doi:10.1016/S0022-2836(03)00896-9 J. Mol. Biol. (2003) 332, 489–503

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins

H. Kaspar Binz, Michael T. Stumpp, Patrik Forrer, Patrick Amstutz and Andreas Plückthun*

Biochemisches Institut
Universität Zürich
Winterthurerstrasse 190
CH-8057 Zürich, Switzerland

*Corresponding author

We describe an efficient way to generate combinatorial libraries of stable, soluble and well-expressed ankyrin repeat (AR) proteins. Using a combination of sequence and structure consensus analyses, we designed a 33 amino acid residue AR module with seven randomized positions having a theoretical diversity of 7.2×10^7. Different numbers of this module were cloned between N and C-terminal capping repeats, i.e. ARs designed to shield the hydrophobic core of stacked AR modules. In this manner, combinatorial libraries of designed AR proteins consisting of four to six repeats were generated, thereby potentiating the theoretical diversity. All randomly chosen library members were expressed in soluble form in the cytoplasm of *Escherichia coli* in amounts up to 200 mg per 1 l of shake-flask culture. Virtually pure proteins were obtained in a single purification step. The designed AR proteins are monomeric and display CD spectra identical with those of natural AR proteins. At the same time, our AR proteins are highly thermostable, with T_m values ranging from 66 °C to well above 85 °C. Thus, our combinatorial library members possess the properties required for biotechnological applications. Moreover, the favorable biophysical properties and the modularity of the AR fold may account, partly, for the abundance of natural AR proteins.

Keywords: ankyrin repeat; combinatorial library; consensus sequence; protein design; scaffold

Introduction

Repeat proteins mediate numerous key protein–protein interactions in nature.[1,2] Their repetitive architecture permits the adaptation of their size and thus their variable and modular binding surface to a target protein, leading to high-affinity interactions. We developed a strategy that exploits this modular architecture for the generation of combinatorial libraries of repeat proteins with novel binding specificities (Figure 1).[3] Our strategy consists of designing a self-compatible repeat module for a given repeat type. In such a repeat module, residues important for maintaining the repeat structure (i.e. framework residues) are defined, while potential target interaction residues are randomized. Such self-compatible repeat modules can then be joined repetitively to yield a stack of repeats. To form repeat domains, the continuous hydrophobic core of this stack is sealed by N and C-terminal capping repeats (Figure 1). We hypothesized that, by using this strategy, libraries of repeat proteins of different lengths with very large and highly diversified interaction surfaces could be generated. Our strategy is thus in clear contrast to the traditional scaffold approach, which would consist of randomizing the surface or loops of a well characterized natural protein whose dimensions are fixed.[4,5]

Ankyrin repeat (AR) proteins constitute a very attractive class of repeat proteins to test our strategy. AR proteins mediate many important protein–protein interactions in virtually all species,[6] and are found intracellularly, extracellularly

Abbreviations used: AR, ankyrin repeat; IMAC, immobilized metal-ion affinity chromatography; PDB, Protein Data Bank.

E-mail address of the corresponding author: plueckthun@bioc.unizh.ch

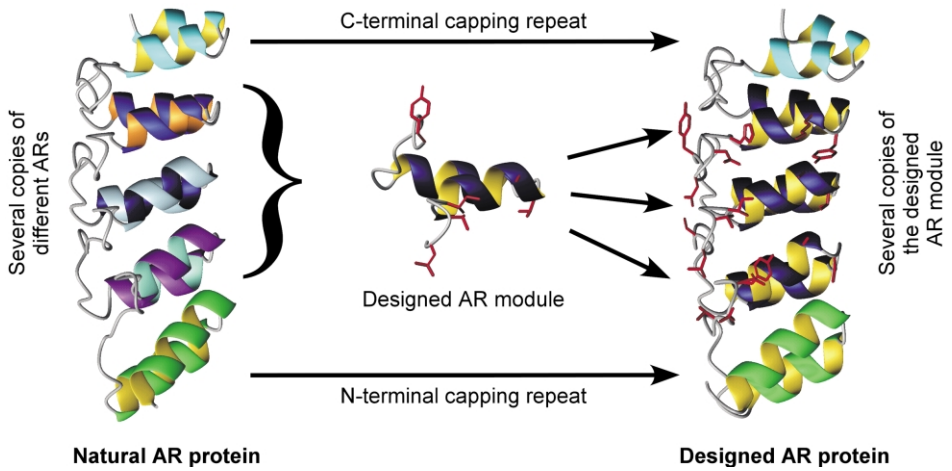

Figure 1. The strategy to generate designed AR protein libraries. From structure and sequence alignments of natural ARs, a self-compatible AR module is designed. This repeat module consists of fixed framework residues and randomized potential interaction residues (shown in red stick mode). Various numbers of this AR module (here three) are then cloned between N and C-terminal capping repeats. By using this strategy, combinatorial libraries of designed AR proteins of varying repeat numbers can be generated. The randomized positions on several adjacent repeats create a large potential interaction surface presented on a rigid AR scaffold. This Figure was created using PDB entries 1AWC[21] and 1MJ0[29] with the help of MOLMOL.[54]

and in membrane-bound form, indicating that these proteins can adapt to many different environments. The fact that there are more than 2000 known AR proteins (>14,000 ARs)[7] underscores their importance in nature. AR domains are composed of stacked ARs, consisting typically of 33 amino acid residues, each forming a β-turn followed by two antiparallel helices and a loop reaching the β-turn of the next repeat (Figure 2).[8,9] Usually, four to six repeats[6] assemble into domains, but the crystal structure of ankyrin R, consisting of 12 ARs in a single domain, was reported recently,[10] indicating that there is virtually no limit to the number of repeats that can fold in one AR domain.

Even though different parts of the surface of AR domains could, in principle, be involved in protein–protein interactions,[10] most AR domains interact with their cognate partners *via* the protruding β-turns and the following α-helices. Typically, several adjacent repeats establish contact. This patch-wise interaction mode leads to high-affinity interactions, exemplified by the mouse GA-binding protein (GABP) β1 binding GABPα with a K_D of 0.78 nM or by IκBα inhibiting the DNA binding of NF-κB with a K_i of 3.1 nM.[11,12]

We implemented our novel strategy on AR proteins and generated combinatorial libraries of designed AR proteins of distinct repeat numbers. Here, we describe the design, construction and analysis of these libraries. The accompanying paper by Stumpp *et al*.[13] describes the implementation of our strategy on leucine-rich repeat proteins, another abundant repeat protein class.

Results

A key step in our strategy (Figure 1) is the design of self-compatible repeat modules.[3] This should ensure the proper stacking of the repeat modules into repeat domains. ARs feature a high degree of sequence and structure conservation and, thus, one way to generate self-compatible repeats is through consensus design. In consensus design, the conserved intra- and interrepeat interactions characteristic for the AR domain fold are implemented into the repeat module. In addition, consensus design may lead to improved repeat stability.[3] We describe here the design of a consensus AR, in which conserved framework residues are fixed and in which potential target interaction residues are randomized. The design is based on sequence and structure consensus analyses.

AR consensus sequence definition using sequence databases

The first important task for our consensus design was to choose an appropriate sequence data set. The SMART database[7] provided a large number of functionally unbiased AR sequences and was therefore chosen as the starting point. Numbering of the positions of the AR consensus was adapted from that used by Sedgwick & Smerdon.[9] The Clustal W[14] alignment of ARs was downloaded from SMART. The data set was further reduced to those sequences matching the length of the 33 amino acid residue consensus described earlier.[6,9,15,16] Only repeats without extra insertions or deletions were considered for the consensus definition. The resulting alignment of 229 ARs yielded consensus A containing residues 3–32 (Figure 3). To further refine consensus A and to define the lacking residues (1, 2 and 33), consensus A was circularly permutated and the lacking residues and those without clear preference (frequency

