doi:10.1016/j.jmb.2007.12.014 J. Mol. Biol. (2008) 376, 1282–1304

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core

Fabio Parmeggiani[1], Riccardo Pellarin[1], Anders Peter Larsen[1], Gautham Varadamsetty[1], Michael T. Stumpp[1], Oliver Zerbe[2], Amedeo Caflisch[1] and Andreas Plückthun[1]*

[1]*Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

[2]*Department of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

Received 3 July 2007; received in revised form 13 November 2007; accepted 5 December 2007
Available online
14 December 2007

Edited by F. E. Cohen

Armadillo repeat proteins are abundant eukaryotic proteins involved in several cellular processes, including signaling, transport, and cytoskeletal regulation. They are characterized by an armadillo domain, composed of tandem armadillo repeats of approximately 42 amino acids, which mediates interactions with peptides or parts of proteins in extended conformation. The conserved binding mode of the peptide in extended form, observed for different targets, makes armadillo repeat proteins attractive candidates for the generation of modular peptide-binding scaffolds. Taking advantage of the large number of repeat sequences available, a consensus-based approach combined with a force field-based optimization of the hydrophobic core was used to derive soluble, highly expressed, stable, monomeric designed proteins with improved characteristics compared to natural armadillo proteins. These sequences constitute the starting point for the generation of designed armadillo repeat protein libraries for the selection of peptide binders, exploiting their modular structure and their conserved binding mode.

Keywords: consensus design; armadillo repeat; hydrophobic core; molten globule; molecular dynamics and minimization

Introduction

In recent years, as an alternative to raising monoclonal antibodies by immunization, recombinant antibodies[1] and an increasing number of other protein scaffolds[2] have been investigated as novel binding molecules. However, neither antibodies themselves nor any of these alternative protein scaffolds were specifically designed to bind peptides. Target-specific binding molecules are, in general, obtained from large protein libraries by *in vitro* selection or, in the case of monoclonal antibodies, through traditional immunization procedures. Both approaches require that, for each target, each new binding molecule is individually generated and characterized for specificity and

*Corresponding author. E-mail address: plueckthun@bioc.uzh.ch.

Present addresses: A.P. Larsen, Department of Biomedical Sciences, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen, Denmark; M.T. Stumpp, Molecular Partners AG, Grabenstrasse 11a, CH-8952 Schlieren, Switzerland.

Abbreviations used: αArm, Armadillo domain of human importin-α1 (residues 83–505); βArm, Armadillo domain of mouse β-catenin (residues 150–665); ANS, 1-anilino-naphthalene-8-sulfonate; C-type, overall consensus repeat; CD, circular dichroism; HA, hemagglutinin tag; HSQC, heteronuclear single quantum coherence; I-type, importin-derived consensus armadillo repeat; IMAC, immobilized metal-ion affinity chromatography; M-type, mutated armadillo repeat obtained by computational approach; MALS, multiangle light scattering; MRE, mean residue ellipticity; NLS, nuclear localization sequence; NOE, nuclear Overhauser enhancement; pD, phage lambda protein D; PDB, Protein Data Bank; SDS-PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis; SEC, size-exclusion chromatography; T-type, catenin/plakoglobin-derived consensus armadillo repeat.

cross-reactivity, making the generation of binders against a large number of peptide targets (e.g., representing a full proteome) an almost prohibitive task.

The aim of the present study was to develop a scaffold for the generation of peptide-specific binding proteins. In more detail, we wanted to develop proteins that were stable under various conditions and with the intrinsic ability to bind peptides in a conserved fashion. To recognize peptides in a sequence-selective manner the specificity of binding should ideally be conferred through specific interactions with the peptide side chains.

Natural peptide-binding scaffolds can be grouped in different classes. Antibodies are known to be able to bind peptides and have been well characterized.[3–6] Although peptide-binding antibodies have certain structural features in common, the mode of binding is not conserved. Thus, the information acquired through studies of antibody–peptide complexes cannot easily be applied to the general design of peptide-binding antibodies or extended to other proteins.

Small adaptor domains (e.g., SH2, SH3, and PDZ)[7] show specific binding to their targets, usually in a conserved binding mode within one family, but their affinity is generally low. The recognition sequence is very short and biased toward certain amino acid types, posttranslational modifications, or free N- or C-termini. While several such domains could be linked together by flexible peptides to recognize longer peptide sequences, a coverage of any arbitrary sequence would still be very difficult since these small domains might not be adaptable to the recognition of any arbitrary sequence. Furthermore, the entropy loss upon binding of such flexibly linked constructs would not necessarily lead to high affinities.

The major histocompatibility complex proteins (MHC I and MHC II)[8] possess a higher intrinsic variability and the ability to recognize a broad range of peptides, but the difficulties in their handling reduce their attractiveness as a scaffold candidate.

Repeat proteins, in particular tetratricopeptide repeats (TPRs),[9] armadillo,[10] and WD40[11] proteins, have been shown to possess an intrinsic ability to bind peptides, taking advantage of their repetitive structure. Thus, for our purpose, a scaffold based on repeat proteins seemed to constitute a promising candidate. For reasons outlined below, we chose the armadillo repeat protein family as the basis for our scaffold candidate.

Armadillo repeat proteins[12,13] are abundant in eukaryotes, where they are involved in a broad range of biological processes (e.g., transcription regulation,[14] cell adhesion,[15] tumor suppressor activity,[16] and nucleocytoplasmic transport[17]). These proteins are characterized by tandem repeats of approximately 42 amino acids that were first discovered in the product of the *Drosophila melanogaster* segmentation polarity gene Armadillo, which is homologous to mammalian β-catenin.[18,19] Armadillo repeat proteins participate in protein–protein interactions, and the armadillo domain is usually involved in the recognition process. The domain forms a right-handed superhelix[20,21] (Fig. 1a), as shown by the crystal structures of β-catenin[22] and importin-α.[23] Every repeat is composed of three α-helices, named H1, H2, and H3 (Fig. 1b), and several repeats stack to form the compact domain. Specialized repeats are present at the N- and C-termini of the protein, protecting the hydrophobic core from solvent exposure (Fig. 1a).

Armadillo repeat proteins are able to bind different types of peptides, yet relying on a conserved binding mode of the peptide backbone. Reported dissociation constant (K_d) values as low as 10–20 nM[24]

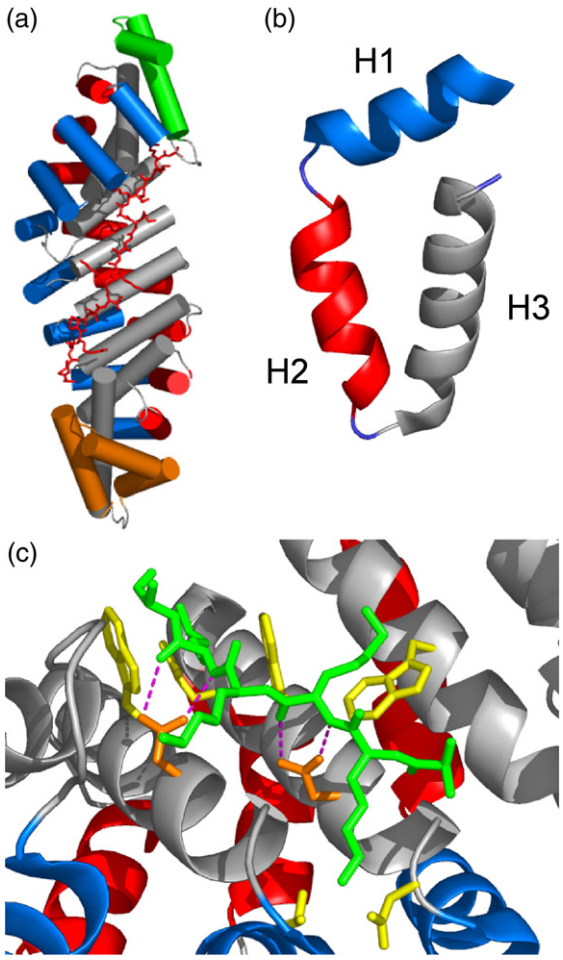


Fig. 1. (a) Structure of *S. cerevisiae* importin-α in complex with nucleoplasmin NLS (PDB ID 1EE5), showing the right-handed superhelical structure typical for armadillo repeat proteins. The cylinders represent the α-helices. The N-terminal repeat is indicated in green, and the C-terminal repeat is shown in orange. The bound peptide is depicted in red in a stick representation. (b) Detail of repeat 6 from 1EE5. The α-helices are represented as ribbons. (c) Detail of the peptide-binding mode. The conserved asparagine residues (in orange) contact *via* hydrogen bonds (purple) the backbone of the peptide, depicted in green. The residues that are responsible for the interactions with the side chains of the target peptide are shown in yellow. In all panels, helix 1 (H1) is indicated in blue, helix 2 (H2) in red, and helix 3 (H3) in gray.

indicate that high affinities can be achieved. Crystal structures of armadillo repeat proteins in complex with bound peptides have revealed that most peptide targets are bound in an extended conformation along the surface, inside the groove formed by the H3 helices. The superhelical armadillo domain winds around the peptide, oriented in the opposite N- to C-terminal direction (Fig. 1a), thus forming a double-helical complex, topologically similar to the DNA double strand. An asparagine residue, conserved in almost every repeat at the C-terminal part of H3, makes hydrogen bonds to the main chain of the target peptide, thereby keeping it in an extended conformation. Additional interactions to the target side chains are provided by neighboring residues, mostly in H3 (Fig. 1c). In a first approximation, each dipeptide unit of the target peptide is specifically recognized by one repeat in the armadillo domain (Fig. 2a).

In theory, the possibility of developing individual repeats that specifically bind a two-amino-acid sequence unit is very attractive. Given that the individual repeats are based on the same optimized scaffold and, thus, compatible with each other, any given number of repeats can be directly stacked to extend the recognition to much longer peptide sequences. In contrast to flexibly linked small adaptor domains mentioned above, armadillo repeats directly stack on each other in a rather rigid manner, allowing binding to uninterrupted longer peptides. This would exploit the specificity of the individual repeats to provide a peptide-binding designed armadillo protein with high and predetermined specificity, governed by the individual repeats. Such an approach (Fig. 2b), using armadillo proteins assembled from previously selected "building blocks," could effectively bypass the current *in vitro* selection procedures for individual peptides. However, this requires such individual peptide-specific repeats first to be developed, using a library-based approach.

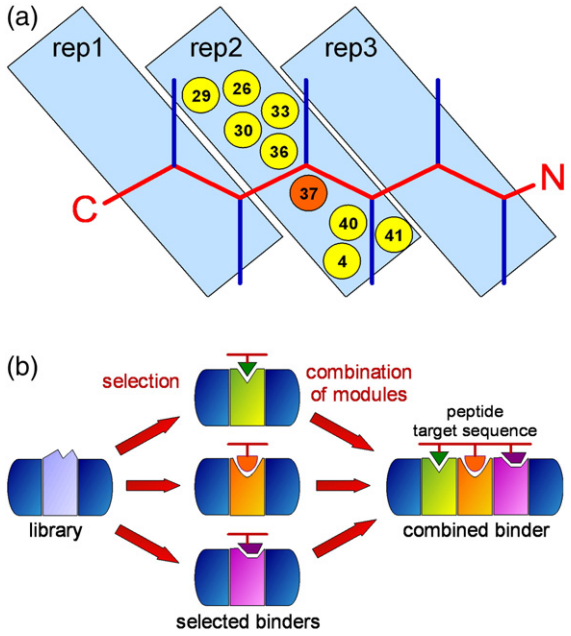


Fig. 2. Binding of target peptides. (a) Schematic drawing of an armadillo repeat protein binding to an extended peptide. The target peptide is bound in an antiparallel orientation to the protein. N and C indicate N- and C-termini of the peptide, which is depicted in red, with the amino acid side chains shown in blue. The residues of armadillo repeats involved in binding occupy specific positions within the single repeat sequences, mostly on helix 3. The position indicated in orange (a conserved Asn) is responsible for the binding of the peptide main chain; the positions in yellow are involved in recognition of the peptide side chains. (b) Designed armadillo repeat proteins potentially allow the selection of single repeats that specifically recognize short sequences. The selected peptide-specific repeats can be then combined to recognize longer peptides without performing additional selections.

In the present study, we have, as a first step, designed armadillo repeat modules based on consensus sequences. Proteins containing different types of modules have been assembled and characterized, initially only leading to stable dimeric proteins or monomeric molten-globule-like proteins. We subsequently used a combination of molecular dynamics and minimization to improve the hydrophobic core packing and convert the consensus-designed armadillo repeat protein with molten-globule-like properties to a monomeric, stable folded protein. Finally, the protein characteristics were evaluated for exploring the possibility of generating a modular peptide-binding scaffold. We succeeded in developing a stable, monomeric consensus protein that can be used now in the generation of peptide-specific individual armadillo repeat proteins.

Results

Armadillo repeat protein design

A consensus design strategy[25] has been applied in order to generate armadillo repeat proteins with high expression levels of soluble protein in *Escherichia coli*, monomeric state, high thermodynamic stability, and absence of cysteines for convenient expression and handling.

