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

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

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

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

* corresponding author

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

Email address of corresponding author: plueckthun@bioc.uzh.ch

Residue Choices in Internal Repeats

Several residues present in the originally defined consensus sequences (Fig. S1) were replaced and some maintained despite their apparent unfavorable properties. The rationale behind these choices is illustrated here. The residues obtained from the consensus sequences were maintained in the positions not mentioned here.

Gln5, present in the consensus type I and C, can potentially form a hydrogen bond with Asp9 of the same repeat, stabilizing H1 (observed in structure 1EE5 [1]). Gly11 is conserved for its crucial role in bending the polypeptide chain between H1 and H2, being compatible with a positive φ angle required at this position. Pro14 is an unusual conserved feature of armadillo repeats, present at a frequency of 62% in the alignment of importin repeats, 35% for catenin/plakoglobin repeats and 50% in the overall alignment. It is located at the beginning of H2, at a position where it is still not necessary to involve the backbone nitrogen in a hydrogen bond; instead of disrupting the secondary structure, it adopts the φ/ψ angles typical for α-helices in the available crystal structures. Asn37 is a well conserved residue in all consensus sequences, due to its critical role in binding to the backbone of target peptides. Trp33 is also involved in binding, specifically in the recognition of target side chains in the case of importin-α proteins. Thus, it appears with high frequency in the importin subfamily and it is therefore present in the overall consensus as well. All the residues described above were maintained.

Further modifications were introduced in the original consensus sequences (Fig S1) to meet the requirements for protein production (e.g. lack of cysteines) or to avoid structural defects (e.g. presence of potential clashes) that could have arisen from a purely sequence-based alignment. Cys30 (18%) and Cys41 (21%) in the catenin consensus were replaced by the second most common amino acid (Ala 12% and His 17%, respectively), to avoid the formation of undesired disulfide bonds that might limit possible future applications. Pro2 (30%), in the importin consensus, was substituted with the approximately equally conserved Glu (26%), as Pro at such position would probably disrupt H1, as indicated by importin-α crystal structures. In the catenin consensus, position 9 shows a preference for long aliphatic side-chains, either non-polar or polar (Leu, Glu, Gln); however, this residue is solvent-exposed, and Leu9 (21%) was substituted by the second most common amino acid Glu (19%). In the importin consensus, Pro15 (19%) (Fig. S1) was derived

from sequences which do not possess a Pro in position 14. A double Pro14-Pro15 never occurs in the observed sequences and it is likely to be extremely destabilizing for H2. Position 15 is usually occupied by small hydrophobic residues in combination with Pro14. Arg also represents a relatively common choice (16%) but it occurs almost exclusively in the second repeat of natural importins. Val was therefore chosen as more general substitution, instead of Arg, due to the slightly higher frequency of occurrence (13%) compared to other residues. The catenin consensus has Gln as most frequent amino acid (27%) at position 18. However, both Arg and Lys are represented almost at the same frequency (25% and 20% respectively), indicating a preference for positively charged residues. Arg was thus chosen at this position due to its higher frequency. Positions 24 and 25 at the joint between loop H2-H3 and H3 show a clear preference for acidic residues in all consensus sequences. However, a pair of negatively charged residues never occurs in the observed sequences, and it could lead to charge repulsion or formation of a negatively charged belt along the whole protein. The most conserved residues in importin and catenin consensus sequences were preserved (Asp24 in the importin consensus (49%) and Glu25 in the catenin/plakoglobin consensus (43%), respectively). To reduce the local negative charge, Asn (11%) was chosen to replace Glu (19%) at position 25 in the importin consensus, because it is the second most frequent polar residue. Asp24 in the catenin/plakoglobin consensus (20%) was substituted by Asn (10%), a choice driven by the similarity to the original residue in a pool of candidates with almost the same frequency (Arg, Asn, Met, Ser, Tyr, Val). In the overall consensus, a preferred position for the negative charge is not pronounced (Asp24 36%, Glu25 27% in type C), and the alternative amino acids have all very low frequency (<10%). The residues selected for replacement were thus chosen to improve the H3 stability: Glu25 was kept in the overall consensus due to its higher helical propensity and Asn24 was introduced to keep structural similarity to the more frequent residue Asp and to take advantage of its propensity as an N-cap residue [2; 3; 4; 5]. Gly was introduced at position 42 for cloning purposes. It is noteworthy that, because of short H3-H1 loops (1-3 residues in general), one position of the loop is very often occupied by a Gly. Taking into account that position 41 is sometimes involved in binding and will then possibly be subjected to

mutation for applications, it is important to keep a constant glycine inside the loop to maintain the required flexibility.

Designed capping repeats

Capping repeats were designed based on type C internal repeat. A detailed description of the residues introduced in the capping repeats is provided here. The original residues of the type C repeat were maintained in the positions which are not described.

