Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy

Pietro Alfarano,[1] Gautham Varadamsetty,[1] Christina Ewald,[2] Fabio Parmeggiani,[1] Riccardo Pellarin,[1] Oliver Zerbe,[2] Andreas Plückthun,[1]* and Amedeo Caflisch[1]*

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

Received 17 May 2012; Revised 25 June 2012; Accepted 25 June 2012
DOI: 10.1002/pro.2117
Published online 5 July 2012 proteinscience.org

Abstract: A multidisciplinary approach based on molecular dynamics (MD) simulations using homology models, NMR spectroscopy, and a variety of biophysical techniques was used to efficiently improve the thermodynamic stability of armadillo repeat proteins (ArmRPs). ArmRPs can form the basis of modular peptide recognition and the ArmRP version on which synthetic libraries are based must be as stable as possible. The 42-residue internal Arm repeats had been designed previously using a sequence-consensus method. Heteronuclear NMR revealed unfavorable interactions present at neutral but absent at high pH. Two lysines per repeat were involved in repulsive interactions, and stability was increased by mutating both to glutamine. Five point mutations in the capping repeats were suggested by the analysis of positional fluctuations and configurational entropy along multiple MD simulations. The most stabilizing single C-cap mutation Q240L was inferred from explicit solvent MD simulations, in which water penetrated the ArmRP. All mutants were characterized by temperature- and denaturant-unfolding studies and the improved mutants were established as monomeric species with cooperative folding and increased stability against heat and denaturant. Importantly, the mutations tested resulted in a cumulative decrease of flexibility of the folded state *in silico* and a cumulative increase of thermodynamic stability *in vitro*. The final construct has a melting temperature of about 85°C, 14.5° higher than the starting sequence. This work indicates that *in silico* studies in combination with heteronuclear NMR and other biophysical tools may provide a basis for successfully selecting mutations that rapidly improve biophysical properties of the target proteins.

Keywords: repeat protein; protein design; structural biology; implicit solvent

Abbreviations: 2D, two-dimensional; ANS, 1-anilino-8-naphthalene sulfonate; ArmRP, Armadillo Repeat Protein; CD, circular dichroism; GdnHCl, guanidine hydrochloride; HSQC, heteronuclear single-quantum coherence; MD, molecular dynamics; NMR, nuclear magnetic resonance; PCR, polymerase chain reaction; RMSD, root mean square deviation; RMSF, root mean square fluctuation; SDS-PAGE, sodium dodecylsulfate polyacrylamide gel electrophoresis; SEC, size-exclusion chromatography.

Additional Supporting Information may be found in the online version of this article.

Pietro Alfarano and Gautham Varadamsetty contributed equally to this work.

Grant sponsor: Swiss National Science foundation; Grant number: SINERGIA 122686.

**Correspondence to:* Andreas Plückthun, Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland. E-mail: plueckthun@bioc.uzh.ch or Amedeo Caflisch, Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland. E-mail: caflisch@bioc.uzh.ch

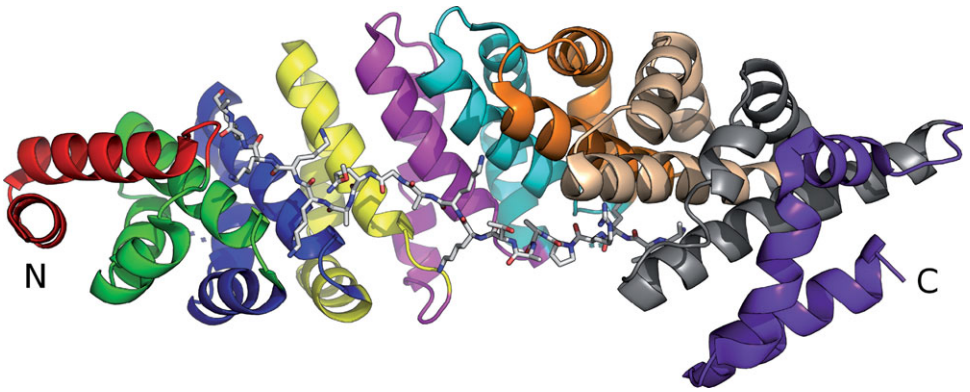

Figure 1. An armadillo repeat protein bound to a peptide. Importin-α (PDB accession code: 1EE5)[21] in complex with a nucleoplasmin NLS peptide is shown. Every repeat is colored differently and the NLS peptide is in stick representation.

Introduction

Molecular recognition is a very important aspect of biochemistry and is involved in almost all biological processes. Consequently, it is also the basis of numerous procedures in biological research and biomedical applications. To extend the applications beyond what is possible with antibodies, a number of different protein scaffolds[1–3] were explored over the past two decades for the generation of designed binding molecules using both rational and combinatorial approaches.

Although many recognition processes involve the mutual recognition of folded proteins, unstructured regions also play an important role. They frequently occur in linkers and termini of folded proteins, and many posttranslational modifications (e.g., phosphorylation, acetylation, methylation, etc.) are usually within extended regions of proteins. The recognition of unstructured regions of proteins has important applications in proteomics, as proteins frequently get denatured or even need to be unfolded by denaturants or detergents for analysis, such as for example for Western blots or protein chips. Additionally, the analysis by mass spectrometry frequently requires a proteolytic digestion, in which the proteins also lose all structural information. The sequence-specific recognition of unfolded proteins or extended regions or termini could thus enable the identification or quantitation of proteins or mutants in a very efficient way, using numerous technologies.

Repeat proteins are made up of several tandem repeats of defined structural units, which create an extended superhelical structure. They are especially attractive for designing binding proteins because of the modular nature of their surface.[2] Several repeat proteins bind peptides, such as HEAT-repeats,[6] Armadillo-repeats[7–10] or TPR-repeats.[11] We found armadillo repeat proteins (ArmRPs) of particular interest, since they bind a peptide in an extended conformation along a continuous surface contributed to by each module,[12] each of which can form contacts to two consecutive amino acids.[13–15]

The ArmRP family received its name when the first member discovered was found to be encoded by the *armadillo* locus, the DNA region that codes for a set of segment polarity genes required during *Drosophila* embryogenesis.[16,17] This protein is now recognized as the *Drosophila* homolog of β-catenin, involved in Wnt signaling.[18–20] Importin-α is another important member of the family, recruiting the nuclear localization sequence (NLS) in the classical import pathway of cargo molecules into the nucleus.

Armadillo repeats are made up of 42 amino acids formed by three α-helices, named H1, H2, and H3. Helix H3 forms multiple contacts with the bound peptide, amongst them hydrogen bonds from a conserved asparagine residue to main chain peptide bonds. Other side chains on the binding surface provide the specificity for the peptide sequence. Internal repeats have a solvent-accessible surface and two buried surfaces, where they contact neighboring flanking repeats (Fig. 1). The first and last repeats, called N- and C-terminal capping repeats (or N- and C-caps for short), respectively, have only one buried surface. In case of ArmRPs, the N-terminal cap is shorter than the other repeats, as the N-cap only begins with helix 2.

Several crystal structures of ArmRPs in complex with different NLSs (a representative one is shown in Fig. 1) revealed that the NLS peptide runs antiparallel to the direction of the importin-α main chain and that the NLS peptide crosses helix H3 at an angle of approximately 45°.[13,14] In a first approximation, the complex of the NLS peptide to the ArmRP can be described as an asymmetric antiparallel double helix.

In our efforts to develop ArmRPs with defined binding specificity we initiated a project aimed at creating an ArmRP of utmost stability that will subsequently serve as the scaffold from which libraries are generated to select for specific peptide binding. Previously, Parmeggiani *et al.*[12] designed artificial ArmRPs derived from a consensus sequence, optimizing the hydrophobic core using a computational

approach. The consensus sequence had been obtained by multiple sequence alignments of single armadillo repeat modules from both the importin-α and the β-catenin families to generate a unique stable internal module sequence.

The aim of the present study was to further improve the stability of these proteins. Prompted by earlier studies,[12] in which NMR spectra showed markedly better spectra for these proteins at very high pH, we investigated positions of potential electrostatic repulsions at neutral pH in the internal repeats. Moreover, we focused on the optimization of the N- and C-terminal caps. Previous work in designed ankyrin repeat proteins (DARPins) had demonstrated a significant influence of cap engineering on the overall stability of the protein.[22]

While this study was carried out, no crystal structure of a designed ArmRP was available, and thus it was based on homology models largely derived from importin-α. We used implicit solvent molecular dynamics (MD) as well as explicit water MD to assess the fluctuations of different regions in the protein, notably the caps. Based on these simulations, mutants were constructed and experimentally tested. Using a systematic approach optimizing the electrostatics of the internal repeats and the sequence of the N- and C-terminal caps, proteins could be constructed that are entirely monomeric, posses melting temperatures as high as 85°C, and display biophysical properties as well as NMR spectra characteristic of well-folded and stable proteins at neutral pH.

Results

The goal of ArmRP engineering is to create a stable scaffold for the generation of libraries as the basis for selecting a new type of sequence-specific peptide binders, where the peptides are bound in an extended conformation. For this purpose it is crucial to create an ArmRP scaffold of utmost stability as a starting point, since we expect that mutations required to achieve binding will inevitably lower the stability of the proteins. We describe here a combination of computational and biophysical approaches to design and characterize the mutants.

Initially, attempts to crystallize the various consensus ArmRP designs had been unsuccessful. Suggestions for modifications of the sequences of the internal repeats came from early NMR studies and homology models. Modifications of the caps were largely derived from MD simulations based on homology models. The MD simulations provided insight into the molecular features that affect structural stability of these proteins. Promising mutants were expressed, and assessed by heteronuclear NMR regarding stability and side chain packing, as well as by thermal and denaturant-induced unfolding observed by optical spectroscopy. The tight interplay of computational techniques with NMR and other biophysical techniques helped to rapidly improve the stability of the ArmRP.

To succinctly describe the proteins with regards to repeat identity and repeat numbers we have introduced a shorthand nomenclature, which should be consulted in Materials and Methods. The protein that was at the start of our studies is termed YM_4A.

Optimization of the internal repeats using heteronuclear NMR spectroscopy

$^{15}N,^{1}H$ heteronuclear NMR spectroscopy represents a suitable tool to investigate the state of folding of small to medium-sized proteins. Although other biophysical data have indicated that YM_4A is a well-folded protein,[12] spectra recorded at close to neutral pH displayed very broad lines, indicating the presence of conformational exchange processes. Interestingly, when the pH was adjusted to a value of 11, most of the peaks in the 2D NMR spectrum appeared well-resolved and narrow. Due to accelerated amide exchange at that pH, however, peaks arising from Gly residues, which are mostly located in loops and hence not protected from exchange, disappear [Fig. 2(c)]. As a result signal dispersion in the ^{15}N dimension is limited.

