doi:10.1016/S0022-2836(03)00897-0 J. Mol. Biol. (2003) 332, 471–487

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Designing Repeat Proteins: Modular Leucine-rich Repeat Protein Libraries Based on the Mammalian Ribonuclease Inhibitor Family

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

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

We present a novel approach to design repeat proteins of the leucine-rich repeat (LRR) family for the generation of libraries of intracellular binding molecules. From an analysis of naturally occurring LRR proteins, we derived the concept to assemble repeat proteins with randomized surface positions from libraries of consensus repeat modules. As a guiding principle, we used the mammalian ribonuclease inhibitor (RI) family, which comprises cytosolic LRR proteins known for their extraordinary affinities to many RNases. By aligning the amino acid sequences of the internal repeats of human, pig, rat, and mouse RI, we derived a first consensus sequence for the characteristic alternating 28 and 29 amino acid residue A-type and B-type repeats. Structural considerations were used to replace all conserved cysteine residues, to define less conserved positions, and to decide where to introduce randomized amino acid residues. The so devised consensus RI repeat library was generated at the DNA level and assembled by stepwise ligation to give libraries of 2–12 repeats. Terminal capping repeats, known to shield the continuous hydrophobic core of the LRR domain from the surrounding solvent, were adapted from human RI. In this way, designed LRR protein libraries of 4–14 LRRs (equivalent to 130–415 amino acid residues) were obtained. The biophysical analysis of randomly chosen library members showed high levels of soluble expression in the *Escherichia coli* cytosol, monomeric behavior as characterized by gel-filtration, and α-helical CD spectra, confirming the success of our design approach.

Keywords: combinatorial library; consensus sequence; leucine-rich repeat; protein design; ribonuclease inhibitor

*Corresponding author

Introduction

The search for protein frameworks with binding properties similar to antibodies was initiated once technologies for synthetic library generation and selection systems became available. In the past decade, many alternative framework libraries have been generated.[1,2] A number of desirable properties for novel designed binding molecules can be defined: Besides high thermodynamic stability and high-level soluble expression in bacteria, the stability of designed binding molecules should be independent of disulfide bonds in order to work in intracellular and extracellular environments. In addition, designed binding molecules should possess a sufficiently large surface suitable for target binding, and they should be tolerant of surface variations. Nevertheless, most scaffolds meet only a fraction of these requirements or fall short of antibodies in terms of affinity or the range of targets recognized.[1,2]

In addition to classical antibody applications, designed binding molecules could be a powerful tool in functional genomics, where specific protein

Abbreviations used: AR, ankyrin repeat; IMAC, immobilized metal-ion affinity chromatography; LRR, leucine-rich repeat; NTA, nitrilotriacetic acid; PDB, Protein Data Bank; RI, ribonuclease inhibitor; TBS, Tris-buffered saline.

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

knockouts will help to elucidate intracellular protein networks. Hence, there is a particular interest in proteins that are stable in the absence of disulfide bonds while binding targets specifically under physiological conditions.[3] Additionally, the absence of unpaired cysteine residues will preclude air oxidation and thus facilitate protein handling. Also, the spatial separation of framework and recognition function would be of crucial importance for the design of binding molecules, so that positions important for binding do not compromise the properties of the framework.

Besides the immunoglobulin framework used in the immune system, nature has adopted repeat proteins as protein scaffolds to support a wide range of protein–protein interactions.[4] To underline the importance of repeat proteins in nature, it should be mentioned that the innate immune system relies mainly on LRR proteins for target recognition.[5] Therefore, we wished to investigate whether LRR proteins could be used to generate a radically different class of designed binding molecules. LRR proteins, like other natural repeat proteins, are composed of consecutive structural units, or repeats, that stack to form elongated protein domains providing large interaction surfaces.[6] Each repeat potentially contributes to target binding by the amino acids in designated surface positions, whereas the framework is stabilized by the amino acids in framework positions, thus achieving a spatial separation of framework and recognition function. Furthermore, repeats are sometimes encoded by single exons,[7,8] and evolution has varied both the number and the position of single repeats within a given domain.[9] To exploit this modularity for the design of binding molecules, we developed a strategy based on self-compatible repeat modules.[10] The implementation of our strategy for ankyrin repeat (AR) proteins is described in the accompanying paper,[11] and is described here for LRR proteins. Libraries generated according to this strategy will allow the unlimited assembly and shuffling of repeat modules. Most importantly, binding proteins of adjustable size can be created, which would open a new dimension in the selection and affinity maturation of binding molecules.

Natural LRR proteins participate in many protein–protein interactions.[12,13] Well-known examples are the mammalian RI,[14] the extracellular domains of Toll-like receptors of the innate immune system,[5] the bacterial internalins,[15] and the plant disease-resistance R proteins.[16] Since the determination of the first crystal structure,[17] it has been clear that the LRR sequence corresponds to a structural motif. The LRR consists of a β-strand connected by variable loops to a helical or extended part (Figure 1(a)), and an initial classification in six LRR subfamilies has been proposed, based on repeat length and specific positions of conserved residues.[4] Many LRRs of one subfamily stack to form a protein domain (Figure 1(b)) with a continuous hydrophobic core maintained by leucine and other hydrophobic amino acid residues. The elongated and curved shape of LRR domains support large interaction areas with unequaled affinities. For example, an area of about 2900 Å^2 is buried when RI binds angiogenin[18] and this large interaction area leads to inhibition constants in the low femtomolar range.[19] Remarkably,