This design procedure was aimed at the generation of self-compatible repeat modules; therefore, consensus sequences were derived from multiple alignments of single armadillo repeats from the Swiss-Prot database.[26] A consensus design strategy has been successfully applied previously to other designed repeat proteins,[27–30] and it is based on consensus design of internal repeats (or internal modules). Special terminal capping repeats (terminal modules) have been generated to protect the hydrophobic core from solvent exposure. The crucial role of capping repeats has been previously shown in studies with designed ankyrin repeat proteins.[31]

The numbering used here to define the positions within the repeats was based on the family align-

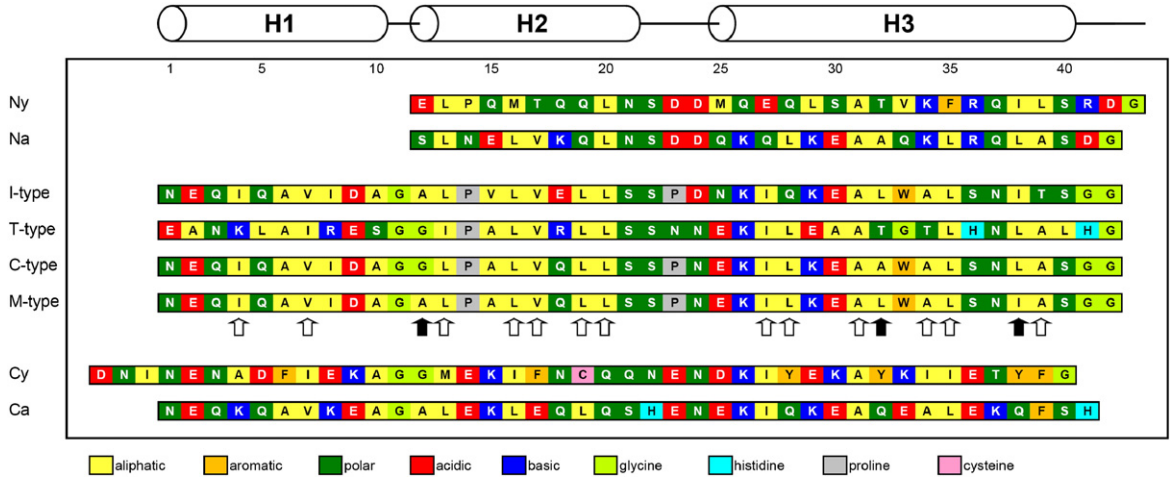


Fig. 3. Sequences of the designed internal modules and capping repeats; the cylinders indicate the helices, and the numbers denote the positions within the single repeats according to the convention introduced. *Ny* is an N-terminal capping repeat derived from importin-α from the *y*east *S. cerevisiae*; *Na* is an *a*rtificial N-terminal capping repeat. *I-type* is the internal module based on sequences from the *i*mportin-α subfamily, *T-type* is the internal module based on sequences from the β-ca*t*enin/plakoglobin subfamily, *C-type* is the overall *c*onsensus based on both subfamilies. *M-type* is the mutant sequence obtained through the computational approach described here. *Cy* is a C-terminal capping repeat derived from importin-α from the *y*east *S. cerevisiae*; *Ca* is an *a*rtificial C-terminal capping repeat. The amino acid color code is indicated below the sequences. The arrows indicate the positions considered in the computational approach. The filled arrows show the positions that differ between the C-type and the finally chosen M-type module.

ment proposed by Andrade *et al.*[32] Position −2 in that work corresponds to position 1 in the numbering used in the present study, where the putative helices H1, H2, and H3 encompass residues 1–10, 12–21, and 25–40, respectively.

Consensus design of internal modules

The initial sequence profile was generated using the family alignment from SMART†[33,34] (data from January 2004) as starting point (the consensus sequence is shown in Supplementary Fig. S1a). We largely followed the steps previously outlined.[27,30] We first removed all sequences lacking annotation in the Swiss-Prot database, especially hypothetical proteins or sequences for which no protein data were available with the exception of indirect evidence by sequence homology. The final set of 319 sequences led to a profile of 40 residues, covering the repeat sequence from H1 to H3 but excluding the loop between H3 and the next repeat. This sequence profile was used for a further search against the Swiss-Prot database. The repeats thus found belong to proteins that fall into different subfamilies of armadillo repeat proteins, as indicated by Andrade *et al.*[32]

Nevertheless, the sequences from different subfamilies might not be compatible. Taking this possibility into account, three final consensus sequences were constructed: one derived from β-catenin/plakoglobin (110 sequences), one from importin-α (133 sequences), and one from the combination of both (243 sequences). No normalization was applied during the calculation of the combined consensus sequence, which would compensate for a slight overrepresentation of importin-α over β-catenin/plakoglobin sequences in our selected set (133 over 110). The automatic alignments, performed with ClustalW[35] (Supplementary Fig. S1b), were manually refined including the loop connecting adjacent repeats (Supplementary Fig. S1c).

Structural information was taken into account to replace the cysteines present in the consensus sequences and reduce possible steric clashes. A more detailed description of additional sequence features and the rationale for amino acid exchanges are provided in the Supplementary Materials. Requirements for the cloning strategy were also considered at this stage, leading to the final module sequences type I (derived from *i*mportin-α subfamily), type T (derived from β-ca*t*enin/plakoglobin subfamily), and type C (*c*ombined consensus between these two subfamilies) (Fig. 3). Positions 7, 16, 17, 19, 20, 31, 34, 35, and 38 are well conserved in all the sequences and are part of the hydrophobic core of the armadillo proteins.

The positions potentially involved in binding of peptides (4, 26, 29, 30, 33, 36, 37, 40, and 41) have been defined based on the analysis of structures of complexes (summarized by Lange *et al.*[36] and Xu and Kimelman[37]) and data from mutation experiments.[38–40] The conserved Asn, responsible for binding to the main chain of the target peptide and at least in part for keeping it in an extended conformation, is located at position 37. Position 4 is part of both the hydrophobic core and the peptide binding site, and thus, the types of residues allowed at this position in a potential library would probably be restricted.

† http://smart.embl.de

Design of capping repeats

N- and C-terminal capping repeats, found in natural armadillo domains, protect the hydrophobic core, as they present a hydrophobic surface to the internal repeat side but a hydrophilic surface to the solvent. Capping sequences have also been considered in the previous design of other repeat proteins.[27,29–31]

The boundaries of armadillo domains have been estimated by limited proteolysis.[22,23] However, they are not clearly defined, partly due to the weak similarity of the terminal repeats to the internal ones. In addition, not all the residues are visible in the crystal structures of importin-α and β-catenin. It is likely that only the visible residues contribute to the armadillo domain, and the additional parts are unstructured and do not strictly belong to the domain. We have defined the N-terminal capping repeat as starting from position 12 (the beginning of H2). In contrast, the C-terminal capping repeat is completely resolved in the x-ray structures, and we defined it to comprise position 1 to position 41, thus including H1 to H3.

The capping repeats have been designed by using two different approaches. In the first, natural capping repeats were adapted to our designed internal repeats. Structural information to ensure compatibility between the capping repeats and the designed internal repeat is a fundamental prerequisite. The importin-α from *Saccharomyces cerevisiae* was considered to be the best candidate for a general capping repeat donor: all our designed modules present a flat surface that can interact with the inner surface of yeast importin-α capping repeats, as judged from molecular models. The *y*east importin-α-derived N-terminal and C-terminal capping repeats were named N*y* and C*y*, respectively.

The N-terminal capping repeat covers the residues from Glu88 to His119 of yeast importin-α. However, the two residues Glu118–His119 were replaced by Asp–Gly (Fig. 3, positions 42 and 43 of Ny) to adapt the terminal loop to the designed modules: glycine is used for assembly of the modules (as its codon overlaps a restriction site) and aspartate keeps a negative charge, which is frequently present at this position in natural proteins, reducing at the same time the helical propensity in the turn region.

The C-terminal capping repeat covers the region from Asn471 to Gly510 in yeast importin-α. However, the loop connecting the last internal repeat with the C-terminal repeat contains additional residues in yeast importin-α, compared to other natural importins. A modified version of this C-terminal capping repeat has thus been generated by introducing three residues (Asp-Asn-Ile) before H1 (Fig. 3, first three residues of Cy). Asn and Ile are naturally present at these positions; Asp has been included to keep a negatively charged loop as observed in several natural sequences while reducing the helical propensity.

In the second approach, two completely *a*rtificial N- and C-terminal capping repeats were designed (named N*a* and C*a*, respectively, and shown in Fig. 3), starting from the type C consensus and substituting the exposed hydrophobic residues with hydrophilic ones. Positions 12, 19, 27, and 34 of the N-terminal capping repeat are occupied by hydrophobic residues in the consensus sequence and were replaced by hydrophilic residues based on structures and frequently occurring residues obtained from alignment of N-terminal capping sequences. In a similar way, positions 8, 13, 17, 20, 28, 32, 35, 38, and 39 of the C-terminal capping repeat were replaced by hydrophilic residues based on structures and frequently occurring residues obtained from alignment of C-terminal capping sequences. A detailed description of the residues introduced in the designed capping repeats is provided in the Supplementary Materials. A second version of the Cy capping repeat was also designed without the three initial residues and with Ala replacing Cys at position 19; no change was observed, compared to Cy, in the level of expression and in the amount of soluble protein, and it will thus not be discussed further.

Assembly, cloning, and expression of designed armadillo repeat proteins

The amino acid sequences of all modules were back-translated to DNA sequences, optimizing the codon usage for expression in *E. coli*. Each module was synthesized, starting from overlapping oligonucleotides (Supplementary Table S1).

The modules were assembled stepwise using type IIS restriction enzymes (Supplementary Fig. S2), following the approach reported by Binz *et al.*[27] The final proteins were named according to the modules that they contain: the name indicates, in order, the type of N-terminal repeat (A for *A*rtificial or N*a*, Y for *Y*east derived or N*y*), the type of internal repeats (type I, type T, or type C) with the number of modules used as a subscript, and the type of C-terminal repeat (A for *A*rtificial or C*a*, Y for *Y*east derived or C*y*): for example, YI_4A contains a *Y*east-derived N-terminal repeat (N*y*), four internal repeats based on *I*mportin consensus (type *I*), and an *A*rtificial C-terminal repeat (C*a*).

Thus, Na or Ny as N-capping modules were combined with T-, I-, or C-type internal modules and Ca or Cy C-terminal modules, leading to 12 possible combinations. The proteins contain only one type of internal module to avoid incompatible surfaces at the interface between repeats. The influence of capping and internal repeats was evaluated by analyzing the expression properties of all the constructs, containing two or four internal repeats. The proteins were expressed in *E. coli* XL1-blue using a pQE30-based expression plasmid, providing an N-terminal $MRGSH_6$ tag for purification. The insert was constructed with a double stop codon (Supplementary Fig. S3). As an example, the DNA and protein sequences of YC_2A are provided in Supplementary Fig. S3.

The highest level of soluble protein expression was obtained when the internal modules were combined with Ny and Ca (Fig. 4a). The Na cap leads to almost undetectable expression in Coomassie-stained polyacrylamide gels, and the presence of Cy resulted in a substantial portion of the protein found in the insoluble fraction after cell lysis. The observed effects of terminal capping repeats were independent of the type and the number of internal modules. However, increasing the number of internal modules enhanced the amount of soluble protein and the absolute amount of protein produced. Remarkably, type T proteins are characterized by a lower apparent mobility in sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE), compared to type I and type C proteins (Fig. 4b).

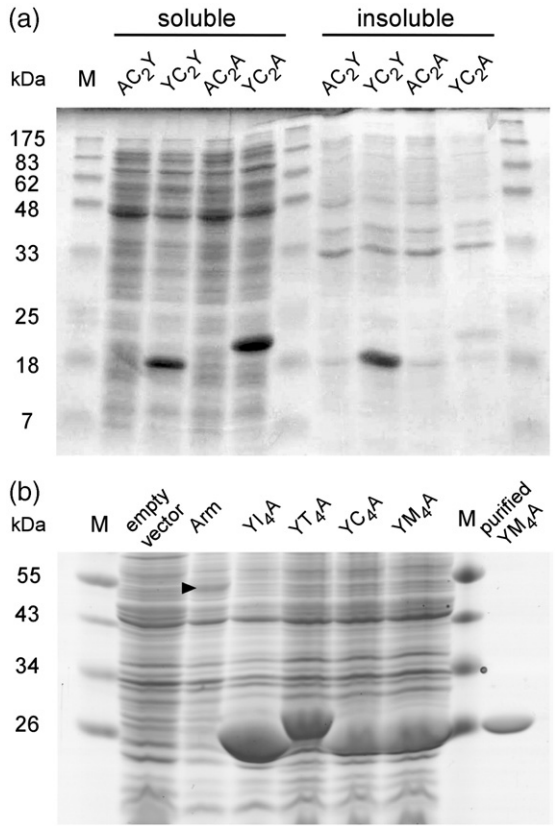


Fig. 4. (a) Influence of capping repeats on expression. Soluble and insoluble fractions of *E. coli* cell extracts are shown in a Coomassie-stained SDS polyacrylamide gel. The proteins contain two internal C-type modules with different combinations of capping repeats. (b) Whole-cell extracts of consensus proteins. The constructs contain Ny and Ca as capping repeats. Cells transformed with the empty vector or with the vector containing the armadillo domain of mouse β-catenin (Arm) were used as control. The proteins can be easily purified in a single step by IMAC, as shown for YM_4A. The expected size is 18 and 27 kDa for proteins containing two or four internal modules, respectively, and 56 kDa for Arm. The triangle indicates the band corresponding to the armadillo domain of β-catenin, which is expressed at much lower yield than the designed proteins. The molecular mass of the marker (M) is indicated in kilodaltons on the left.

Protein purification and characterization: Comparison with natural armadillo domains

Proteins containing the combination of Ny, Ca, and two, four, or eight internal repeats have been chosen for biophysical characterization and evaluation of the properties of type I, type T, and type C modules. The results are summarized in Table 1.

The purification by immobilized metal-ion affinity chromatography (IMAC) in a single step provided up to 100 mg of pure protein from 1 l of bacterial culture (Fig. 4b). No sign of precipitation or degradation was detected by spectrophotometry and SDS-PAGE in protein solutions stored for up to 1 month at 4 °C in the IMAC elution buffer.

The natural human importin-α1 (Swiss-Prot P52294) and the mouse β-catenin (Swiss-Prot Q02248) were also expressed using the same pQE30-based plasmid. The importin contains 10 armadillo repeats and the catenin 12, including the capping repeats in the count in both cases. Human importin-α1 gave the highest yield among the importin-α family members tested (data not shown) after a two-step coupled IMAC–ion exchange purification and, together with mouse β-catenin, was used for the comparison with the designed armadillo repeat proteins.

Importin-α1 and β-catenin, despite their elongated shape, elute at the volume expected from their molecular weight in gel filtration on a Superdex 200 column, and the monomeric state was confirmed for both proteins by multiangle light scattering (MALS) measurements (Table 1).