We suspected that the pH dependence of the NMR spectrum may be attributed to the titration of Lys residues, for which side chain pK_as are typically about 10.5. Accordingly, at pH lower than 10, the ε-amino group is charged, resulting in unfavorable side-chain packing. Two Lys residues at positions 26 and 29 in each repeat are arranged such that they may form repulsive interactions between the repeats (Fig. 3). A series of mutants in which, in each repeat, either one of these two Lys residues (data not shown) or both were replaced by Gln [Fig. 2(d)] indicated that the best spectra were obtained when both Lys residues were replaced. The comparison of spectra of YM_4A (containing both Lys, thus KK-type) [Fig. 2(a)] and $Y\overline{M}_4A$ (both mutated to Gln, thus QQ type) [Fig. 2(d)] at pH 8.0 clearly illustrates the much-improved properties of the QQ-mutant (see Materials and Methods for nomenclature). We would like to emphasize here that no assignments were required at this stage of NMR analysis. In subsequent work the QQ-mutant was used as the scaffold for further optimizations.

MD simulations suggest mutations at the N-cap and C-cap that result in improved protein stability

A series of MD simulations was carried out to provide suggestions for additional mutations aimed at improving the general stability of the scaffold. An initial explicit water simulation with the model of the original YM_4A (KK-type) provided evidence that the overall fold was preserved during the trajectory.

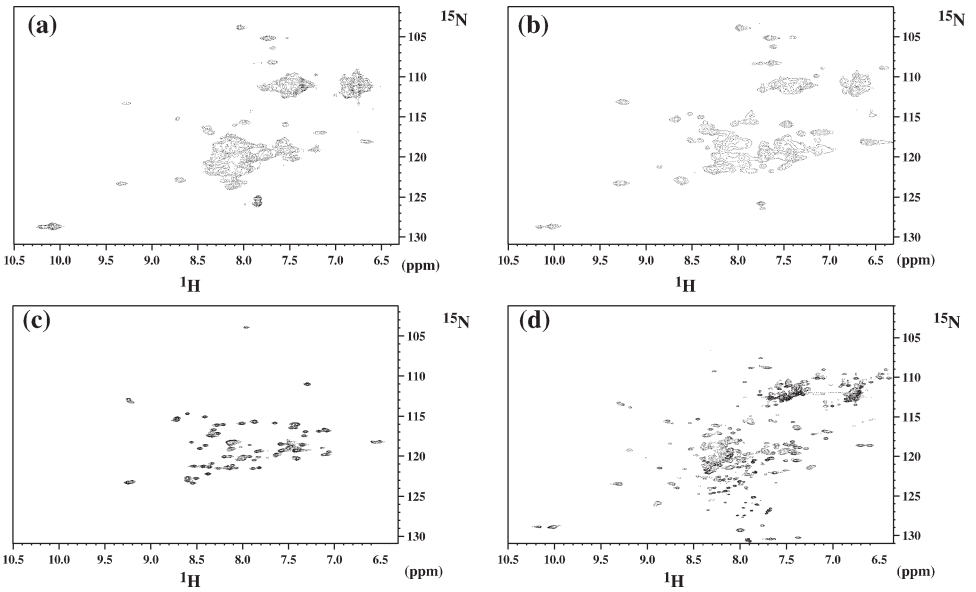

Figure 2. (a) to (c): Representative [^{15}N,^{1}H]-HSQC spectra of YM_4A recorded at various values of pH: Top left: pH 8.0, top right pH 9.0 and bottom left pH 11.0. Panel (d) displays the spectrum of $Y\overline{M}_4A$, (= YM_4A with the mutations K26Q,K29Q in every repeat = QQ type) at pH 8.0 for comparison.

However, two water molecules permeated the interface R4/C-cap close to a buried glutamine (Q240) (Fig. 4). In the crystal structure of β-catenin (2BCT), this position is occupied by a methionine (M662), which is also buried. Furthermore, the C-cap displayed higher conformational instability than the internal repeats in both the implicit and explicit solvent simulations (Supporting Information Figs. S2, S3, and S4). These simulation results were used to suggest the Q240L mutation C-cap, but the Met mutant was also tested experimentally (see below). At the same time, the simulation suggested to mutate the solvent-exposed F241 to glutamine (Supporting Information Fig. S2). The C-cap containing the mutations Q240L and F241Q is termed "A_{II}".

As the NMR data indicated that the QQ-mutant displays better side chain packing at neutral pH, a QQ-model ($Y\overline{M}_4A$) was derived from the KK model (YM_4A). The RMSF plot of the KK-model showed high flexibility in both the N- and the C-cap (see Supporting Information Fig. S2). To reduce the flexibility of the N-cap, three mutations were introduced in the QQ-model (Supporting Information Fig. S5). Their positions in the sequence are shown in

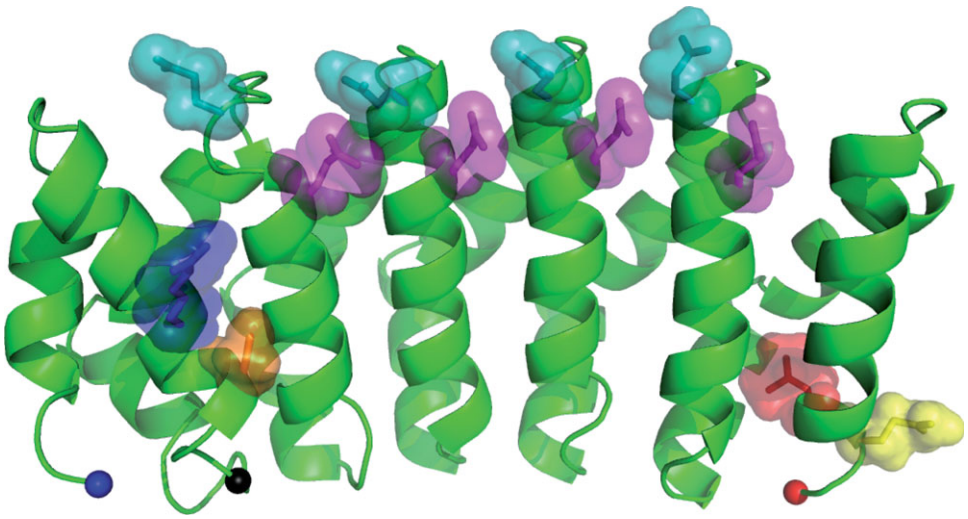

Figure 3. $Y_{II}\overline{M}_4A_{II}$ (QQ-type) model displaying the location of the stabilizing mutations as sticks. In the N-cap: R24 (blue) and S27 (orange), the deletion site of R32 is marked by a black ball. In the four internal repeats: Q59, Q62, Q101, Q104, Q143, Q146, Q185, and Q188; the glutamine at position 26 of each repeat is depicted in cyan, the one at position 29 in magenta. In the C-cap: L240 (red) and Q241 (yellow). The locations of the N- and C-terminus of the protein are marked by blue and red balls, respectively.

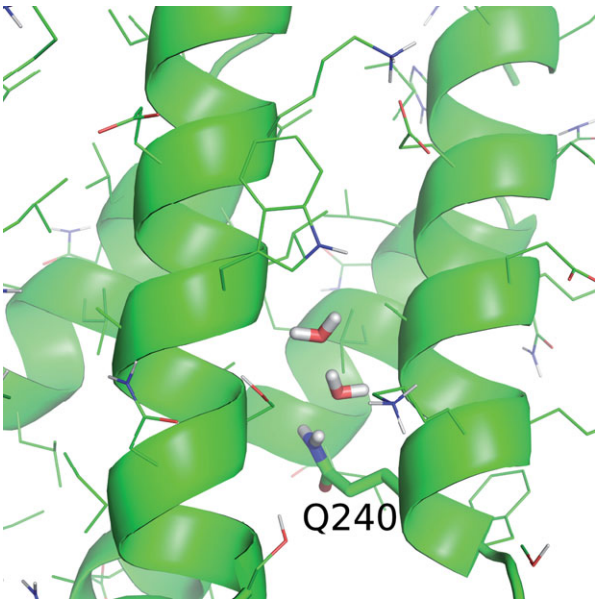

Figure 4. Water molecules permeate into the R4/C interface. In the explicit water simulation of YM_4A, water molecules permeate into the hydrophobic surface between the fourth internal repeat and the C-cap, close to buried Q240.

Supporting Information Figure S1. The V24R mutation was introduced to favor an inter-repeat salt bridge with E64, and to remove the solvent exposed V24. The R27 side chain was replaced by Ser, as found in the internal repeats at this position. The loop connecting the N-cap with the first repeat is one residue longer than the ones between the internal repeats (Supporting Information Fig. S1). RMSF analysis showed that the backbone R32 is highly flexible. Hence, this residue was deleted to match the length of the loops between internal repeats. The N-cap with all three mutations is termed "Y_{II}" (cf. nomenclature in Materials and Methods).

The mutations investigated *in silico* are summarized in Table I. To assess the effects of the mutations on the flexibility of the whole protein, the quasiharmonic entropy was calculated (see Material and Methods section). This quantity can be interpreted as an approximation of the configurational entropy. A reduced value corresponds to a reduction of the flexibility and thus to an increase of structural stability. Surprisingly, the average value of the entropy of the $Y\overline{M}_4A$ model is not significantly lower than that of the YM_4A model [Fig. 5(a)]. In contrast, the $Y\overline{M}_4A$-Q240L mutation provides a significant reduction of entropy. Also, the $Y_{II}\overline{M}_4A_{II}$ model, which contains mutations in the N- and C-caps, has the lowest entropy among all variants, in agreement with the NMR spectra (*vide infra*) and biophysical analysis. Furthermore, to investigate the local effect of these mutations, the entropy of the N-cap/R1, R1/R2, R2/R3, R3/R4, and R4/C-cap pairs was calculated [Fig. 5(b)]. The trend found when comparing the total entropy of $Y\overline{M}_4A$ and $Y_{II}\overline{M}_4A_{II}$ is reproduced: the quasiharmonic entropy of $Y_{II}\overline{M}_4A_{II}$ is lower than that of $Y\overline{M}_4A$ for all the repeat pairs. Interestingly, when comparing the results for $Y_{II}\overline{M}_4A$ (mutated only in the N-cap) and $Y\overline{M}_4A$-Q240L (mutated only in the C-cap by a single point mutation), the entropy reduction is mainly localized in the mutated capping repeats themselves, without affecting the internal repeats. An overall decrease of whole and local entropy throughout the whole protein is observed only for $Y_{II}\overline{M}_4A_{II}$, in which both caps are modified.