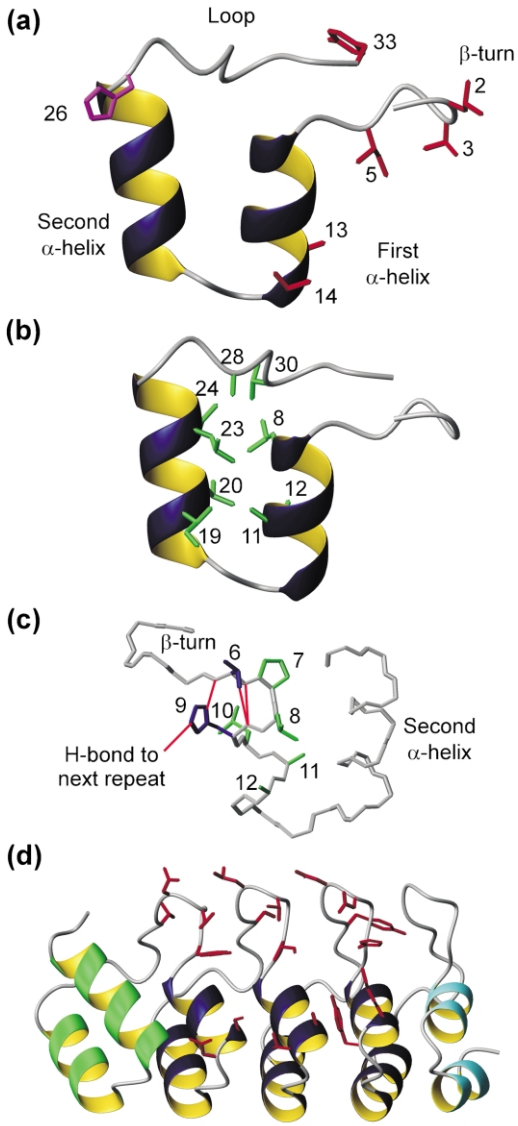

Figure 2. Crystal structure of the unselected N3C library member E3_5 (PDB entry: 1MJ0)[29] illustrated with MOLMOL.[54] (a) Potential interaction residues of the middle AR module (residues 77–109) are displayed in red on the AR framework in ribbon representation. The potential interaction residues are located in the β-turn and the concave surface of the L-shaped repeat. The partly randomized framework position 26 is displayed in magenta. The structural elements of the AR are labeled. (b) Hydrophobic framework residues and alanine residues pointing into the core of the middle AR module are colored in green on the AR in ribbon representation. (c) A rotated view of this middle AR module, which shows more clearly the TPLHLAA motif (residues 6–12) of the first α-helix with its characteristic H-bond pattern. Hydrophobic residues and alanine residues are colored in green, Thr6 and His9 are colored in blue and H-bonds are colored in red. The loop has been cut in this representation. (d) Crystal structure of E3_5 displaying a large potential interaction surface built by the randomized positions (shown in red stick mode). The N and C-terminal capping repeats and the internal repeat modules are colored in green, light blue and dark blue, respectively.

≤30%) were chosen from the alignment of repeats of AR proteins with known structure. The resulting consensus B (Figure 3) was subjected to a BLAST[17] search against GenBank.[18] The resulting top 200 BLAST hits were manually aligned and analyzed, yielding consensus C (Figure 3). An alignment of 2220 AR sequences stored in the PFAM database[19] confirmed the choice of the most frequent amino acids of consensus C (threshold 30%).

AR consensus refinement using structural data

To extend our sequence database analyses, we decided to include structural data for the final refinement leading to consensus D (Figure 3). The structural analysis included the ten AR protein 3D structures 1YCS,[8] 1AP7,[20] 1AWC,[21] 1A5E,[22] 1IKN,[23] 1NFI,[24] 1MYO,[25] 1IHB,[26] 1DCQ[27] and 1SW6.[28] In a first step, the PDB files were used to define potential target interaction residues and framework residues.

Potential target interaction positions

From 3D structures of complexes of AR domains with target proteins (1AWC, 1YCS, 1IKN, 1NFI), target interaction residues were identified using NACCESS† by analyzing the change in solvent-accessible surface area of AR domain residues upon complex formation. Interactions mostly involve the β-turns and the first α-helices of the AR proteins, i.e. positions 2, 3, 5, 13, 14 and 33 of the repeats. In consensus D (Figure 3), these positions were permitted to contain any amino acid except glycine, proline (both structurally unfavorable) or cysteine (may form unwanted disulfide bonds). All other positions in the consensus were defined as framework residues, and we thus tried to assign defined amino acids to these positions.

Framework positions

Positions 1 and 4 were defined as Asp and Gly, respectively, as these residues are frequent (Asp1 37%, Gly4 75% in BLAST search) and form a network of H-bonds extending over consecutive β-turns.[29] Furthermore, position 4 asks for a positive ϕ angle, and is thus best accommodated by Gly. The motif TPLHL for positions 6 to 10 is highly conserved in natural ARs. Thr6 forms several H-bonds to His9 (Figure 2), Pro7 breaks into α-helix 1 and is in the hydrophobic core. Leu8 lies in the hydrophobic core, pointing towards helix 2 and towards the next repeat. In addition to making H-bonds to Thr6, His9 establishes H-bond contact to Ala32 and to the next repeat (randomized position 5; Figure 2). Leu10 points towards helix 1 of the previous repeat, and probably stabilizes the interface between two repeat modules. Since Leu10 is the least conserved

† http://wolf.bms.umist.ac.uk/naccess/

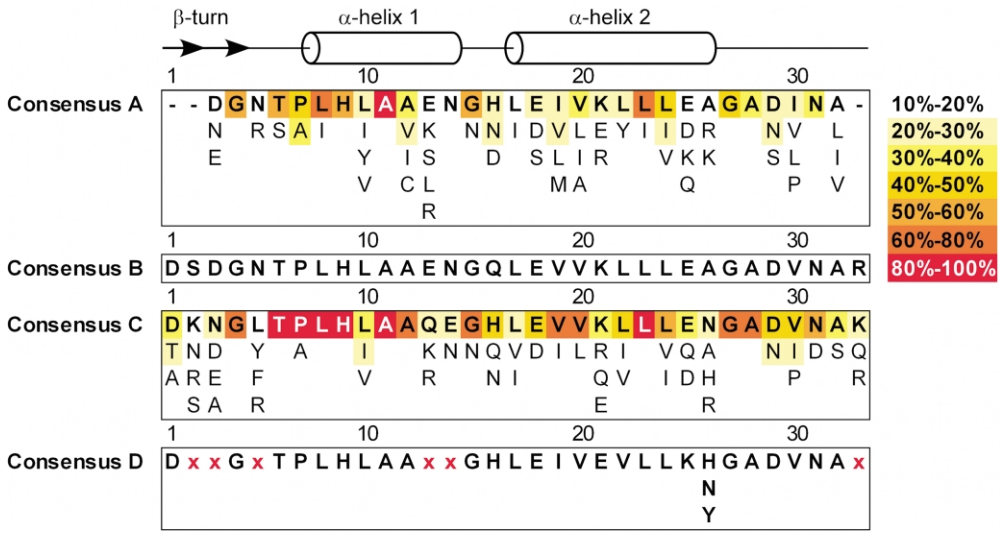

Figure 3. Stepwise definition of the AR consensus used in the present study. The amino acid frequency color code is indicated in the panel. For orientation, the secondary structure elements are indicated above the sequences. Consensus A was derived from an alignment of 229 ARs of the SMART database. It contained only residues 3–32 of the AR consensus. The sequence of consensus B was derived from consensus A, where lacking or non-conserved (cut-off ≤30%) residues were substituted by residues resulting from an alignment of repeats of AR proteins with known structure. A circularly permutated form of consensus B (starting from residue V20) was submitted for a BLAST search against GenBank. Through the circular permutation, residues 1, 2 and 33 could be analyzed and the consensus could be refined. Consensus C was derived from the BLAST search with consensus B. Structure-based considerations (see Results) led from consensus C to consensus D, the final sequence of the designed AR module. In consensus D, the potential target interaction residues are highlighted in red.

residue in the TPLHL motif (30% frequency in the BLAST search), and since it is, in part, surface exposed, it could, in principle, have been considered as potential target interaction residue. The TPLHL motif represents the N-terminal helix-cap of the first α-helix. The small hydrophobic helix-formers Ala11 and Ala12 are important for the overall shape of ARs (Figure 2). Their small size allows the repeat to be conical in form, narrow at the bottom and wide at the top.[10] Gly15 breaks out of helix 2 and has a positive ϕ angle. His16 is semi-buried and forms side-chain H-bonds to helix 1 of the previous repeat as well as backbone H-bonds to Ala11, Ile19 and Val20. Position 17 was not clear from database statistics (Leu being most prevalent with 26% in the BLAST search). From structural considerations, however, it is likely that a leucine residue would stabilize the repeat interface and could be the initiator of helix 2. For these reasons, and because of its high α-helical propensity,[30] we chose Leu at this position. Helix 2 is amphipathic and contributes to the overall shape of the AR. Larger amino acid residues in the hydrophobic core are situated closer to the loop, leading to increasing helix-helix distances. Several positions in helix 2 were not well defined in the consensus from database analyses and were thus assigned using other decision parameters. Glu18 was chosen because it occurs repeatedly in the cdk4 inhibitor p18 (chain B of PDB entry 1IHB)[26] and can thus apparently be tolerated multiple times. Similarly, Glu21 occurs repeatedly in GABPβ1 (chain B of PDB entry 1AWC, here called 1AWC_B).[21] Since positions 18 and 21 are well separated in space, both negative charges should be tolerated. Ile19 was chosen because it fits similarly as Val, but has a higher α-helical propensity.[30] Ile19 is part of the hydrophobic core (Figure 2), and so is Val20, which was defined from sequence analysis. Leu22, Leu23 and Leu24 constitute a rather conserved patch in the upper part of the hydrophobic core (Figure 2). However, position 22 was chosen to be Val, as this occurs repeatedly in GABPβ1. Lys25 was chosen, because it has the opposite charge of Glu21 and, as the latter, occurs repeatedly in GABPβ1. Arg could have also been chosen in this position. Position 26 was ambiguous, although Asn was most abundant. Ala was prominent in the alignment; however, to control the distance of the repeats, the amino acid at this position should fill enough space and should probably be polar. His was an alternative but there was the danger of creating a charge belt Lys25/His26 across the repeat domain. Tyr was another alternative suggested by GABPβ1. Finally, a combination of His, Tyr and Asn was chosen, since these amino acids can be encoded by the HAC codon. Gly27 breaks out of helix 2 having a positive ϕ angle and initiates the loop. Ala28 points into the hydrophobic core to anchor the loop (Figure 2), as do Val30 and Ala32. Asp29 and Asn31 are important for H-bond networking and keeping contact between the consecutive repeats.[29]

The final consensus D is displayed in Figure 3.