The N-terminal designed capping repeat (Na) goes from position 12 to 42 and includes only H2 and H3. Positions 12, 19, 27, 34 are occupied by hydrophobic residues in the consensus and had to be replaced by hydrophilic residues based on structures and common residues obtained from alignment of N-terminal capping sequences. Ser12 provides the N-terminal helix cap of H2. Asn14 substitutes the more common proline, providing a polar residue with a relatively short side chain. Glu15 can interact with Ser12 in a helix and can additionally stabilize it [6]. Lys18 can form a salt bridge with Glu15, stabilizing the helix, and, in general, a long polar residue is required at this position. Gln19 provides a hydrophobic part for interaction with the neighboring internal repeat as well as a polar moiety for solvent exposure. Asn21 is common at this position and it has a good propensity as helix C-capping residue. Asp23 and Asp24 are conserved as a charged couple in several N-terminal capping repeats. Gln25 is well conserved, polar and with high helical propensity. Gln27 provides a hydrophobic part for interaction with the neighboring internal repeat as well as a polar moiety for solvent exposure. Gln33, well conserved among capping repeats, substitutes the Trp, present at high frequency in internal modules because involved in binding in importins. Lys34 is present at moderate frequency in N-terminal capping repeats, among other polar residues. Arg36 has a high frequency of occurrence and seems to be able to interact with Trp33 present in the importin and overall consensus. Gln37 has a long side chain typical for residues at this position, and, instead of the more common lysine, avoids the formation of a positively charged spot in combination with Arg36. Asp41 maintains the negative charge often present in this position and breaks the helix. Gly42 was introduced to add flexibility and for further module assembly.

The C-terminal designed capping repeat (Ca) includes all three helices. Positions 8, 13, 17, 20, 28, 32, 35, 38, 39 are occupied by hydrophobic residues in the consensus and had to be replaced by hydrophilic residues based on structures and common residues obtained from alignment of C-terminal capping sequences. Lys4 can potentially contribute to the hydrophobic core with the long aliphatic part of the side chain, while contacting the solvent with the positively charged amino-group. Lys8 is present to avoid the formation of a cluster of negative charges that would be formed if a conserved glutamate was used at this position, while keeping a high helical propensity and a long side chain. Glu9 is a highly conserved residue. Ala12 can potentially interact with the hydrophobic core. Glu14 is a common polar residue with high helical propensity as substitute for proline. Lys15 is also a common polar residue with relatively high helical propensity. Leu13 occupies a former hydrophobic core position, but it was retained for its high helical propensity and its ability to interact with Phe39. Glu17 corresponds to a core position in an internal repeat; the hydrophobic residue was substituted with this frequently occurring hydrophilic amino acid with high helical propensity. Gln20 is the more frequent polar residue used to substitute the conserved leucine present in the internal repeats. Positions from 21 to 23 are not clearly defined, showing strong conservation in the catenin/plakoglobin subfamily (maybe for functional reasons) and higher degree of variability in importins. The most conserved residues from the importin subfamily were thus chosen to occupy these positions. Gln28 can provide hydrophobic interactions and a polar side chain, and represents a better choice compared to a conserved tyrosine in the importin subfamily and an alanine in the catenin/plakoglobin subfamily. Gln32 provides high helical propensity and a polar side chain. The presence of frequent aromatic residues at this position does not seem to have a structural reason, judging from the crystal structures. Glu33 is one of the charged residues often found at this position and it has high helical propensity. Glu36 has high frequency in importins where this position is occupied by acidic residues, while in the catenin/plakoglobin subfamily phenylalanine and tyrosine are present. The aromatic residues have probably a functional role, but in our case a charged residue constitutes the better choice, because of the exposed position. Lys37 was chosen to replace the conserved asparagine in the internal repeats. Gln38 was chosen to replace hydrophobic residues, always present at

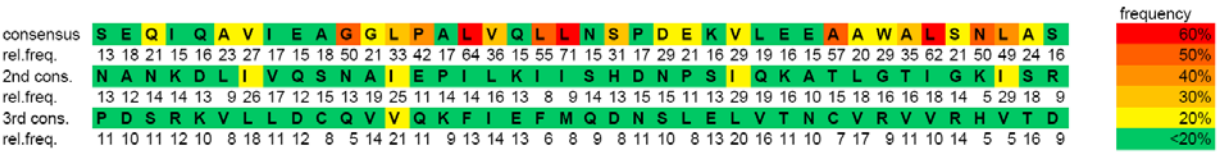
this position both in internal and capping repeats, while providing a polar moiety in contact with the solvent. Phe39 is conserved in several capping repeats. From the available structures, it seems to be important for sealing of the hydrophobic core of importins and for compactness of C-terminal capping repeat via interaction with Leu13. His41 has been added as capping residue to stabilize H3.

References

1. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure Fold Des* **8**, 329-38.
2. Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A. R. (1992). Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J Mol Biol* **227**, 544-59.
3. Penel, S., Hughes, E. & Doig, A. J. (1999). Side-chain structures in the first turn of the alpha-helix. *J Mol Biol* **287**, 127-43.
4. Doig, A. J. & Baldwin, R. L. (1995). N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci* **4**, 1325-36.
5. Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science* **240**, 1648-52.
6. Harper, E. T. & Rose, G. D. (1993). Helix stop signals in proteins and peptides: the capping box. *Biochemistry* **32**, 7605-9.