Similar conclusions can be drawn from a comparison of the root mean square fluctuations (RMSF) [Fig. 5(c)] between the mutants and the wild-type YM_4A. Mutations at the N- and C-cap measurably reduce the local flexibility of the backbone at the mutation sites. Interestingly, the QQ mutation in the internal repeats, introduced in the $Y\overline{M}_4A$ model, reduces the flexibility of the wild type N-cap. Moreover, $Y_{II}\overline{M}_4A$ and $Y_{II}\overline{M}_4A_{II}$ models have a comparable flexibility, which is lower than that calculated for the YM_4A and $Y\overline{M}_4A$-Q240L models. These observations support the robustness of the method.

To validate the results of the implicit solvent simulations, three independent 80 ns MD simulations with explicit solvent were run for the YM_4A, $Y\overline{M}_4A$, and $Y_{II}\overline{M}_4A_{II}$ models. Therein, the representative of the most populated cluster obtained from the implicit solvent simulations served as starting conformation. The RMSF profiles along the sequence show similar flexibility for implicit and explicit solvent simulations (Supporting Information Figs. S3 and S4). Moreover, the three simulations seem to have converged as they individually yield similar RMSF profiles.

To further assess the conformational flexibility, global and local entropies were calculated. Similarly to the implicit water simulations, the global entropy plot [Fig. 6(a)] reveals that $Y_{II}\overline{M}_4A_{II}$ is more rigid than $Y\overline{M}_4A$ and YM_4A. The partial entropy

Table I. *Mutants Investigated by MD Simulations*

Format	Name	Residue 26,29[a]	Mutations
NR_4C	YM_4A	KK	–
NR_4C	$Y\overline{M}_4A$	QQ[b]	K to Q at 60, 63, 102, 105, 144, 147, 186, 189[c]
NR_4C	$Y_{II}\overline{M}_4A$	QQ	QQ + V24R, R27S, ΔR32
NR_4C	$Y\overline{M}_4A$–Q240L	QQ	QQ + Q240L
NR_4C	$Y_{II}\overline{M}_4A_{II}$	QQ	QQ + V24R, R27S, ΔR32, Q240L, F241Q

[a] For the residue numbering of internal repeats see Supporting Information Fig. S1.
[b] These eight mutations are collectively called QQ, and the repeat is termed $\overline{M}$. In combination with a Y_{II} cap these positions are shifted to positions 59, 62, 101, 104, 143, 146, 185, 188 due to the deletion of R32.
[c] Numbering of the entire protein.

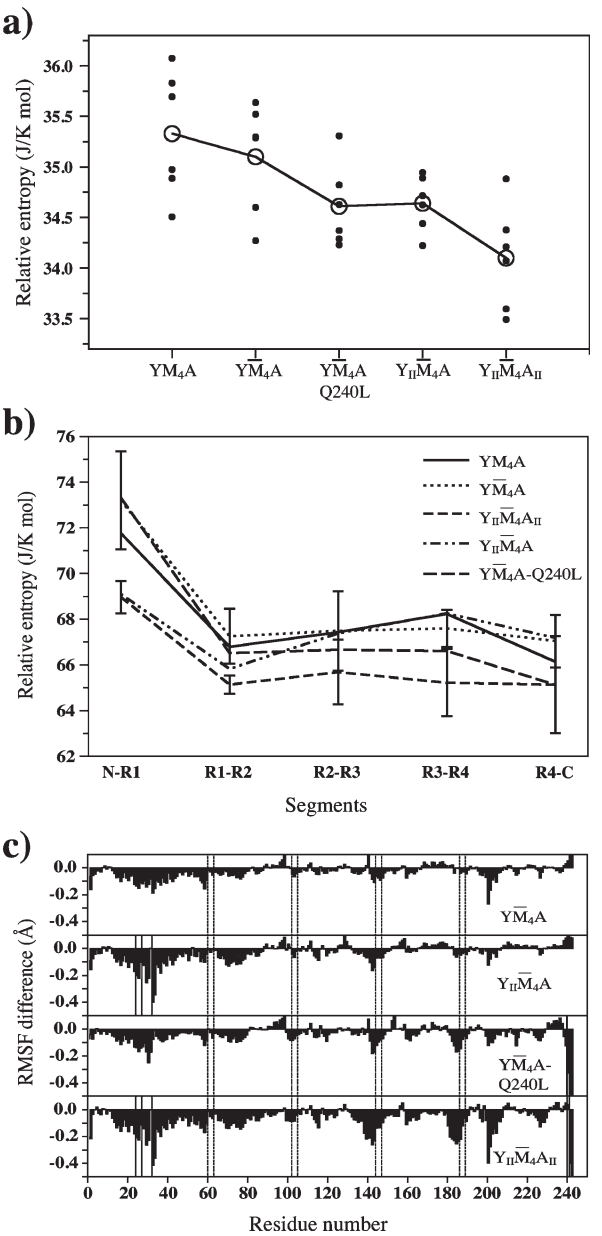

Figure 5. Analysis of implicit solvent MD simulations. Panel (a) displays the per-residue quasiharmonic entropy of YM_4A variants. The small filled circles are the results from single MD trajectories and the bigger open circles present their averages. The entropy values are normalized to the number of residues to allow comparing the models with and without the deletion ΔR32 in the N-cap ($Y_{II}\overline{M}_4A$ and $Y_{II}\overline{M}_4A_{II}$). The mutation labels are used as in Table I. Panel (b) displays changes in quasiharmonic entropy of repeat pairs due to mutations. Error bars represent the standard deviation. Error bars are only shown for $Y\overline{M}_4A$ and $Y_{II}\overline{M}_4A_{II}$ simulations. The entropy values are normalized to the number of residues in the repeat to allow comparison with the ΔR32 deletion mutants. Panel (c) displays differences in RMSFs of the various $Y\overline{M}_4A$ cap variants, using the RMSF of the YM_4A model as reference. Negative values indicate lower fluctuations relative to the reference. The Lys to Gln mutations introduced in the internal repeats (YM_4A to $Y\overline{M}_4A$ mutations) are indicated by vertical dotted lines, while mutations at the N-cap and C-cap are indicated by vertical solid lines.

[Fig. 6(b)] shows a trend similar to the one observed in the implicit solvent simulations [Fig. 5(b)]. However, the average conformational flexibility of the YM_4A-model in the N-cap/R1 repeat pair is lower than for $Y\overline{M}_4A$ or $Y_{II}\overline{M}_4A_{II}$. This result is in disagreement with the implicit solvent simulations, where the flexibility of the N-cap/R1 pair of YM_4A is higher than the one of $Y_{II}\overline{M}_4A_{II}$ [Fig. 5(b)]. This

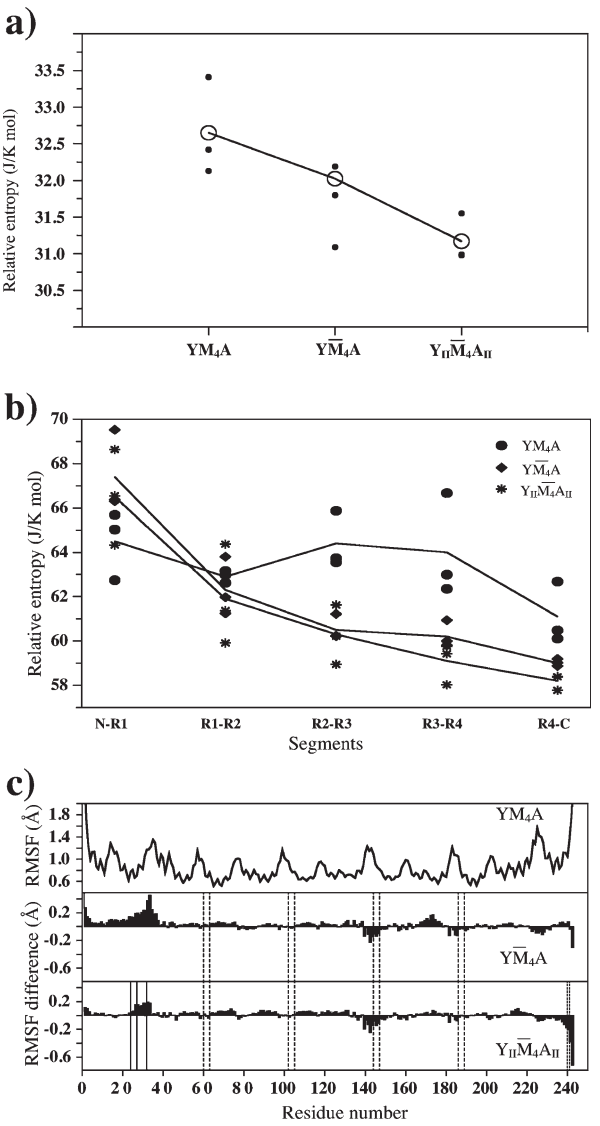

Figure 6. Analysis of explicit water MD simulations. Panel (a) Per-residue quasiharmonic entropy derived from explicit water simulations. Three explicit water simulations were run per model. The small filled circles are the individual calculations and the open circles represent the averages. Panel (b) Effect of the mutations on the per-residue quasiharmonic entropy of repeat pairs in explicit water simulations. Panel (c) RMSF comparison of explicit water simulations. The topmost plot is the RMSF plot of the YM_4A-model. Below, the RMSF difference to the YM_4A-model is plotted for every $Y\overline{M}_4A$ mutant. Negative values indicate lower fluctuations than for the YM_4A-model. Locations of the Lys to Gln mutations in $\overline{M}$ are indicated by dashed lines, mutations in the N-cap and C-cap by solid lines.