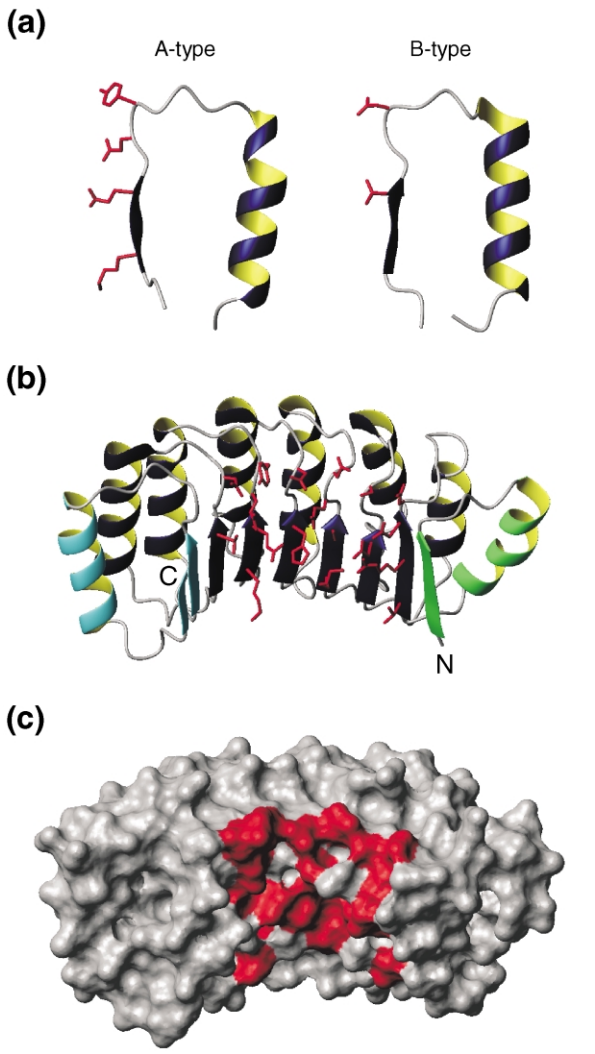

Figure 1. Model of a designed LRR protein based on human RI (PDB: 1A4Y[18]). The side-chains of amino acid residues at randomized positions are highlighted in red. (a) A-type (amino acid residues 142–169 of human RI) and B-type (residues 170–198) RI-like LRR are shown in ribbon representation. (b) Ribbon representation of a designed N3C LRR protein. The abbreviation N3C refers to a protein composed of an N-terminal capping repeat, followed by three double repeat modules and a C-terminal capping repeat. RIs would correspond to N7C molecules in our nomenclature. The α-helix and the β-strand of the N-terminal capping repeat are shown in green, whereas the secondary structure elements of the six internal repeats are shown in dark blue, and those of the C-terminal capping repeat are shown in light blue. (c) Surface representation of the same molecule as in (b). Randomized positions are colored in red, all other positions are in grey. The Figure was generated with MOLMOL.[45]

the interaction relies on 13 independent patches recruited from very distant parts of the RI. Taken together, all these characteristics urged us to investigate whether LRRs are suitable to build binding protein libraries.

We chose the mammalian RI family as our starting point, since it has been relatively well studied and the three-dimensional structures of human RI in complex with angiogenin[18] as well as that of pig RI with and without bound RNase A[17,20] have been determined. Moreover, the cytoplasmic occurrence and the outstanding affinities of RI make the LRR proteins very attractive for the design of intracellular binding molecules. To design novel binding molecules on the basis of mammalian RIs, we asked the following questions: How can we design a self-compatible LRR module that can be assembled into an LRR domain? What other structural elements are required to stabilize proteins assembled from self-compatible LRR modules? Can these proteins be expressed in soluble form in *Escherichia coli*? How does the number of repeat modules influence expression, solubility and stability?

Results

The design, construction and biophysical analysis of LRR protein libraries based on the mammalian RI family is described. Human RI has 460 amino acid residues,[21] contains altogether 16 wedge-shaped LRRs, and forms 1:1 complexes with pancreatic RNases.[22] Since the biophysical properties of LRR proteins have not been studied extensively, and since it is not clear which residues are crucial for folding, stability and binding specificity, we first analyzed the available amino acid sequences and crystal structures of the mammalian RI family as a basis for our library design.

General considerations for the design of self-compatible LRR modules

The RI-type LRRs are somewhat special in the LRR family, since the repeats are relatively long (28 or 29 amino acid residues *versus* 20–26 in other LRR subfamilies) and since two types of repeats are strictly alternating (Figure 2(a)). The two alternating repeats have been termed A-type and B-type[17] and do not seem to occur by themselves. The secondary structure elements of both repeat types are similar, with a β-strand encompassing positions 2–8 of the LRR, followed by a loop and an α-helix encompassing positions 14–27 (Figures 2(a) and 3). Since we wanted to construct LRR proteins where repeat modules could be exchanged, added or deleted easily, we reasoned that a self-compatible repeat module corresponding to the RI-type LRR should be designed (for a more detailed discussion, see Forrer *et al.*[10]). Therefore, we had to analyze whether this self-compatible repeat module would correspond to one LRR or a double repeat module of both A-type and B-type LRR. The comparison of both LRRs revealed obvious differences of A-type and B-type RI-like LRR, for example, the amino acid at position 10 (Figure 2(a)), where a sharp turn terminates the β-strand (Figure 3). Whereas the A-type carries a cysteine residue, the B-type has asparagine, which is most common across all LRR subfamilies. Nevertheless, both residues have been described to participate in the asparagine ladder of LRR proteins, a hydrogen bond network supposedly stabilizing adjacent LRRs.[23] Furthermore, structural inspection of all known RI-like LRRs revealed differences of A and B-type in the helical part (Figures 1(a) and 2(a)): Whereas the helix in the B-type is straight and parallel to the preceding β-strand, the first turn of the helix in the A-type is tighter and the second part of the helix is inclined with respect to the β-strand (Figure 1(a)). Additionally, different hydrogen bonds are found: For example, the side-chain of serine or threonine in position 13 is bonded to the main-chain nitrogen atom of the amino acid residue in position 15 in the A-type LRR. In the B-type LRR, the side-chain of aspartate in position 14 is hydrogen bonded to the side-chain of the serine or threonine residue in position 13 of the following A-type LRR. In position 16 of the B-type, glycine is 100% conserved, since there is no space for a side-chain, whereas in the A-type repeat also bigger side-chains are found. Position 17 is 100% cysteine in the A-type, whereas the B-type shows a number of small hydrophobic residues. Position 23 of the B-type again shows high preference for glycine, whereas in the A-type small hydrophobic amino acids are found; again steric reasons seem the most likely explanation. Finally, the positions of the helix-terminating proline residues are obviously different. In the A-type, position 28 is often occupied by proline, connecting the A-type LRR directly to the B-type (Figure 3(a)) and leading to an overall inclined α-helix relative to the β-strand. The situation in the B-type, however, is different: proline is found mostly in position 27 followed by glycine in position 28, which results in an α-helix axis parallel with the β-strand, because the kink is initialized from an earlier part of the last turn in the α-helix. Position 29 is found only in the B-type and bridges the gap from the relatively distant α-helix to the following β-strand (Figure 3(b)).

All these differences suggested to incorporate both the A-type and the B-type repeat into a double LRR module with 57 amino acid residues for the LRR design. We think that the interrepeat interfaces of the A-type and B-type repeats are compatible with only the alternating type. In other words, only the combination of the two repeat types leads to self-compatible building blocks, which can stack to form protein domains. To achieve self-compatibility, the consensus sequence of the 57 amino acid repeats was determined, which should filter out specific repeat incompatibilities that might have arisen during evolution.

