M., Ferrari, E., and Hoch, J.A. (1983). Cloning structural gene sacB, which codes for exoenzyme levansucrase of Bacillus subtilis: expression of the gene in Escherichia coli. J. Bacteriol. *153*, 1424–1431.

Haigis, M.C., Haag, E.S., and Raines, R.T. (2002). Evolution of ribonuclease inhibitor by exon duplication. Mol. Biol. Evol. *19*, 959–963.

Hatzfeld, M. (1999). The armadillo family of structural proteins. Int. Rev. Cytol. *186*, 179–224.

Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. *4*, 435–447.

Huber, A.H., and Weis, W.I. (2001). The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. Cell *105*, 391–402.

Interlandi, G., Wetzel, S.K., Settanni, G., Plückthun, A., and Caflisch, A. (2008). Characterization and further stabilization of designed ankyrin repeat proteins by combining molecular dynamics simulations and experiments. J. Mol. Biol. *375*, 837–854.

Johnsson, N., and Varshavsky, A. (1994). Split ubiquitin as a sensor of protein interactions in vivo. Proc. Natl. Acad. Sci. USA *91*, 10340–10344.

Jørgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J. Chem. Physiol. *79*, 926–935.

Kajander, T., Cortajarena, A.L., Main, E.R., Mochrie, S.G., and Regan, L. (2005). A new folding paradigm for repeat proteins. J. Am. Chem. Soc. *127*, 10188–10190.

Keller, R.L.J. (2004). The Computer Aided Resonance Assignment Tutorial. http://cara.nmr-software.org/downloads/3-85600-112-3.pdf.

Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T.B., Montelione, G.T., and Baker, D. (2012). Principles for designing ideal protein structures. Nature *491*, 222–227.

Lee, J., and Blaber, M. (2011). Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. Proc. Natl. Acad. Sci. USA *108*, 126–130.

Linge, J.P., Williams, M.A., Spronk, C.A., Bonvin, A.M., and Nilges, M. (2003). Refinement of protein structures in explicit solvent. Proteins *50*, 496–506.

Ludwig, C., Schwarzer, D., Zettler, J., Garbe, D., Janning, P., Czeslik, C., and Mootz, H.D. (2009). Semisynthesis of proteins using split inteins. Methods Enzymol. *462*, 77–96.

Lutz, R., and Bujard, H. (1997). Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Res. *25*, 1203–1210.

MacDonald, B.T., Tamai, K., and He, X. (2009). Wnt/beta-catenin signaling: components, mechanisms, and diseases. Dev. Cell *17*, 9–26.

Madhurantakam, C., Varadamsetty, G., Grütter, M.G., Plückthun, A., and Mittl, P.R. (2012). Structure-based optimization of designed Armadillo-repeat proteins. Protein Sci. *21*, 1015–1028.

Makarov, D.E., Keller, C.A., Plaxco, K.W., and Metiu, H. (2002). How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proc. Natl. Acad. Sci. USA *99*, 3535–3539.

Marcotte, E.M., Pellegrini, M., Yeates, T.O., and Eisenberg, D. (1999). A census of protein repeats. J. Mol. Biol. *293*, 151–160.

Marfori, M., Mynott, A., Ellis, J.J., Mehdi, A.M., Saunders, N.F., Curmi, P.M., Forwood, J.K., Bodén, M., and Kobe, B. (2011). Molecular basis for specificity of nuclear import and prediction of nuclear localization. Biochim. Biophys. Acta *1813*, 1562–1577.

Mello, C.C., and Barrick, D. (2004). An experimentally determined protein folding energy landscape. Proc. Natl. Acad. Sci. USA *101*, 14102–14107.

Miozzari, G.F., and Yanofsky, C. (1978). Translation of the leader region of the Escherichia coli tryptophan operon. J. Bacteriol. *133*, 1457–1466.

Muñoz, V., and Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. Nat. Struct. Biol. *1*, 399–409.

Muona, M., Aranko, A.S., Raulinaitis, V., and Iwaï, H. (2010). Segmental isotopic labeling of multi-domain and fusion proteins by protein trans-splicing in vivo and in vitro. Nat. Protoc. *5*, 574–587.