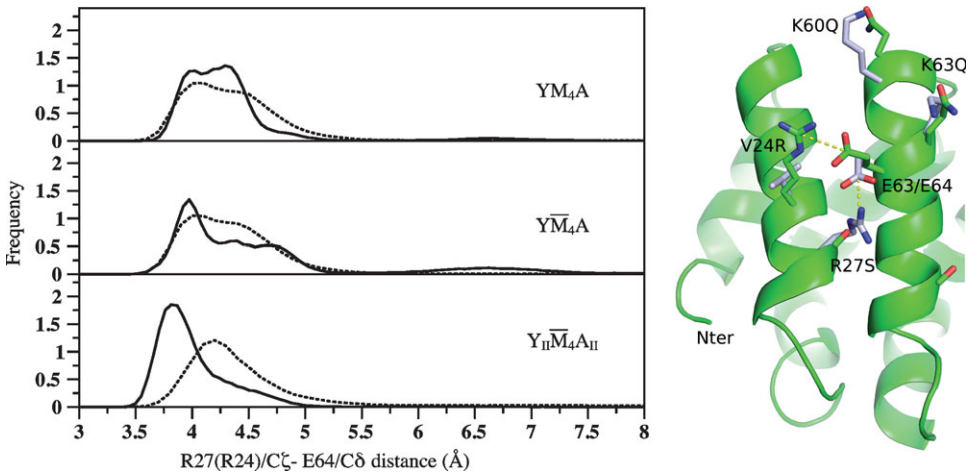

Figure 7. Left: Distance distribution of the salt bridge between the N-cap (R27 in YM_4A and $Y\overline{M}_4A$, R24 in $Y_{II}\overline{M}_4A_{II}$) and the first repeat (E64 in YM_4A and $Y\overline{M}_4A$, E63 in $Y_{II}\overline{M}_4A_{II}$). The solid and dotted lines refer to the explicit and implicit solvent simulations, respectively. The salt bridge distance distribution in the case of R27 (top and middle) is less peaked than in the case of R24 (bottom). Right: Ribbon model of the N-cap and the first internal repeat with the salt bridge and mutations. The conformation of $Y_{II}\overline{M}_4A_{II}$ used for starting explicit solvent simulations is depicted in green and side chains from YM_4A are colored in white after superposition of the N-cap/R1 segments. The salt bridges between R27 and E64 (in YM_4A and $Y\overline{M}_4A$) and R24 and E63 (in $Y_{II}\overline{M}_4A_{II}$) are indicated by dashed lines. The side chains of K60 and K63 are shown to illustrate the two residues of the first repeat mutated to glutamines in the $Y\overline{M}_4A$ and $Y_{II}\overline{M}_4A_{II}$ models.

discrepancy, as well as the slight increase in the flexibility of the N-cap [Fig. 6(c)], is in part a consequence of limited sampling in the explicit solvent simulations. For the other repeat dimers flexibility decreases as $YM_4A > Y\overline{M}_4A, > Y_{II}\overline{M}_4A_{II}$, in agreement with the implicit solvent simulations.

It is interesting to analyze the effects of the double mutation R27S and V24R introduced in the Y_{II} cap on the stability of the salt bridges engaged by residue E64. In the original N-cap of YM_4A we observed that E64 strongly interacts with R27. In Y_{II}, as a result of the structural proximity of the newly introduced arginine and the deletion of R27, the salt bridge is formed with R24 (Fig. 7 right). We measured the stability of the salt bridges as the ratio of MD snapshots where the distance between the Arg-Cζ and the Glu-Cδ is lower than 4 Å. The comparison between the frequency histograms calculated for the three mutants (Fig. 7 left) reveals that the salt bridge introduced in the $Y_{II}\overline{M}_4A_{II}$ sequence is more stable than the original one. It is worth noting that this result is more pronounced in the explicit than in the implicit solvent simulations. The treatment of the long-range electrostatic interactions and solvation effects are more accurate in the explicit solvent calculations, which may have an influence on the salt-bridge distance range, considering the relatively high solvent exposure of the two side chains involved in the salt bridge.

Biophysical characterization of the M- and $\overline{M}$-type proteins

Our investigations aimed at constructing very stable consensus ArmRPs have included analysis of the internal repeats, as well as the capping repeats. When changing the Lys residues at position 26 and 29 of the original M repeat (KK-type) to Gln (individually or collectively, but always in all repeats), we found that the QK version led to aggregating molecules and was not pursued further. Both KQ- and QQ-types displayed improved NMR spectra, with the QQ-type (the $\overline{M}$ repeat) showing the strongest effects [see Fig. 2(d) above]. We thus concentrated on comparing molecules containing $\overline{M}$-type repeats with those based on the original M-type. This was done in the context of many different cap combinations, which will be discussed below.

To compare the biophysical properties of different ArmRP variants, we carried out expression and solubility tests, CD spectroscopy, thermal and chemical denaturation, and [^{15}N,^{1}H]-HSQC NMR analysis. All variants were completely soluble and in this respect comparable with the wild-type protein YM_4A. Expression in *E. coli* XL1-blue at 37°C yielded up to 100 mg/L of soluble protein, with similar results for all variants. Immobilized metal-ion affinity chromatography (IMAC) purification yielded pure protein in a single step, as judged by SDS-PAGE (15%). The expected molecular mass values were confirmed by mass spectroscopy.

The CD spectra of all IMAC-purified protein samples [Figs. 8(a,e) and 9(a), and Supporting Information Fig. S7] display the expected α-helical secondary structure with minima at 222 nm and 208 nm. The mean residue ellipticity (MRE) of the mutants is similar, but those stabilized in the C-cap show a slightly more pronounced peak at 208 nm (Table II).

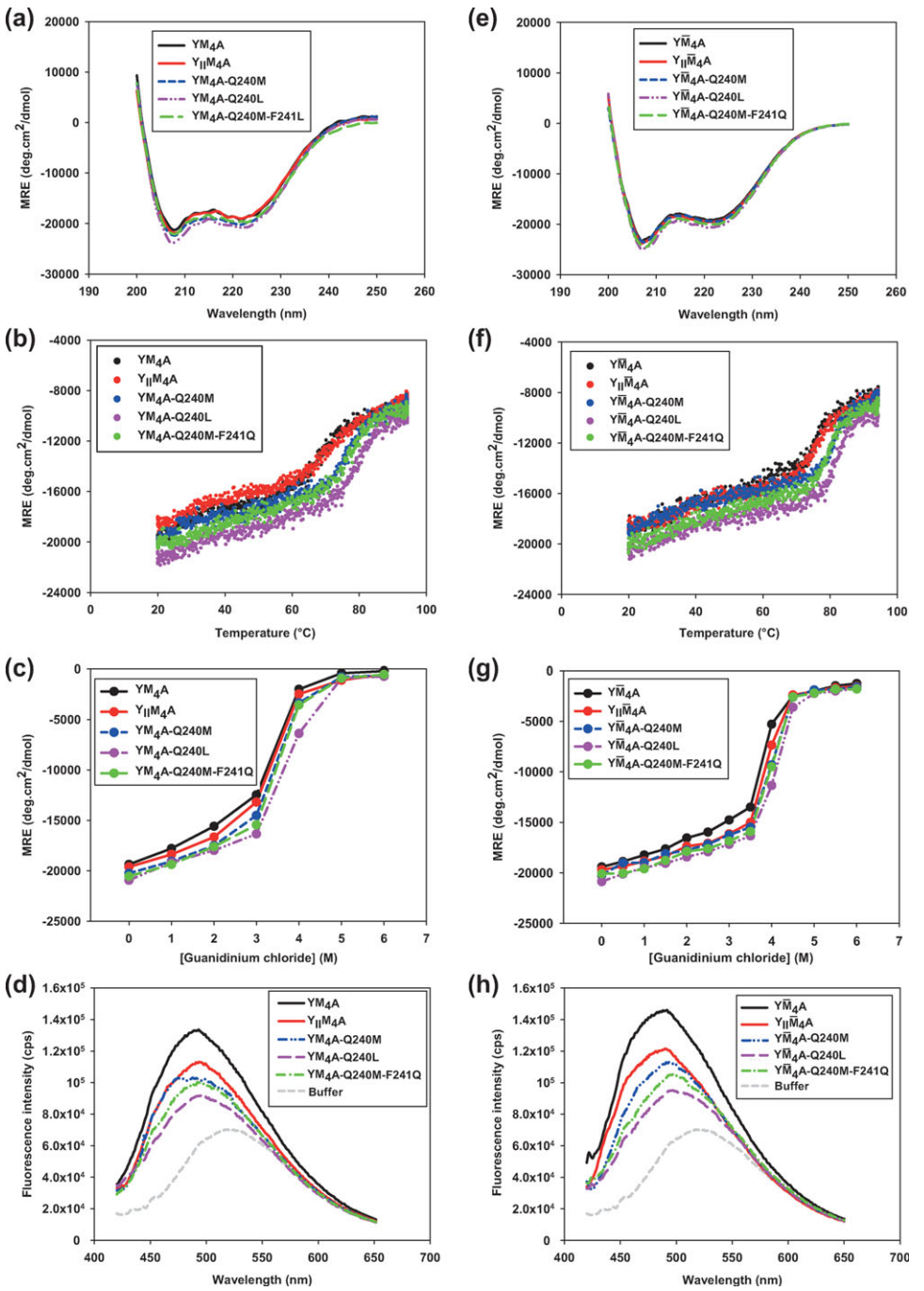

Figure 8. Biophysical characterization of designed ArmRP with different consensus repeats M and $\overline{M}$ and cap variants. (a–d) YM_4A (KK in the internal repeats) and (e–f) $Y\overline{M}_4A$ (QQ in the internal repeats). Identical cap variants have been constructed for both types of internal repeats: Y and Y_{II} for the N-cap; A, A-Q240L, A-Q240M, and A-Q240M-F241Q for the C-cap, as indicated in the figure legends. (a),(e) CD spectra; (b),(f) thermal denaturation curves; (c,g) GdnHCl-induced denaturation curves. The denaturation experiments were followed by CD. The values of MRE at 222 nm are reported. (d),(h) ANS binding. The values without buffer subtractions are shown. The protein concentration was 10 μM.

The CD signal at 222 nm was chosen to monitor thermal and denaturant-induced unfolding. At 10 μM protein concentration, heat denaturation was completely reversible for all proteins (data not shown).

Since both the M and $\overline{M}$ variants with four internal repeats were modified with analogous capping repeats, we could directly compare the influence of the charge repulsion on a variety of biophysical parameters (Fig. 8 and Table II). For all investigated $Y\overline{M}_4A$ constructs, regardless of the caps, the melting temperature is 4–5°C higher than for the corresponding YM_4A constructs. Similarly, the midpoint of GdnHCl denaturation is 0.2M–0.4M higher. This indicates that the removal of the charge repulsion within the internal ArmR is clearly stabilizing the protein. These results also mean that the effect of