When checking pairs, triplets and quadruplets of amino acids of consensus D, all combinations occurred at least once in natural ARs, except the two quadruplets in the HLEIV sequence motif. Taken as a whole, consensus D was designed to encode self-compatible repeat modules that are built from conserved framework residues and randomized potential target interaction residues.

Design of capping repeats

Like other repeat proteins (see the accompanying paper[13]), natural AR domains have specialized terminal repeats (capping repeats) that function to seal the hydrophobic core of a stack of ARs (Figure 4). While the internal ARs have two hydrophobic repeat–repeat interfaces, the capping repeats have only one such interface and the exposed surface is hydrophilic (Figure 4). We thus reasoned that these capping repeats are needed in order to form a stable, well-folded AR domain, and we included capping repeats in our strategy.[3]

We decided to adapt naturally occurring capping ARs to our designed modules. The choice of appropriate capping repeats was based on two criteria: (i) the structure had to be known; and (ii) they had to be as compatible and thus as homologous as possible to our designed repeat module. When joining four, five or six designed repeat modules *in silico* and subjecting these sequences to BLAST searches against the PDB, mouse GABPβ1 was always the best hit. We thus decided to adapt the capping repeats of GABPβ1 for our design. While residues 1–26 of the N-terminal capping repeat, which form the two anti-parallel α-helices, were taken directly from 1AWC_B, the loop sequence GAPFT was changed to GADVNA. There were two reasons for this change: (i) modeling suggested that the bulky GAPFT loop did not sterically match the consensus GADVNA loop of the neighboring repeat. Phe in the GAPFT loop serves as a spacer between the N-terminal repeat and the second repeat of GABPβ1 and appears in combination with Ala in position 26 of the second repeat, where our design had adopted bigger residues. (ii) For cloning purposes, the end of the loop had to contain the sequence Asp-Val. The finally chosen sequence for the N-terminal capping repeat was therefore: DLGKKLLE AARAGQDDEV RILMAN GADV.

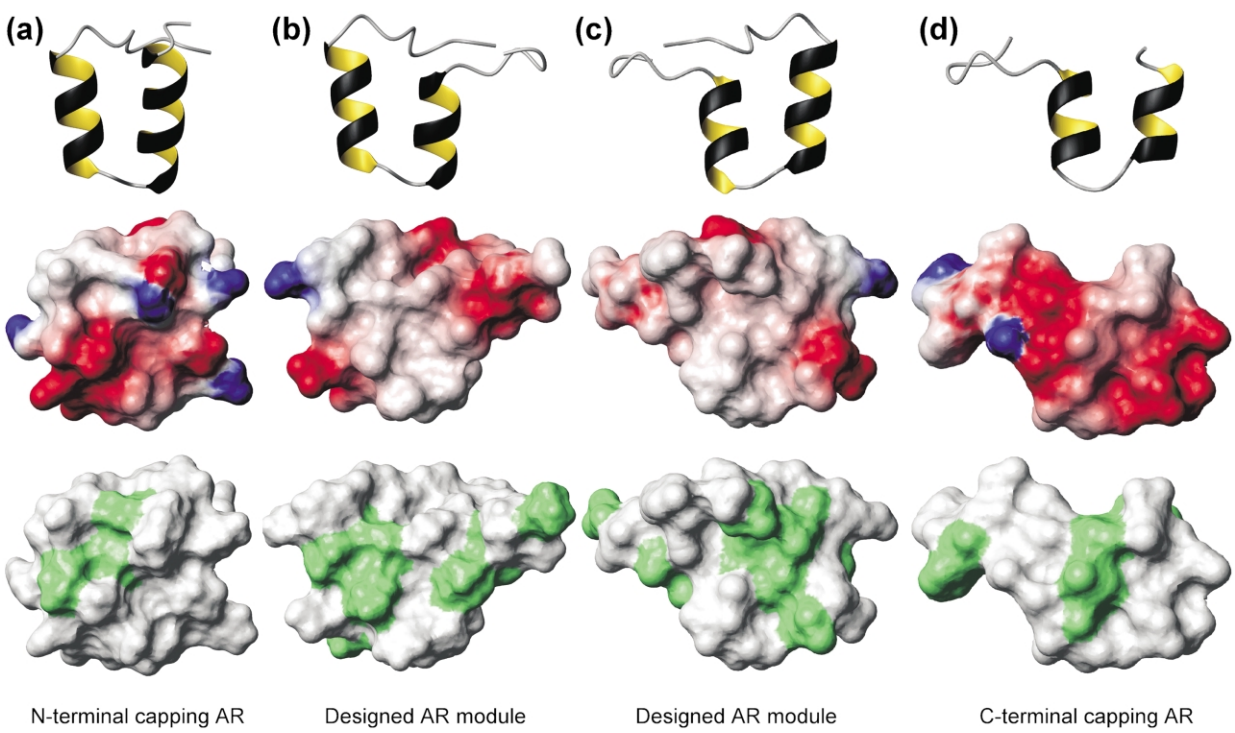

Figure 4. Charge distribution and hydrophobicity of AR surfaces. The charge distribution and hydrophobicity of the middle AR module of E3_5 (PDB entry: 1MJ0),[29] an unselected N3C library member, is compared to the N and C-terminal capping repeats of the same molecule. (a) The N-terminal capping repeat, seen in a lateral view from the N terminus, i.e. from the "outside" of the protein. (b) The middle repeat, seen from the same direction as in (a), i.e. exposing an otherwise buried surface. (c) The middle repeat, seen in a lateral view from the C terminus, i.e. an otherwise buried surface is shown. (d) The C-terminal capping AR seen in the same view as (c), i.e. from the "outside" of the protein. For orientation, the ARs are represented in ribbons on top. In the middle, charge representations are given with negative charges in red and positive charges in blue. The hydrophobicity is illustrated in the bottom row with hydrophobic side-chains in green. The solvent-exposed capping repeats have surfaces that are more charged than the repeat–repeat interfaces of the middle repeat. Likewise, larger hydrophobic areas characterize the repeat–repeat interfaces of the middle repeat in (b) and (c) compared to smaller hydrophobic patches of the solvent-exposed capping repeats in (a) and (d). The Figure and the charge calculations were made using MOLMOL.[54]

The C-terminal capping repeat (Figure 4) consisted of amino acid residues 129 to the end of 1AWC_B. However, the β-turn had to be adapted to our design, and was changed from the sequence SKFC to DKFG. Ser was replaced by Asp to fit the consensus, and Cys was changed to Gly as in the consensus to prevent problems in oxidizing environments. The chosen sequence was: VNAQ DKFGKT AFDISIDNGN EDLAEILQ.

PHD secondary structure prediction[31] of a construct consisting of the N-terminal capping AR, three AR modules and the C-terminal capping AR was in accordance with our design. For this analysis, the randomized positions and position 26 of the AR modules were not defined (i.e. residues submitted as x). In a prediction of degradation by scanning for PEST[32] sequences and by an analysis with PEPTIDESORT of the GCG package,[33] a construct consisting of the N-terminal capping AR, three AR modules and the C-terminal capping AR showed results equivalent to GABPβ1 (in this case, the randomized positions and position 26 of the AR modules were defined using the corresponding residues of GABPβ1).

Assembly of DNA libraries encoding designed AR domains

The peptide sequences of consensus D (Figure 3) and the N and C-terminal capping repeats were backtranslated into DNA using codons optimal for *Escherichia coli* expression.[33] Multiple copies of a single base were prevented if possible. The codons of the randomized positions (2, 3, 5, 13, 14 and 33) of the designed AR module were encoded by trinucleotides, since they efficiently restrict variability by encoding library positions with a defined mixture of specific base triplets.[34] Using this strategy, we allowed A, D, E, H, K, N, Q, R, S, T with 7% probability each, and F, I, L, M, V, W, Y with 4.3% probability each. The randomized framework position 26 was defined by the degenerate codon HAC, which codes for His, Tyr or Asn.

The modular structure of repeat domains suggests assembling them in a stepwise fashion. Hence, the capping repeats and the designed AR module were assembled separately. The constant N and C-terminal capping repeats were PCR-assembled and subcloned individually into pPANK (see Materials and Methods). The designed AR modules were PCR-assembled and subcloned for sequence analysis; five of eight modules showed no error. To construct DNA cassettes encoding whole AR domains, PCR-assembled designed repeat modules were ligated stepwise to the previously PCR-assembled N-terminal capping repeat by using type IIs restriction enzymes (Figure 5). By this strategy, DNA pools encoding the N-terminal capping repeat and two (N2), three (N3) or four (N4) designed AR modules were obtained. These ligation products were then cloned into a vector containing the cloned C-terminal capping repeat to obtain DNA encoding full-length proteins (Figure 5). The full-length proteins were termed N2C, N3C and N4C, reflecting their content of two, three or four repeat modules, respectively, between the N and the C-terminal capping repeats (resulting in four, five or six repeats in total).