On the other hand, the designed proteins show elution volumes corresponding to higher-than-expected apparent molecular masses in size-exclusion chromatography (SEC) (Table 1). MALS indicates that the I- and T-type proteins are probably present as a mixture of dimers and monomers in solution. The main peak (Fig. 5a) corresponds to the dimeric form, and this value is reported in Table 1. At high concentration (2–4 mg/ml), I- and T-type proteins are present as a mixture of oligomers. In contrast, monomeric and oligomeric fractions of C-type proteins YC_4A (Fig. 5a) and YC_8A (data not shown) can be separated, up to the highest concentration tested (4 mg/ml). However, the fractions of YC_4A and YC_8A, shown by MALS to be monomeric, elute earlier than expected for proteins of comparable size. The smaller YC_2A represents the only exception: independent of the concentration, the MALS-calculated mass values are always intermediate between monomer and dimer. A decrease in pH to 7 favors the formation of oligomeric species of I- and T-type proteins. C-type proteins are, in contrast, unaffected by pH (data not shown).

The circular dichroism (CD) spectra (Fig. 5b) indicate the presence of significant α-helical secondary structure content for all proteins, particularly for the I-type proteins. For I- and C-type consensus repeats, the absolute value of mean residue ellipticity (MRE) and the helical content generally increase

Table 1. Biophysical properties of designed and natural armadillo repeat proteins

Construct	Residues (repeats)[a]	p*I*[b]	MW_{calc} (kDa)[b]	Oligomeric state[c]	MW_{obs} (kDa)[d]	$MW_{obs/calc}$[e]	CD_{222} (MRE)[f]	Helical content (%)[g]	Observed T_m (°C)[h]
YI_2A	169 (4)	5.2	18.6	Dimer	64.6	1.7	−13,000	63	~55
YI_4A	253 (6)	4.8	27.4	Dimer	116.1	2.1	−19,500	80	~69
YI_8A	421 (10)	4.6	44.9	Dimer	148.8	1.7	−22,600	85	>85
YT_2A	169 (4)	6.3	18.6	Dimer	141.2	3.8	−7100	23	~56
YT_4A	253 (6)	6.5	27.3	Dimer	219.6	4.0	−10,100	40	~75
YT_8A	421 (10)	6.7	44.8	Dimer	229.7	2.6	−9400	35	~83
YC_2A	169 (4)	5.4	18.4	Mixture	59.1	n.d.	−9100	45	n.d.
YC_4A	253 (6)	5.1	26.9	Monomer	50.0	1.9	−12,100	49	n.d.
YC_8A	421 (10)	4.8	44.0	Monomer	76.7	1.7	−20,000	62	n.d.
YM_4A	253 (6)	5.1	27.1	Monomer	32.2	1.2	−18,800	87	~70
αArm[i]	435 (10)	5.5	48.2	Monomer	42.9	0.9	−14,300	54	~43
βArm[j]	528 (12)	8.7	57.6	Monomer	52.6	0.9	−16,800	60	~58

n.d. indicates that the value has not been determined due to either an inhomogeneous sample (oligomeric state of YC_2A) or lack of cooperative transition in thermal denaturation (YC_2A, YC_4A, and YC_8A).

[a] The number of residues includes the $MRGSH_6$ tag; the number of repeats includes capping repeats.

[b] p*I* and molecular weight calculated from the sequence; masses were confirmed by mass spectrometry.

[c] Oligomeric state as indicated by multiangle static light scattering.

[d] Observed molecular weight as determined in SEC.

[e] Ratio between observed and calculated molecular weight, taking into account the oligomeric state (Os): $MW_{obs/calc} = MW_{obs}/(Os \cdot MW_{calc})$.

[f] Mean residue ellipticity at 222 nm expressed as $deg \cdot cm^2/dmol$.

[g] Helical content estimated with the program CDpro.[41]

[h] T_m observed in thermal denaturation by CD.

[i] Armadillo domain of human importin-α1.

[j] Armadillo domain of mouse β-catenin.

with the number of internal repeats; in contrast, the helical content is almost constant for T-type proteins (Supplementary Fig. S4). The values of helical content were calculated using the program CDpro[41] and are indicated in Table 1.

The CD signal at 222 nm was chosen to monitor stability against thermal and denaturant-induced unfolding. I- and T-type proteins show a cooperative transition, while no transition was observed in C-type proteins (Fig. 5c). The midpoint of transition during thermal denaturation (T_m) increases with the number of repeats, for example, from approximately 70 °C for YI_4A to more than 80 °C for YI_8A (Table 1). Importin-α1 and β-catenin, containing 8 and 10

Fig. 5. Biophysical characterization of designed and natural armadillo repeat proteins. (a) SEC and MALS of designed armadillo repeat proteins containing four internal modules and of importin-α1. YI_4A, YT_4A, and YC_4A show apparent molecular weights higher than the globular proteins with the same calculated mass (about 27 kDa). The broad peaks shown by YI_4A and YT_4A are due to a mixture of dimers and monomers, as indicated by the molecular mass determined by light scattering. The highest point of the peak corresponds to the dimeric fraction. In the case of YC_4A, the first peak eluted contains probably a mixture of oligomers with high molecular masses. The monomeric peak after separation remains monomeric and was further characterized. The importin-α1 (αArm) is a monomer as indicated by LS and elutes at the expected volume. The data were obtained with a Superdex 200 column. The elution was followed by absorbance at 280 nm for YC_4A, YI_4A and αArm; YT_4A does not possess any residue absorbing significantly at 280 nm; thus, the elution was followed at 230 nm. V_0 indicates the void volume of the column. Alcohol dehydrogenase (ADH; MW = 150 kDa), bovine serum albumin (BSA; MW = 66 kDa), carbonic anhydrase (CA; MW = 29 kDa), and aprotinin (Apr; MW = 6.5 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by the arrows. (b) CD spectra of I-type, T-type, and C-type proteins containing four internal modules. The natural armadillo domains of human importin-α1 (αArm) and mouse β-catenin (βArm) are indicated by open and filled circles, respectively. The values are reported as MRE. (c) Thermal denaturation curves. A comparison between designed armadillo repeat proteins containing four or eight internal modules is shown, from the top, for I-type, T-type, and C-type proteins. αArm and βArm are displayed in the bottom panel. The denaturation was followed by CD. The MRE at 222 nm is reported. (d) Thermal denaturation and renaturation of designed armadillo repeat proteins. From the top, YI_4A, YT_4A, and YC_4A are shown. For comparison, the bottom graph shows the irreversible denaturation of αArm. βArm shows a similar irreversible denaturation (data not shown). The denaturation was followed by CD. The values of MRE at 222 nm were normalized by setting the initial and the final values of the denaturation curves as 0 and 1, respectively. (e) Guanidinium-chloride-induced denaturation of armadillo repeat proteins containing eight internal modules. Comparison of YI_8A, YT_8A, and YC_8A with αArm. The denaturation was followed by CD. The values of MRE at 222 nm were normalized by setting the initial and the final values of the denaturation curves as 0 and 1, respectively. (f) Emission spectra of ANS in the presence of designed armadillo repeat proteins. YI_4A, YT_4A, and YC_4A are compared to αArm and βArm. I- and T-type proteins show fluorescence levels in the same range as natural proteins; in contrast, the fluorescence emission for C-type proteins is significantly higher and increases with the number of repeats. The values without buffer subtractions are shown. αArm was measured in a separate experiment and scaled according to the values of YC_4A present in both sets of experiments. Similar results were obtained with proteins containing two or eight internal repeats.

internal repeats, respectively, have lower midpoints of transition, even when compared with designed proteins with only 4 internal repeats (Table 1). It should be noted that the designed proteins retain a significant percentage of secondary structure at 95 °C and that the thermal unfolding is almost

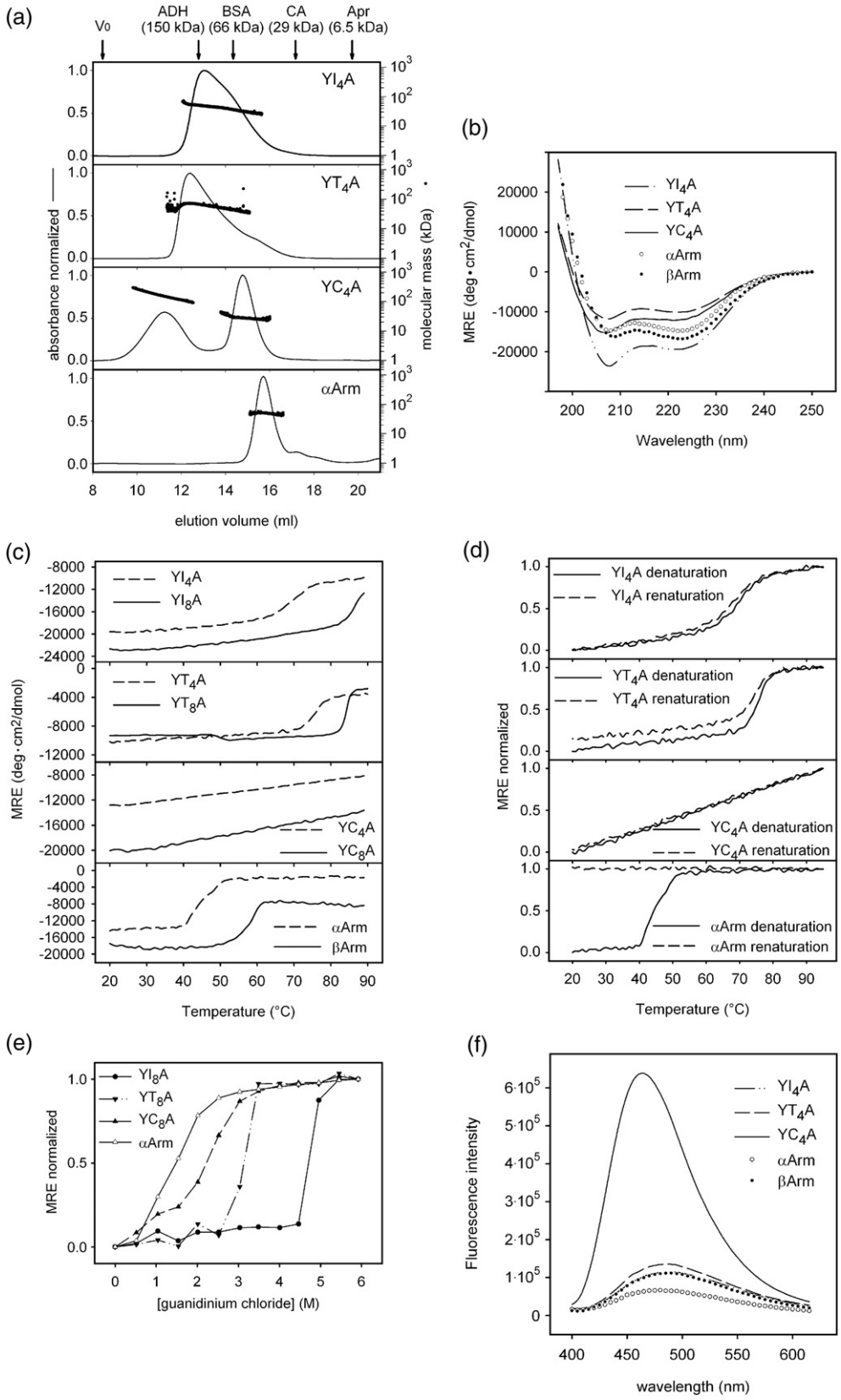


Fig. 5 (*legend on previous page*)

completely reversible, in contrast to natural armadillo proteins that cannot refold after thermal unfolding (Fig. 5d); YT_8A is the only designed armadillo repeat protein whose thermal unfolding is irreversible (data not shown).

We also investigated unfolding induced by guanidinium chloride. A direct comparison between natural and designed armadillo repeat proteins composed of 10 repeats (Fig. 5e) reveals for importin-α1 (αArm), with a midpoint of transition of 1.4 M guanidinium chloride, a lower stability than that for YI_8A and YT_8A, with approximately 4.8 and 3.2 M as midpoints of transition, respectively. YC_8A shows a gradual loss of secondary structure, especially at low concentrations of denaturant, apparently similar to αArm. Data from urea-induced unfolding experiments confirm the gradual loss of secondary structure for C-type proteins with increasing denaturant concentration. Natural armadillo domains show a stable pretransition baseline in unfolding induced by the weaker denaturant urea (data not shown).

The three types of consensus proteins (C-, I-, and T- type) also show a different behavior in 1-anilino-naphthalene-8-sulfonate (ANS) binding experiments. ANS is a fluorescent dye sensitive to the hydrophobic environment.[42] C-type proteins bind ANS strongly, suggesting the presence of an accessible hydrophobic core, while I- and T-type proteins show ANS binding in the same low range as the natural armadillo repeat proteins (Fig. 5f).

Thermal and guanidinium-induced denaturation and ANS results indicate that I- and T-type proteins share many characteristics with proteins with stable folds. Based on MALS data, however, I-type proteins are mainly present as dimers. T-type proteins show even higher deviations of elution behavior in SEC, and remarkably, the helical content does not seem to be significantly affected by the number of internal repeats, in contrast to the T_m value. C-type proteins, though monomeric, are characterized by strong ANS binding, an elution volume smaller than expected for a monomeric protein in SEC, and lack of cooperativity in thermal and chemical denaturation. These features, similar to some extent to the properties of molten globules,[43] indicate that the C-type proteins are probably not folded in a well-packed conformation, even though the expected secondary structure is detected by CD. Nonetheless, we chose the C-type proteins as the basis for our further investigations.