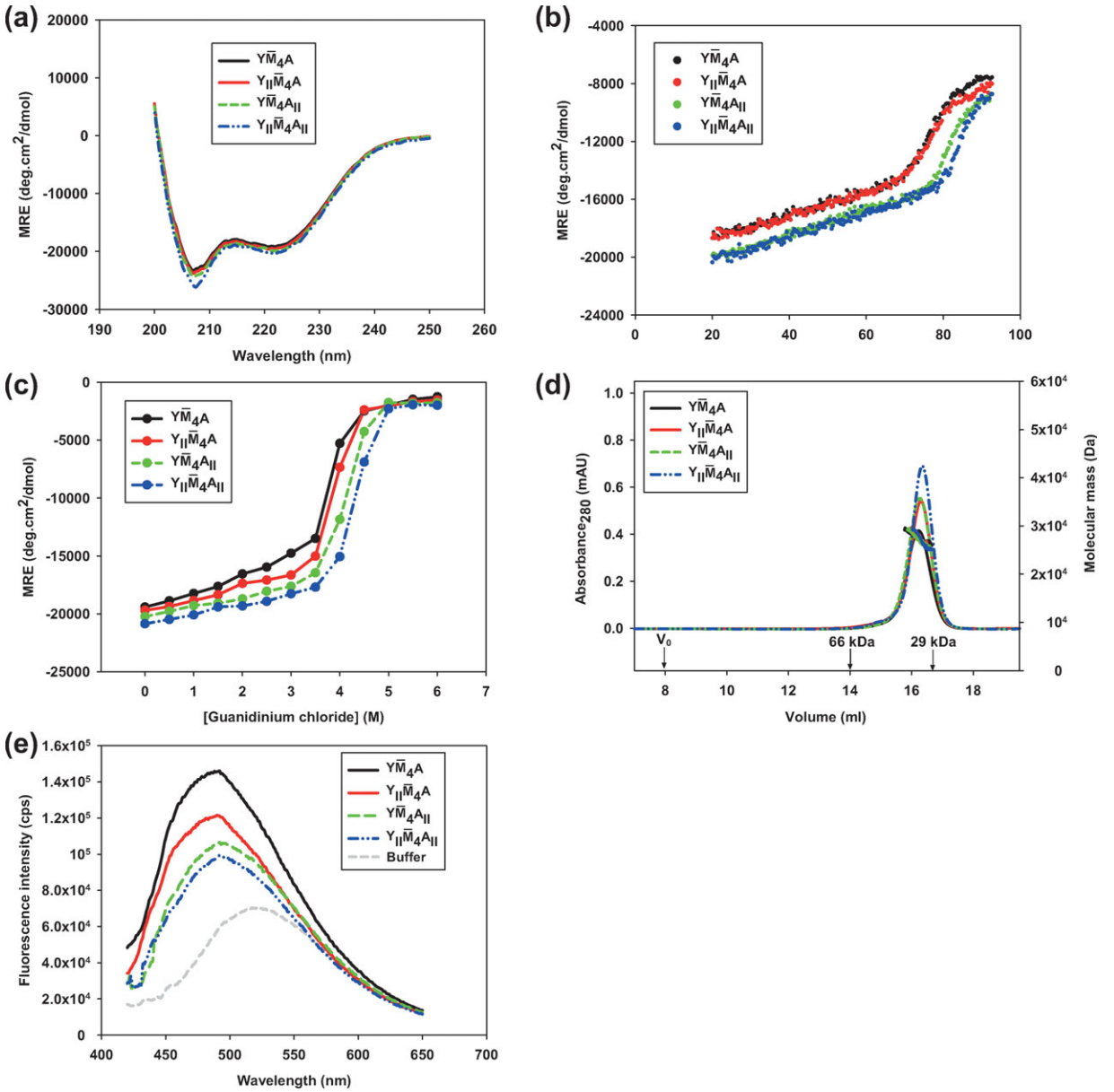

Figure 9. Biophysical characterization of designed ArmRPs $Y\overline{M}_4A$ and its cap variants ($Y_{II}\overline{M}_4A$, $Y\overline{M}_4A_{II}$, and $Y_{II}\overline{M}_4A_{II}$). (a) CD spectra, (b) thermal denaturation curves and (c) GdnHCl-induced denaturation curves. The denaturation experiments were followed by CD. The values of MRE at 222 nm are reported. (d) SEC and MALS of designed ArmRPs. The absorbance at 280 nm from SEC is shown on the left *y*-axis, the calculated MW from MALS on the right *y*-axis. V_0 indicates the void volume of the column. Bovine serum albumin (MW = 66 kDa), and carbonic anhydrase (MW = 29 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by the arrows. (e) ANS binding. The values without buffer subtractions are shown. The protein concentration was 10 μ*M* in a–c and e and 30 μ*M* in d.

the cap variants is quite independent of the internal repeats, thus offering two independent and additive measures to increase stability in ArmR proteins.

We also compared the hydrophobicity of the proteins by evaluating the binding to the fluorescent dye 1-anilino-8-naphthalene sulfonate (ANS) that binds to solvent-exposed hydrophobic patches or to pockets of molten-globule state proteins.[23] When comparing YM_4A and $Y\overline{M}_4A$ constructs with the same caps, ANS binding was very similar [Fig. 8(d,h)]. The caps themselves, however, do influence ANS binding (see below).

Biophysical characterization of various cap mutants allows identifying mutants with much improved stability

The MD simulations have suggested a set of mutations in the caps that should increase the stability of the protein. For validation the new cap variants were constructed in designed ArmRP with both types of internal repeats, YM_4A and $\overline{M}_4A$. The proteins were expressed, purified and characterized biophysically. All three N-cap mutations were introduced at once to create the second generation "Y_{II}" N-cap: V24R, R27S, and ΔR32 (a deletion mutant).

Table II. *Biophysical Properties of Designed ArmRPs with Different Capping Repeats*

Constructs[a]	Type	Residues (repeats)[b]	pI[c]	MW_{calc} (kDa)[d]	Oligom. State[e]	MW_{obs} (kDa)[f]	$MW_{obs/calc}$[g]	CD_{222} (MRE)[h]	T_m (°C)[i]	ΔT_m (°C)[j]	CD GdnHCl (M)[k]
YM_4A	M	253 (6)	4.5	27.1	Monomer	32.3	1.22	−19255	71.0	0	3.50
$Y_{II}M_4A$	M	252 (6)	4.5	27.0	n.d.	n.d.	n.d.	−19007	72.0	1.0	3.55
YM_4A-Q240M	M	253 (6)	4.5	27.1	n.d.	n.d.	n.d.	−20192	76.0	5.0	3.65
YM_4A-Q240L	M	253 (6)	4.5	27.1	n.d.	n.d.	n.d.	−20763	79.0	8.0	3.80
YM_4A-Q240M-F241Q	M	253 (6)	5.1	27.1	n.d.	n.d.	n.d.	−19577	76.5	5.5	3.65
$Y\overline{M}_4A$-Q240M	$\overline{M}$	253 (6)	4.5	27.1	Monomer	32.5	1.2	−19476	80.5	9.5	4.10
$Y\overline{M}_4A$-Q240L	$\overline{M}$	253 (6)	4.5	27.1	Monomer	32.5	1.2	−20457	82.5	11.5	4.20
$Y\overline{M}_4A$-Q240M-F241Q	$\overline{M}$	253 (6)	5.1	27.1	Monomer	32.2	1.19	−20018	81.0	10.0	4.10
Cap combinations											
$Y\overline{M}_4A$	$\overline{M}$	253 (6)	4.5	27.1	Monomer	32.3	1.19	−19162	76.0	5.0	3.70
$Y_{II}\overline{M}_4A$	$\overline{M}$	252 (6)	4.5	27.0	Monomer	31.6	1.17	−19553	77.5	6.5	3.80
$Y\overline{M}_4A_{II}$	$\overline{M}$	253 (6)	4.5	27.1	Monomer	31.4	1.16	−19921	83.0	12.0	4.25
$Y_{II}\overline{M}_4A_{II}$	$\overline{M}$	252 (6)	4.5	26.9	Monomer	31.2	1.16	−20401	85.5	14.5	4.40
$Y\overline{M}_3A$	$\overline{M}$	211 (5)	4.7	22.8	Monomer	27.5	1.21	−17714	70.0	–	2.80
$Y_{II}\overline{M}_3A$	$\overline{M}$	210 (5)	4.6	22.6	Monomer	28.0	1.24	−18096	72.0	–	2.90
$Y\overline{M}_3A_{II}$	$\overline{M}$	211 (5)	5.2	22.7	Monomer	27.1	1.19	−18579	76.5	–	3.40
$Y_{II}\overline{M}_3A_{II}$	$\overline{M}$	210 (5)	4.6	22.6	Monomer	27.5	1.22	−19015	77.0	–	3.60

[a] Constructs in boldface have been studied by MD simulations (see Table I).
[b] The number of residues includes the $MRGSH_6$ tag; the number of repeats includes capping repeats.
[c] Isoelectric point (pI).
[d] Molecular weight calculated from the sequence; masses were confirmed by mass spectrometry.
[e] Oligomeric state as indicated by multiangle static light scattering.
[f] Observed molecular weight as determined by SEC.
[g] Ratio between observed and calculated molecular weight $MW_{obs/calc}$.
[h] Mean residue ellipticity at 222 nm expressed as deg·cm^2/dmol.
[i] T_m observed in thermal denaturation measured by CD.
[j] Difference in T_m relative to YM_4A.
[k] Midpoint of transition in GdnHCl-induced denaturation measured by CD.

Mutations in the C-cap were investigated individually (Q240L or Q240M), and as double mutant (Q240L/F241Q, denoted as A_{II}) (Tables I and II). All mutations were tested in the context of the M and the $\overline{M}$ series, and some mutations were also tested in the $Y\overline{M}_3A$ format (Table II).

The thermal stabilities of the cap variants were compared with the respective precursor proteins YM_4A [see Fig. 8(b)] and $Y\overline{M}_4A$ [Fig. 8(f)]. All proteins displayed a significant slope prior to the main transition and an indication for a cooperative denaturation step at higher temperature.

The modified N-cap in $Y_{II}\overline{M}_4A$ results in a T_m of 77.5°C that is 1.5°C above the transition midpoint of $Y\overline{M}_4A$ wild-type (i.e., T_m = 76°C), suggesting that the N-cap engineering was successful, although its contribution to overall stability is only modest [Table II, Fig. 8(b,f)].

For the C-cap, the replacement of Gln-240 by a hydrophobic residue resulted in a significant increase in stability to 80°C or 82.5°C for $Y\overline{M}_4A$-Q240M or $Y\overline{M}_4A$-Q240L, respectively, compared with $Y\overline{M}_4A$ (Table II). Stability can be further improved by additionally mutating Phe-241 to Gln, with $Y\overline{M}_4A$-Q240M-F241Q and $Y\overline{M}_4A$-Q240L-F241Q (also called $Y\overline{M}_4A_{II}$) displaying transition temperatures of 81 or 83°C, respectively.

We also investigated unfolding induced by GdnHCl [Fig. 8(c,g)]. All proteins displayed cooperative denaturation in these equilibrium-unfolding experiments. The transition point for the curves shifted to higher GdnHCl concentrations for the C-cap mutants Q240M, Q240L, and Q240M-F241Q, both in the YM_4A and the $Y\overline{M}_4A$ format. On the other hand, the transition of constructs with the original Y N-cap was almost identical with those carrying the Y_{II} N-cap, again both in the YM_4A and the $Y\overline{M}_4A$ format [Fig. 8(c,g)]. The most significant shift in the transition midpoint was observed for the Q240L mutation in the C-cap, and this could again be improved further by additionally mutating Phe-241, to result in Q240L-F241Q (also called YM_4A_{II} or $Y\overline{M}_4A_{II}$).