With the seven randomized positions per designed AR module, the theoretical diversity amounts to $3 \times 17^6 = 7.2 \times 10^7$ per repeat. The N2C and N3C libraries will thus have theoretical diversities of $(3 \times 17^6)^2 = 5.2 \times 10^{15}$ and $(3 \times 17^6)^3 = 3.8 \times 10^{23}$, respectively.

Sequence analysis of unselected library members

Having cloned libraries of AR proteins of distinct repeat numbers, we wanted to assess their quality at the DNA level. Analysis of single library members should reveal possible sequence bias. DNA sequencing showed that eight of 14 N2C constructs, six of 19 N3C and four of 19 N4C were correct at the DNA level (i.e. no frameshift, no stop codon, correct framework residue codons and correct trinucleotide codons). The percentage of correct clones decreased with increasing repeat number, as expected. Sequencing of 28 erroneous constructs revealed that 11 errors were located in the N-terminal capping AR, which was generated by assembly PCR using standard oligonucleotides. This high error rate is probably due to the lower quality of these standard oligonucleotides compared to the oligonucleotides containing trinucleotide mixtures. The remaining mutations were located in the designed repeat modules. Four of the 28 errors resulted in frameshifts.

Each designed repeat module of 33 amino acid residues contained six randomized positions, which were encoded by trinucleotides.[34] In total, 777 randomized positions containing trinucleotides were sequenced, showing an approximately random distribution of the codons. As in the leucine-rich repeat protein libraries used by Stumpp *et al.*,[13] Asn was overrepresented (12.1% *versus* 7% expected). Glu, Gln, Arg and Trp were underrepresented (3.1%, 2.3%, 4.4% and 2.3% found *versus* 7%, 7%, 7% and 4.3% expected, respectively). The other codons were found at a frequency less than 2% different from the expected value (see Materials and Methods). Twelve mutations were observed that were not encoded by trinucleotides (one amber, two Gly, nine other amino acids), but were most probably accumulated during the extensive PCR. Apart from these, no undesired codon (Cys, Gly, Pro, stop) was found in the trinucleotide positions. In rare cases (0.4%), entire trinucleotides were missing. The seventh randomized position, consensus framework position 26, was occupied by 30% His, 30% Asn and 40% Tyr (128 positions sequenced). Hence, no clear sequence bias was detectable at any randomized

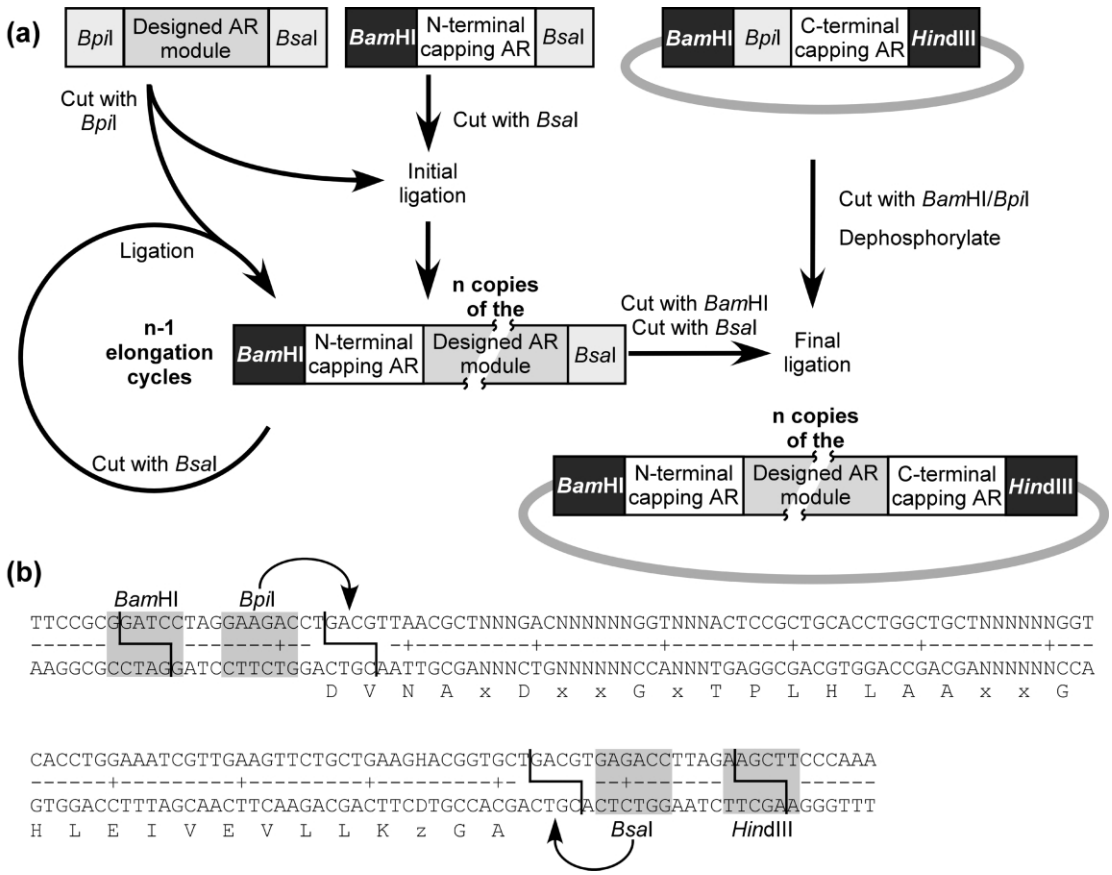

Figure 5. Assembly of designed AR domains at the DNA level and DNA sequence of a designed AR module. (a) The N-terminal capping AR and the designed AR modules are generated by assembly PCR. The N-terminal capping AR is ligated to the first AR module using the type IIs restriction enzymes *Bpi* I and *Bsa* I. To the resulting N1 molecule, more AR modules can be ligated step by step, yielding N2, N3, N4 and longer molecules. Once the desired number of AR modules is connected to the N-terminal capping AR, the construct can be cloned into a vector containing the C-terminal capping AR. By this strategy, AR domains of N1C, N2C, N3C, N4C and longer can be generated. The use of type IIs restriction enzymes ensures the seamless junction of the repeats in a directional manner. Type II restriction sites are represented by black boxes, type IIs restriction sites by light grey boxes. The AR module is represented as a grey box, the N and C-terminal ankyrin capping modules as white boxes. (b) The assembly PCR product of a single AR module is shown. The restriction enzyme recognition sites are shown as grey boxes and the cutting sites are indicated with continuous lines. Note that the DNA recognition sites of the type IIs restriction enzymes are distant from their cleavage site, and thus these sites are lost upon cleavage.

position. Importantly, 75% of the designed AR modules were correct.

Biophysical characterization of randomly chosen library members

We wanted to validate both our strategy and our AR library design by the biophysical analysis of unselected library members, i.e. randomly chosen constructs with correct DNA sequences. The analysis consisted of expression and solubility tests, CD spectroscopy and thermal denaturation. Furthermore, equilibrium unfolding and crystallography were performed.[29]

A first expression screening revealed that all of the above library members that were correct at the DNA level could indeed be expressed in soluble form in large amounts in *E. coli*. The corresponding proteins ran at the expected molecular mass position during SDS-15% PAGE. Six of the correct clones, named E2_5 and E2_17 (N2C library members), E3_5 and E3_19 (N3C library members) and E4_2 and E4_8 (N4C library members) were chosen randomly to be analyzed further. Expression at 37 °C (Figure 6) yielded up to 200 mg/l of soluble protein. Immobilized metal-ion affinity chromatography (IMAC) purification yielded pure protein in a single step, as judged from SDS-15% PAGE (Figure 6). The molecular mass values of the proteins were confirmed by mass spectroscopy. At 10 mg/ml in TBS_{150} (pH 8.0; see Materials and Methods), the proteins remained soluble and did not aggregate over several weeks at 4 °C. An IMAC purification of E3_5 gave sufficiently pure material to successfully determine its structure by X-ray crystallography (Figure 2).[29]

Following SDS-15% PAGE, additional bands were occasionally observed with slightly higher or lower apparent molecular mass than expected. Gel-filtration, mass spectroscopy (data not shown)

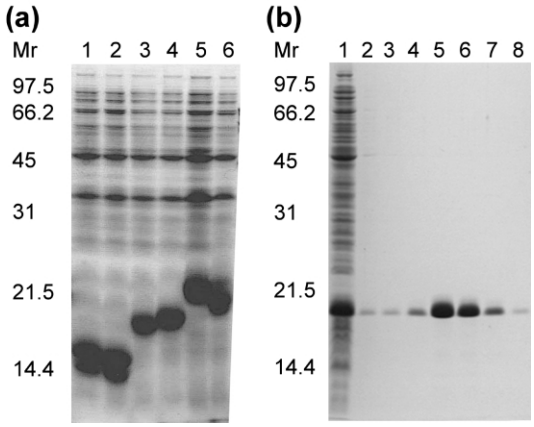

Figure 6. Expression and purification of unselected AR protein library members. (a) Crude extracts of *E. coli* XL1-Blue expressing six consensus AR proteins (see Materials and Methods). Proteins were expressed for four hours and the cell lysates were analyzed by SDS-15% PAGE (lane 1, E2_5; 2, E2_17; 3, E3_5; 4, E3_19; 5, E4_2; 6, E4_8). (b) Single-step IMAC purification of E3_5, an unselected N3C library member (lane 1, column flow-through of the overloaded column; 2, last 1 ml of column wash; 3–8, elution fractions). The size marker is indicated in kDa.