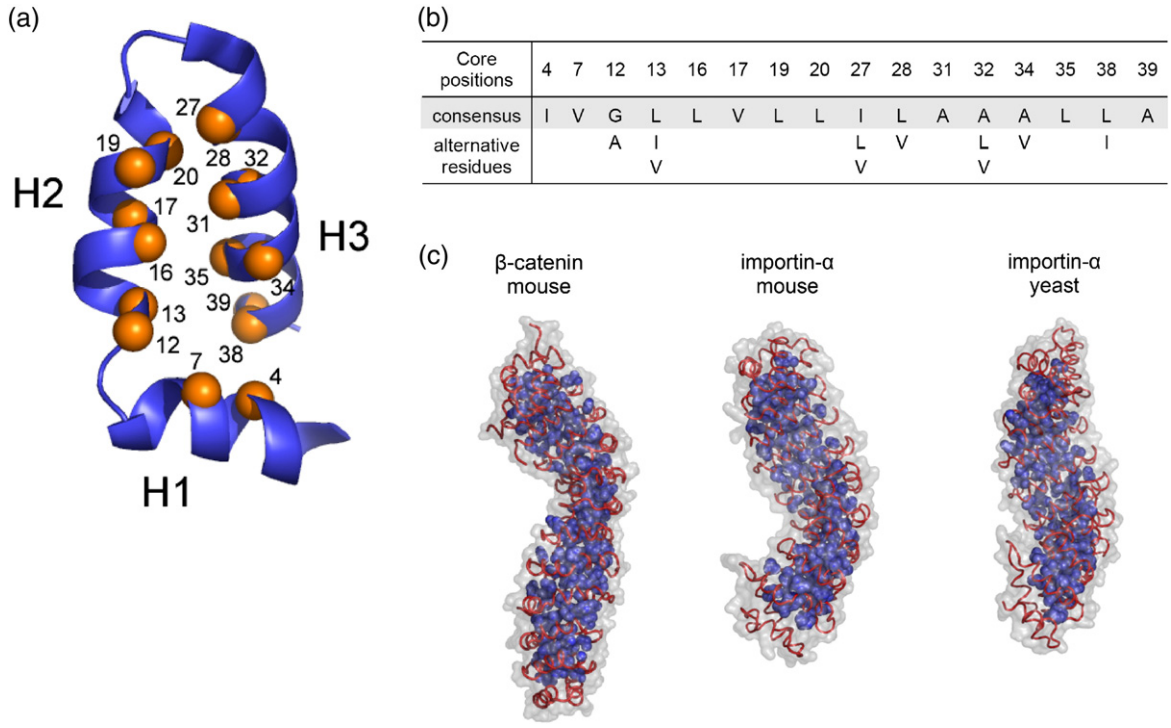
Consensus design improvement: Substitutions in the hydrophobic core

Due to the lack of conserved interrepeat hydrogen bonds and salt bridges, the tertiary structure of natural armadillo repeat proteins holds together mainly through nonpolar interactions. If the packing is not ideal, alternative conformations may become accessible. As a consequence, the molten-globule-like features of C-type proteins could be due to nonoptimal packing of the hydrophobic core.

The modular architecture of designed armadillo repeat proteins suggests that the computational search for a sequence leading to stable packing of the hydrophobic core might be achievable by considering a single repeat. However, the repeat can assume its correct conformation only in the context of a complete protein. It was, therefore, necessary to use the known structures of natural armadillo domains (comprising 400 to 500 residues) as templates for the sequence search.

The use of available algorithms (self-consistent mean field, dead-end elimination, genetic algorithm, and Monte Carlo search) for structures as large as armadillo domains has so far not been reported, despite recent achievements (reviewed by Butterfoss and Kuhlman[44]); such approaches would be, however, seriously compromised by the computational load and probably not even be possible in the case of dead-end elimination, as suggested by Voigt *et al.*[45] Therefore, we used here a different approach to treat a system of such size: information from sequence alignments was used to reduce the complexity in terms of variable positions and allowed residue types. The selected mutants were ranked according to energy values obtained by rotamer sampling. The method allows, in a simple way, to identify a number of hydrophobic core mutant sequences, which are likely to represent an improvement of the original C-type sequence.

The 16 positions contributing to the hydrophobic core in each repeat (Figs. 3 and 6a) were defined by having a solvent-accessible surface corresponding to less than 5% of the total residue surface, as determined by a probe with 1.4 Å radius. The final choice was made after visual inspection of the structures. The number of mutations was restricted to the most frequently occurring aliphatic amino acids at each position, based on the sequence alignment, while keeping the most conserved positions constant. Using these criteria, only 7 positions out of the 16 forming the hydrophobic core of a single repeat were allowed to vary and to host two or three different residue types (Fig. 6b). Mutants were modeled starting from three different backbones to average the influence of single structures out. Therefore, the structures of three different proteins {mouse β-catenin [Protein Data Bank (PDB) ID 2BCT],[22] yeast importin-α (PDB ID 1EE4),[46] and mouse importin-α (PDB ID 1Q1T[47])} were chosen to generate all the mutants (Fig. 6c). Model structures were constructed by substituting the core positions of every internal repeat with either the residues present in the C-type consensus or the aforementioned mutations (Fig. 6b). The initial rotamer conformations were randomly assigned. The noncore residues of the original structures were kept. In each structure, every repeat of the protein carries the same mutations. Structures corresponding to all the 432 combinations of allowed mutations, including also the set of residues of the original



(b)

Core positions	4	7	12	13	16	17	19	20	27	28	31	32	34	35	38	39
consensus	I	V	G	L	L	V	L	L	I	L	A	A	A	L	L	A
alternative residues			A	I V					L V	V		L V	V		I	

Fig. 6. (a) Hydrophobic core positions in a single repeat are indicated with orange spheres and the corresponding numbers. (b) Amino acids, in a single-letter code, allowed at the core positions during the calculations. The original amino acids present in the type C consensus are highlighted in gray. The total number of different combinations in each repeat is 432 ($2^4 \times 3^3$). The number of mutants is also 432 because the same mutation pattern was applied to all repeats in each protein. (c) Armadillo domains used as starting structures for the models of the mutants: murine β-catenin (PDB ID 2BCT) and importin-α from mouse (PDB ID 1Q1T) and *S. cerevisiae* (PDB ID 1EE4). The backbone trace is shown in red, and the protein surface is indicated in gray. The side chains belonging to the hydrophobic core residues, which correspond to the parts allowed to move freely during the simulation, are depicted in blue.

C-type consensus, were generated and subjected to energy minimization.

A sequence of heating–quench cycles (Fig. 7), followed by energy minimization, resulted in a series of structures and corresponding energy values that were used to generate the final ranking of the mutants (Supplementary Table S3). A detailed description of the rotamer sampling procedure is provided in Materials and Methods. Mutants with a hydrophobic core volume lower than the original consensus, calculated with values reported by Chothia,[48] were not included in the final ranking to reduce the number of false positives that might arise due to underpacking of the core (see Discussion).

Gene assembly, expression, and characterization of selected hydrophobic core mutants

Among the 30 top-ranked single repeat mutant sequences, 18 were selected for experimental validation. The best-ranking mutant sequence with low core volume was also selected to challenge the initial choice of a core volume filter during the ranking process (Table 2 and Supplementary Table S3). The influence of mutant repeats on the protein properties was experimentally evaluated in the format of proteins containing four identical internal repeats and Ny and Ca as capping repeats (Fig. 3). The original reference consensus sequence is thus YC_4A. The proteins were named with a progressive number, from mut1 to mut18; mut19 contains the sequence with low core volume.

The assembly of single repeats from oligonucleotides and the stepwise ligations were performed as described above, and the proteins were expressed and purified by IMAC in a single step with yields comparable to those obtained for YC_4A, that is, up to 100 mg/l of bacterial culture.

The experimental comparison was carried out by using CD, SEC, and binding of ANS. All the mutants share a similar CD spectrum with the original consensus but are characterized by a general increase in MRE at 222 nm, indicating a higher percentage of α-helical secondary structure. The increased elution volume of the mutants indicates a higher compactness of the proteins (Fig. 8) and correlates well with a decreased ANS binding. The mutant mut1, being a dimer, represents the only outlier, while all the other mutants are monomers, as indicated by MALS. Some of the core mutants carry additional mutations (indicated in Table 2), which were unintentionally introduced during the gene synthesis. Most of these mutations are located in the loops or at the surface of the helices and, thus, have probably only a small

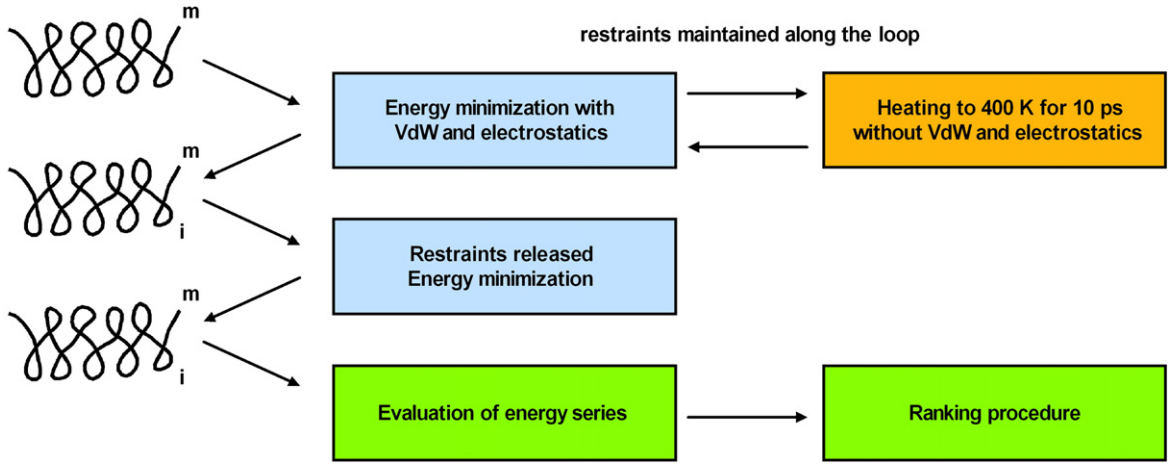


Fig. 7. Schematic diagram of the computational procedure for the evaluation of the hydrophobic core mutants. *m* indicates a particular mutant, and *i* is 1 of the 100 conformations of the mutant *m* obtained after each minimization step in the recursive sampling procedure. VdW, van der Waals interactions.

influence, if any, on the stability of the hydrophobic core; furthermore, they are present only in a single repeat out of four, reducing their overall contribution to protein properties.

Mutants mut2, mut3, mut4, mut7, mut11, mut12, and mut13 showed the best combination of low ANS binding and compactness, as judged by SEC, and were thus selected for further characterization by thermal denaturation. The mutant mut7 shows a significantly increased cooperativity during unfolding, compared to YC_4A and the other mutants (Supplementary Fig. S5).

The internal module corresponding to mut7, which was named M-type, contains three point mutations compared to the initial consensus sequence (Fig. 3). The mutant protein mut7, renamed YM_4A, is a stable monomer at several salt and protein concentrations, such as YC_4A; however, dimer formation of YM_4A was observed at pH 7 at high protein concentrations (5 mg/ml). No sign of precipitation or degradation was detected in protein solutions stored for up to 1 month at 4 °C in the IMAC elution buffer. The values for the biophysical properties examined are reported in Table 1.

The direct comparison of YC_4A and YM_4A is shown in Fig. 9. The [^{15}N,^{1}H]-heteronuclear single quantum coherence (HSQC) NMR spectra of YM_4A were recorded at pH 7, 8, 9, 10, and 11. YC_4A spectra

Table 2. Hydrophobic core of the selected mutants

	Hydrophobic core residues															
	4	7	**12**	**13**	16	17	19	20	**27**	**28**	31	**32**	**34**	35	**38**	39
C-type	I	V	**G**	**L**	L	V	L	L	**I**	**L**	A	**A**	**A**	L	**L**	A
mut1	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut2	–	–	**A**	**–**	–	–	–	–	**V**	**–**	–	**L**	**–**	–	**–**	–
mut3	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut4	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**–**	**–**	–	**I**	–
mut5	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**–**	**V**	–	**–**	–
mut6	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**–**	**–**	–	**–**	–
mut7	–	–	**A**	**–**	–	–	–	–	**–**	**–**	–	**L**	**–**	–	**I**	–
mut8	–	–	**A**	**V**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut9	–	–	**A**	**I**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**–**	–
mut10	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**V**	**–**	–	**–**	–
mut11	–	–	**A**	**–**	–	–	–	–	**V**	**–**	–	**L**	**V**	–	**–**	–
mut12	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**L**	**V**	–	**–**	–
mut13	–	–	**A**	**I**	–	–	–	–	**L**	**–**	–	**L**	**V**	–	**–**	–
mut14	–	–	**A**	**–**	–	–	–	–	**L**	**–**	–	**V**	**V**	–	**–**	–
mut15	–	–	**A**	**–**	–	–	–	–	**–**	**–**	–	**–**	**–**	–	**–**	–
mut16	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**L**	**–**	–	**I**	–
mut17	–	–	**A**	**V**	–	–	–	–	**L**	**–**	–	**L**	**V**	–	**I**	–
mut18	–	–	**A**	**I**	–	–	–	–	**–**	**–**	–	**–**	**–**	–	**–**	–
mut19	–	–	**–**	**–**	–	–	–	–	**L**	**–**	–	**–**	**–**	–	**–**	–
I-type	–	–	**A**	**–**	–	–	–	–	**–**	**Q**	–	**L**	**–**	–	**I**	T

The numbers indicate the positions in the single repeat (cf., Fig. 3). The hydrophobic core positions subjected to mutation (12, 13, 27, 28, 32, 34, and 38) are indicated in boldface. The amino acids present at each position are reported as single-letter code. "–" indicates no difference with respect to C-type consensus. As a comparison, in the last row, the sequence corresponding to the I-type consensus is shown. An Ala→Thr mutation occurs in mut6 at position 15 in repeat 3, in mut9 at position 31 in repeat 4, in mut10 at position 12 in repeat 1, and in mut17 at position 15 in repeat 1. mut8 has a mutation Gly→Val at position 42 in repeat 4.