Similar to the results for heat denaturation, equilibrium denaturation by GdnHCl revealed that the influence of the N-cap engineering is rather minor (cf. YM_4A with $Y_{II}M_4A$ or $Y\overline{M}_4A$ with $Y_{II}\overline{M}_4A$), whereas the effect of the C-cap mutation is very significant, with the single mutation Q240L increasing the midpoint of $Y\overline{M}_4A$ from 3.7*M* to 4.2*M* GdnHCl (Table II), and the double mutation present in $Y\overline{M}_4A_{II}$ even to 4.25 M GdnHCl.

The purified proteins differ slightly in their running behavior when analyzed by SDS-PAGE (Supporting Information Fig. S6). Remarkably, the C-cap mutation Q240L and the double mutations Q240L-F241Q present in the A_{II} cap are characterized by a higher mobility in SDS-PAGE, both in the context of

the original M-type and of the $\overline{M}$-type, whereas N-cap mutations have a smaller effect. This faster running behavior suggests a higher compactness of these proteins and/or an incomplete unfolding by SDS.

The consensus-designed YM_4A and $Y\overline{M}_4A$ and their cap variants display different behavior in ANS binding experiments. The difference between the curves of corresponding constructs differing only by the Q240L mutation [Fig. 8(d,h)] indicates that this mutation in the C-cap reduces the hydrophobic solvent-exposed surface or accessible interface. The mutation probably stabilizes the hydrophobic core indicated by the increase in the midpoint of transition both in thermal and GdnHCl- induced denaturation [Fig. 8(b–c, f–g)].

Biophysical characterization of cap combinations

Having established that the Y_{II} N-cap and the A_{II} C-cap variants result in the highest improvements in stability, it became of interest to test whether the observed effects are additive or even synergistic. We thus generated the combinations $Y_{II}M_4A_{II}$ and $Y_{II}\overline{M}_4A_{II}$ and investigated their properties in more detail.

The stability of the combined cap mutant $Y_{II}\overline{M}_4A_{II}$ was assessed by thermal and GdnHCl-induced denaturation [Fig. 9(b,c)]. $Y_{II}\overline{M}_4A_{II}$ possesses a melting temperature of $T_m = 85.5°C$ [Fig. 9(b) and Table II]. When compared with the variant with the original N-cap ($Y\overline{M}_4A_{II}$), the increase in stability is 2.5°C, or 8°C compared with the variant with the original C-cap ($Y_{II}\overline{M}_4A$). This demonstrates that most of the additional stability is contributed by the engineered C-cap. These data also reveal that the cap improvement is additive to a first approximation, suggesting negligible cooperative interactions throughout the whole protein. In summary, when the engineered Y_{II}- and A_{II}- caps are combined, an increase in the melting point by almost 10°C is observed, compared with $Y\overline{M}_4A$, and almost 15°C are obtained relative to the original YM_4A ($T_m = 71°C$), demonstrating the success of our engineering efforts (Table II).

In the GdnHCl-induced unfolding experiments of $Y\overline{M}_4A_{II}$ and $Y_{II}\overline{M}_4A_{II}$ the transition point for the curves are shifted to higher GdnHCl concentrations, compared with $Y\overline{M}_4A$, whereas the transition of $Y_{II}\overline{M}_4A$ was almost superimposable with that of $Y\overline{M}_4A$ [Fig. 9(c)]. The highest shift in the transition point was observed for $Y_{II}\overline{M}_4A_{II}$, consistent with the data obtained in temperature-induced unfolding (Table II). Again, the effect was only modest for N-cap engineering ($Y\overline{M}_4A \rightarrow Y_{II}\overline{M}_4A$ and $Y\overline{M}_4A_{II} \rightarrow Y_{II}\overline{M}_4A_{II}$ shifted by 0.1*M* or 0.15*M* GdnHCl, respectively), and more pronounced for C-cap engineering ($Y\overline{M}_4A \rightarrow Y\overline{M}_4A_{II}$ and $Y_{II}\overline{M}_4A \rightarrow Y_{II}\overline{M}_4A_{II}$ shifted by 0.55*M* or 0.6*M* GdnHCl, respectively), and the effects were again additive to a first approximation.

The difference between the curves of $Y\overline{M}_4A$ and $Y_{II}\overline{M}_4A_{II}$ in the ANS binding experiments demonstrates that the cap mutations reduce the solvent-exposed hydrophobic surface [Fig. 9(e)]. SEC-MALS analysis displayed single symmetric peaks for all variants, and the determined mass indicates a monomeric state [Fig. 9(d)]. The smaller elution volume than for the globular proteins of the standard (Table II) is thus almost certainly due to the elongated shape of the molecules. Similar trends and results (Supporting Information Fig. S7) were observed when the cap mutations were introduced into $Y\overline{M}_3A$, and are summarized in Table II.

Considering the inherent error in the stability measurements, the data are consistent with a fairly constant gain in stability while going from $Y\overline{M}_3A$ to $Y\overline{M}_4A$, independent of the caps. In summary, we could increase the stability of designed ArmRPs by four *additive* components: by engineering the N-cap, the C-cap, and electrostatics of the internal modules ($M \rightarrow \overline{M}$), and by increasing the number of internal repeats.

Heteronuclear NMR allows to rank $Y\overline{M}_3A$ and $Y\overline{M}_4A$ cap mutants according to their conformational stability

The potential of (heteronuclear) NMR to judge the conformational stability of proteins has been increasingly exploited in the course of structural genomics projects.[24] In this study, 1D ^1H NMR spectra of all proteins were recorded (data not shown) in order to preliminarily evaluate the influence of different mutations or combinations of mutations in the capping repeats of $Y\overline{M}_3A$ and $Y\overline{M}_4A$ with respect to conformational rigidity. Wild-type consensus proteins and their mutants were ranked according to signal dispersion in the amide- and methyl-region as well as the linewidth of their proton resonances. A subset of these, namely the original consensus proteins $Y\overline{M}_3A$ and $Y\overline{M}_4A$, and the improved cap mutants Y_{II} and A_{II} described above (Table II), which all appeared to be well structured in 1D proton NMR spectra, were expressed in uniformly ^{15}N-labeled form and analyzed using [^{15}N,^1H]-HSQC spectra. Since preliminary work (data not shown) had revealed that the single Gln mutants (QK and KQ for pos. 26 and 29 in the M-repeats) displayed less favorable properties, they were not further pursued here.

The repetitive nature of the sequence and the inherently reduced signal dispersion in purely α-helical proteins is expected to result in limited signal dispersion (Fig. 10). This feature is seen particularly well in the center of the spectrum (see the region between 7.9 and 8.4 ppm in the ^1H dimension in Fig. 10). Due to overlap of peaks fewer than the

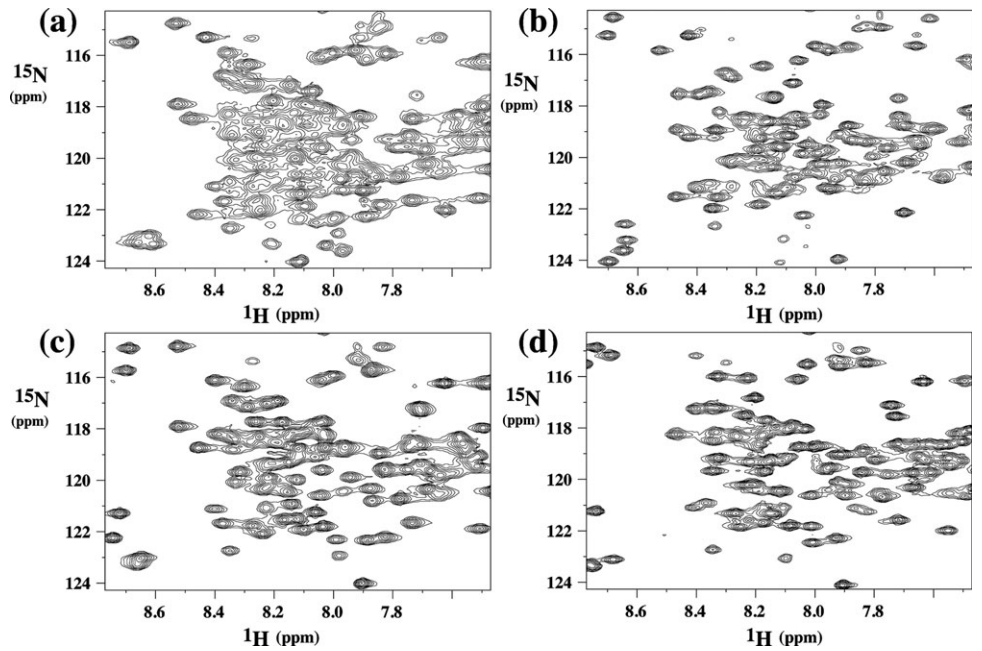

Figure 10. [^{15}N, ^{1}H]-HSQC spectra of designed ArmRP Y$\overline{M}_3$A (a) and its cap variants $Y_{II}\overline{M}_3A$ (b), $Y\overline{M}_3A_{II}$ (c), and $Y_{II}\overline{M}_3A_{II}$ (d) at pH 7.4. All spectra were recorded at 310 K in 50 m*M* phosphate buffer and 150 m*M* NaCl. The protein concentration was 0.5 m*M*.

expected number of peaks were usually observed, for example, for $Y_{II}\overline{M}_3A_{II}$ 170 out of the expected 192 cross-peaks were visible. Nevertheless, signal dispersion is remarkably good, and significantly further improved in the cap mutants, when compared with the original Y$\overline{M}_3$A and Y$\overline{M}_4$A. The line widths suggest that all proteins are monomeric, in agreement with results obtained by size-exclusion chromatography and MALS experiments [Fig. 9(d) and Supporting Information Figure]. Interestingly, the effects due to the C-cap mutations Q240L and F241Q (A_{II}) again are stronger than those of the N-cap mutations (V24R, R27S, and the deletion of R32; Y_{II}), a feature that was also observed in the MD simulations and in the biophysical characterization of the mutants. The combination of N- and C-cap mutations displays a synergistic effect, resulting in the best signal dispersion and comparably narrow lines for $Y_{II}\overline{M}_3A_{II}$ (Fig. 10).

Spectra for the $\overline{M}_4$ series displayed similar trends although the increase in line width due to the larger size was significant (Supporting Information Fig. S8). Again, the results for the $Y_{II}\overline{M}_4A_{II}$ construct are consistent with the observations from equilibrium unfolding studies and the MD simulations.

Discussion

Engineering of proteins for increased stability is a prerequisite for using them as a starting point for randomization, as is needed in the creation of libraries for the selection of binding molecules. Although we have used consensus engineering initially[12] and have already applied a computationally guided optimization of the hydrophobic core of the internal ArmRs, the stability of the resulting proteins was still unsatisfactory.