and dynamic light-scattering[29] could not confirm the presence of any protein species other than the expected one. Extensive boiling in SDS and SDS/urea buffer did, however, change the pattern and the relative band intensity in SDS-PAGE analyses,

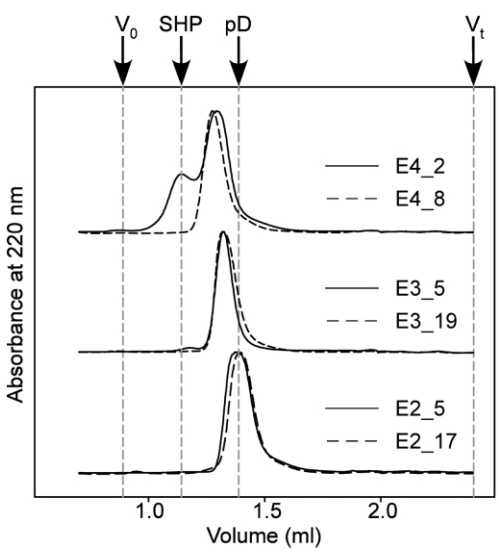

Figure 7. Size-exclusion chromatography of designed AR proteins. The chromatograms of N2C (E2_5 and E2_17), N3C (E3_5 and E3_19) and N4C (E4_2 and E4_8) molecules are shown. All molecules are monomeric, except E4_2, which is a mixture of monomer and (presumably) dimer. The void volume ($V_0 = 0.95$ ml), the total volume ($V_t = 2.4$ ml) and the molecular mass standards (phage protein D with an apparent mass of 17.6 kDa; phage protein SHP, a trimer with an apparent mass of 50.2 kDa)[55] are indicated by broken gray lines in the graph.

Table 1. Table 1 Biophysical data of designed AR proteins of varying length

Protein	CD_{222} (MRE)[a]	MW_{calc} (kDa)[b]	MW_{obs} (kDa)[c]	T_m (°C)[d]	ΔG (kcal/mol)[e]
E2_5	−11,600	14.4	19	79	11.4 ± 0.7
E2_17	−10,300	14.4	18	70	9.5 ± 0.6
E3_5	−11,300	17.7	23	>85	14.8 ± 2.0
E3_19	−12,000	17.8	24	66	9.6 ± 0.5
E4_2[f]	−9400	21.2	30	85	–
E4_8	−11,900	21.3	29	79	21.1 ± 1.3

[a] Mean residue ellipticity (deg cm^2 $dmol^{-1}$) at 222 nm.
[b] As calculated from the sequence.
[c] As determined by gel-filtration.
[d] As determined by thermal unfolding observing the CD signal at 222 nm.
[e] Data from Ref. 29.
[f] E4_2 is a mixture between a monomer and (presumably) a dimer. The monomer value is listed in MW_{obs}.

suggesting that the multiple bands correspond to different conformers, which are stable during SDS-PAGE.

Size-exclusion chromatography showed that five of the six designed AR proteins were monomeric, and only a single protein species was observed (Figure 7; Table 1). The sixth protein, E4_2, could be purified as a monomer. However, it turned into a mixture of monomer and (presumably) dimer over time at 4 °C (Figure 7). The molecular mass values obtained from the gel-filtration studies are given in Table 1. The observed molecular mass is always slightly higher (by a factor of 1.25 to 1.42) than the calculated value, which can be interpreted to reflect the elongated shape of AR domains in combination with a flexible N-terminal tail (MRGS-HHHHHHGS), which leads to increased hydrodynamic radii of the molecules. In addition, the monomeric state of E3_5 was confirmed by its

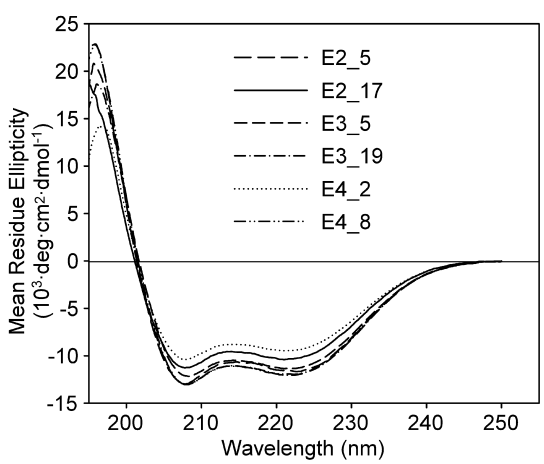

Figure 8. Circular dichroism spectra of designed AR proteins. The spectra of two unselected members from the N2C (E2_5 and E2_17), N3C (E3_5 and E3_19) and N4C (E4_2 and E4_8) library are shown. All proteins exhibit spectra and α-helical content identical with those of natural AR proteins.

crystal structure.[29] We therefore conclude that the majority of the proteins are stable monomers.

The CD spectra of the IMAC-purified protein samples were determined (Figure 8; Table 1). The recorded spectra can be superimposed on the spectra of natural AR proteins such as myotrophin,[35] notch,[36] p19,[37] and an engineered form of p16, p16(Δ1-8)-His.[38] The secondary structure composition of our designed AR proteins thus corresponds to natural AR proteins. In the case of E4_2, oligomerization may influence the spectrum. Combining the CD data with the findings from the protein expression and gel-filtration experiments, we conclude that the designed AR proteins form soluble, monomeric domains having an AR domain fold as designed. For E3_5, the AR domain fold was confirmed by X-ray crystallography (Figure 2).[29]

Thermal stability

To assess the thermal stability of the randomly chosen AR protein library members, heat denaturation was measured by observing the CD signal at 222 nm. All proteins showed cooperative unfolding while exhibiting considerable heat resistance (Figure 9; Table 1). The midpoints of the cooperative transitions in physiological buffer were between 66 °C and more than 85 °C (Table 1). For E4_2, a discontinuity in the CD signal in the pre-transition baseline was observed, probably due to a shift of the monomer/dimer mixture towards a single molecular species. The midpoint of denaturation of E3_5 could not be determined, since the post-transition baseline was not reached after heating the sample to 95 °C. The heat denaturation was only partly reversible for all proteins. The observed high degree of thermal stability reflects the high degree of thermodynamic stability of the designed AR proteins.[29] In GdmCl equilibrium unfolding experiments, the proteins showed cooperative unfolding with midpoints from 2.9 M to 5.1 M GdmCl. Assuming two-state unfolding, ΔG values of unfolding from 9.5 kcal/mol to 21.1 kcal/mol were calculated (Table 1).[29]

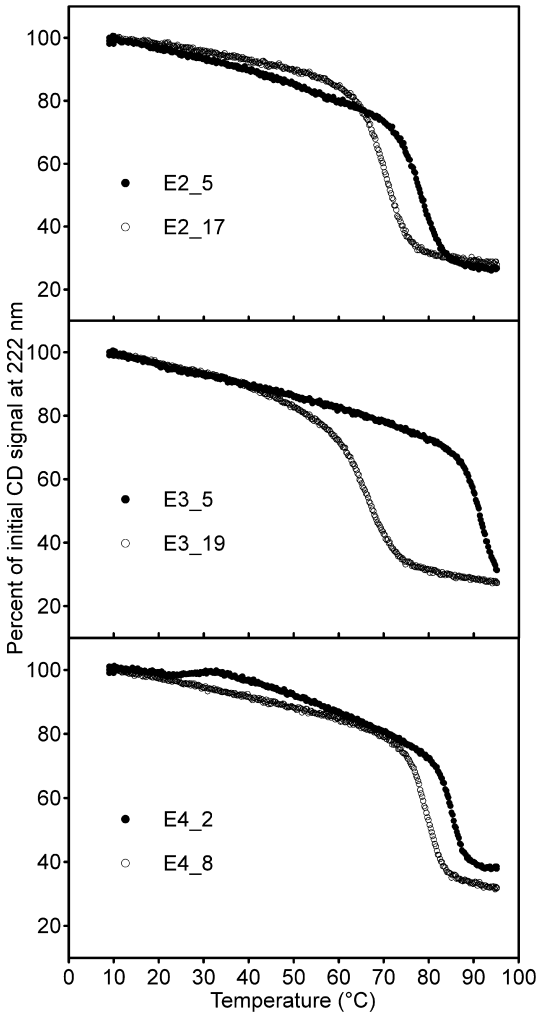

Figure 9. Thermal denaturation of designed AR proteins. Six unselected members of designed AR protein libraries were measured, two each from the N2C (E2_5 and E2_17), N3C (E3_5 and E3_19) and N4C (E4_2 and E4_8) libraries. The denaturation was monitored by observing the CD signal at 222 nm (see Materials and Methods). The CD signal is represented as a percentage of the initial CD signal at 10 °C. Note that this representation makes no assumption about the pre-transition or post-transition baseline.