(a)

initial consensus from SMART

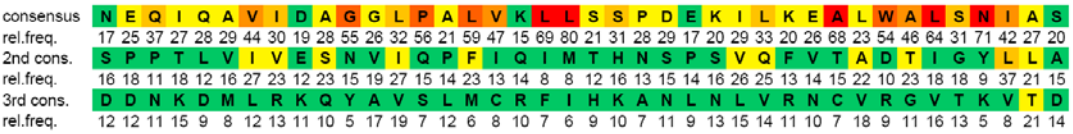


(b)

importin consensus

catenin/plakoglobin consensus

overall consensus



(c)

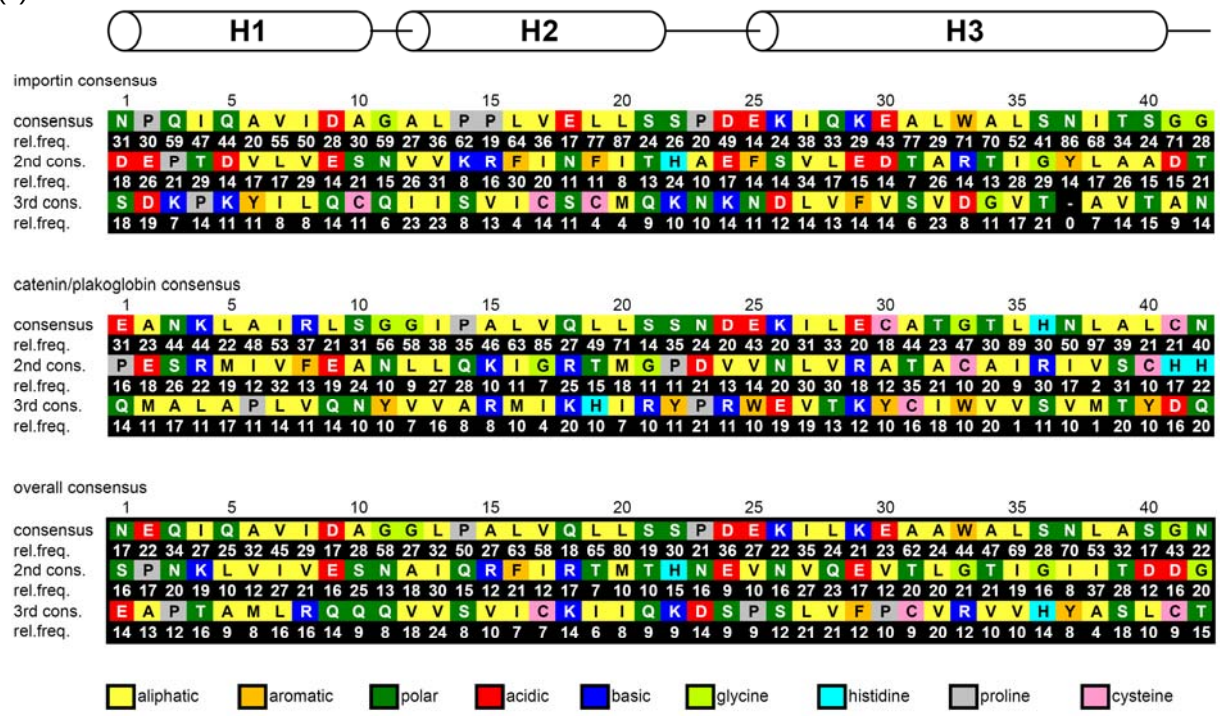


Fig. S1 Consensus sequences derived from multiple alignments. (a) Consensus sequence obtained from alignment of SMART armadillo seed sequences. (b) Consensus sequences of importin and catenin/plakoglobin subfamilies and overall consensus; sequences used for the alignment were retrieved using a profile based on SMART sequences. The sequences are limited to 40 residues and do not contain the loop between adjacent repeats. Amino acids are colored according to their relative frequency. (c) Consensus sequences of importin and catenin/plakoglobin subfamilies and overall consensus after manual refinement of the alignment. The cylinders indicate the putative α-helices and the numbers denote the positions inside the single repeats according to the conventions introduced. The residues are colored according to amino acid type, as indicated at the bottom.
For each position of thesequences, the most frequent, the second and the third most frequent type of residue are indicated, with the relative frequency expressed as percentage.