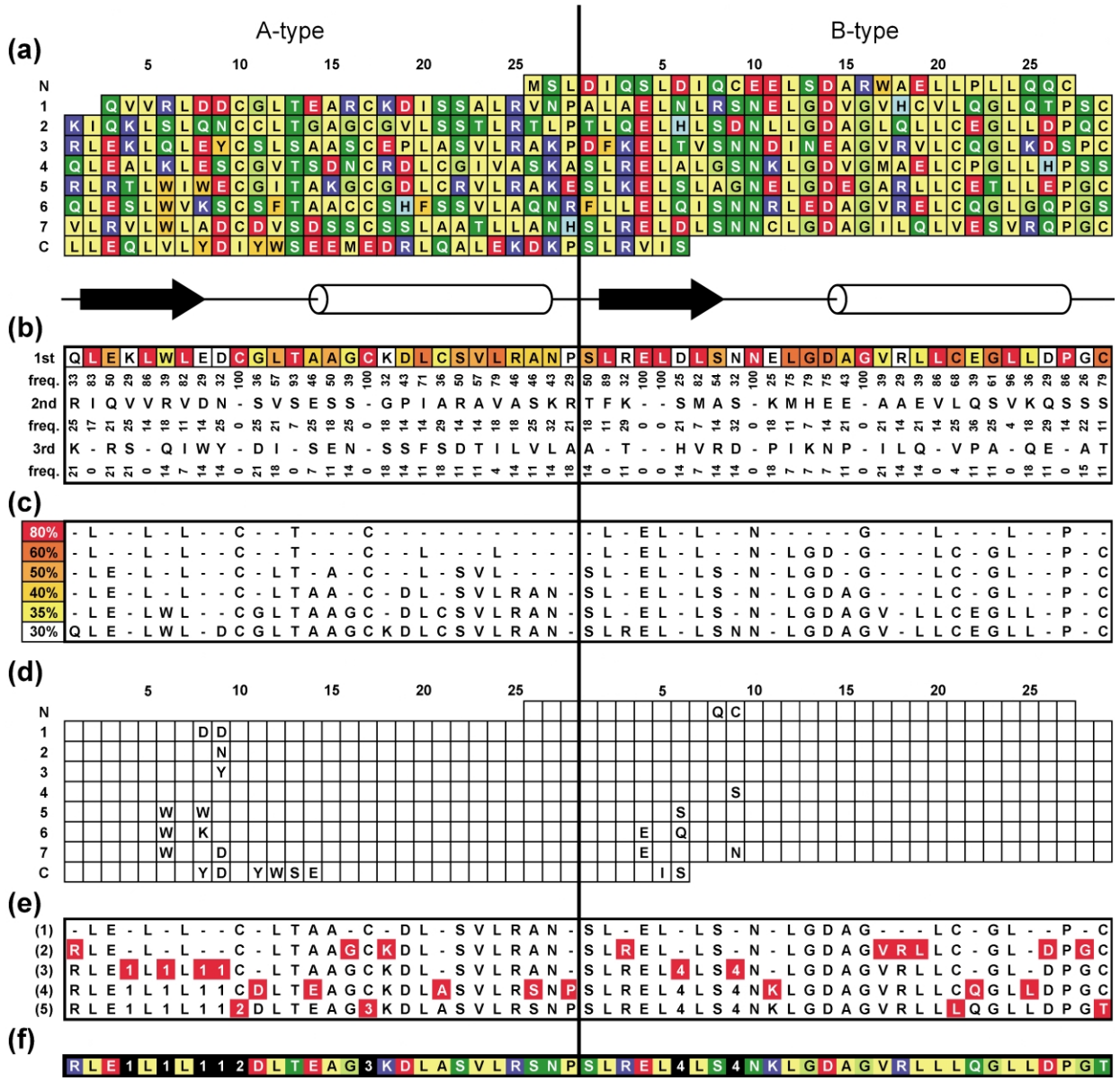

Figure 2. Structure-based LRR alignment of human RI and statistical analysis of the mammalian RI family consensus. (a) LRR alignment of human RI colored according to amino acid type (yellow, aliphatic; orange, aromatic; green, polar; red, acidic; blue, basic; light green, glycine). The N-terminal capping repeat is followed by seven double repeats and the C-terminal capping repeat as indicated. The LRR secondary structure elements are shown below the sequence. (b) Occurrence (as percentage) of the three most frequent amino acid residues found in the statistical analysis of human, pig, rat and mouse RIs aligned to fit the LRR consensus.[17] Amino acid residues occurring in more than 80% of the respective repeat positions are shown on a red background, less frequent positions are shown in decreasing intensities of orange as indicated in (c), which shows consensus LRR sequences determined at different threshold values as indicated at the left. Positions with less than the indicated frequency are marked by a dash, e.g. positions 4, 8, and 28 of the A-type RI-like LRR and positions 6, 11, 18, 26, and 28 of the B-type RI-like LRR in case of less than 30% conservation. Many differences between the two repeat types become apparent, e.g. positions 13 and 14 of the two LRR types. (d) Overview on all residues of human RI interacting with angiogenin[18] indicated in the LRR alignment. (e) Design of a sequence motif encoding an LRR double repeat module with 57 amino acid residues (for details, see the text). Consensus (1) shows the sequence motif for a threshold of 40% frequency of one amino acid. Consensus (2) shows the sequence motif after the identification of the conserved type of amino acid at positions 1, 16, and 18 of the A-type LRR and positions 2, 17, 18,19, 26, and 28 of the B-type LRR (positions are highlighted in red). Consensus (3) shows the sequence motif after the analysis of interacting residues used by RI to bind its target proteins. Consensus (4) shows the sequence motif based on additional structural analysis of non-conserved positions. Consensus (5) shows the sequence motif after replacement of all conserved cysteine residues. The digits denote mixtures of amino acids of the following compositions: 1 stands for D, E, N, Q, R, K, S, Y, W; 2 for N, S, T; 3 for L, M, V; and 4 for G, D, N, S, T, H. (f) Consensus sequence adopted for the construction of the designed LRR proteins following sequence and structural analyses. Colors are as in (a), randomized positions are shown in black. The digits are abbreviations for mixtures of amino acids as explained in Table 2.