Module-wise elongation of AR domains

The N3C library member E3_5 was used to demonstrate the feasibility of a module-wise elongation of AR domains by single repeats. Using PCR cloning, we elongated the N3 part of E3_5 to N5C and N6C domains. Again, the corresponding proteins were expressed in large amounts and could be purified in a single IMAC purification step. Moreover, the proteins exhibited CD spectra identical with those of the other designed AR proteins. In size-exclusion chromatography, two N5C proteins were monomeric with an apparent molecular mass of 39.2 kDa (expected 24.7 kDa) and 37.7 kDa (expected 24.7 kDa), respectively. One N6C protein was monomeric with an apparent molecular mass of 42.7 kDa (expected 28.2 kDa). One N5C protein was a mixture between oligomers and a monomer with an apparent molecular mass of 38.1 kDa (expected 24.7 kDa). One N6C protein had a monomer peak at 44.2 kDa (expected 28.2 kDa) but showed some aggregation.

Discussion

The designed AR proteins possess very favorable biophysical properties

We have developed a novel strategy for constructing combinatorial repeat protein libraries.[3] Here, we have applied this strategy to AR proteins. The accompanying paper by Stumpp *et al.*[13] shows

the application of this strategy to leucine-rich repeat proteins. Using the strategy described here, we were able to generate combinatorial libraries of AR proteins of varying length. The analysis of unselected library members revealed that our design leads to AR proteins with very favorable biophysical properties. We focused our analysis on N2C, N3C and N4C library members, since most natural AR proteins possess repeat numbers in this range. Like natural AR proteins, our designed AR proteins can be expressed in large amounts (Figure 6) but, while the expression of natural AR proteins often results in the formation of inclusion bodies, our designed proteins are expressed in soluble form in the cytoplasm of *E. coli* and remain soluble and folded over weeks at 4 °C. The designed AR proteins are monomeric (Figure 7; Table 1) and, as indicated by CD spectroscopy, they exhibit secondary structure compositions indistinguishable from those of natural AR proteins (Figure 8; Table 1). The crystal structure of the N3C library member E3_5 was determined and it was shown that the designed protein has an AR domain fold.[29]

In thermal denaturation, unselected library members showed cooperative unfolding with midpoints of the cooperative transition ranging from 66 °C to above 85 °C (Figure 9; Table 1). Natural AR proteins that have been tested denature around or below 50 °C, as indicated by CD measurements of notch variants (N4C and N5C) and myotrophin (N2C).[35,39] Thermal denaturation midpoints depend very much on experimental conditions such as protein concentration, buffer composition and the temperature ramp, i.e. the kinetics of unfolding and aggregation. However, the differences in stability between natural and designed AR proteins are large enough to indicate that our designed AR proteins are considerably more stable. The thermal denaturation data reflect the high-level thermodynamic stabilities measured by denaturant-induced equilibrium unfolding of the unselected AR proteins presented here (Table 1).[29]

We were able to design a self-compatible AR module, which could be cloned in various numbers between designed capping ARs, leading to well-expressed, soluble, folded and stable AR domains.

Consensus design of AR proteins

Besides its importance for the self-compatibility of AR modules, our consensus design resulted in very stable AR proteins (Figure 9 and Table 1).[29] These results are consistent with effects of previous consensus design approaches. Consensus strategies have been used to generate enzymes with improved thermostability[40] and to improve antibody stability.[41–43] Stability is not a main selection criterion in the evolution of proteins, once a threshold stability is reached that allows the protein to fulfil a function.[44] The most stable variants may not necessarily be implemented by naturally occurring sequences, but they may be encoded by consensus or "canonical" sequences.[43] We used extensive structural criteria to finally decide on the consensus sequence. We observed a remarkable gain in stability, suggesting that the juxtaposition of AR modules had a synergistic effect.

When examining the crystal structure of E3_5 (Figure 2), we were able to pinpoint several features of the designed AR proteins that could explain this increased stability, such as the absence of irregularities or the presence of extended H-bond networks.[29] These stability findings are similar to data published recently by Mosavi *et al.*,[45] who analyzed full consensus AR proteins. These full consensus AR proteins are based on a slightly different consensus sequence compared to ours. Differences in the two consensus sequences are mainly at positions where they considered sequence data for consensus definition, while we used structural decision parameters (framework residue positions Val19, Lys21, Leu22, Glu25, Ala26 vs. Ile19, Glu21, Val22, Lys25, and a mixture of His, Asn and Tyr at position 26 in our molecules; see Results). The structures of their full consensus proteins are nearly identical with the structure of our designed ARs in E3_5 ($RMSD_{C\alpha}$ 0.51 Å compared to PDB entry 1N0Q). Another important difference is the presence of capping repeats (Figure 4) sealing the stack of designed ARs in our molecules compared to the full consensus AR proteins described by Mosavi *et al.*[45] Experimentally, the proteins described by Mosavi *et al.*[45] are more soluble at acidic pH (pH 4–5) than at neutral pH, while our proteins are soluble and stable under physiological conditions (20–50 mM Tris–HCl (pH 6.5–8.5), 50–500 mM NaCl). Both the differences in consensus sequence and the presence of capping ARs in our constructs might lead to this altered behavior. Nevertheless, both studies show that the AR framework is intrinsically very stable. This stability could, in part, account for the abundance of AR proteins in nature. Main and co-workers[46] recently reported the consensus design of tetratricopeptide repeat proteins of different repeat numbers. They also observed high thermal stability of the consensus designed proteins.

Module-wise assembly of AR proteins

The modular nature of AR proteins suggests assembling repeat domains module-wise. We used type IIs restriction enzymes for this purpose, which allow the cloning of repeats in a directional manner, independent of the repeat sequence (Figure 5). The advantage of type IIs restriction enzymes is the freedom of choice of the ligation site. It could, in principle, be placed in any part of the repeat module except for the randomized positions. We did not want to affect the capping repeat structure and sequence; therefore, the ligation site could not be placed in any of the α-helices or in

the short loop connecting the helices. The only remaining possibility was the loop connecting the second α-helix with the β-turn of the next repeat.

Similar to the approach described in the accompanying paper,[13] the chosen cloning strategy using type IIs restriction enzymes (Figure 5) allows a number of evolution strategies that are amenable only to repeat proteins. For example, repeats might be shuffled, added or subtracted. In nature, the IκBα/Bcl-3 pair gives an example of such a repeat extension, where the repeat number reflects the different binding properties of these molecules.[47] We have shown the feasibility of this extension approach for the designed AR proteins by elongating E3_5 (N3C) to N5C or N6C.

In addition to repeat shuffling and module-wise addition or subtraction of single repeats, other evolution strategies are amenable to our molecules. Alterations from the 33 amino acid residue consensus, similar to what is observed in single repeats of IκBα,[23,24] Swi6[28] or the INK4 family members, could be used for the improvement of selected binding molecules *via in vitro* evolution or rational design. Another evolution strategy could involve increasing or decreasing the AR domain curvature, which can be achieved by varying specific framework residues.[10]

Designed AR proteins in biotechnology

Our findings show that we have libraries of well-behaved AR proteins for use in selection procedures. The proteins are expressed in large amounts, they are soluble, monomeric and stable under physiological conditions, they are cysteine-free and allow a great variety of amino acids in the randomized positions. Therefore, these molecules exactly match the requirements for novel scaffolds to be used for the generation of novel binding proteins. The stability of the consensus designed AR proteins is sufficiently high that some losses in stability can be tolerated during the course of directed evolution of the designed AR proteins.

In our AR domains, the potential target interaction residues are located in the β-turn and the first α-helix of each repeat module, creating a large and modular interaction surface (Figure 2).[3,29] This interaction mode extends and combines previous concepts in the field of combinatorial libraries. Usually, either flexible loops (e.g. Knappik *et al*.[41]) or rigid, flat surfaces (e.g. Nord *et al*.[48]) were randomized, but not a combination of turns and helices, which constitute a continuous surface that can be extended by adding more repeats.

We have recently used designed N2C and N3C AR protein libraries in ribosome display[49,50] selections against various globular proteins. Specific nanomolar binders were obtained, which prove the success of designed AR proteins as novel scaffolds for molecular recognition. A detailed analysis of these experiments will be published elsewhere. Because of their favorable biophysical properties, designed AR proteins could ideally serve as recognition molecules on protein chips. Similarly, the absence of intracellular aggregation or misfolding and the absence of cysteine residues would allow these proteins to be used as intracellular protein binders or enzyme inhibitors. In this regard, designed AR proteins could be an attractive and more stable alternative to intrabodies.[51]

Conclusions

We have successfully implemented our novel strategy harnessing the modular nature of repeat proteins for the generation of designed AR protein libraries. Through sequence and structure consensus analyses, we designed an AR module composed of fixed framework positions and randomized potential interaction positions. AR domains were generated by cloning two, three or four designed modules between N and C-terminal capping repeats. All tested proteins exhibit very favorable biophysical properties. They can be expressed in soluble form in large amounts and they can be purified easily. They are monomeric and show CD spectra indistinguishable from those of natural AR proteins. Furthermore, they are exceptionally resistant to heat denaturation. These findings suggest that the abundance of natural AR proteins is, at least in part, based on the exceptional properties of the AR framework, a stable and modular protein–protein interaction motif. Our findings show that we can build modular and stable proteins with randomized surfaces that may be used to create novel binding molecules. The modular structure of repeat proteins will allow completely new evolution strategies that are not feasible with classical scaffolds.