Herein we have developed a method in which stability of proteins is improved using a rational approach that results in the expression of only a few mutants but nevertheless very effectively increased the stability. The approach uses MD simulations based on homology models of the repeat proteins to provide important information for suggesting the mutations. Furthermore, heteronuclear NMR helped to detect a charge repulsion problem in the internal repeats that resulted in destabilization of the protein and improper side-chain packing. In general, NMR was useful to correctly rank the stability of proteins even in the absence of any backbone assignments.

Improvements were obtained by removal of electrostatic repulsions within the internal repeats. However, cap re-engineering guided by MD simulations made the largest contribution. NMR measurements and a variety of biophysical measurements confirmed that the newly designed N-and C-cap mutants are significantly more stable and better structured. The largest increase in stability is due to modifications of the C-cap, and in particular to the Q240L mutation, as demonstrated by thermal and chemical denaturation experiments [Fig. 8(b,c,f,g)]. Furthermore, NMR measurements confirm that the newly designed Y_{II} and A_{II} mutants (as present in $Y_{II}\overline{M}_4A_{II}$ and $Y_{II}\overline{M}_3A_{II}$) are significantly more stable and better structured than the corresponding initial constructs Y$\overline{M}_3$A and Y$\overline{M}_4$A. The reduced line width observed in the [^{15}N,^{1}H]-HSQC spectra is most likely due to better packing of side chains. Hence,

the NMR data are in good agreement with predictions from MD simulations and results from thermal and chemical unfolding experiments and the ANS-binding behavior of the tested proteins. The more stable A_{II} cap therefore "couples" better to the rest of the protein. In summary, the weak link in the artificially designed original C-cap has been strengthened by our engineering, inspired by MD simulations.

Apparently, the better packing of the C-cap against internal repeats due to this modification prevents local unfolding events that may eventually trigger complete unfolding. This observation is supported by results from our previous study of proteins with Ankyrin repeats, in which we observed a similar influence of the stability of capping repeats on the overall protein stability.[22,25,26] In fact, the Ising model predicts that the stability of these proteins arises from mutual stabilization of neighboring repeats,[25–28] and these effects are therefore expected to propagate throughout the entire protein. In principle, the stabilization should be similar for all repeats, but our experience has shown that the potential for optimization is the largest for the capping repeats. A very important result of this study is that both caps could be re-engineered *independently*, and that improvements resulting from the modified caps were additive (and perhaps synergistic) to a first approximation. In MD simulations a lower flexibility of the internal repeats was seen only when both caps are mutated, otherwise stabilization remained a local effect.

This work highlights several strategies for improving the stability of repeat proteins. Similar to the original work by Parmeggiani,[12] where the hydrophobic core of the internal repeats has been optimized, here the hydrophobic core of the caps was improved. Additionally, electrostatic repulsions in the internal repeats were found to be a main contributor, as shown in the conversion of the M-type modules to the $\overline{M}$-type modules. In the caps an additional attractive interaction has conferred more rigidity. Because of the modular nature of repeat proteins, the improvements in the internal repeats and the caps can easily be combined. Finally, the very simple addition of more internal increased the stability of the repeat proteins.

In summary, this work has brought consensus-designed ArmRPs through various generations of engineering to a point that they can now form the basis of libraries for the construction of sequence-specific peptide binders. Evolved ArmRP, based on the $Y\overline{M}_4A_{II}$ sequence and engineered for binding to neurotensin, allowed their successful study by NMR due to the much-improved stability of these mutations. In contrast, initial work on mutants based on the YM_4A design was unsuccessful because the derived proteins rapidly oligomerized and/or precipitated (data not shown). Our experience therefore underlines the value of optimizing the basic skeleton before introducing mutations for ligand binding. The present work indicates that this optimization process can be guided and accelerated by computational studies.

Materials and Methods

Nomenclature

The consensus armadillo proteins investigated here consist of an N-terminal capping repeat, derived from *y*east importin-α, termed "Y." It is followed by several consensus repeats, which are termed "M" and have been described previously,[12] and their number in the protein is given as a subscript. Finally the protein contains a C-terminal capping repeat, which was *a*rtificially designed,[12] termed "A." A protein with four internal repeats is thus called YM_4A. To indicate improved (e.g., second generation) versions of the capping repeats, the caps are labeled with a roman numeral, for example, $Y_{II}M_4A_{II}$.

The sequences of Y-type N-cap, M-type internal repeat and A-type C-cap are shown in Supporting Information Figure S1. In the present study, we have also investigated the effect of mutating Lys26 and Lys29 to Gln (numbering of individual repeats), individually or in combination. This was always done for every repeat in a protein at once. We thus refer to these two residues in the single-letter code: the original M-type internal repeat[12] with Lys26 and Lys29 is thus referred to as the KK-type, and from this QK, KQ and QQ have been generated. Thus, the YM_4A (QQ-type) sequence carries mutations of lysine residues 60, 63, 102, 105, 144, 147, 186, and 189 (numbering based on the whole protein). To abbreviate the nomenclature further, we refer to the original M-repeat (KK-type) as "M" and the newly engineered M-repeat (QQ-type) as "$\overline{M}$." Thus, the proteins would be termed, for example, YM_4A and $Y\overline{M}_4A$.

When it is necessary to specify an individual internal repeat, R*i* stands for the *i*th internal repeat, and R*i*-R*j* stands for the repeat pair composed by the *i*th to *j*th repeats.

MD simulations

Langevin dynamics simulations were performed at 300 K using the program CHARMM[29] and the implicit solvent FACTS.[30] The protein was modeled according to the united atom CHARMM PARAM19 force field.[31] The protonation state of the side chains was chosen to reproduce pH 7.4 of the CD and NMR experiments: aspartate and glutamate side chains as well as the C-terminal carboxyl group were negatively charged, lysine and arginine side chains together with the N-terminal amino group were positively charged and histidine residues were kept

neutral. All bonds between hydrogen and heavy atoms were constrained using SHAKE,[32] allowing an integration step of 2 fs. Different initial random velocities were assigned to every simulation. Unless differently specified, each simulation consisted of three phases: 0.2 ns heating, followed by 0.4 ns equilibration, and 30 ns production. About 10.5 h on a core of a XEON 5410 Quadcore CPU running at 2.33 GHz are required for a 1 ns trajectory of the KK model (nearly 2220 atoms).

Explicit solvent MD simulations were performed at 300 K using the program CHARMM. The protein was modeled according to the all-hydrogen CHARMM force field (PARAM22 with CMAP correction)[33,34] and TIP3P water model[35] with the same protonation state discussed above. The protein was inserted into a water-filled orthorhombic box whose dimensions were determined such that each atom of the protein had at least 13 Å distance from the boundary. Chloride and sodium ions were added to neutralize the total charge of the system at a concentration of 200 m*M*. To avoid finite-size effects, periodic boundary conditions were applied. Different initial random velocities were assigned to every simulation. Coulombic and van der Waals interactions were calculated up to a cutoff distance of 12 Å, whereas long-range electrostatic effects were accounted for by the Particle Mesh Ewald summation method.[36] The temperature was kept constant by the Nosé-Hoover thermostat,[37,38] whereas the pressure was held constant at 1 atm by applying the Langevin piston. Hydrogens were constrained with SHAKE,[32] allowing an integration step of 2 fs. Lookup tables[39] for the calculation of pairwise nonbonded interactions (van der Waals and Coulomb) were used to increase efficiency.

Clustering of trajectories

Clustering was applied to the MD snapshots (saved every 20 ps) to obtain the most populated conformers for iterative restarting of implicit solvent MD simulations. The first nanosecond of every trajectory was discarded. Pairs of snapshots were compared using the positional root mean square deviation (RMSD) upon optimal structural overlap, and clustering was performed by the Leader algorithm as implemented in the trajectory analysis program Wordom.[40]

The conformations of contiguous repeat pairs were clustered as follows: the N-terminal cap and the first internal repeat (N-cap/R1); the last internal repeat and the C-terminal cap (R4/C-cap); and all the internal repeats (Rn/Rn+1). As the pairs R1/R2, R2/R3, and R3/R4 are topologically identical, the conformations of the internal repeat pairs (R1/R2, R2/R3, and R3/R4) were collected together to increase the statistics and generate a single model for the internal repeat pair. Structures were clustered using the RMSD of C_α atoms (except for the first two residues for the N-cap and the last residue of the C-cap) and Cγ atoms to account for the side chain orientation in the hydrophobic core. We excluded Cγ atoms of lysine, glutamine, asparagine, glutamate, and arginine residues because they are usually exposed to the solvent. Based on visual inspection of the structural dispersion of the most populated clusters, we selected a cutoff for RMSD clustering of 1.5 Å. For each cluster found its representative was extracted as the structure with the lowest RMSD from all the other cluster members.

Trajectory analysis

RMSD and root mean square fluctuation (RMSF) were calculated using as reference structures, respectively, the starting structure used in the dynamics and the structures averaged over 2 ns trajectory segments.

The quasiharmonic entropy was computed from the covariance matrix of the atomic fluctuations[41] using the trajectory analysis program Wordom.[40] Global entropies, calculated on all Cα atoms, were normalized by the number of residues in order to compare models of different lengths (e.g., YM_4A and $Y\overline{M}_4A$ have 243 residues, whereas their variants $Y_{II}M_4A_{II}$ and $Y_{II}\overline{M}_4A_{II}$ have 242 residues). Local entropies were calculated for a subset of atoms spanning individual repeat dimers (i.e., N-cap/R1, R1/R2, R2/R3, R3/R4, and R4/C-cap).

Model generation

The initial armadillo model was derived from three homology models built with Insight II (Accelrys Inc.) by mapping the YM_4A (KK type) sequence onto the crystallographic structure of three natural ArmRPs: yeast karyopherin (importin-α), mouse importin-α, and murine β-catenin (PDB accession codes: 1EE4, 1Q1T, and 2BCT, respectively). A single implicit solvent MD simulation was run for each homology model, whereas for further generation models, six MD simulations were run (data not shown).

The optimization of the initial position of hydrogens and subsequent energy minimization were performed with the CHARMM PARAM19 united atom force field with distance-dependent dielectric function. Loops connecting α-helices were relaxed through four minimization cycles consisting of 100 iterations of steepest descent and 200 steps of conjugate gradient algorithms with gradually decreasing harmonic restraints on the C_α atoms of the helices (i.e., force constants of 10, 5, 1, and 0.1 kcal mol^{-1} $Å^{-2}$).

The system was further optimized using the implicit solvent model FACTS[30] without restraints by 100 steps of steepest descent and 200 iterations of conjugate gradient, followed by an adopted basis Newton-Raphson minimizer, until an energy gradient of 0.02 kcal mol^{-1} $Å^{-1}$ was reached.