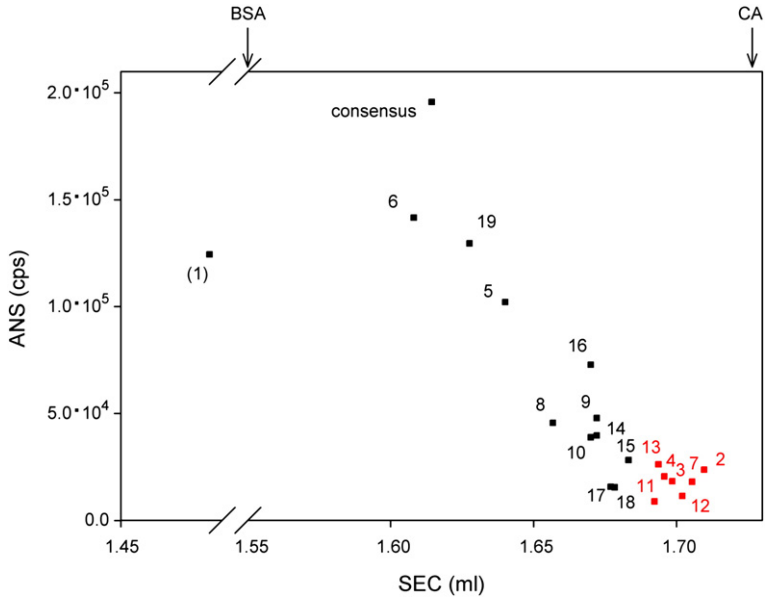


Fig. 8. Experimental evaluation of hydrophobic core mutants: elution volumes in SEC and fluorescence emission upon ANS binding. The numbers refer to the mutants reported in Table 1. *Consensus* indicates the protein containing four C-type internal repeats (YC_4A). All the proteins have a molecular mass of approximately 27 kDa. *mut1* (in parentheses) elutes before the consensus and the other mutants because of its dimeric state. All other mutants were shown to be monomeric by MALS. Peak values from absorbance at 280 nm in SEC and from fluorescence intensity are plotted. Errors in the measurements have been estimated with a subset of six proteins and two different preparations, leading to an average standard deviation of 0.01 ml for SEC and an average percentage error of 4% for ANS fluorescence intensity. As reference, carbonic anhydrase (CA; MW = 29 kDa) and bovine serum albumin (BSA; MW = 66 kDa) elute at 1.73 and 1.55 ml, respectively. The mutants depicted in red were selected for further characterization.

were collected at pH 6, 7, and 8. An increase in pH increases the line broadening of YC_4A but decreases it for YM_4A. Nevertheless, the overall dispersion is conserved for each protein at different pH values (data not shown). The YM_4A spectrum recorded at pH 11 and the YC_4A spectrum recorded at pH 6 are shown in Fig. 10. Amide proton frequencies of YC_4A are generally limited to the random-coil range (7.5–8.5 ppm), whereas many cross peaks of YM_4A are located outside this range. Moreover, the line widths from signals of YC_4A are slightly larger than those from signals of YM_4A. Increased line widths due to conformational exchange processes as well as limited signal dispersion are characteristic features of molten globule states of proteins.[49,50] Although no attempts have been made to assign the $^{15}N,^{1}H$ correlation map, $^{15}N\{^{1}H\}$-nuclear Overhauser enhancement (NOE) data were recorded to characterize internal backbone dynamics[51] and to probe for increased rigidity of YM_4A (data not shown). All detected amide moieties of YM_4A are characterized by $^{15}N\{^{1}H\}$-NOEs larger than 0.6, indicating well-folded segments, whereas for YC_4A, all the values are smaller than 0.3, many of which have negative NOEs, indicating a large flexibility. Thus, the NMR measurements confirm the molten-globule-like characteristics of YC_4A and the folded state properties of YM_4A.

Binding assay as functionality test

YM_4A and YC_4A share with natural importins a considerable number of residues critical for binding to nuclear localization sequences (NLSs), which are the natural ligands of importin-α proteins. Therefore, the designed proteins might retain some binding properties toward NLS. The NLS from the SV40 large T antigen[52] is considered a prototype sequence: it has been extensively studied in the literature and constitutes the reference point for the evaluation of NLS binding.[47]

The NLS from SV40 large T antigen (SPKKKRKVE) was expressed as a fusion protein with phage lambda protein D (pD), biotinylated, and immobilized on NeutrAvidin-coated plates. Being of similar size, the hemagglutinin tag (YPYDVPDYA, here referred to as HA), also fused to protein D, was used as a negative control. ELISA experiments (Fig. 11) reveal that both YM_4A and YC_4A bind specifically to the NLS and that the binding can be competed by a free NLS peptide in solution. However, the unspecific binding of YM_4A to HA and NeutrAvidin is reduced in comparison to YC_4A.

In summary, even though the high concentrations of protein and competing peptide indicate a rather weak affinity, YM_4A was able to specifically recognize the same target as the natural armadillo repeat proteins and to reduce the unspecific binding observed for YC_4A, further validating the design process.

Discussion

Consensus design

Consensus design has been successfully applied in this work to generate designed armadillo repeat proteins. Similar to leucine-rich repeat proteins,[30] but in contrast to ankyrin repeat proteins[27] and tetratricopeptide repeats,[29] different subfamilies can be clearly defined in the case of armadillo repeat proteins, based on sequences and available structures. Out of 42 signature positions, 12 are char-

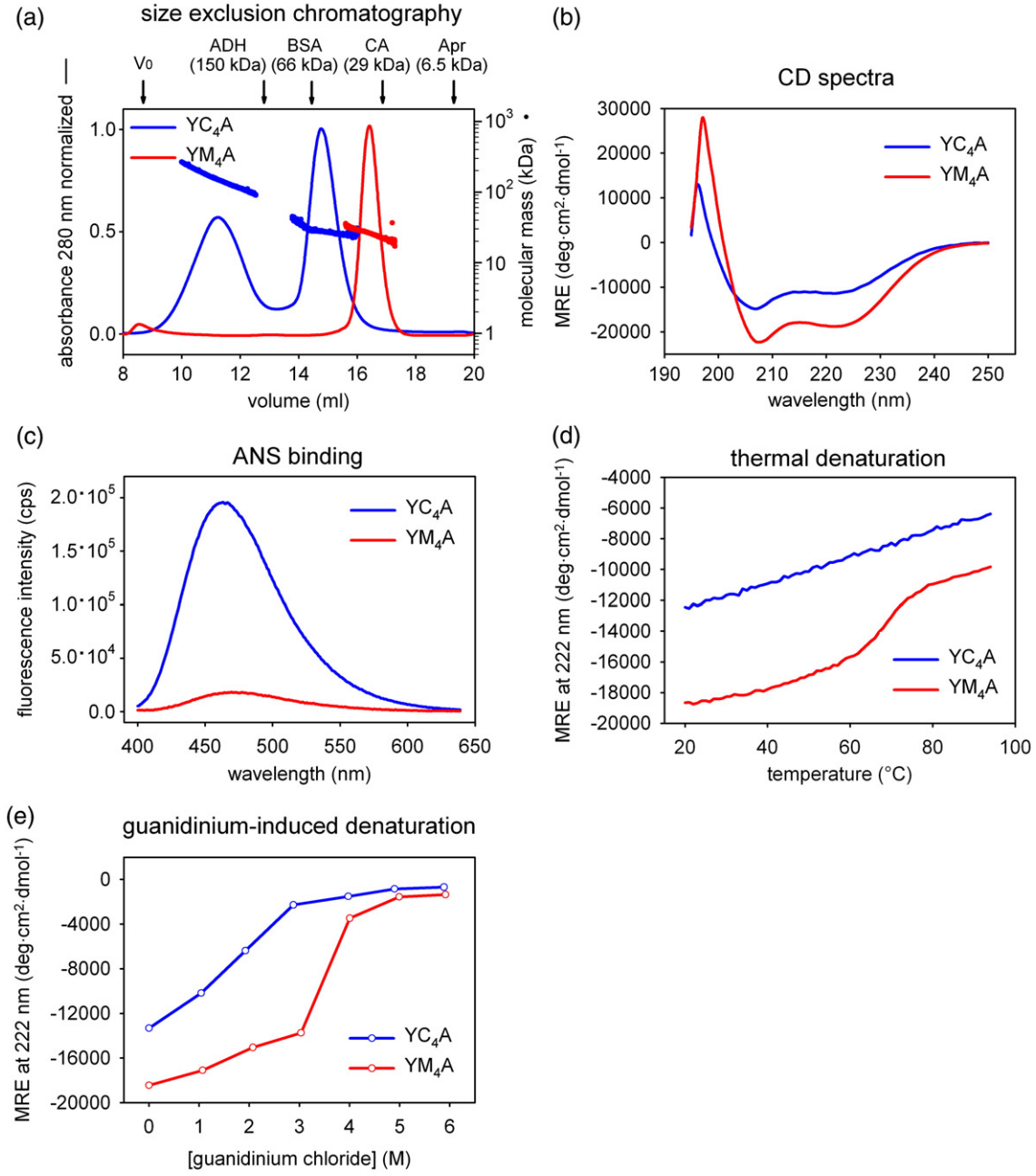


Fig. 9. Comparison between YC_4A, in blue, and YM_4A, in red. SEC (a) was performed with samples directly after IMAC purification. MALS data are also shown. The chromatogram of YC_4A displays one peak corresponding to the monomer (on the right) and one corresponding to oligomeric fractions (on the left). CD spectroscopy (b) shows an increase in ellipticity for YM_4A. ANS binding (c) is drastically reduced for YM_4A to levels typical of natural armadillo repeat proteins; the data shown refer to values after buffer subtraction. Thermal denaturation (d) and guanidinium-induced denaturation (e) indicate the presence of a cooperative unfolding transition, characteristic for native-like proteins, for YM_4A. V_0 indicates the column void volume. Alcohol dehydrogenase (ADH; MW = 150 kDa), bovine serum albumin (BSA; MW = 66 kDa), carbonic anhydrase (CA; MW = 29 kDa), and aprotinin (Apr; MW = 6.5 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by arrows.

acteristic for armadillo repeats, but the conservation at other positions is relatively low.[32] To obtain a more reliable and informative consensus, we deemed it necessary to analyze the subfamilies independently. The use of closely related sequences should also improve the self-compatibility between designed repeats. At the time of the initial sequence design, only members of importin-α and β-catenin/plakoglobin subfamilies were known to be peptide binders and had crystal structures available. As a consequence, only the repeats from proteins belonging to these subfamilies were thus chosen for the calculation of the consensus, to avoid interference from other subfamilies of unknown structure that could negatively affect the final sequences. Indeed, the later publication of the structure of plakophilin[53] (a member of the p120 subfamily) revealed an unexpected shape with a pronounced bend in the

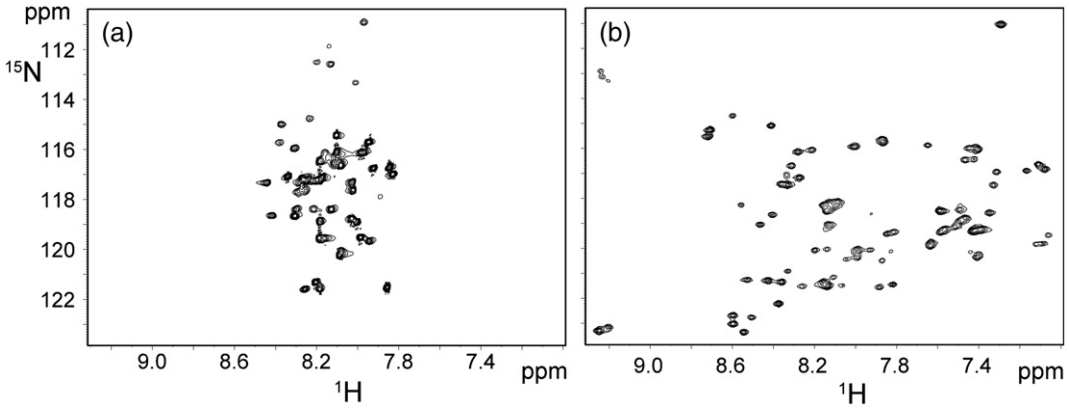


Fig. 10. [^{15}N,^{1}H]-HSQC spectra of designed armadillo repeat proteins: YC_4A (a) at pH 6 and YM_4A (b) at pH 11. Both spectra were recorded at a temperature of 310 K in 20 mM Tris–HCl and 30 mM NaCl. The protein concentration was 0.6 mM.

middle of the domain, supporting, *a posteriori*, the initial choice of sequence restriction to the subfamilies mentioned above.

An overall consensus was, however, also realized to take into account the possible combination of sequences belonging to importin-α and β-catenin/plakoglobin subfamilies. An obvious concern regarding the combination of these two subfamilies was that the overall consensus (type C) might be too similar to the importin consensus (type I) due to the slight overrepresentation of importin sequences in the original set. After the exclusion of the positions involved in binding, highly conserved for functional reasons especially in the importin subfamily and thus preserved also in the overall consensus, the C- and I-type repeats share 74% identity and 82% similarity, while C- and T-types have corresponding values of 70% identity and 87% similarity. The values indicate that the overall consensus is thus not significantly biased toward the importin consensus in the "framework" positions, that is, the positions not responsible for binding. The positions involved in peptide binding will be randomized in the library design and thus do not play a role in these considerations. Nevertheless, despite the similarity between I-, T-, and C-type modules, we always used only one type of consensus modules in every repeat protein tested, to provide a constant interface between the repeats and to be able to correlate the protein properties with the types of modules.

The capping repeats represented a second key point in the protein design. As observed for designed ankyrin repeat proteins,[31,54] capping repeats can dramatically increase *in vivo* folding yield and prevent aggregation. We found that an N-terminal capping repeat derived from yeast importin-α (Ny) and an artificial C-terminal capping repeat (Ca), designed by replacing exposed hydrophobic residues, give the highest expression yield of soluble protein in *E. coli*. Remarkably, we could find a single combination of capping repeats that allowed us to analyze the properties connected to the types of internal modules.

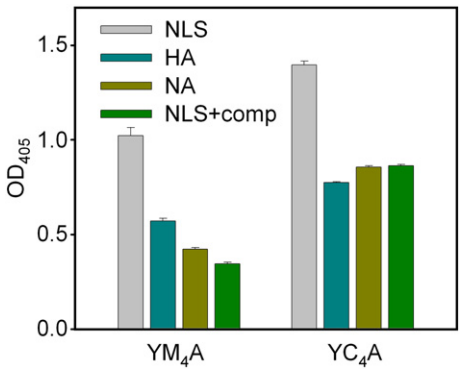


Fig. 11. ELISA of YM_4A and YC_4A. YM_4A binds specifically to immobilized SV40 large T antigen NLS. Immobilized hemagglutinin tag peptide (HA) and NeutrAvidin (NA) are negative controls. Binding to NLS can be competed by addition of NLS peptide (SPKKKRKVE) in solution at a concentration of 10 μM (NLS+comp). Experiments were performed in duplicate with YM_4A and YC_4A at a concentration of 1 μM.