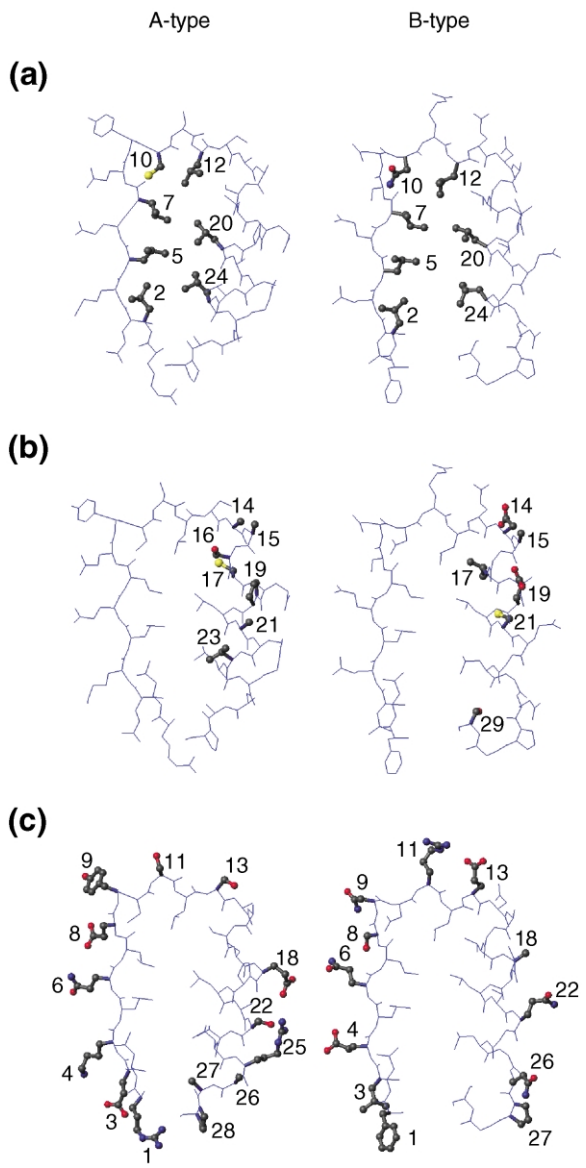

Figure 3. Structural comparison of A-type and B-type RI-like LRRs of human RI. Amino acid residues 142–169 of human RI were used as a representative A-type LRR, whereas amino acid residues 341–369 were used as a representative B-type LRR. The orientation of the LRRs is as in Figure 1(a). (a) Side-chains of amino acids mediating intrarepeat interactions are shown in ball and stick representation. Both A-type and B-type rely on hydrophobic amino acids in positions 2, 5, 7, 12, 20, and 24. Position 10 is occupied by Cys in the A-type and Asn in the B-type. (b) Side-chains of amino acid residues mediating interrepeat interactions are shown in ball and stick representation. Note that positions 16, 23, 25, and 28 of the B-type RI-like LRR are occupied by glycine and therefore no side-chain is present. Different amino acid preferences between A-type and B-type LRR are obvious in positions 14, 16, 17, 19, 21, 23, and 29. Note, for example, the conserved Cys at positions 17 of the A-type, where Val is preferred in the B-type. (c) Side-chains of amino acid residues exposed to the surrounding solvent are shown in ball and stick representation. The helix terminating proline residues in position 28 of the A-type and position 27 of the B-type induce the kink connecting to the following repeat. The Figure was generated with MOLMOL.[45]

The consensus so obtained should facilitate the definition of positions amenable to randomization for the desired binding function. Furthermore, consensus protein design has improved expression yields and stability of immunoglobulins and other proteins in previous studies.[24–26] The steps of the mammalian RI consensus calculation are described below.

A mammalian RI consensus

Four protein sequences of the RIs from man, pig, rat, and mouse (Swiss-Prot accession numbers: P13489, P10775, P29315, and GenBank accession number AAK68859, respectively) were used to derive an RI-type LRR consensus. Despite extensive database searches, no other RI-like LRR sequence was found at the time of the initial alignments. The sequence identities of all four RIs are above 74% (similarity: above 79%) for each pair, mouse and rat being the closest homologs with 91% identity (similarity: 93%). Initial alignments to fit the hallmark LRR motif $LxxLxLxx^{N}/_{C}xL$ resulted in 14 internal LRR repeats for each sequence.[4] None of the mammalian RI-like LRR displays an insertion or deletion, which makes the alignment straightforward. The internal repeats are flanked at both termini by different LRR sequences, which we call capping repeats. The N-terminal capping repeat contains 30 amino acid residues and the C-terminal capping repeat contains 34 amino acid residues in the case of the human RI (Figure 2(a)). Since these capping repeats clearly deviate from the internal repeats, they were treated separately in the following analysis (see Capping repeats, below). The 14 internal LRR repeats showed the expected alternating pattern of seven A-type and seven B-type RI repeats (Figure 2(a)). To obtain the consensus for the 57 residue A-type/B-type double repeat module, the mammalian RI protein sequences were again aligned internally (e.g. human RI, Figure 2(a)) and the frequency of each amino acid residue at each position was calculated over all four protein sequences (Figure 2(b)).

In the following, the position of amino acids in the A-type or B-type RI-like LRR are denoted by Ax or Bx, respectively, where x refers to the previously defined numbering (shown in Figures 2 and 3) within a LRR.[17] Among the positions with less than 40% frequency of one residue type, the following positions were readily defined by identifying the conserved type of amino acid: Position A1: 25% of Arg and 21% of Lys together argued for a positively charged amino acid, so Arg was chosen. Position A16: 39% Gly together with 25% Ser argued for a small amino acid, so Gly was chosen. Position A18: 32% Lys, 7% Arg, 14% Glu indicated that charged amino acids would be preferred, so Lys was chosen. Position B3: 32% Arg and 29% Lys argued for a positively charged amino acid, so Arg was chosen. Position B17: 39% Val, 21% Ala, and 21% Ile suggested strongly that

aliphatic amino acids were preferred, so Val was chosen. Position B18: 29% Arg and 7% Lys pointed to a positively charged amino acid, so Arg was chosen. Position B19: 39% Leu together with 7% each of Ala and Val suggested that hydrophobic amino acids were preferred, so Leu was chosen. Position B26: 29% Asp, 29% Gln, 11% Glu, and 7% Asn suggested that charged or polar amino acids were preferred, so Asp was chosen. Position B28: 26% Gly and 22% Ser indicated that a small amino acid would be preferred, so Gly was chosen.