Nabuurs, S.B., Nederveen, A.J., Vranken, W., Doreleijers, J.F., Bonvin, A.M., Vuister, G.W., Vriend, G., and Spronk, C.A. (2004). DRESS: a database of REfined solution NMR structures. Proteins *55*, 483–486.

Otting, G., and Wüthrich, K. (1990). Heteronuclear filters in two-dimensional [^{1}H,^{1}H]-NMR spectroscopy: combined use with isotope labelling for studies of macromolecular conformation and intermolecular interactions. Q. Rev. Biophys. *23*, 39–96.

Parmeggiani, F., Pellarin, R., Larsen, A.P., Varadamsetty, G., Stumpp, M.T., Zerbe, O., Caflisch, A., and Plückthun, A. (2008). Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J. Mol. Biol. *376*, 1282–1304.

Richards, F.M. (1958). On the enzymatic activity of subtilisin-modified ribonuclease. Proc. Natl. Acad. Sci. USA *44*, 162–166.

Schmidt, E., and Güntert, P. (2012). A new algorithm for reliable and general NMR resonance assignment. J. Am. Chem. Soc. *134*, 12817–12829.

Sedgwick, S.G., and Smerdon, S.J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. Trends Biochem. Sci. *24*, 311–316.

Shekhawat, S.S., and Ghosh, I. (2011). Split-protein systems: beyond binary protein-protein interactions. Curr. Opin. Chem. Biol. *15*, 789–797.

Shen, Y., and Bax, A. (2013). Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J. Biomol. NMR *56*, 227–241.

Söding, J., and Lupas, A.N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. BioEssays *25*, 837–846.

Tewari, R., Bailes, E., Bunting, K.A., and Coates, J.C. (2010). Armadillo-repeat protein functions: questions for little creatures. Trends Cell Biol. *20*, 470–481.

Thirumalai, D., Klimov, D.K., and Dima, R.I. (2003). Emerging ideas on the molecular basis of protein and peptide aggregation. Curr. Opin. Struct. Biol. *13*, 146–159.

Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J. (2005). GROMACS: fast, flexible, and free. J. Comput. Chem. *26*, 1701–1718.

Varadamsetty, G., Tremmel, D., Hansen, S., Parmeggiani, F., and Plückthun, A. (2012). Designed Armadillo repeat proteins: library generation, characterization and selection of peptide binders with high specificity. J. Mol. Biol. *424*, 68–87.

Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J., and Laue, E.D. (2005). The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins *59*, 687–696.

Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K., and Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. J. Mol. Biol. *376*, 241–257.

Wetzel, S.K., Ewald, C., Settanni, G., Jurt, S., Plückthun, A., and Zerbe, O. (2010). Residue-resolved stability of full-consensus ankyrin repeat proteins probed by NMR. J. Mol. Biol. *402*, 241–258.

Wu, H., Hu, Z., and Liu, X.Q. (1998). Protein trans-splicing by a split intein encoded in a split DnaE gene of Synechocystis sp. PCC6803. Proc. Natl. Acad. Sci. USA *95*, 9226–9231.

Xu, W., and Kimelman, D. (2007). Mechanistic insights from structural studies of beta-catenin and its binding partners. J. Cell Sci. *120*, 3337–3344.

Yamazaki, T., Otomo, T., Oda, N., Kyogoku, Y., Uegaki, K., Ito, N., Ishino, Y., and Nakamura, H. (1998). Segmental isotope labeling for protein NMR using peptide splicing. J. Am. Chem. Soc. *120*, 5591–5592.

Zhang, O., Kay, L.E., Olivier, J.P., and Forman-Kay, J.D. (1994). Backbone 1H and 15N resonance assignments of the N-terminal SH3 domain of drk in folded and unfolded states using enhanced-sensitivity pulsed field gradient NMR techniques. J. Biomol. NMR *4*, 845–858.

Zimm, B.H., and Bragg, J.K. (1959). Theory of the phase transition between helix and random coil polypeptide chains. J. Chem. Phys. *31*, 526–535.