Design and synthesis of DNA encoding designed ArmRPs, protein expression and purification

Individual modules for the KK-type were assembled from overlapping primers (Supporting Information Table 1) as described previously[12] and cloned into a vector. Subsequently, to form proteins with identical internal modules, the single modules were PCR-amplified from the vectors and assembled as described.[12] Point mutations at position 26 and/or 29 (KK, QK KQ and QQ) were introduced into the M-type consensus using site-directed mutagenesis (QuikChange, Stratagene). The modules were then digested from the vector with the type IIS restriction enzymes *Bpi*I and *Bsa*I and directly ligated together with similarly assembled original Y and A caps as described previously.[12] *BamH*I and *Kpn*I restriction sites were used for insertion of the whole genes into the vector pPANK and the plasmids were sequenced. For a more detailed description of the cloning procedure see the Methods in Supporting Information.

Protein purification

All unlabeled ArmRP variants were expressed in *E. coli* XL1-blue, and purified as described previously.[12] Proteins for NMR studies were produced in the *E. coli* strain M15 (Qiagen) additionally containing the plasmid pREP4 (encoding *lacI*). Cells were grown in minimal medium with ^{15}N-ammonium chloride as the sole nitrogen source. The medium was supplemented with trace metals, 150 μ*M* thiamine and 30 μg/ml kanamycin and 100 μg/ml ampicillin. Expression and purification by IMAC and gel filtration were performed as described previously.[12] Protein size and purity were assessed by 15% SDS-PAGE, stained with Coomassie PhastGel Blue R-350 (GE Healthcare, Switzerland). The expected protein masses were confirmed by SDS-PAGE and mass spectroscopy. Elution fractions from IMAC were passed over a desalting column (PD-10, GE Healthcare) to remove imidazole from the elution buffer.

Circular dichroism spectroscopy

All CD measurements were performed on a Jasco J-810 spectropolarimeter (Jasco, Japan) using a 0.5 mm or 1 mm circular thermo cuvette. 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 band width of 1 nm. Each spectrum was recorded three times and averaged. Measurements were performed at room temperature unless stated differently. The CD signal was corrected by buffer subtraction and converted to mean residue ellipticity (MRE). Heat denaturation curves were obtained by measuring the CD signal at 222 nm with temperatures increasing from 20 to 95°C (data pitch, 1 nm; heating rate, 1°C/min; response time, 10 s; bandwidth, 1 nm). GdnHCl-induced denaturation measurements were performed after overnight incubation at 20°C with increasing concentrations of GdnHCl (99.5% purity, Fluka) in phosphate-buffered saline (pH 7.4).

ANS fluorescence spectroscopy

The fluorophore 1-anilino-naphthalene-8-sulfonate (ANS) binds to exposed hydrophobic patches or pockets in proteins, thereby increasing its fluorescence intensity. The measurements were performed at 20°C by adding ANS (final concentration 100 μ*M*) to 10 μ*M* of purified protein in 20 m*M* Tris·HCl, 50 m*M* NaCl, pH 8.0. The fluorescence signal was recorded using a PTI QM-2000-7 fluorimeter (Photon Technology International). The emission spectrum from 400–650 nm (1 nm/s) was recorded with an excitation wavelength of 350 nm. For each sample, three spectra were recorded and averaged.

Size exclusion chromatography and multiangle light scattering

The mass and oligomeric state of selected ArmRP was determined using a liquid chromatography system (Agilent LC1100), Agilent Technologies, Santa Clara, CA) coupled to an Optilab rEX refractometer and a miniDAWN three-angle light-scattering detector (both Wyatt Technology, Santa Barbara, CA). For protein separation a 24 ml Superdex 200 10/30 column (GE Healthcare Biosciences, Pittsburg, PA) was run at 0.5 ml/min in PBS. Typically, 50 μl of solution containing 30 μM protein was injected. Analysis of the data was performed using the ASTRA software (version 5.2.3.15; Wyatt Technology).

NMR spectroscopy

Buffers used for NMR measurements of the internal repeat module optimization (KK- to QQ-type) contained 20 m*M* deuterated Tris·HCl, 30 m*M* NaCl, and the pH was adjusted to pH values of pH 8–11 using NaOH. All cap variants were analyzed in PBS buffer containing 150 m*M* NaCl and 50 m*M* sodium phosphate at pH 7.4. Proteins were concentrated to 0.5–1.0 m*M* for NMR measurements.

Proton-nitrogen correlation maps were derived from [^{15}N,^{1}H]-HSQC experiments[42] utilizing pulsed-field gradients for coherence selection and quadrature detection[43] and incorporating the sensitivity enhancement element of Rance and Palmer.[42,43] 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 2.1 and calibrated relative to the proton water resonance at 4.63 ppm, from which the ^{15}N scale was calculated indirectly using the conversion factor of 0.10132900.

References

1. Binz HK, Amstutz P, Plückthun A (2005) Engineering novel binding proteins from nonimmunoglobulin domains. Nat Biotechnol 23:1257–1268.
2. Boersma YL, Plückthun A (2011) DARPins and other repeat protein scaffolds: advances in engineering and applications. Curr Opin Biotechnol 22: 849–857.
3. Lofblom J, Frejd FY, Ståhl S (2011) Non-immunoglobulin based protein scaffolds. Curr Opin Biotechnol 22: 843–848.
4. Clonis YD (2006) Affinity chromatography matures as bioinformatic and combinatorial tools develop. J Chromatogr A 1101:1–24.
5. Spisak S, Guttman A (2009) Biomedical applications of protein microarrays. Curr Med Chem 16:2806–2815.
6. Andrade MA, Petosa C, O'Donoghue SI, Müller CW Bork P (2001) Comparison of ARM and HEAT protein repeats. J Mol Biol 309:1–18.
7. Hatzfeld M (1999) The armadillo family of structural proteins. Int Rev Cytol 186:179–224.
8. Marfori M, Mynott A, Ellis JJ, Mehdi AM, Saunders NF, Curmi PM, Forwood JK, Boden M, Kobe B (2011) Molecular basis for specificity of nuclear import and prediction of nuclear localization. Biochim Biophys Acta 1813:1562–1577.
9. Tewari R, Bailes E, Bunting KA, Coates JC (2010) Armadillo-repeat protein functions: questions for little creatures. Trends Cell Biol 20:470–481.
10. Xu W, Kimelman D (2007) Mechanistic insights from structural studies of beta-catenin and its binding partners. J Cell Sci 120:3337–3344.
11. Cortajarena AL, Regan L (2006) Ligand binding by TPR domains. Protein Sci 15:1193–1198.
12. Parmeggiani F, Pellarin R, Larsen AP, Varadamsetty G, Stumpp MT, Zerbe O, Caflisch A, Plückthun A (2008) Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J Mol Biol 376:1282–1304.
13. 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.
14. 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.
15. Huber AH, Weis WI (2001) The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. Cell 105: 391–402.
16. Perrimon N, Mahowald AP (1987) Multiple functions of segment polarity genes in Drosophila. Dev Biol 199: 587–600.
17. Wieschaus E, Riggleman R (1987) Autonomous requirements for the segment polarity gene armadillo during Drosophila embryogenesis. Cell 49:177–184.
18. MacDonald BT, Tamai K, He X (2009) Wnt/beta-catenin signaling: components, mechanisms, and diseases. Dev Cell 17:9–26.
19. Mason DA, Stage DE, Goldfarb DS (2009) Evolution of the metazoan-specific importin alpha gene family. J Mol Evol 68:351–365.
20. Moroianu J, Blobel G, Radu A (1996) Nuclear protein import: Ran-GTP dissociates the karyopherin alpha-beta heterodimer by displacing alpha from an overlapping binding site on beta. Proc Natl Acad Sci USA 93: 7059–7062.
21. 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.
22. Interlandi G, Wetzel SK, 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 373:837–854.
23. Slavik J (1982) Anilinonaphthalene sulfonate as a probe of membrane composition and function. Biochim Biophys Acta 694:1–25.
24. Montelione GT, Arrowsmith C, Girvin ME, Kennedy MA, Markley JL, Powers R, Prestegard JH, Szyperski T (2009) Unique opportunities for NMR methods in structural genomics. J Struct Funct Genomics 10: 101–106.
25. Kramer MA, Wetzel SK, Plückthun A, Mittl PR, Grütter MG (2010) Structural determinants for improved stability of designed ankyrin repeat proteins with a redesigned C-capping module. J Mol Biol 404:381–391.
26. Wetzel SK, Ewald C, Settanni G, Jurt S, Plückthun A, Zerbe O (2010) Residue-resolved stability of full-consensus ankyrin repeat proteins probed by NMR. J Mol Biol 402:241–258.
27. Zimm BH, Bragg JK (1959) Theory of the phase transition between helix and random coil polypeptide chains. J Chem Phys 31:526–535.
28. Aksel T, Barrick D (2009) Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. Methods Enzymol 455:95–125.
29. Brooks BR, Brooks CL, III, Mackerell AD, Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S (2009) CHARMM: the biomolecular simulation program. J Comp Chem 30:1545–1614.
30. Haberthür U, Caflisch A (2008) FACTS: fast analytical continuum treatment of solvation. J Comp Chem 29: 701–715.
31. Brooks BR, Bruccoleri RE, Olafson BD, Swaminathan S, Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 4:187–217.
32. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical-integration of cartesian equations of motion of a system with constraints—molecular-dynamics of *n*-alkanes. J Comput Phys 23:327–341.
33. MacKerell AD, Jr, Bashford D, Bellott M, Dunbrack Jr RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616.
34. Mackerell AD, Jr., Feig M, Brooks CL, III (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comp Chem 25:1400–1415.
35. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79: 926–935.
36. Darden T, York D Pedersen L (1993) Particle Mesh Ewald—an n.log(n) method for Ewald sums in large systems. J Chem Phys 98:10089–10092.
37. Hoover WG (1985) Canonical dynamics: equilibrium phase-space distributions. Phys Rev A 31:1695–1697.
38. Nosé S (1984) A unified formulation of the constant temperature molecular-dynamics methods. J Chem Phys 81:511–519.

39. Nilsson L (2009) Efficient table lookup without inverse square roots for calculation of pair-wise atomic interactions in classical simulations. J Comp Chem 30: 1490–1498.
40. Seeber M, Felline A, Raimondi F, Muff S, Friedman R, Rao F, Caflisch A, Fanelli F (2011) Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. J Comp Chem 32:1183–1194.
41. Andricioaei I, Karplus M (2001) On the calculation of entropy from covariance matrices of the atomic fluctuations. J Chem Phys 115:6289–6292.
42. Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69:185–189.
43. Keeler J, Clowes RT, Davis AL, Laue ED (1994) Pulsed-field gradients: theory and practice. Methods Enzymol 239:145–207.