Protein properties

Data from the artificial repeat proteins previously designed[28–30] indicate that biophysical properties often correlate with the number of internal repeats. Indeed, this behavior was also observed for designed armadillo repeat proteins.

I-type proteins

I-type proteins show that helical secondary structure content, thermal stability, and resistance to guanidinium-induced denaturation increase with the number of repeats, pointing in the same direction as data from other artificial repeat proteins.[27,29,30,55] A helical content of approximately 80% for YI_4A and YI_8A (Table 1) corresponds to the expected theoretical value from the design and is even higher than the values observed for natural armadillo domains. Low ANS binding and clearly defined transitions in thermal and guanidinium-

induced denaturation indicate that the I-type module can lead to native-like molecules, and the elevated midpoint of denaturation points toward a superior thermodynamic stability compared to natural proteins. At the same time, thermal denaturation is almost completely reversible. The thermal denaturation was employed here as a qualitative method to assess the stability of the designed proteins and to compare them to their natural counterparts. A detailed thermodynamic analysis requires, however, further investigation of the folding mechanism, which is probably more complex than a simple two-state transition and possibly also described by the Ising model as in the case of other designed repeat proteins.[56,57]

I-type proteins could, thus, be good candidates as scaffold for peptide-binding molecules. However, their predominant dimeric state constitutes a disadvantage during selection and characterization of binding properties due to possible avidity effects. Even considering that the I-type proteins are dimers, the SEC data indicate an elution volume still larger than expected, which could be interpreted as a result of an elongated shape. It is noteworthy that natural armadillo proteins do not show a higher-than-expected apparent mass in gel filtration (Table 1).

T-type proteins

T-type proteins share several native-like characteristics with I-type proteins, such as the presence of a compact hydrophobic core inaccessible to solvent, as suggested by ANS binding levels that are as low as those of natural armadillo repeat proteins, and the transitions observed in thermal and guanidinium-induced denaturation. The reversibility of thermal denaturation in T-type proteins is, however, less complete than that in I-type proteins and completely lost in the case of YT_8A. The helical content of T-type proteins is approximately independent of the number of repeats and generally lower than that in natural armadillo repeat proteins. The gel filtration results are, however, similar to I-type proteins, with apparent molecular masses higher than expected by more than a factor of 3 on average, already taking the dimeric state into account (Table 1). Due to the native-like properties of T-type proteins, the increase in hydrodynamic radius could be interpreted as an effect of a rodlike shape. Despite the different behavior in gel filtration, T-type proteins are therefore more similar in their biophysical characteristics to native armadillo proteins than to an idealized scaffold. When applying a strategy of protein assembly from preselected modules, which represents one of the aims of a general modular peptide binder, the scaffold properties should ideally change in a regular and predictable way, when adding modules, without altering the general characteristics. However, this is not observed for YT_8A, where the reversibility in thermal unfolding is completely lost.

C-type proteins

C-type proteins, in contrast to the other designed armadillo proteins, do not show a clear transition in thermal or guanidinium-induced denaturation but a gradual loss of secondary structure, and ANS binding results indicate the presence of an accessible hydrophobic core. Thus, C-type proteins are probably not completely folded but rather are in a molten-globule-like state. The secondary structure is present, as indicated by CD, but the proteins retain a high level of flexibility due to the lack of a fixed tertiary structure. The high apparent molecular masses observed in gel filtration, where MALS indicate monomeric states, might then be interpreted as a consequence of the intrinsic flexibility of the polypeptide chain. The molten-globule-like characteristics of C-type proteins represent a serious limitation in library generation, where framework stability and tolerance to mutations are desired. From the point of view of the design, however, the observation of a molten-globule-like state for armadillo repeat proteins built from overall consensus (type C) modules could suggest either an insufficient stability of each repeat or an inadequate interaction between them, supporting the initial choice of restricting the design to specific subfamilies.

Molten globule stabilization

The initial consensus-based approach led to stable dimers (I- and T-types) or molten-globule-like monomeric proteins (C-type). The further possible design steps to obtain a stable monomer were either the disruption of the interaction in the dimer or the stabilization of the C-type proteins. However, no information was available concerning the dimerization interface and the residues involved; both surface interaction and domain swap were conceivable as mechanisms of dimer formation. The improvement strategy would thus have to involve systematic point mutations of several single residues and combination of residues, with the risk that the disruption of the dimer will simply lead to a stability loss or even a molten globule state. For disrupting a dimer, the improvement strategy would have to consist of systematic mutations of single residues or combination of residues without a structural hint to select mutations.

We chose instead an alternative approach, focused on the stabilization of the hydrophobic core of the molten-globule-like C-type proteins, using a computational approach. As mentioned above, the molten-globule-like state suggested inadequate inter- or intra-repeat interactions at the hydrophobic core level and, hence, insufficient packing of the core. Our results show that the introduction of only three point mutations in the hydrophobic core of C-type repeats was sufficient to convert a molten-globule-like protein with four internal repeats to a stable conformation. This strongly argues that the packing of the hydrophobic core was indeed the critical parameter for obtaining a stable fold.

The underpacking of the core may be one of the reasons for the molten globule behavior[58] of the C-type proteins. Two of the mutations (Gly to Ala at position 12 of the repeat and Ala to Leu at position 32) increase the calculated volume of the hydrophobic core, bringing it close to the average value of natural repeats. These residues are also the most common among the 50 highest-ranking sequences with a frequency of 72% for Ala12 and 50% for Leu32. The third mutation (Leu to Ile at position 38) probably reduces the local flexibility by limiting the number of available rotamers. Such a restriction can help to lock the hydrophobic core in a unique conformation, and this positive effect could overcome the disadvantage of having a residue with low helical propensity, such as isoleucine. However, as observed for several mutants, the contributions from the single residues are not additive and the core packing is the result of a particular combination of residues.

Strikingly, the M-type repeat has, among the 432 mutants screened *in silico*, the core sequence closest to the I-type repeat (Table 2). The only two core residues that differ between M- and I-type repeats (Gln28 and Thr39) were not included in the set of possible mutations in our computational approach. The protein YI_4A, derived from I-type modules, shows characteristics very similar to YM_4A, apart from its dimeric state. Therefore, the particular core sequence obtained for both types of repeats represents a reliable solution for core packing, considering that it has been obtained by consensus design (for I-type) and simulation (for M-type). The hydrophobic core is probably rather stable, and we may speculate that the dimerization observed for I-type proteins takes place most likely via surface interaction instead of domain swap. The introduction of surface point mutations could then possibly lead to the formation of stable monomers.

YM_4A

The observed biophysical characteristics indicate that YM_4A represents a significant improvement of the original consensus sequence. YM_4A is almost as compact as globular proteins with similar molecular weight, as judged from elution volumes in SEC, and only marginally binds ANS, with values in the range observed for natural armadillo repeat proteins. The thermal and guanidinium-induced denaturation curves have sigmoidal profiles, indicating the presence of a cooperative unfolding, a hallmark of natural globular proteins.

NMR

NMR spectra provide further indications of the folded structure of YM_4A. Due to the repetitive nature of the sequence, it is *a priori* not clear how many peaks should be expected, but, even in the absence of specific assignments and considering the effects of symmetry, most of the peaks are probably present. However, Gly residues, usually observed in a characteristic region of the correlation map, are missing, most likely due to the highly accelerated amide proton exchange at pH 11 for residues outside the regions of secondary structure. Nevertheless, the presence of most peaks at the elevated pH indicates that the majority of amide moieties are protected from solvent access.

Although it was not possible to assign the spectra, the $^{15}N\{^1H\}$-NOE data indicate that almost all peaks in the proton–nitrogen correlation spectrum correspond to amide moieties with motional properties similar to those of residues from stably folded secondary structural elements. Hence, the NMR measurements suggest that YM_4A at pH 11 can be considered as a well-folded globular protein, whereas YC_4A shows characteristics of a molten globule. YM_4A, at pH lower than 10, displays broader lines, without affecting the signal dispersion, indicating that under those conditions, good side-chain packing is probably disturbed by the presence of an ionized group. A large range of pH values was also tested for YC_4A, yet without leading to any improvement in the dispersion of the signals in $[^{15}N,^1H]$-HSQC spectra or narrowing of the line width. Hence, electrostatic interactions are not dominating the molten globule properties of YC_4A. These observations rather indicate that subtle effects of side-chain packing are involved and that proper side-chain packing is achieved only in YM_4A, which presumably requires a neutral state of one group that is charged at neutral pH but uncharged at pH 11. As the lines are sharper at basic pH, lysine residues are the candidates for causing this effect, because of the possible repulsive interaction with the lysines in the neighboring repeats when both are charged, as observed in the molecular models.

Peptide binding

The binding to the SV40 large T antigen NLS observed in ELISA confirms the interpretation of the biophysical data. Unspecific binding has been reduced in YM_4A, compared to the original molten-globule-like YC_4A, as observed in binding to NeutrAvidin and to the hemagglutinin tag and in the competition experiment.

Even though no design effort was made in the present work for binding to a target peptide, YM_4A and YC_4A do show a weak but specific binding, indicating a correct disposition of the residues involved in interactions with the peptide. Glu30, Trp33, and Asn37 in the M-type repeat correspond to the residues responsible for binding to NLS in natural importin-α proteins. These residues are present in YC_4A and YM_4A due to the high conservation in the importin-α sequences, which were used in the original consensus design. The competition with soluble peptide strongly suggests the presence of specific interactions rather than a merely electrostatic binding phenomenon.

Further experiments will be needed to clarify the binding of consensus-designed armadillo repeat proteins. Nevertheless, the results already achieved indicate a correct structure. Armadillo repeat proteins based on M-type repeats can thus be used as scaffold for library generation and selection.

Evaluation of the computational method

A rotamer sampling method was chosen to identify, from a large pool of candidates, armadillo repeat proteins with improved core packing. The approach was devised for use with large proteins, up to 500 residues in our case. Despite recent advances,[59] such complexity is still not easily treatable by the available methods for core repacking, which proceed through a cycle of mutation, selection of residue conformation, and energy minimization. In terms of computational load, the search for a sequence with minimal energy is highly demanding, or even not affordable at all, for large proteins. In contrast, a simple evaluation of the potential energy of protein models after energy minimization is rather unreliable. The introduction of point mutations in the hydrophobic core requires the rearrangement of the core side chains to optimize the core packing. This task is, however, not fulfilled by a simple energy minimization, especially when the energy barrier between rotamers is too high to be overcome and only the nearest local minima for the side chains are reached (e.g., for tightly packed side chains). However, our aim was not to find the conformation at the global minimum but to estimate the packing efficiency of given mutants. A random sampling, helped by the partial removal of the energy barriers and followed by statistical analysis, is thus a feasible procedure for evaluating the packing of each mutant protein.

Though being a simplified approach, it was still necessary to reduce the complexity of the system. The choice of candidate sequences was based on information derived from sequence alignment and interactions in the structures. This approach is not exhaustive, but it restricts the search space to the most promising mutants according to criteria that are independent of the computational method.

Nevertheless, some further restrictions were required to keep the computational load within reasonable boundaries. Calculations without solvation terms are computationally less expensive and can be applied in our case, where the mutations correspond only to aliphatic–aliphatic substitutions in the hydrophobic core and are not expected to influence solute–solvent energy contributions. When restraining backbone atoms, the influence of the core mutations on the surface electrostatics is negligible and the electrostatic contribution to the potential energy does not vary upon mutation.

A second crucial issue was the choice of three different armadillo structures as starting points for the generation of the mutant models to avoid a result biased by the use of a single structure. Additionally, the introduction of multiple backbone templates and backbone flexibility has already been shown to improve the quality of the sequence search.[60,61] In the present work, we did not attempt to build a new backbone including the information from multiple structures, but we allowed "flexibility" by using three starting structures and fixing the coordinates of the backbone atoms using harmonic restraints. The use of three structures per mutant also helped to enhance the signal-to-noise ratio of the ranking, as a high rank for all three armadillo structural backgrounds is generally required to obtain a high overall rank.

Sequences with low core volume were excluded from the final ranking to decrease the number of false positives in the pool selected for experimental characterization (Supplementary Table S3). One example of such a low core volume sequence is the original C-type consensus, which is highly ranked; its core volume was set as the lower volume threshold for the discrimination of the mutants. On a fixed backbone, a reduced core volume allows side chains to assume nearly ideal values of bond lengths, angles, and dihedrals, leading to a significant reduction of the total energy and to an artificially high rank. An increase in flexibility of the backbone could also be the source of artifacts: the cavities in the hydrophobic core of low volume mutants can be compensated by compressing the backbone structure and bringing the side chains close enough to take advantage of the van der Waals interactions. However, a reduction of the backbone flexibility could be detrimental for mutants slightly more overpacked than the natural structures, which would not be able to reach low energy values without backbone adjustments.

Homology models of C-type proteins, based on the armadillo crystal structures, indicated the likely presence of small cavities in the hydrophobic core, suggesting that underpacking is one of the possible reasons for the molten globule state. It is thus unlikely that proteins with core volume lower than the original C-type consensus can provide better packing. Furthermore, mut19, the highest-ranked mutant with low core volume, displays ANS binding and SEC properties closest to the original overall consensus sequence, confirming the validity of our selection filter based on core volume. No threshold level was set for possible overpacking cases: the maximal value of core volume among the considered mutants was still in the range of the average repeat volume calculated for the reference structure of murine importin-α (PDB ID 1Q1T; Supplementary Table S3).

On the experimental side, the use of ANS binding and SEC to discriminate mutants is rather qualitative but can represent an efficient and relatively fast method for screening. A good overall indication of the quality of our method is given by the fact that all the monomeric mutants analyzed show an improvement compared to the original overall consensus.