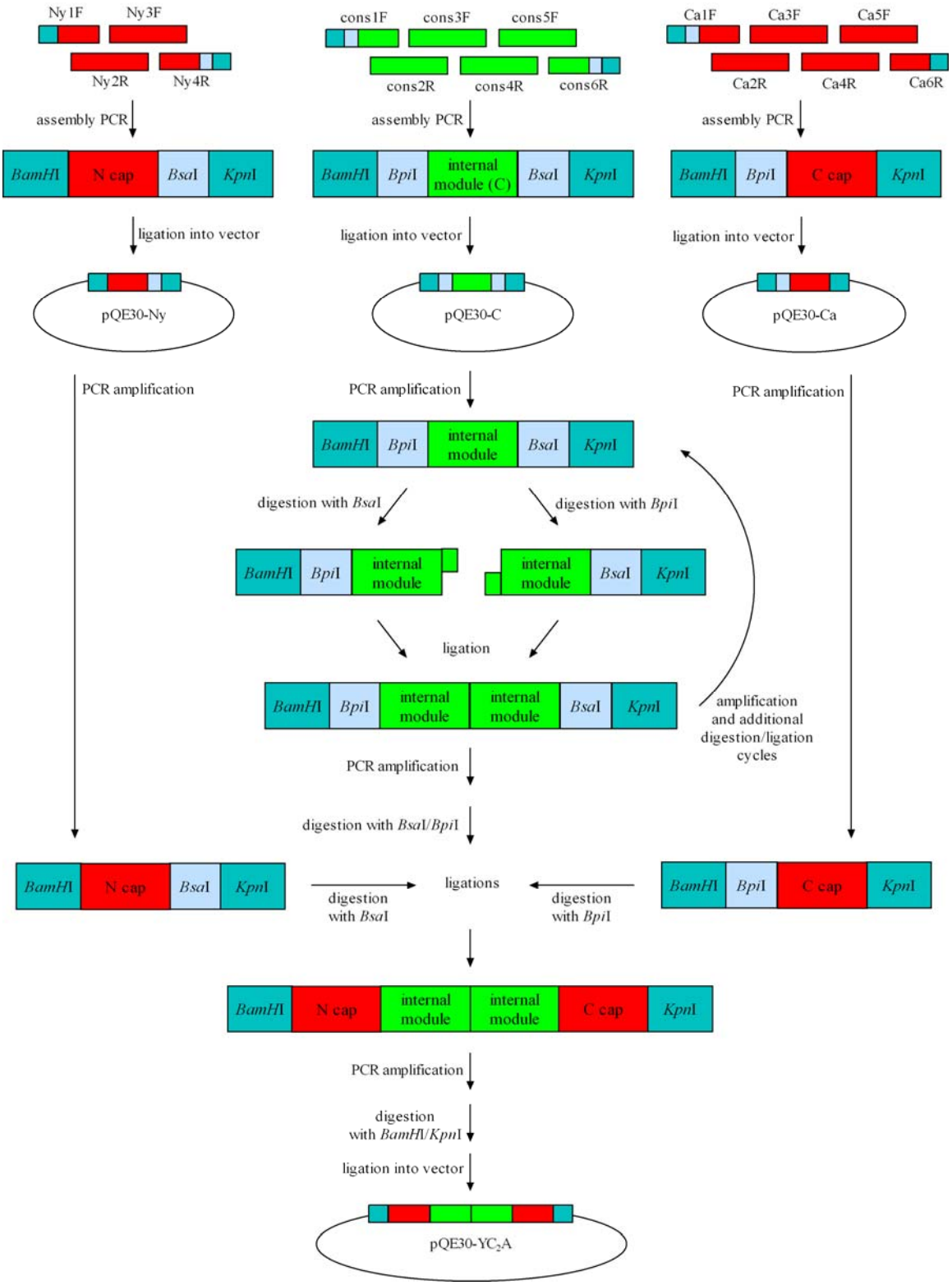


Fig. S2 Scheme of the assembly strategy for designed armadillo repeat protein constructs at the DNA level. Oligonucleotides are assembled to an internal or terminal capping module by PCR. The single modules contain external restriction sites for *BamH*I and *Kpn*I for insertion in the vector and sites for the type IIS restriction enzymes *Bsa*I and *Bpi*I for ligation of the modules. The construction of YC_2A from the internal module C, the N-terminal cap Ny and the C-terminal cap Ca is shown as an example.