Material and Methods

In silico analysis

We used the SMART†,[7] the GenBank‡,[18] and the PDB§ [52] databases for our analyses. Clustal W‖ [14] and BLAST¶ [17] were used for alignments. Structural modeling was done with InsightII (Accelrys, USA). NACCESS helped to identify target interaction residues from structures of complexes. PHD prediction[a] [31] was used for secondary structure prediction. PEST[b] [32] and PEPTIDESORT of GCG (Accelrys, USA)[33] were used to

† http://smart.embl-heidelberg.de
‡ http://www.ncbi.nlm.nih.gov/Genbank/
§ http://www.pdb.org
‖ http://www.ch.embnet.org/software/ClustalW.html
¶ http://www.ncbi.nlm.nih.gov/blast/
a http://cubic.bioc.columbia.edu/predictprotein/
b http://www.at.embnet.org/embnet/tools/bio/PESTfind/

compare designed and natural AR proteins. GCG was used for designing the DNA sequence.

General molecular biology

Unless stated otherwise, all experiments were performed as described.[53] Enzymes and buffers were from New England Biolabs (USA) or Fermentas (Lithuania). The cloning and production strain was *E. coli* XL1-Blue (Stratagene, USA). The cloning and protein expression vector was pPANK, a pQE30 (QIAgen, Germany) derivative lacking the *Bbs* I and *Bsa* I sites. pPANK was generated *via* PCR-cloning using the oligonucleotides Bbs I (5′-TGATTTCTCGAGGTGTAGTCGAAAGGGCCTCGTG-3′), Bsa I (5′-GCAATGATACCGCGAGAACCACGCTCA CCGGC-3′) and Avr2 (5′-CCGCCGCTCTAGAGGGAAA CCTAGGGCTGCCTCGCGCG-3′) and pQE30 as template. The oligonucleotides *Bbs* I and *Bsa* I were used to generate a PCR product, which was used as primer in a second PCR reaction together with oligonucleotide Avr2. The resulting PCR product was *Xho* I/*Xba* I-digested and ligated to the *Xho* I/*Xba* I promoter fragment of pQE30.

Synthesis of DNA encoding AR proteins

Oligonucleotides incorporating mixed trinucleotides as building blocks[34] were from MorphoSys AG (Germany). INT1:

5′-CTGACGTTAACGCTNNNGACNNNNNNGGTN NNACTCCGCTGCACCTGGC-3′ and INT2:

5′-ACTCCGCTGCACCTGGCTGCTNNNNNNGGTC ACCTGGAAATCG-3′.

NNN represents a mixture of trinucleotides encoding the amino acids A, D, E, H, K, N, Q, R, S, T (7% each) and F, I, L, M, V, W, Y (4.3% each). Standard oligonucleotides were from Microsynth (Switzerland).

INT3: 5′-AACGTCAGCACCGTDCTTCAGCAGAAC TTCA ACGATTTCCAGGTGACC-3′; D represents any of the nucleotides A, G or T).

INT4: 5′-AGCAGCCAGGTGCAGCGGAGT-3′.

INT5: 5′-TTCCGCGGATCCTAGGAAGACCTGACGT TAAC GCT-3.

INT6: 5′-TTTGGGAAGCTTCTAGAAGACAACGT CAGCAC CGT-3′.

INT6a: 5′-TTTGGGAAGCTTCTAAGGTCTCACGT CAGCAC CGT-3′.

INT6b: 5′-TTTGGGAAGCTTCTAAGGTCTC-3′.

EWT1: 5′-TTCCGCGGATCCGACCTGGGTAA GAAACTGCT GGAAGCTGCTCGTGCTGGTCAGGAC GACGAAG-3′.

EWT2: 5′-AACGTCAGCACCGTTAGCCATCAGGA TACGAA CTTCGTCGTCCTGACC-3′.

EWT3: 5′-TTCCGCGGATCCGACCTGGG-3′.

TEN3: 5′-TTCCGCGGATCCG-3′.

WTC1: 5′-CTGACGTTAACGCTCAGGACAAATTCG GTAAG ACCGCTTTCGACATCTCCATCGACAACGG TAACGA GG-3′.

WTC2: 5′-TTGCAGGATTTCAGCCAGGTCCTCGT TACCGTT GTC-3′.

WTC3: 5′-TTTGGGAAGCTTCTATTGCAGGATTTCA GC-3′.

The AR modules were generated by assembly PCR using oligonucleotides INT1, INT2, INT3, INT4, INT5 and INT6a, and Vent® Polymerase (one minute annealing at 50 °C; standard buffer with a final concentration of 5.5 mM $MgSO_4$). A subset of the resulting PCR product was cloned *via* *Bam* HI/*Hin*dIII into pPANK and sequenced using standard techniques. The AR module sequence is shown in Figure 5.

The N-terminal capping AR was prepared by assembly PCR using oligonucleotides EWT1, EWT2, TEN3 and INT6. The resulting DNA was cloned *via* *Bam* HI/*Hin*dIII into pPANK. The DNA sequence was verified using standard techniques. The C-terminal capping AR was prepared similarly, but by using oligonucleotides WTC1, WTC2, WTC3 and INT5.

The ligation of the DNA encoding an AR protein from single AR modules and AR capping repeats is represented schematically in Figure 5. To clone DNA encoding AR proteins, the PCR-assembled N-terminal capping AR (using oligonucleotides EWT1, EWT2, TEN3 and INT6a) was cut with *Bsa* I and ligated to a *Bpi* I-cut AR module. The ligation product, termed N1 (where N denotes the N-terminal capping repeat and the digit is the number of randomized repeat modules), was PCR-amplified using oligonucleotides EWT3 and INT6b. The amplified product was cleaved again with *Bsa* I. The subsequent ligation to *Bpi* I-cut AR modules started a new cycle of elongation, which was repeated until the desired number of AR modules was added to the N-terminal capping AR (termed N2, N3, N4 etc.). DNA corresponding to PCR-amplified N2, N3 and N4 were then cut with *Bam* HI/*Bsa* I and ligated to a *Bam* HI/*Bpi* I-cut pPANK containing the C-terminal capping AR (Figure 5). This yielded cloned DNA molecules encoding N2C, N3C and N4C AR protein libraries (where N denotes the N-terminal capping repeat, the digit is the number of randomized repeat modules and C is the C-terminal capping repeat).

An unselected N3C library member (named E3_5, see below) was used as template for repeat protein elongation. Using the oligonucleotides EWT3 and INT6a, fragments corresponding to N, N1, N2 and N3 were generated by PCR. The N3 fragment was isolated and then reamplified using oligonucleotides EWT3 and INT6b. Elongation of the N3 to N5 and N6 fragments and cloning to N5C and N6C molecules was carried out as described above.

Screening for protein expression and DNA sequencing

An SDS-15% PAGE screening of single unselected clones was performed using 10 ml cultures. A stationary overnight culture (5 ml of LB, 1% (w/v) glucose, 100 mg/l of ampicillin; 37 °C) was used to inoculate the cultures (1 ml of inoculum in 9 ml of the above medium). After one hour, protein expression was induced using 200 μM IPTG and cultures were incubated for five hours. In parallel, all screened clones were subjected to DNA sequence analysis. Two unselected clones of the libraries N2C (clone names: E2_5 and E2_17), N3C (E3_5 and E3_19) and N4C (E4_2 and E4_8) were chosen for subsequent protein analyses.

Protein expression and purification

The N2C, N3C, N4C, N5C and N6C clones were expressed as follows: 25 ml of stationary overnight cultures (LB, 1% glucose, 100 mg/l of ampicillin; 37 °C) were used to inoculate 1 l cultures (same medium). At $A_{600} = 0.7$, the cultures were induced with 300 μM IPTG and incubated for four hours. Samples were analyzed by SDS-15% PAGE (Figure 6). The cultures were

centrifuged and the resulting pellets were resuspended in 40 ml of TBS_{500} (50 mM Tris–HCl (pH 8.0), 500 mM NaCl) and sonicated. The lysate was recentrifuged and glycerol (10% final concentration) and imidazole (20 mM final concentration) were added to the resulting supernatant. Proteins were purified over a Ni-nitrilotriacetic acid column (2.5 ml column volume) according to the manufacturer's instructions (QIAgen, Germany; Figure 6).

Size-exclusion chromatography

IMAC-purified proteins were analyzed on a Superdex 75 gel-filtration column (Amersham Pharmacia Biotech, USA) using a Pharmacia SMART system at a flow-rate of 60 μl/minute and with TBS_{150} (50 mM Tris–HCl (pH 7.4), 150 mM NaCl) as running buffer (Figure 7).

CD spectroscopy

Circular dichroism spectra were recorded in 10 mM sodium phosphate buffer (pH 6.5), 100 mM NaCl, using 10 μM purified protein on a Jasco J-715 instrument (Jasco, Japan). The CD signal was converted to mean residue ellipticity using the concentration of the sample determined spectrophotometrically at 280 nm under denaturing conditions (Figure 8).

Heat denaturation was performed in 20 mM sodium phosphate (pH 7.4), 200 mM NaCl with 10 μM protein and a temperature shift from 10 °C to 95 °C within 120 minutes. CD data were collected at 222 nm every 20 seconds with a bandwidth of 2 nm and 16 seconds response time (Figure 9).