At this point, the suitability of the adopted decisions was checked manually to verify the compatibility of the chosen amino acids. Only positions A18 and B18 seemed to be questionable, because electrostatic repulsion between the chosen Lys and Arg might occur. However, since this combination of positive charges was found also in double repeats 3–5 of human RI (Figure 2(a)), and the negatively charged Asp had been chosen at position A19, these choices were kept. Taken together, consensus (2) was obtained (Figure 2(e); newly defined positions are highlighted in red). For the remaining 12 non-conserved positions, a more detailed analysis including structural and functional data was undertaken.

Target interaction positions

A detailed description of the interacting residues of RI in complex with RNase A or angiogenin has been published,[18] and we performed the internal alignment of these amino acid residues (Figure 2(d)). The alignment indicated that positions A6, A8, and A9 and positions B4, B6 and B9 are used frequently for RNase binding. All these positions are part of the β-sheet region and have side-chains projecting from the β-sheet (Figure 3(c)). The calculated frequencies of each amino acid of these interaction positions in the mammalian RI family revealed that Trp at position A6 occurred with the highest frequency (39%), whereas all other amino acids were found with a frequency of 32% or less. The choice of a threshold of 40% conservation for the first consensus therefore most likely did not include any target interaction position. Among the positions with 30% or less conservation, position A4 was included in the target interaction positions, since it exhibits Lys (29%), Val (25%), Ser (21%), Thr (18%), Glu (4%), and Ala (4%), suggesting that all residue types can occur at this position. Position B4, however, which also interacts with RNase A, was not included, since Glu is 100% conserved and might have a structural importance. The decision, which amino acids to allow in each position, was guided by the following findings. Positions A6, A8, and A9 are almost always occupied by big, polar or charged amino acid residues, whereas in positions B6 and B9 a tendency to small amino acid residues was observed. For example, position B6 is occupied by Asp (25%), Ser (21%), His (14%), Gln (14%), Thr (7%), Val (7%) and Ala, Asn, Cys (less than 4% each). No structural role of the side-chains of these target interaction positions was obvious. Guided by the distribution of binding residues observed in mammalian RI, we used positions A4, A6, A8, A9 and B6, B9 for randomization (Figure 1(a)). The use of trinucleotide mixtures[27] for oligonucleotide synthesis allowed us to freely define which amino acids should be used in which position. As suggested by mammalian RI, we used mainly polar, charged and aromatic residues in the A-type target interaction positions and smaller residues in the B-type target interaction positions (Figure 2(e)). For better orientation, a surface plot of a space-filling model of an N3C library member is shown (Figure 1(c)). The abbreviation N3C refers to a designed LRR protein composed of an N-terminal capping repeat, followed by three double repeat modules and a C-terminal capping repeat. Note that this nomenclature is different from the designed AR proteins,[11] since we count a double repeat as "one" in the case of the designed LRR proteins. Taken together, consensus (3) was obtained (Figure 2(e); newly defined positions are highlighted in red). The theoretical diversity of one double repeat module is 2.13×10^6, an N2C library thus has a theoretical diversity of above 10^{12}, the theoretical diversity for an N4C library would even be above 10^{25}.

Defining non-conserved positions

The arbitrarily chosen threshold of 40%, which resulted in consensus (1), defined 36 of 57 amino acid residues for the double LRR module (Figure 2(c)). A further nine positions were defined on the basis of a conserved type of amino acid (consensus (2)) and another six positions were used for randomization, as they might be used for target interaction (consensus (3)). For the remaining six undefined positions a detailed structural analysis was undertaken, since no conserved sequence pattern was obvious, and they had not been implicated in target interaction. However, since not all known RI complexes have been crystallized, amino acids at non-conserved positions may still have a function in other protein–protein interactions. Another possibility is that the residue type at these positions simply has no particular structural or functional role. Because extremely high affinities were reported for mammalian RIs, we followed the hypothesis that sufficient target binding was accomplished *via* the β-sheet surface alone. As a consequence, we decided to choose one particular amino acid in each of the remaining positions based on structural considerations. In position 11 of both A-type and B-type, the amino acid side-chain is exposed to the solvent, so that a hydrophilic amino acid should be preferred. Since none of the residues in position 11 displayed inter-repeat side-chain H-bonds, A11 was arbitrarily set to Asp and B11 to Lys. This mimics the situation in double repeat 7 of human RI (Figure 2(a)), where Asp378 and Arg351 occupy these positions.

Position A21 had 36% Cys, 29% Ala, and 18% Ser in the RI consensus. The side-chains are located at the interface between the hydrophobic core and surface and mediate interrepeat contacts (Figure 3(b)). Alanine was chosen at position A21, since it should be a good compromise and has a high helical propensity. Position A28 had 29% Pro, 21% Arg, and 18% Ala in the RI consensus. Since the α-helix in the A-type LRR is often terminated by proline, this amino acid was adopted for position A28. Position B22 had 39% Glu and 36% Gln. Since the positions B18 and B26 above and below in the α-helix were already occupied by the charged amino acids Arg and Asp, Gln was chosen to prevent electrostatic repulsion. Position B25 had 36% Leu, 18% Lys, and 18% Gln. Again, the side-chain of the amino acid at this position was at the interface between hydrophobic core and surface. Since B26, however, was already set to Asp, and the combination of Leu at B25 and Glu or Asp at B26 occurred frequently in human RI, position B25 was chosen to be Leu. Further structural inspection of all crystal structures for conserved hydrogen bonds or salt-bridges did not suggest further changes. Two positions, however, were modified to increase the polarity of the surface of the designed LRR proteins. In position A14, where 46% Ala, 25% Glu, and 7% Ser were found, Glu was chosen, since no steric limitation was obvious. This increases the number of negative charges and lowers the pI, thus mimicking the situation in mammalian RI (Table 1). Similarly, position A26 (46% Ala, 25% Ser, and 11% Val) was set to Ser, since the side-chain clearly points to the solvent. Taken together, consensus (4) was obtained (Figure 2(e); newly defined positions are highlighted in red).

Replacing the cysteine residues

The proteins of the mammalian RI family possess a large number of unpaired cysteine residues (32 in human RI).[18] Positions A10 and A17 of the RI-type LRR consensus are occupied exclusively by Cys, and positions B21 and B29 still have about 70% Cys (Figure 2(b)). These cysteine residues of RI seem to be involved in RI inactivation by oxidation, and this mechanism was shown to lead to protein degradation in cell culture.[28] It is not known whether there is an additional structural importance for the Cys conservation. Since we set out to create repeat proteins for uses independent of the redox potential, we wanted to replace all cysteine residues.