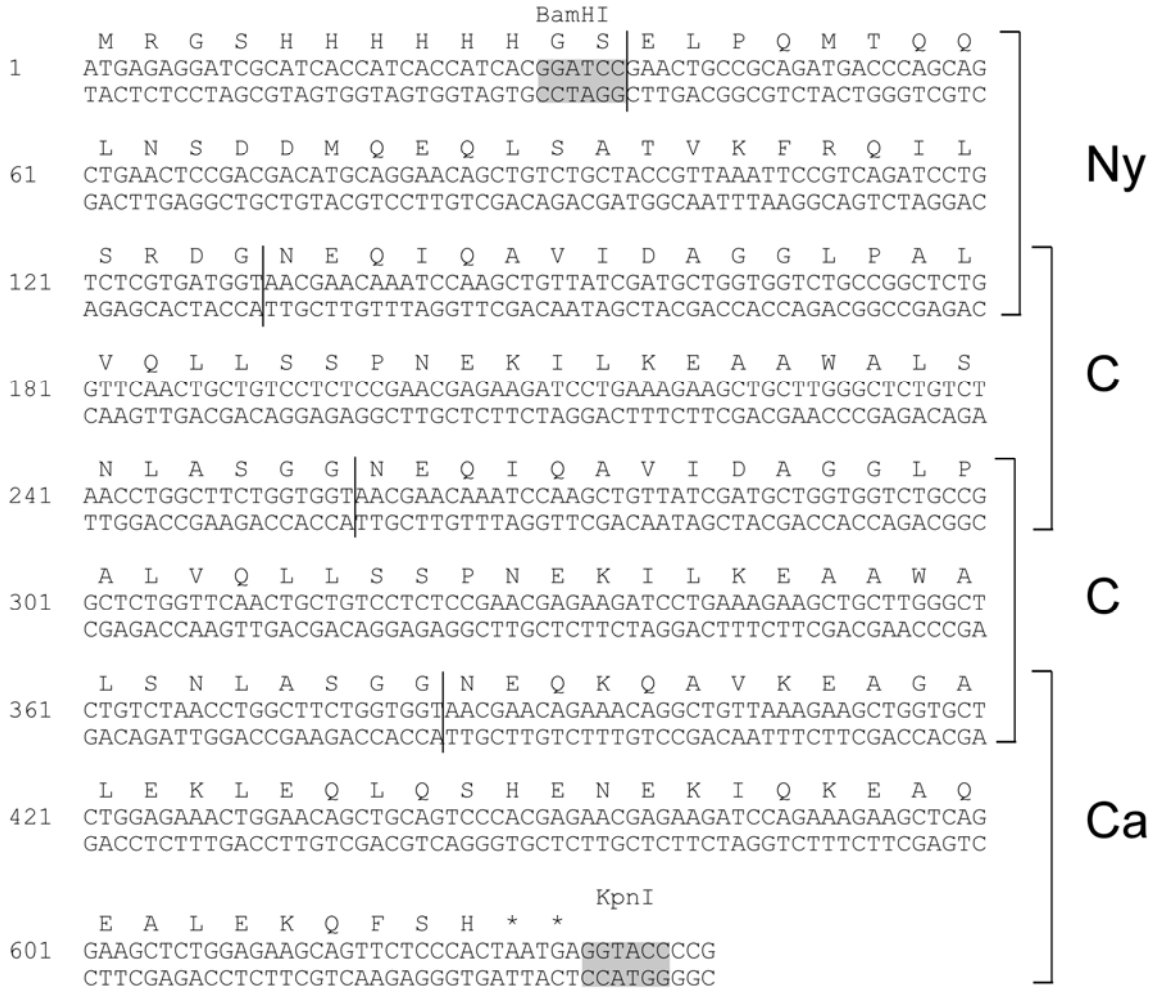


Fig. S3 Sequence of the designed armadillo repeat protein YC_2A. The translated amino acid sequence is shown on the top of the DNA sequence as single letter code. The bars indicate the separations between the modules and between the $MRGSHis_6$ tag provided by the vector and the N-terminal capping module. The modules are indicated on the right. The restriction sites used for cloning are highlighted in gray. The star (*) indicates the presence of a stop codon.

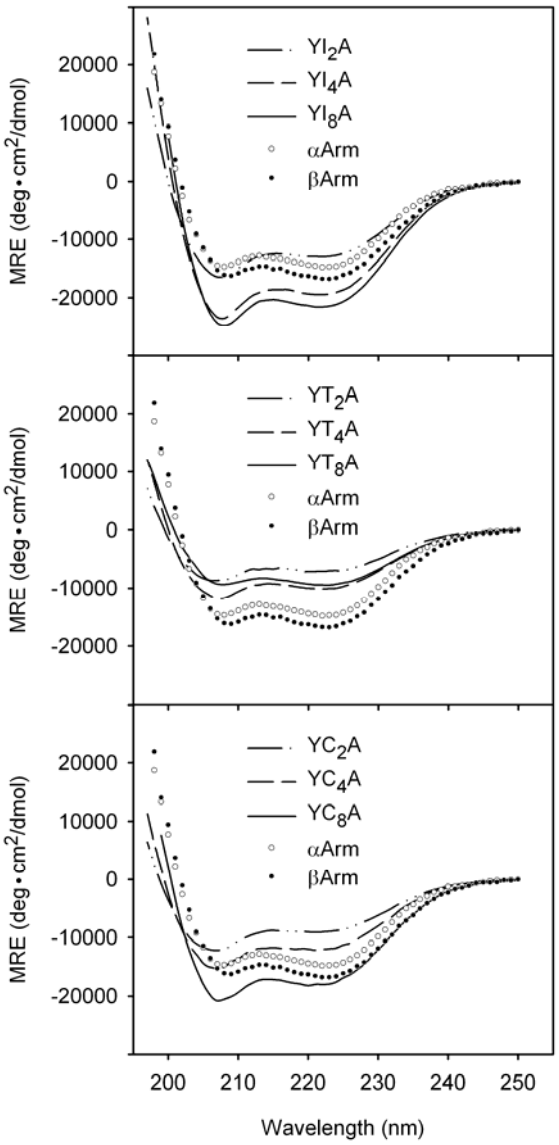


Fig. S4 Circular dichroism (CD) spectra of designed consensus armadillo repeat proteins. From the top, I-type, T-type and C-type proteins containing 2, 4 or 8 internal modules are shown. The CD spectra of the natural armadillo domains of human importin-α1 (α Arm) and mouse β-catenin (βArm) are indicated by empty and filled circles, respectively. The values are reported as mean residue ellipticity (MRE).