Data Bank accession numbers

The DNA and amino acid sequences of proteins E2_5, E2_17, E3_5, E3_19, E4_2 and E4_8 have been deposited in GenBank[18] with accession numbers AY195851, AY195852, AY195853, AY195854, AY195855 and AY195856, respectively. The DNA sequence of pPANK has been deposited in GenBank[18] (accession number AY327140).

Acknowledgements

We thank the members of the Plückthun laboratory for valuable discussions, Dr Annemarie Honegger for EXCEL® macros as well as Dr David L. Zechel and Dr Casim A. Sarkar for critical reading of the manuscript. We thank MorphoSys AG for the trinucleotide-containing oligonucleotides. H.K.B. was supported by a pre-doctoral fellowship of the Roche Research Foundation. M.T.S. was the recipient of an FCI and a BMBF pre-doctoral scholarship. This work was supported by the Swiss National Centre of Competence in Research (NCCR) in Structural Biology and the Swiss Cancer Research grant KFS 1055-09-2000.

References

1. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131.
2. Kobe, B. & Kajava, A. V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* **25**, 509–515.
3. 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 Letters*, **539**, 2–6.
4. Nygren, P.-Å. & Uhlén, M. (1997). Scaffolds for engineering novel binding sites in proteins. *Curr. Opin. Struct. Biol.* **7**, 463–469.
5. Skerra, A. (2000). Engineered protein scaffolds for molecular recognition. *J. Mol. Recognit.* **13**, 167–187.
6. Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins: Struct. Funct. Genet.* **17**, 363–374.
7. Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R. *et al.* (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucl. Acids Res.* **30**, 242–244.
8. Gorina, S. & Pavletich, N. P. (1996). Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. *Science*, **274**, 1001–1005.
9. Sedgwick, S. G. & Smerdon, S. J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem. Sci.* **24**, 311–316.
10. Michaely, P., Tomchick, D. R., Machius, M. & Anderson, R. G. (2002). Crystal structure of a 12 ANK repeat stack from human ankyrinR. *EMBO J.* **21**, 6387–6396.
11. Suzuki, F., Goto, M., Sawa, C., Ito, S., Watanabe, H., Sawada, J. & Handa, H. (1998). Functional interactions of transcription factor human GA-binding protein subunits. *J. Biol. Chem.* **273**, 29302–29308.
12. Malek, S., Huxford, T. & Ghosh, G. (1998). IκBα functions through direct contacts with the nuclear localization signals and the DNA binding sequences of NF-κB. *J. Biol. Chem.* **273**, 25427–25435.
13. 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.
14. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
15. Breeden, L. & Nasmyth, K. (1987). Similarity between cell-cycle genes of budding yeast and fission yeast and the Notch gene of *Drosophila*. *Nature*, **329**, 651–654.
16. Lux, S. E., John, K. M. & Bennett, V. (1990). Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins. *Nature*, **344**, 36–42.
17. 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.
18. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. (2000). GenBank. *Nucl. Acids Res.* **28**, 15–18.
19. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1: 1313

multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260–262.
20. Luh, F. Y., Archer, S. J., Domaille, P. J., Smith, B. O., Owen, D., Brotherton, D. H. *et al.* (1997). Structure of the cyclin-dependent kinase inhibitor p19Ink4d. *Nature*, **389**, 999–1003.
21. Batchelor, A. H., Piper, D. E., de la Brousse, F. C., McKnight, S. L. & Wolberger, C. (1998). The structure of GABPα/ankyrin repeat heterodimer bound to DNA. *Science*, **279**, 1037–1041.
22. Byeon, I.-J.L., Li, J., Ericson, K., Selby, T. L., Tevelev, A., Kim, H. J. *et al.* (1998). Tumor suppressor $p16^{INK4a}$: determination of solution structure and analyses of its interaction with cyclin-dependent kinase 4. *Mol. Cell*, **1**, 421–431.
23. Huxford, T., Huang, D. B., Malek, S. & Ghosh, G. (1998). The crystal structure of the IκBα/NF-κB complex reveals mechanisms of NF-κB inactivation. *Cell*, **95**, 759–770.
24. Jacobs, M. D. & Harrison, S. C. (1998). Structure of an IκBα/NF-κB complex. *Cell*, **95**, 749–758.
25. Yang, Y., Nanduri, S., Sen, S. & Qin, J. (1998). The structural basis of ankyrin-like repeat function as revealed by the solution structure of myotrophin. *Structure*, **6**, 619–626.
26. Venkataramani, R., Swaminathan, K. & Marmorstein, R. (1998). Crystal structure of the CDK4/6 inhibitory protein $p18^{INK4c}$ provides insights into ankyrin-like repeat structure/function and tumor-derived $p16^{INK4}$ mutations. *Nature Struct. Biol.* **5**, 74–81.
27. Mandiyan, V., Andreev, J., Schlessinger, J. & Hubbard, S. R. (1999). Crystal structure of the ARF-GAP domain and ankyrin repeats of PYK2-associated protein β. *EMBO J.* **18**, 6890–6898.
28. Foord, R., Taylor, I. A., Sedgwick, S. G. & Smerdon, S. J. (1999). X-ray structural analysis of the yeast cell cycle regulator Swi6 reveals variations of the ankyrin fold and has implications for Swi6 function. *Nature Struct. Biol.* **6**, 157–165.
29. Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Plückthun, A. & Grütter, M. G. (2003). Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc. Natl Acad. Sci. USA*, **100**, 1700–1705.
30. O'Neil, K. T. & DeGrado, W. F. (1990). A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science*, **250**, 646–651.
31. Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**, 525–539.
32. Rogers, S., Wells, R. & Rechsteiner, M. (1986). Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science*, **234**, 364–368.
33. Womble, D. D. (2000). GCG: The Wisconsin Package of sequence analysis programs. *Methods Mol. Biol.* **132**, 3–22.
34. Virnekäs, B., Ge, L., Plückthun, A., Schneider, K. C., Wellnhofer, G. & Moroney, S. E. (1994). Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucl. Acids Res.* **22**, 5600–5607.
35. Mosavi, L. K., Williams, S. & Peng, Z.-y. (2002). Equilibrium folding and stability of myotrophin: a model ankyrin repeat protein. *J. Mol. Biol.* **320**, 165–170.
36. Zweifel, M. E. & Barrick, D. (2001). Studies of the ankyrin repeats of the *Drosophila melanogaster* Notch receptor. 1. Solution conformational and hydrodynamic properties. *Biochemistry*, **40**, 14344–14356.
37. Zeeb, M., Rosner, H., Zeslawski, W., Canet, D., Holak, T. A. & Balbach, J. (2002). Protein folding and stability of human CDK inhibitor $p19^{INK4d}$. *J. Mol. Biol.* **315**, 447–457.
38. Zhang, B. & Peng, Z.-y. (2000). A minimum folding unit in the ankyrin repeat protein $p16^{INK4}$. *J. Mol. Biol.* **299**, 1121–1132.
39. Zweifel, M. E. & Barrick, D. (2001). Studies of the ankyrin repeats of the *Drosophila melanogaster* Notch receptor. 2. Solution stability and cooperativity of unfolding. *Biochemistry*, **40**, 14357–14367.
40. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. (2000). The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta*, **1543**, 408–415.
41. Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G. *et al.* (2000). Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* **296**, 57–86.
42. Ohage, E. & Steipe, B. (1999). Intrabody construction and expression. I. The critical role of V_L domain stability. *J. Mol. Biol.* **291**, 1119–1128.
43. Steipe, B., Schiller, B., Plückthun, A. & Steinbacher, S. (1994). Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* **240**, 188–192.
44. Wang, Q., Buckle, A. M. & Fersht, A. R. (2000). Stabilization of GroEL minichaperones by core and surface mutations. *J. Mol. Biol.* **298**, 917–926.
45. 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.
46. 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 (Camb.)*, **11**, 497–508.
47. Michel, F., Soler-Lopez, M., Petosa, C., Cramer, P., Siebenlist, U. & Müller, C. W. (2001). Crystal structure of the ankyrin repeat domain of Bcl-3: a unique member of the IκB protein family. *EMBO J.* **20**, 6180–6190.
48. Nord, K., Gunneriusson, E., Ringdahl, J., Ståhl, S., Uhlén, M. & Nygren, P-Å. (1997). Binding proteins selected from combinatorial libraries of an α-helical bacterial receptor domain. *Nature Biotechnol.* **15**, 772–777.
49. Hanes, J. & Plückthun, A. (1997). *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc. Natl Acad. Sci. USA*, **94**, 4937–4942.
50. Amstutz, P., Forrer, P., Zahnd, C. & Plückthun, A. (2001). *In vitro* display technologies: novel developments and applications. *Curr. Opin. Biotechnol.* **12**, 400–405.
51. Cattaneo, A. & Biocca, S. (1999). The selection of intracellular antibodies. *Trends Biotechnol.* **17**, 115–121.
52. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K. *et al.* (2002). The Protein Data Bank. *Acta Crystallog. sect. D*, **58**, 899–907.
53. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edit., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

54. Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.
55. Yang, F., Forrer, P., Dauter, Z., Conway, J. F., Cheng, N., Cerritelli, M. E. *et al.* (2000). Novel fold and capsid-binding properties of the lambda-phage display platform protein gpD. *Nature Struct. Biol.* **7**, 230–237.

Edited by J. Thornton

(Received 23 April 2003; received in revised form 9 July 2003; accepted 10 July 2003)