As outlined above, the amino acid at position A10 is known to participate in the asparagine ladder.[23] Except for the cysteine-containing LRR subfamily, which also carries Cys at position 10, all other LRR subfamilies have Asn as the most frequent amino acid at position 10. Other residues found frequently at position 10 include Ser or Thr (cf. the LRR protein with the GenBank accession number CAA99846). Thus, we decided to use the mixture of Asn, Ser, and Thr for position A10 of the designed LRR module. Position A17 is involved in interrepeat stacking and participates in the continuous hydrophobic core. The comparison of residues found in position A17 with those in position B17, which is structurally similar (Figure 3(b)), showed that mostly large hydrophobic residues are used. Hence, we opted for a mixture of Leu, Val, and Met to replace the cysteine residues in position A17. In positions B21 and B29, the second most frequently occurring amino acids were chosen: Leu (25%) at position B21 and Thr (11%) at position B29. Taken together, consensus (5) was obtained (Figure 2(e); newly defined positions are highlighted in red). Consensus (5) is also shown colored according to amino acid type (Figure 2(f)) to facilitate comparison with the LRRs of human RI shown in Figure 2(a).

Capping repeats

When the DNA sequence of pig RI was first described, the authors noticed that 15 internal LRRs were flanked by short differing terminal amino acid stretches.[29] Nevertheless, the structure of pig RI revealed that the molecule was composed of 16 LRRs and an additional β-strand at the C terminus. This apparent contradiction could be resolved by defining the starting point of an LRR

Table 1. Biophysical properties of designed LRR proteins and mammalian RIs

Name	N3C	N4C	N5C	N6C	Human RI[a]	Pig RI[a]	Rat RI[a]	Mouse RI[a]
Length (amino acid residues)	244	301	358	415	461	456	456	456
Number of LRRs[b]	8	10	12	14	16	16	16	16
Molecular weight[c] (Da)	26,981	33,363	39,610	45,557	49,973	49,023	49,905	49,816
Extinction coefficient[c] (M^{-1} cm^{-1})	15,220	22,190	23,470	17,780	39,900	34,470	41,440	41,440
pI	5.46	4.67	4.93	5.05	4.54	4.59	4.47	4.51
Arg + Lys	29	33	42	50	40	36	40	41
Asp + Glu	36	49	56	62	62	58	63	61
Net charge	−7	−16	−14	−12	−22	−22	−23	−20
Cys	0	0	0	0	32	30	30	30

[a] All mammalian RIs correspond to N7C, i.e. have 16 LRRs.
[b] Counting all LRRs. Note that in our nomenclature "N3C" refers to a LRR protein with two capping repeats and three double repeat modules, i.e. 8 LRRs.
[c] Calculated with PEPTIDESORT (Wisconsin Package Version 10.2, Genetics Computer Group, USA).

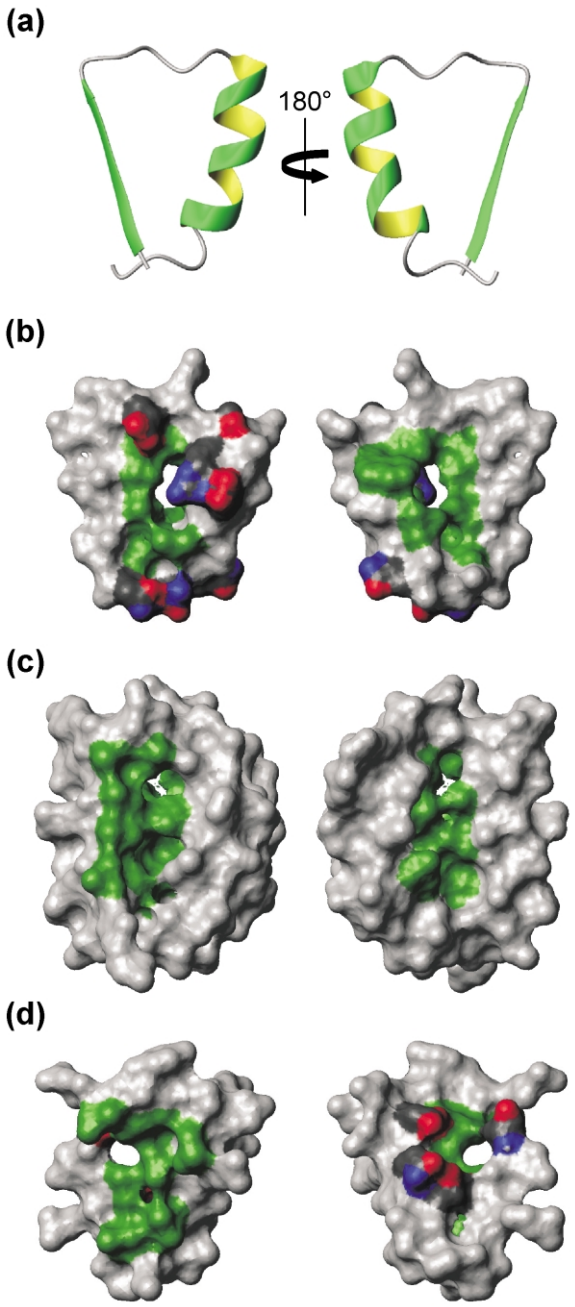

Figure 4. Comparison of surfaces or interrepeat interfaces of internal and capping LRRs of human RI. All representations shown on the left are oriented so that the β-strand is on the left as in (a). To visualize the other faces, each representation has been rotated by 180° around the vertical axis as indicated. Side-chains of amino acid residues belonging to the hydrophobic core are colored in green. Residues atypical for otherwise conserved LRR positions have been colored according to atom type. (a) A ribbon representation of the N-terminal capping repeat. (b) Surface of the N-terminal capping repeat (corresponds to amino acid residues 5–29 of human RI) viewed from the surrounding solvent (left) and viewed from inside the repeat stack (right). The side-chains of residues deviating from the consensus (here, Gln5, Glu12, Arg18, Gln27, and Gln28) are colored according to atom type (N, blue, O, red) and show the adaptation of the N-terminal LRR to the hydrophilic surrounding. (c) Surface of an internal 57 amino acid residue double repeat module (composed of amino acid residues 142–198 of human RI) viewed from the N-terminal side (left) and viewed from the C-terminal side (right). (d) Surface of the C-terminal capping repeat (amino acid residues 435–460) viewed from inside the repeat stack (left) and viewed from the surrounding solvent (right). The side-chains of residues deviating from the consensus (here: Glu443, Gln447, Glu450, and Ser460) are colored according to atom type (N, blue, O, red) and show the adaptation of the C-terminal LRR to the hydrophilic surrounding. The Figure was generated with MOLMOL.[45]

to be at the beginning of each β-strand,[17] so that the short differing terminal peptides became complete but specialized capping LRRs. The optimal alignment of the N and C-terminal LRRs of human RI to the internal LRRs is shown in Figure 2(a) and the boundaries between capping repeats and internal repeats are indicated. The three-dimensional structures of the terminal LRRs suggest the classification of the N-terminal LRR as B-type RI LRR homologue, whereas the C-terminal LRR is homologous to A-type RI LRRs. This classification is also in agreement with the strict alternation of A and B-type LRR.