Conclusions

This work focused on the generation of designed armadillo repeat proteins for the construction of a general modular peptide-binding scaffold. An initial consensus-based design led to well-expressed and stable but dimeric proteins or molten globules. A stable, well-expressed monomeric protein was obtained using a force field-based approach for the stabilization of the hydrophobic core of the molten globule variant.

In a library perspective, a monomeric protein allows a better evaluation of the binding properties, without the influence of possible avidity effects, which can be critical in the discrimination between similar target peptides. The mutations to be introduced to generate a library will only affect surface residues, leaving the hydrophobic core untouched except for one position (position 4 may contribute to both the hydrophobic core and the binding site). Therefore, the favorable characteristics of the designed proteins will probably be kept for most library members and selected specific binders.

Materials and Methods

Sequence analysis and modeling

SMART†,[33,34] Swiss-Prot‡,[26] and PDB§[62] were used as the starting databases for our analysis. GCG (Wisconsin Package Version 10.3, Accelrys Inc., San Diego, CA), BLAST‖,[63,64] and ClustalW¶[35] were used for sequence retrieval and alignment. Structure analysis and modeling were performed with Swiss-Pdb Viewer[a],[65] MOLMOL[b],[66] PyMOL[c] (DeLano Scientific LLC, San Francisco, SA), and INSIGHT II (Accelrys Inc.). Vector NTI (Invitrogen) was used for vector and oligonucleotide design.

General molecular biology methods

Unless stated otherwise, experiments were performed according to Sambrook and Russell.[67] Vent Polymerase (New England Biolabs, USA) was used for all DNA amplifications. Enzymes and buffers were from New England Biolabs or Fermentas (Lithuania). The cloning and production strain was *E. coli* XL1-blue (Stratagene, USA). Competent cells were prepared according to Inoue *et al.*[68] The *E. coli* strain M15 (Qiagen, Germany), containing the plasmid pREP4, was used for the production of ^{15}N-labeled proteins for NMR experiments. The cloning and protein expression vectors were pQE30 (Qiagen, Switzerland) and pPANK (GenBank accession number AY327140). From this, the vector pPANK-NyCa was constructed by cloning of the capping repeats Ny and Ca. pPANK-NyCa contains the BsaI and BpiI restriction sites between the capping repeats for cloning purposes. Note that the inserts were constructed with a double stop codon (see Supplementary Fig. S3). pPANK-NyCa was used to clone the internal repeats for N8C and core mutant proteins. pQE30 and derivatives such as pPANK carry an $MRGSH_6$ tag at the N-terminus of the proteins. The DNA sequences corresponding to the NLS and HA peptides were inserted in the vector pAT223 (GenBank accession number AY327138) and expressed as fusion proteins to pD. The produced proteins consist of N-terminal Avi tag, pD, His_6 tag, and the peptide at the C-terminus. The plasmid pBirAcm (Avidity, USA), encoding *E. coli* biotin-protein ligase BirA, was used for *in vivo* biotinylation of pD peptides.

Cloning of designed armadillo repeat proteins

Oligonucleotides were purchased from Microsynth AG (Balgach, Switzerland). A complete list of all oligonucleotides used is given in Supplementary Table S1. An approach similar to the one described by Binz *et al.*[27] was adopted for gene assembly (Supplementary Fig. S2). All single repeat modules were assembled from oligonucleotides by assembly PCR. The single modules of the core mutants were assembled using the combinations of oligonucleotides indicated in Supplementary Table S2. As an example, for the C-type consensus, pairs of partially overlapping oligonucleotides (1–2, 3–4, and 5–6) were annealed and the double strand was completed by PCR. Then, 2 μl from these PCR reaction mixtures was combined as template for a second assembly reaction in the presence of oligonucleotides 1 and 6. All the oligonucleotides were used at a final concentration of 1 μM. The annealing temperature was 47 °C for the first reaction and 50 °C for the second. Thirty PCR cycles were performed with an extension time of 30 s. The same procedure was applied for the other internal and capping repeats. Only four oligonucleotides were used for the N-terminal capping repeats. BamHI and KpnI restriction sites were used for the direct insertion of the modules into the plasmid pQE30. The single modules were PCR amplified from the vectors, using external primers pQE_f_1 and pQE_r_1 (Qiagen, Switzerland). Neighboring modules were digested with the type IIS restriction enzymes BpiI and BsaI and directly ligated together. The genes coding for the whole proteins were assembled by stepwise ligation of the internal and capping modules. BamHI and KpnI restriction sites were used for insertion of the whole genes into the vector pQE30 and the plasmids were sequenced. For pD-peptide fusions, oligonucleotides encoding both strands of the peptide sequences and containing the restriction sites for BamHI and HindIII were mixed and heated to 95 °C for 10 min and then cooled to 4 °C to allow annealing of the two strands. The double-stranded DNA fragments were subsequently digested with BamHI and HindIII and ligated into the plasmid pAT223.

Natural armadillo domain constructs

The armadillo domain of mouse β-catenin (βArm; residues 150–665) was amplified from the cloned β-catenin gene[22] (a generous gift from W.I. Weis, Stanford University, USA) using oligonucleotides AcatFOR and AcatREV, digested with BamHI and KpnI, and inserted into pQE30. The armadillo domain of human importin-α1 (αArm;

‡ http://www.expasy.org
§ http://www.pdb.org
‖ http://www.ncbi.nlm.nih.gov/blast
¶ http://www.ebi.ac.uk/clustalw
[a] http://www.expasy.org/spdbv/
[b] http://hugin.ethz.ch/wuthrich/software/molmol/
[c] http://pymol.sourceforge.net

residues 83–505) was amplified from a vector containing the importin gene (named importin-α5 in the original publication,[69] a generous gift from M. Köhler, Ostseeklinik Damp, Germany) using oligonucleotides IMAF5 and IMAR5, digested with BamHI and KpnI, and inserted into pQE30. Both proteins carry an N-terminal $MRGSH_6$ tag, as do the designed armadillo repeat proteins.

Protein expression and purification

E. coli XL1-blue cells were transformed with the respective plasmid and grown in LB medium containing 1% (w/v) glucose and 50 μg/ml of ampicillin at 37 °C with vigorous shaking. Expression was induced by IPTG (final concentration of 0.5 mM) when the culture reached $OD_{600} = 0.6$. After 3 h of expression, cells were harvested by centrifugation. For *in vivo* biotinylation of pD peptides that contain an N-terminal Avi tag, cells were cotransformed with pBirAcm and pAT223 (carrying the pD-peptide constructs) and grown in medium containing 30 μg/ml of chloramphenicol and 50 μg/ml of ampicillin. Before induction with IPTG, biotin was added to the medium to a final concentration of 50 μM, according to Cull and Schatz.[70]

Protein purification was performed at 4 °C. Cells were resuspended in 50 mM Tris–HCl and 500 mM NaCl (pH 8.0) and lysed in a French pressure cell (SLM Instruments, USA) at a pressure of 1200 psi. The lysis mixture was further homogenized by sonication (Branson, USA). Insoluble material was pelleted by centrifugation at 20,000*g* for 30 min. The supernatant was purified by IMAC with Ni–NTA material (Qiagen), equilibrated with buffer containing 50 mM Tris–HCl, 500 mM NaCl, 10% (v/v) glycerol, and 20 mM imidazole (pH 8.0). Columns were washed extensively with the equilibration buffer and then proteins were eluted with an elution buffer identical with the equilibration buffer but also containing 250 mM imidazole. β-Catenin was expressed and purified under the same conditions.

For importin-α1, the expression was carried out at 25 °C for 6 h and the cell pellet was resuspended in lysis buffer containing 50 mM Tris–HCl, 500 mM NaCl, 10% glycerol, 5 mM β-mercaptoethanol, and 10 mM imidazole (pH 8.0). IMAC purification was performed as indicated above using the same buffers with the addition of 5 mM β-mercaptoethanol. Samples were then dialyzed overnight against 50 mM Tris–HCl and 2 mM DTT (pH 8.0) and applied to a POROS HQ anion-exchange column, equilibrated with running buffer (50 mM Tris–HCl, pH 8.0), using the BioCAD 700 E Perfusion Chromatography Workstation (Applied Biosystems, Germany). The column was then washed with 50 mM Tris–HCl and 20 mM NaCl (pH 8.0), and the samples eluted with a gradient from 20 mM to 1 M NaCl. Protein size and purity were assessed by 15% SDS-PAGE, stained with Coomassie PhastGel Blue R-350 (GE Healthcare, Switzerland).

The expected mass of all the studied proteins was confirmed by mass spectrometry. Protein concentrations were determined by absorbance at 235 and 280 nm using molecular masses and extinction coefficients calculated with the tools available at the ExPASy proteomics server‡ and by the bicinchoninic acid assay (Pierce).

SEC and MALS

Analytical SEC was carried out either on an Ettan LC system using a Superdex 200 PC 3.2/30 column (flow rate 70 μl/min) or on an ÄKTA explorer chromatography system using a Superdex 200 10/30 GL column (flow rate, 0.5 ml/min) (GE Healthcare). Phosphate buffer (50 mM phosphate and 150 mM NaCl, pH 7.4) and two Tris-based buffers (20 mM Tris–HCl and 50 mM NaCl, pH 8.0, or 50 mM Tris–HCl and 500 mM NaCl, pH 8.0) were used. The armadillo domain of β-catenin was soluble only at 150 or 500 mM salt concentration. The armadillo domain of importin-α1 was analyzed in phosphate buffer (50 mM phosphate, 500 mM NaCl, and 5 mM DTT, pH 7.4). The core mutants were analyzed in buffer containing 20 mM Tris–HCl and 50 mM NaCl, pH 8.0. MALS measurements were performed with a miniDAWN light-scattering detector and an Optilab refractometer (Wyatt Technologies, USA) coupled to the ÄKTA system. Molecular weight estimates were calculated using the ASTRA 4.73.04 software package (Wyatt Technologies).

CD spectroscopy

CD measurements were performed on a Jasco J-810 spectropolarimeter (Jasco, Japan) using a 0.5-mm cylindrical thermocuvette. CD spectra were recorded from 190 to 250 nm with a data pitch of 1 nm, a scan speed of 20 nm/min, a response time of 4 s, and a bandwidth of 1 nm. Each spectrum was recorded three times and averaged. Measurements were performed at 20 °C. The CD signal was corrected by buffer subtraction and converted to MRE. Heat denaturation curves were obtained by measuring the CD signal at 222 nm with temperature increasing from 20 to 95 °C (data pitch, 1 nm; heating rate, 1 °C/min; response time, 4 s; bandwidth, 1 nm). Data were processed as described above. Guanidinium-induced denaturation measurements were performed after overnight incubation at 20 °C with increasing concentrations of guanidinium chloride (99.5% purity, Fluka), and the data were collected and processed as described above. Measurements of designed armadillo repeat proteins were performed in 20 mM Tris–HCl and 50 mM NaCl (pH 8.0). CD spectra and denaturation curves of the armadillo domain of β-catenin were measured in 50 mM Tris–HCl and 500 mM NaCl (pH 8.0). CD spectra and denaturation curves of importin-α1 were measured in 50 mM phosphate, 500 mM NaCl, and 5 mM DTT (pH 7.4). CD spectra were analyzed using CDpro.[41] Among the algorithms available in CDpro, CDSSTR was chosen for the analysis, with the reference protein set SDP48 (IBasis = 7).

ANS binding

ANS fluorescence was measured using a PTI QM-2000-7 fluorimeter (Photon Technology International, USA). The measurements were performed at 20 °C in 20 mM Tris–HCl, 50 mM NaCl, and 100 μM ANS (pH 8.0) using purified proteins at a final concentration of 10 μM. ANS binding to the armadillo domain of β-catenin was measured in 50 mM Tris–HCl, 500 mM NaCl, and 100 μM ANS (pH 8.0) to avoid possible aggregation problems. The emission spectrum from 400 to 650 nm (1 nm/s) was recorded with an excitation wavelength of 350 nm. For each sample, three spectra were recorded and averaged.

Rotamer sampling of hydrophobic core mutants

A computational approach at the atomic level of detail was used to optimize the hydrophobic core. The approach

uses cycles of energy minimization and heating by molecular dynamics to sample favorable arrangements of the buried side chains and estimate the packing efficiency of residues in the hydrophobic core of a given mutant. The number of possible mutations in each repeat is 432 (Fig. 6b). Three x-ray structures were chosen as starting models to improve sampling: importin-α from *S. cerevisiae* (PDB ID 1EE4[46]) and mouse (PDB ID 1Q1T[47]), consisting each of 8 internal repeats and 2 capping repeats, and murine β-catenin (PDB ID 2BCT[22]), which consists of 10 internal repeats and 2 capping repeats. The original capping repeats of the three structures were substituted with Ny and Ca capping repeats (Fig. 3) in the models. Each mutation was modeled by deleting the side chains at the core positions of each repeat and substituting them with the new side chains with random rotamer conformations; the resulting structure was minimized to eliminate clashes. Three models (from the three initial structures) were prepared for each of the 432 combinations of allowed mutations. All the repeats in each model were designed to have the same mutation pattern.

The extended atom approximation (param19) of the CHARMM force field[71] with a distance-dependent dielectric function was used for both energy minimization and heating by short molecular dynamics runs. All the side-chain atoms not directly in contact with core residue atoms (i.e., those more than 5 Å away, in the initial conformation, from any atom of the 16 core residues of each repeat) and all the backbone atoms were restrained using a harmonic potential with a force constant of 1.0 kcal·mol^{-1}·Å^{-2}. As a consequence, only the side-chain rotatable bonds of the core residues were fully flexible. The system was further minimized in the presence of the harmonic potential. A heating–quench protocol was iterated 100 times for each of the 432 mutants and each of the three protein models (Fig. 7). The first step was a 10-ps heating to 400 K, omitting the nonbonding energy terms (i.e., van der Waals and electrostatics). The second step was a minimization including all energy terms. The aim of the heating phase was to shuffle the flexible side-chain rotamers. The absence of nonbonding energy terms and the high temperature granted a more efficient exploration of the energy landscape. After the heating step, the minimization was used to reach the nearest minimum of the total potential energy. The coordinates were stored at the end of each minimization, and a total of 100 conformations were generated for each mutant. These conformations were further minimized without the aforementioned restraints, and the potential energy was evaluated for ranking.