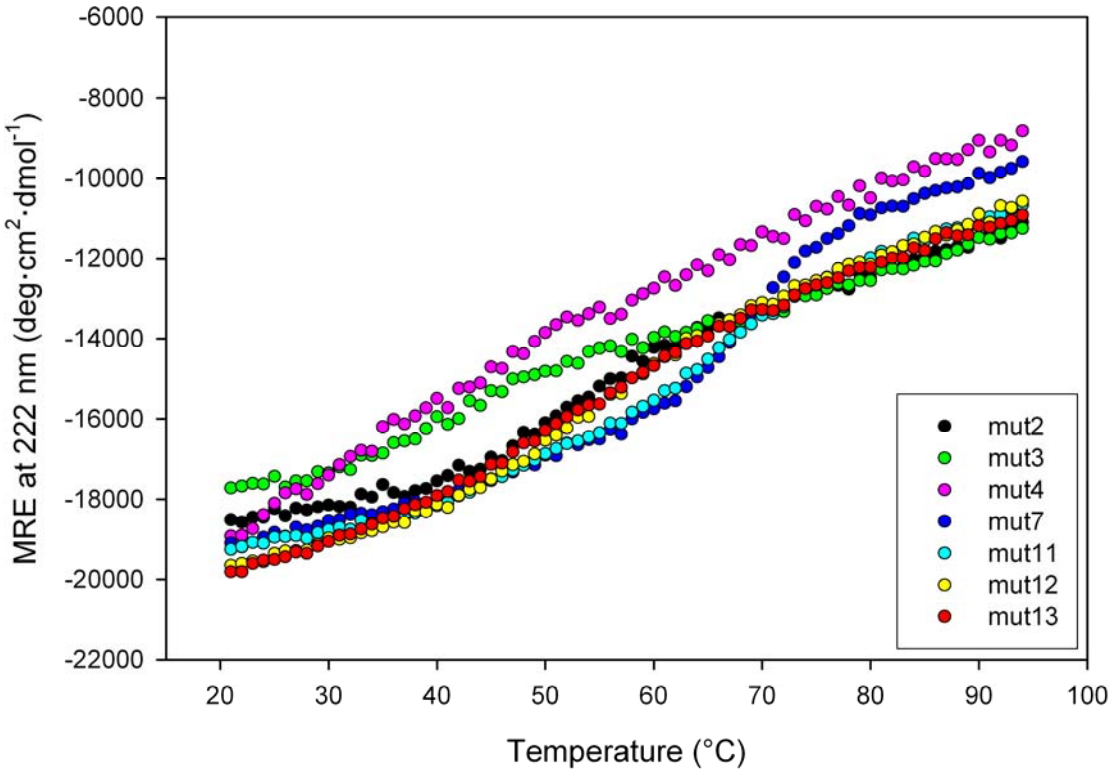


Fig. S5 Thermal denaturation of hydrophobic core mutants. The most promising mutants, based on size exclusion chromatography and ANS binding experiments, are shown. The CD signal at 222 nm is reported as mean residue ellipticity (MRE) as a function of temperature. Remarkably, all the mutants have similar ellipticity at 20°C. mut7 is characterized by the steepest and largest transition.

Table S1: Oligonucleotides used for the assembly and cloning of designed and natural armadillo repeat protein genes

name	sequence 5'-3' direction	description (for=forward, rev=reverse)
AcatFOR	CGGGATCCACACGTGCAATTCCTG	for β-catenin mouse
AcatREV	GCGGTACCATTAGTCCTCAGACATTCGG	rev β-catenin mouse
IMAF5	CGGGATCCCATCACTTCTGACATGATTGAG	for importin-α1 human
IMAR5	GCGGTACCATTACCCGAAGTAATGCTCAATAAG	rev importin-α1 human
pQE_f_1	CGGATAACAATTTCACACAG	forward primer for pQE vectors
pQE_r_1	GTTCTGAGGTCATTACTG	reverse primer for pQE vectors
Ny1F	CCAGGGATCCGAACTGCCGCAGATGACCCAGCAGCTGAACTCTG	for assembly Ny module and amplification
Ny2R	CGGTAGCAGACAGCTGTTCCTGCATGTCGTCAGAGTTCAGCTGCTGGG	rev assembly Ny module
Ny3F	GAACAGCTGTCTGCTACCGTTAAATTCCGTCAGATCCTGTCTCGTGATGG	for assembly Ny module
Ny4R	TTCCTGGTACCCTAAGGTCTCAACCATCACGAGACAGGATCTG	rev assembly Ny module and amplification
Na1F	CCAGGGATCCTCTCTGAACGAACTGGTTAAACAGCTGAACTCCG	for assembly Na module and amplification
Na2R	CTGAGCAGCTTCTTTCAGCTGTTTCTGGTCGTCGGAGTTCAGCTGTTTAACCAG	rev assembly Na module
Na3F	CAGCTGAAAGAAGCTGAAAGAAGCTGCTCAGAAACTGCGTCAGCTGGCTTCCGATGG	for assembly Na module
Na4R	TTCCTGGTACCCTAAGGTCTCAACCATCGGAAGCCAGCTG	rev assembly Na module and amplification
Cy1F	CCAGGGATCCTAGGAAGACCTTGGTGACAACATCAACG	for assembly Cy module and amplification
Cy2R	GCCACCAGCCTTCTCGATGAAGTCCGCGTTCTCGTTGATGTTGTCACCAAGG	rev assembly Cy module
Cy3F	CGAGAAGGCTGGTGGCATGGAGAAGATCTTCAACTGCCAGCAGAACG	for assembly Cy module
Cy4R	GCTTTCTCGTAGATCTTGTCGTTCTCGTTCTGCTGGCAGTTG	rev assembly Cy module
Cy5F	CGACAAGATCTACGAGAAAGCTTACAAGATCATCGAAACCTACTTCGGC	for assembly Cy module
Cy6R	TTCCTGGTACCTCATTAGCCGAAGTAGGTTTCGATG	rev assembly Cy module and amplification
CyM1F	CCAGGGATCCTAGGAAGACCTTGGTAACGAGAACGCGG	for assembly Cm module and amplification
CyM2R	GCCACCAGCCTTCTCGATGAAGTCCGCGTTCTCGTTACCAAGG	rev assembly Cm module
CyM3F	CGAGAAGGCTGGTGGCATGGAGAAGATCTTCAACGCTCAGCAGAACG	for assembly Cm module
CyM4R	GCTTTCTCGTAGATCTTGTCGTTCTCGTTCTGCTGAGCGTTG	rev assembly Cm module