Our classification of capping LRRs reveals important differences of capping repeats and internal repeats (Figures 2(a) and 4). Several hydrophobic positions of the internal repeats are occupied by polar amino acids in the capping repeats. In human RI, the side-chain of glutamate 12 (at LRR position 10) of the N-terminal LRR shields part of the hydrophobic core (Figure 4(b)) from the solvent. Similarly, Arg18 and Gln27 (at LRR positions 16 and 25, respectively) cover part of the hydrophobic core (Figure 4(b)). For the C-terminal repeat, Glu443, Gln447, and Glu450 (at LRR positions 17, 21, and 24, respectively) are obviously different from the A-type RI-like LRR and constitute a charged shield for the hydrophobic core. Arg457 and the C-terminal Ser460 (at LRR positions 3 and 6, respectively) of the additional terminal β-strand complete this arrangement. The structural importance of these capping repeats lies, most probably, in the shielding of the hydrophobic core from the surrounding solvent and thus the stabilization of the domain.

For the library design, the following changes of the capping repeats were engineered. In the case of the N-terminal capping repeat, the most obvious difference of human RI compared to the other mammalian sequences was a five residue extension following the initial methionine residue. At the DNA level, this corresponds to a 15 bp duplication,[21] and structural inspection revealed an extended first β-strand.[18] Since this extension is apparently not necessary for function but exposes two hydrophobic amino acid residues to the solvent, we decided to remove this extension of the first β-strand. Furthermore, the two cysteine residues in the N-terminal capping repeat were

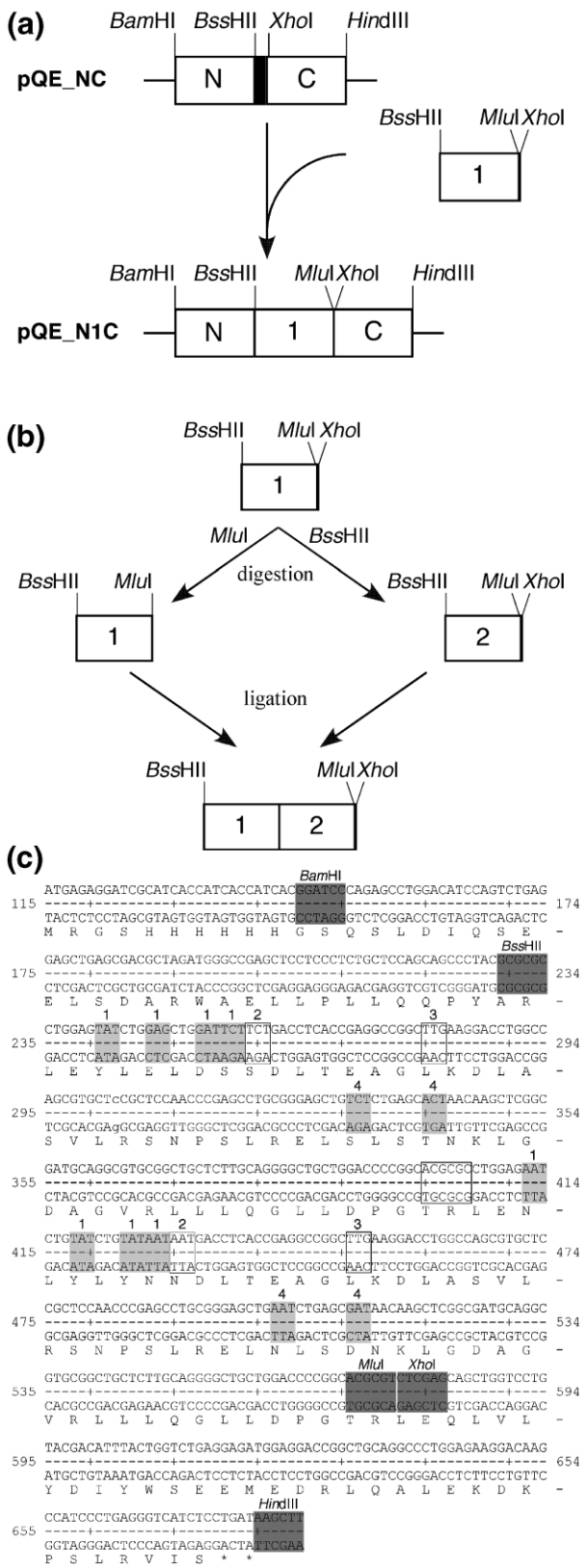

Figure 5. Cloning strategy and final DNA sequence of a designed N2C LRR protein. (a) Plasmid pQE_NC, containing the coding sequence of both capping repeats, was digested with restriction endonucleases *Bss* HII and *Xho* I, and ligated to a PCR-assembled library of designed LRR modules to yield plasmid library pQE_N1C. The same cloning strategy was adopted for LRR libraries encoding multiple repeat modules. (b) Stepwise modular assembly of the LRR library. The first step of the ligation strategy to obtain multiple repeat modules is shown. The PCR-assembled LRR module was digested separately with either *Mlu* I or *Bss* HII. Note that compatible overhangs are generated, which cannot be digested after ligation in the desired orientation. Thus, a stepwise directional elongation of the designed LRR proteins is possible at the DNA level. (c) The DNA sequence and the translated protein sequence of a randomly picked N2C LRR library member is shown. The recognition sequences of the restriction endonucleases *Bam* HI, *Bss* HII, *Mlu* I, *Xho* I, and *Hin*dIII are highlighted by dark grey boxes. The sequence ACGCGC created by the ligation of compatible ends generated by *Bss* HII and *Mlu* I is boxed at position 400. Randomized positions of the first repeat module (position 232–402) and second repeat module (position 403–573) are boxed and numbered. Light grey boxes indicate positions of randomizations for target binding, open boxes indicate framework randomizations to explore optimized protein structure. Digits 1–4 denote which trinucleotide mixture[27] was used for the synthesis of the degenerate oligonucleotides (see Table 2 for detailed composition).