The CHARMM potential energy is:

$$E = E_{\text{bonding}} + E_{\text{vdw}} + E_{\text{elec}} \quad (1)$$

where E_{bonding} is the sum of bond, angle, improper, and dihedral potential terms; E_{vdw} is the van der Waals energy; and E_{elec} is the coulombic energy. The E_{elec} term was neglected for ranking, because it is insensitive to aliphatic-to-aliphatic mutations in the extended atom representation. Moreover, because of the restraints, the restricted flexibility of the backbone polar groups results in a noisy coulombic energy. Therefore, a reduced potential energy was used for ranking. For each starting structure, the energy value for the conformation i of mutant m is:

$$E_i^m = E_{i,\text{bonding}}^m + E_{i,\text{vdw}}^m \quad (2)$$

As each conformation has a different potential energy, median, first percentile (of most favorable values), and minimum energies were extracted from the energy series of the 100 conformers to characterize each mutant: these values were used to make three independent ranks. At the end of this procedure, for each of the three initial structures, three rank numbers (corresponding to median, first percentile, and minimum ranking) were assigned to each mutant, and these nine rank numbers were summed. Finally, this sum was used for the overall rank of the mutant (Supplementary Table S3). The combination of multiple structures and different scoring criteria (i.e., median, first percentile, and minimum) was used to take into account, in an approximate way, the limited sampling. The central processing unit time required for each starting model of a mutant was approximately 5 h for importin structures and 7 h for catenin structures on a single processor of a 2800-MHz Opteron dual core. The total calculation time of approximately 8000 h was distributed over 150 central processing units.

NMR

Proteins for NMR studies were produced using *E. coli* strain M15 (Qiagen) containing the plasmid pREP4 growing in minimal medium with ^{15}N-labeled ammonium chloride as the only nitrogen source. The medium was supplemented with trace metals, 150 μM thiamin, and 30 μg/ml kanamycin. Expression and purification by IMAC and gel filtration were performed as described. The buffers used for NMR measurements contained 20 mM deuterated Tris–HCl and 30 mM NaCl (pH values of 6, 7, 8, 9, 10, or 11). YC_4A and YM_4A were concentrated to 0.6 mM for NMR measurements.

Proton–nitrogen correlation maps were derived from [^{15}N,^{1}H]-HSQC experiments[72] utilizing pulsed-field gradients for coherence selection and quadrature detection[73] and incorporating the sensitivity enhancement element of Rance and Palmer.[51,74] The ^{15}N{^{1}H}-NOE data were measured using a proton-detected version of the ^{15}N{^{1}H} steady-state heteronuclear Overhauser effect.[75] All experiments were recorded on a Bruker AV 700-MHz spectrometer equipped with a triple-resonance cryoprobe at 310 K. Spectra were processed and analyzed in the spectrometer software TOPSPIN 1.3 and calibrated relative to the water resonance at 4.63 ppm proton frequency, from which the ^{15}N scale was calculated indirectly.

ELISA

Biotinylated pD-peptide fusion proteins were immobilized on NeutrAvidin-coated plates after IMAC purification using 200 μl of 10-μM protein solutions and 1 h incubation time. One hundred microliters of 1 μM armadillo repeat proteins was incubated for 1 h. Binding was detected with an anti-$MRGSH_4$ antibody (Qiagen), a secondary anti-mouse immunoglobulin G alkaline phosphatase conjugate (Sigma), and *p*-nitrophenylphosphate (Fluka). Absorbance at 405 nm was measured using a Perkin Elmer HTS 7000 Plus plate reader. A buffer solution containing 50 mM Tris–HCl, 150 mM NaCl, and 0.5% bovine serum albumin (pH 7.4) was used for all the proteins and for the blocking steps. Washing after each step was carried out with $TBST_{150}$ (Tris–HCl 50 mM, NaCl 150 mM, and 0.05% Tween 20, pH 7.4). All steps were carried out at 4 °C. Development with 4-nitrophenylphosphate and readout were performed at room temperature.

Acknowledgements

The authors want to thank W.I. Weis, M. Köhler, and E. Conti for kindly providing the plasmids containing the natural armadillo repeat protein genes. We thank Dr. P. Kolb for valuable suggestions, Dr. A. Honegger for EXCEL macros, and the other members of the Plückthun laboratory for fruitful discussions. The calculations were performed on Matterhorn, a Beowulf Linux cluster at the Informatikdienste of the University of Zürich. We thank C. Bolliger, Dr. T. Steenbock, and Dr. A. Godknecht for installing and maintaining the Linux cluster. F. Parmeggiani was the recipient of a predoctoral fellowship from the Roche Research Foundation. F.P. and G.V. are members of the Molecular Life Science Ph.D. program. This work was supported by the Swiss National Center of Competence in Research (NCCR) in Structural Biology and in part by a Discovery grant from the Kommission für Technologie und Innovation (KTI).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.12.014

References

1. Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **23**, 1105–1116.
2. Binz, H. K., Amstutz, P. & Plückthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* **23**, 1257–1268.
3. Almagro, J. C. (2004). Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J. Mol. Recognit.* **17**, 132–143.
4. MacCallum, R. M., Martin, A. C. & Thornton, J. M. (1996). Antibody–antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* **262**, 732–745.
5. Marchalonis, J. J., Adelman, M. K., Robey, I. F., Schluter, S. F. & Edmundson, A. B. (2001). Exquisite specificity and peptide epitope recognition promiscuity, properties shared by antibodies from sharks to humans. *J. Mol. Recognit.* **14**, 110–121.
6. Wilson, I. A., Ghiara, J. B. & Stanfield, R. L. (1994). Structure of anti-peptide antibody complexes. *Res. Immunol.* **145**, 73–78.
7. Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 259–288.
8. Esteban, O. & Zhao, H. (2004). Directed evolution of soluble single-chain human class II MHC molecules. *J. Mol. Biol.* **340**, 81–95.
9. Blatch, G. L. & Lässle, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein–protein interactions. *BioEssays*, **21**, 932–939.
10. Coates, J. C. (2003). Armadillo repeat proteins: beyond the animal kingdom. *Trends Cell Biol.* **13**, 463–471.
11. Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. (1999). The WD repeat: a common architecture for diverse functions. *Trends Biochem. Sci.* **24**, 181–185.
12. Peifer, M., Berg, S. & Reynolds, A. B. (1994). A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell*, **76**, 789–791.
13. Hatzfeld, M. (1999). The armadillo family of structural proteins. *Int. Rev. Cytol.* **186**, 179–224.
14. Harris, T. J. & Peifer, M. (2005). Decisions, decisions: beta-catenin chooses between adhesion and transcription. *Trends Cell Biol.* **15**, 234–237.
15. Anastasiadis, P. Z. & Reynolds, A. B. (2000). The p120 catenin family: complex roles in adhesion, signaling and cancer. *J. Cell Sci.* **113**, 1319–1334.
16. Nathke, I. S. (2004). The adenomatous polyposis coli protein: the Achilles heel of the gut epithelium. *Annu. Rev. Cell Dev. Biol.* **20**, 337–366.
17. Goldfarb, D. S., Corbett, A. H., Mason, D. A., Harreman, M. T. & Adam, S. A. (2004). Importin alpha: a multipurpose nuclear-transport receptor. *Trends Cell Biol.* **14**, 505–514.
18. Wieschaus, E., Nüsslein-Volhard, C. & Jürgens, G. (1984). Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. 3. Zygotic loci on the X-chromosome and 4th chromosome. *Wilhelm Roux's Arch. Dev. Biol.* **193**, 296–307.
19. Riggleman, B., Wieschaus, E. & Schedl, P. (1989). Molecular analysis of the armadillo locus: uniformly distributed transcripts and a protein with novel internal repeats are associated with a *Drosophila* segment polarity gene. *Genes Dev.* **3**, 96–113.
20. Groves, M. R. & Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383–389.
21. 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.
22. Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). Three-dimensional structure of the armadillo repeat region of beta-catenin. *Cell*, **90**, 871–882.
23. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell*, **94**, 193–204.
24. Catimel, B., Teh, T., Fontes, M. R., Jennings, I. G., Jans, D. A., Howlett, G. J. *et al.* (2001). Biophysical characterization of interactions involving importin-alpha during nuclear import. *J. Biol. Chem.* **276**, 34189–34198.
25. 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.
26. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788.
27. 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.
28. Mosavi, L. K., Minor, D. L., Jr & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA*, **99**, 16029–16034.
29. Main, E. R., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. (2003). Design of stable alpha-helical arrays from an idealized TPR motif. *Structure*, **11**, 497–508.

30. Stumpp, M. T., Forrer, P., Binz, H. K. & Plückthun, A. (2003). Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J. Mol. Biol.* **332**, 471–487.
31. Interlandi, G., Wetzel, S. K., Settanni, G., Plückthun, A. & 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.
32. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309**, 1–18.
33. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
34. Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, D257–D260.
35. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
36. Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E. & Corbett, A. H. (2007). Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J. Biol. Chem.* **282**, 5101–5105.
37. Xu, W. & Kimelman, D. (2007). Mechanistic insights from structural studies of beta-catenin and its binding partners. *J. Cell Sci.* **120**, 3337–3344.
38. von Kries, J. P., Winbeck, G., Asbrand, C., Schwarz-Romond, T., Sochnikova, N., Dell'Oro, A. *et al.* (2000). Hot spots in beta-catenin for interactions with LEF-1, conductin and APC. *Nat. Struct. Biol.* **7**, 800–807.
39. Hoffmans, R. & Basler, K. (2004). Identification and *in vivo* role of the Armadillo–Legless interaction. *Development*, **131**, 4393–4400.
40. Leung, S. W., Harreman, M. T., Hodel, M. R., Hodel, A. E. & Corbett, A. H. (2003). Dissection of the karyopherin alpha nuclear localization signal (NLS)-binding groove: functional requirements for NLS binding. *J. Biol. Chem.* **278**, 41947–41953.
41. Sreerama, N. & Woody, R. W. (2004). Computation and analysis of protein circular dichroism spectra. *Methods Enzymol.* **383**, 318–351.
42. Slavik, J. (1982). Anilinonaphthalene sulfonate as a probe of membrane composition and function. *Biochim. Biophys. Acta*, **694**, 1–25.
43. Ptitsyn, O. B. (1995). Molten globule and protein folding. *Adv. Protein Chem.* **47**, 83–229.
44. Butterfoss, G. L. & Kuhlman, B. (2006). Computer-based design of novel protein structures. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 49–65.
45. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **299**, 789–803.
46. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure*, **8**, 329–338.
47. Fontes, M. R., Teh, T., Toth, G., John, A., Pavo, I., Jans, D. A. & Kobe, B. (2003). Role of flanking sequences and phosphorylation in the recognition of the simian-virus-40 large T-antigen nuclear localization sequences by importin-alpha. *Biochem. J.* **375**, 339–349.
48. Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254**, 304–308.
49. Baum, J., Dobson, C. M., Evans, P. A. & Hanley, C. (1989). Characterization of a partly folded protein by NMR methods—studies on the molten globule state of guinea-pig alpha-lactalbumin. *Biochemistry*, **28**, 7–13.
50. Dyson, H. J. & Wright, P. E. (1998). Equilibrium NMR studies of unfolded and partially folded proteins. *Nat. Struct. Biol.* **5**, 499–503.
51. Palmer, A. G. (2001). NMR probes of molecular dynamics: overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 129–155.
52. Kalderon, D., Roberts, B. L., Richardson, W. D. & Smith, A. E. (1984). A short amino acid sequence able to specify nuclear location. *Cell*, **39**, 499–509.
53. Choi, H. J. & Weis, W. I. (2005). Structure of the armadillo repeat domain of plakophilin 1. *J. Mol. Biol.* **346**, 367–376.
54. Mosavi, L. K. & Peng, Z. Y. (2003). Structure-based substitutions for increased solubility of a designed protein. *Protein Eng.* **16**, 739–745.
55. Main, E. R., Stott, K., Jackson, S. E. & Regan, L. (2005). Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proc. Natl Acad. Sci. USA*, **102**, 5721–5726.
56. Kajander, T., Cortajarena, A. L., Main, E. R., Mochrie, S. G. & Regan, L. (2005). A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.* **127**, 10188–10190.
57. Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K. & Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full consensus ankyrin repeat proteins. *J. Mol. Biol*. In press. doi:10.1016/j.jmb.2007.11.046
58. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.
59. Schueler-Furman, O., Wang, C., Bradley, P., Misura, K. & Baker, D. (2005). Progress in modeling of protein structures and interactions. *Science*, **310**, 638–642.
60. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305–318.
61. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185–190.
62. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
63. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
64. McGinnis, S. & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25.
65. Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
66. Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics*, **14**, 51–55, 29–32.
67. Sambrook, J. & Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual*, 3rd edit., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
68. Inoue, H., Nojima, H. & Okayama, H. (1990). High efficiency transformation of *Escherichia coli* with plasmids. *Gene*, **96**, 23–28.

69. Köhler, M., Speck, C., Christiansen, M., Bischoff, F. R., Prehn, S., Haller, H. *et al.* (1999). Evidence for distinct substrate specificities of importin alpha family members in nuclear protein import. *Mol. Cell. Biol.* **19**, 7782–7791.
70. Cull, M. G. & Schatz, P. J. (2000). Biotinylation of proteins *in vivo* and *in vitro* using small peptide tags. *Methods Enzymol.* **326**, 430–440.
71. 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.
72. Bodenhausen, G. & Ruben, D. J. (1980). Natural abundance N-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.* **69**, 185–189.
73. Keeler, J., Clowes, R. T., Davis, A. L. & Laue, E. D. (1994). Pulsed-field gradients: theory and practice. *Methods Enzymol.* **239**, 145–207.
74. Kay, L. E., Keifer, P. & Saarien, T. (1992). Pure absorption gradient enhanced heteronuclear single-quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc.* **114**, 10663–10665.
75. Noggle, J. H. & Schirmer, R. E. (1971). *The Nuclear Overhauser Effect: Chemical Applications,* Academic Press, New York.