Ca1F	CCAGGGATCCTAGGAAGACCTTGGTAACG AACAGAAACAGGC	for assembly Ca module and amplification
Ca2R	GTTTCTCCAGAGCACCAGCTTCTTTAACA GCCTGTTTCTGTTCGTTACC	rev assembly Ca module
Ca3F	GCTGGTGCTCTGGAGAAACTGGAACAGCT GCAGTCCCACGAG	for assembly Ca module
Ca4R	CCTGAGCTTCTTTCTGGATCTTCTCGTTC TCGTGGGACTGCAGC	rev assembly Ca module
Ca5F	GATCCAGAAAGAAGCTCAGGAAGCTCTGG AGAAGCAGTTCTCCC	for assembly Ca module
Ca6R	TTCCTGGTACCTCATTAGTGGGAGAACTG CTTCTCCAG	rev assembly Ca module and amplification
imp1F	CCAGGGATCCTAGGAAGACCTTGGTAACG AACAGATCC	for assembly importin module and amplification
imp2R	ACCGGCAGAGCACCAGCGTCGATAACAGC CTGGATCTGTTCGTTACCAAGG	rev assembly importin module
imp3F	CTGGTGCTCTGCCGGTTCTGGTTGAACTG CTGTCCTCTCCGGAC	for assembly importin module
imp4R	CCACAGAGCTTCTTTCTGGATCTTGTTGT CCGGAGAGGACAGCAG	rev assembly importin module
imp5F	TCCAGAAAGAAGCTCTGTGGGCTCTGTCT AACATCACTTCTGGTGGTTGAGACC	for assembly importin module
imp6R	TTCCTGGTACCCTAAGGTCTCAACCACCA GAAGTG	rev assembly importin module and amplification
cat1F	CCAGGGATCCTAGGAAGACCTTGGTGAAG C	for assembly catenin module and amplification
cat2R	CCACCAGATTCACGGATAGCCAGTTTGTT AGCTTCACCAAGGTCTTCC	rev assembly catenin module
cat3F	CTATCCGTGAATCTGGTGGTATCCCGGCT CTGGTTCGTCTGCTGTCCTC	for assembly catenin module
cat4R	TAGCAGCTTCCAGGATCTTCTCGTTGTTA GAGGACAGCAGACGAACC	rev assembly catenin module
cat5F	AAGATCCTGGAAGCTGCTACTGGCACTCT GCACAACCTGGCTCTGCATGGTTGAG	for assembly catenin module
cat6R	TTCCTGGTACCCTAAGGTCTCAACCATGC AGAGCC	rev assembly catenin module and amplification
cons1F	CCAGGGATCCTAGGAAGACCTTGGTAACG AACAAATCC	for assembly consensus module and amplification
cons2R	AGCCGGCAGACCACCAGCATCGATAACAG CTTGGATTTGTTCGTTACCAAGG	rev assembly consensus module
cons3F	GGTGGTCTGCCGGCTCTGGTTCAACTGCT GTCCTCTCCGAACG	for assembly consensus module
cons4R	CCAAGCAGCTTCTTTCAGGATCTTCTCGT TCGGAGAGGACAGC	rev assembly consensus module
cons5F	CCTGAAAGAAGCTGCTTGGGCTCTGTCTA ACCTGGCTTCTGGTGGTTGAG	for assembly consensus module
cons6R	TTCCTGGTACCCTAAGGTCTCAACCACCA GAAGCCAG	rev assembly consensus module and amplification
2A-rev	AGCCGGCAGAGCACCAGCATCGATAACAG CTTGGATTTGTTCGTTACCAAGG	rev hydrophobic core mutants assembly
2AVrev	AGCCGGAACAGCACCAGCATCGATAACAG CTTGGATTTGTTCGTTACCAAGG	rev hydrophobic core mutants assembly
2AIrev	AGCCGGGATAGCACCAGCATCGATAACAG CTTGGATTTGTTCGTTACCAAGG	rev hydrophobic core mutants assembly

3A-for	GGTGCTCTGCCGGCTCTGGTTCAACTGCT GTCCTCTCCGAACG	for hydrophobic core mutants assembly
3AVfor	GGTGCTGTTCCGGCTCTGGTTCAACTGCT GTCCTCTCCGAACG	for hydrophobic core mutants assembly
3AIfor	GGTGCTATCCCGGCTCTGGTTCAACTGCT GTCCTCTCCGAACG	for hydrophobic core mutants assembly
4LLrev	CCACAGAGCTTCTTTCAGCAGCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
4LVrev	CCAAACAGCTTCTTTCAGCAGCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
4VLrev	CCACAGAGCTTCTTTCAGAACCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
4L-rev	CCAAGCAGCTTCTTTCAGCAGCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
4-Lrev	CCACAGAGCTTCTTTCAGGATCTTCTCGT TCGGAGAGGACAGC	rev hydrophobic core mutants assembly
5L--for	CTGAAAGAAGCTCTGTGGGCTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5V--for	CTGAAAGAAGCTGTTTGGGCTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5-V-for	CTGAAAGAAGCTGCTTGGGTTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5LV-for	CTGAAAGAAGCTCTGTGGGTTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5VV-for	CTGAAAGAAGCTGTTTGGGTTCTGTCTAA CCTGGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5--Ifor	CTGAAAGAAGCTGCTTGGGCTCTGTCTAA CATCGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5L-Ifor	CTGAAAGAAGCTCTGTGGGCTCTGTCTAA CATCGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
5LVIfor	CTGAAAGAAGCTCTGTGGGTTCTGTCTAA CATCGCTTCTGGTGGTTGAG	for hydrophobic core mutants assembly
6Irev	TTCCTGGTACCCTAAGGTCTCAACCACCA GAAGCGAT	rev hydrophobic core mutants assembly and amplification

Table S2 : oligonucleotides used in assembly of single internal modules

Mutant	oligo #1	oligo #2	oligo #3	oligo #4	oligo #5	oligo #6
mut1	cons1F	cons2R	cons3F	4LLrev	5L--for	cons6R
mut2	cons1F	2A-rev	3A-for	4VLrev	5L--for	cons6R
mut3	cons1F	2A-rev	3A-for	4LLrev	5L--for	cons6R
mut4	cons1F	2A-rev	3A-for	4L-rev	5--Ifor	6Irev
mut5	cons1F	2A-rev	3A-for	4L-rev	5-V-for	cons6R
mut6	cons1F	2A-rev	3A-for	4L-rev	cons5F	cons6R
mut7	cons1F	2A-rev	3A-for	4-Lrev	5L-Ifor	6Irev
mut8	cons1F	2AVrev	3AVfor	4LLrev	5L--for	cons6R
mut9	cons1F	2AIrev	3AIfor	4LLrev	5L--for	cons6R
mut10	cons1F	2A-rev	3A-for	4LVrev	5V--for	cons6R
mut11	cons1F	2A-rev	3A-for	4VLrev	5LV-for	cons6R
mut12	cons1F	2A-rev	3A-for	4LLrev	5LV-for	cons6R
mut13	cons1F	2AIrev	3AIfor	4LLrev	5LV-for	cons6R
mut14	cons1F	2A-rev	3A-for	4LVrev	5VV-for	cons6R
mut15	cons1F	2A-rev	3A-for	cons4R	cons5F	cons6R
mut16	cons1F	cons2R	cons3F	4LLrev	5L-Ifor	6Irev
mut17	cons1F	2AVrev	3AVfor	4LLrev	5LVIfor	6Irev
mut18	cons1F	2AIrev	3AIfor	cons4R	cons5F	cons6R
mut19	cons1F	cons2R	cons3F	4L-rev	cons5F	cons6R

Summary of Table S3

Table S3 shows the summary of the results from the simulated annealing approach for the 432 mutants (left column) and the original structures (in the upper part).

In the top rows, energy values for the original crystal structures after simulated annealing are indicated. 2bct, 1q1t, 1ee4 are the original structures of mouse β-catenin, mouse importin-α and yeast importin-α, respectively. Catm, impm, impy are the starting models derived from the original structures where the capping repeats have been replaced by the designed capping repeats Ncap and Ccap.

In the leftmost column (#), each hydrophobic core mutant is identified by a number, from 0 to 431 and the mutants are ordered according to the total rank (second leftmost column). These values do not take into account electrostatic contributions. The value used for the ranking (m1m) is the sum of the ranks of median (me), first percentile (1) and minimum (mi) in all the three starting structures. A second value (m1) is given, which denotes the sum of median and first percentile among the three structures. Individual ranks are also reported for median (me), first percentile (1), minimum (mi) referring to each starting structure. On the right of the ranks, the corresponding potential energy values (expressed as kcal/mol) are indicated for median, first percentile (1^{st} perc.) and minimum in each structure.

For each mutant the hydrophobic core residues are indicated as a change, compared to the original C-type consensus: the dashes indicate no change. The amino acids are indicated in single letter code. The positions not mutated are indicated in gray.

The average volume for core residues of internal repeats is indicated for mouse β-catenin, mouse importin-α, yeast importin-α (corresponding PDB ID are indicated) and for the C-type consensus. In the case of mutants, the core volume is expressed as difference to the core volume of C-type consensus sequence: V(mut)-V(cons). Volumes were calculated according to Chothia [1] in units of $Å^3$.

The mutants selected for expression and characterization are indicated left of the first column with the corresponding name. The first eight mutants, with a ranking value better than the C-type consensus (#264) were selected. Other high ranking mutants, even if not present in the very top group, were selected because of particularly interesting sequences or results in the ranking process:

mut5 (#2): moderate volume increase, comparison with #8, same composition, only position 27 and 28 exchanged.

mut6 (#0): equivalent to consensus among the first 50 top mutants. Always good rank, only high median in importin mouse structure; it was thus considered to be a good candidate

mut7 (#53): most similar to importin consensus sequence

mut10 (#8): moderate volume increase, comparison with #8, same composition, only position 27 and 28 exchanged.

mut13 (#78): high volume and several β-branched residues

mut14 (#10): high volume and several β-branched residues

mut15 (#48): simplest mutation with volume gain (G->A) compared to consensus

mut16 (#221): good rank, only high median in catenin mouse structure; it was thus considered to be a good candidate

mut17 (#151): high volume and several β-branched residues

mut18 (#120): good rank, only high median in catenin mouse structure; it was thus considered to be a good candidate

mut19 (#216): highest ranking mutant with low core volume

References

1. Chothia, C. (1975). Structural invariants in protein folding. *Nature* **254**, 304-8.