replaced. Cys12 (at LRR position 9) in human RI was changed to serine, because the side-chain was found to be exposed to the solvent. Cys30 was replaced by proline to fit the consensus position 27 of the designed LRR sequence motif. Since all mammalian RI are lacking four residues between the N-terminal LRR and the first internal LRR (Figure 2(a)) compared to the consensus LRR modules, a four residue insertion fulfilling the structural requirements for the adjacent consensus repeat module was sought. The insertion was designed according to the B-type LRR consensus, because the N-terminal capping repeat is homologous to B-type LRR repeats: Thus, the choice of the first amino acid of the insertion was based on the analysis of position B28. In this position, a very low level of conservation is found, but it is clear that the side-chain points to the solvent and has no steric limitation, so tyrosine was chosen. For the second insertion at the end of the N-terminal capping repeat, the choice was based on the analysis of position B29. Cysteine, serine or threonine do occur, but alanine was chosen, since it was compatible with the cloning strategy and no negative structural effect was to be expected. The alignment of all mammalian C-terminal capping repeats showed only three homologous variations (alignment not shown). None of these variations was considered important for the design and the human C-terminal RI capping repeat sequence was used without further modification.

Design verification, gene synthesis and cloning

The designed protein sequences were assembled *in silico* and compared to mammalian RIs. Human RI, a protein with 16 repeats and thus corresponding to N7C, showed 57% similarity to and 52%

identity with the designed N7C protein; however, if all 56 randomized positions were identical with human RI, 68% similarity and 64% identity would be obtained. The biophysical properties such as charge and pI were comparable to all mammalian RIs (Table 1). Having designated the amino acids at all positions, the protein sequence was back-translated into a corresponding DNA sequence by choosing codons known to support strong expression in *E. coli*.[30] The resulting DNA sequence was checked for the absence of restriction endonuclease sites of the 6 bp cutter category. Where necessary, the DNA sequence was modified with alternative codons.

Our cloning strategy relied on insertion of the designed LRR libraries into plasmid pQE30 (QIAgen) *via* the *Bam* HI and *Hin*dIII restriction endonuclease sites. First, the DNA fragments corresponding to the N-terminal LRR capping repeat and the C-terminal LRR capping repeat were cloned into the *Bam* HI and *Hin*dIII sites of pQE30. This resulted in the plasmid pQE_NC containing all restriction endonuclease sites required for the cloning of the library modules (Figure 5(a)). For the stepwise ligation of LRR modules, the restriction endonucleases *Bss* HII and *Mlu* I were chosen, since they were compatible with the designed amino acid sequence and since they create compatible overhangs upon digestion. The designed LRR modules were PCR-assembled from partly randomized oligonucleotides of 50–72 bp length with overlapping regions of 15–18 bp (see Materials and Methods). Separate digestion of repeat modules with either *Bss* HII or *Mlu* I was followed by ligation to give repeat module dimers (Figure 5(b)). The ligation product was again digested with either *Bss* HII or *Mlu* I and ligated until the desired number of repeat modules was obtained. Finally, DNA libraries of various lengths were digested with *Bss* HII and *Xho* I, and ligated into the similarly treated pQE_NC plasmid to give plasmid libraries pQE_N1C, pQE_N2C, or pQE_N4C. The plasmid libraries of lengths N3C, N5C, and N6C were obtained by extension of the N2C and N4C libraries by one or two repeat modules. The coding sequence of an N2C library member and the corresponding amino acid sequence are shown (Figure 5(b)).

After transformation, single colonies were picked randomly and the DNA sequence of the plasmids was determined. The DNA sequences of 18 unselected library members of lengths N1C, N2C, and N4C (equivalent to a total of 50 repeat modules) were determined and 13 full-length clones were obtained. The remaining five clones had six single-base deletions resulting in premature stop codons. Importantly, about 80% of all sequenced LRR modules were without frameshift. The designed randomizations showed the desired amino acid residues with a trend to slightly higher asparagine incorporations. The percentages of amino acids found are given in Table 2. In mixture 1, a percentage of 12% was anticipated for all amino acids except Tyr and Trp, where 6% were anticipated. Asn was found in a significantly higher proportion, all other amino acids were in the range of the expectations. The deviations are probably not significant, due to the small number of analyzed sequences. In mixture 2, values of 50% for Asn and 25% each for Ser and Thr were anticipated and found in the analyzed sequences. Similarly, mixture 3 was anticipated to contain 50% of Leu and 25% each for Met and Val, which was also found. Finally, mixture 4 was anticipated to contain 17% of each of the six designated amino acids. Again, Asn was significantly higher, but the remaining distribution was as anticipated. In the case of the AR protein library, a similar trend was observed.[11] For further analyses, four clones devoid of frameshifts and stop codons were chosen. The chosen lengths of N3C–N6C are especially interesting, since we want to create binding molecules with sufficiently large target interaction areas.

Expression and purification of randomly chosen library members

Designed repeat proteins of all LRR libraries (N1C–N6C) ranging from 130 to 415 amino acid residues were expressed at 37 °C in XL1-Blue cells following standard procedures. Cells were

Table 2. Library composition of 18 randomly chosen LRR library members

Mixture[a]	D	E	G	H	K	L	M	N	Q	R	S	T	V	W	Y
1	11	9			13			25	5	5	15			2	14
2								52			18	26			
3						48	28						24		
4	6		9	14				37			22	12			

The experimentally determined frequency of each amino acid occurrence at randomized positions is given as a percentage. No value indicates that this amino acid was not included in the mixture and indeed, only the amino acids encoded by the respective mixture were found. A total of 199 codons (one deletion was observed) was analyzed for mixture 1, 50 codons were analyzed for each of mixtures 2 and 3, and 100 codons were analyzed for mixture 4. In mixture 1, the expected frequencies were 13% for each amino acid except Trp and Tyr (each 6%). In mixture 2, the expected frequencies were 50% for Asn and 25% for each Ser and Thr. In mixture 3, the expected frequencies were 50% for Leu and 25% for each Met and Val. In mixture 4, the expected frequencies were 17% for all six amino acid residues.

[a] Denotes one of the four oligonucleotide mixtures used for the randomization of the LRR module. The locations in the LRR sequence motif are given in Figures 2(f) and 5(c).

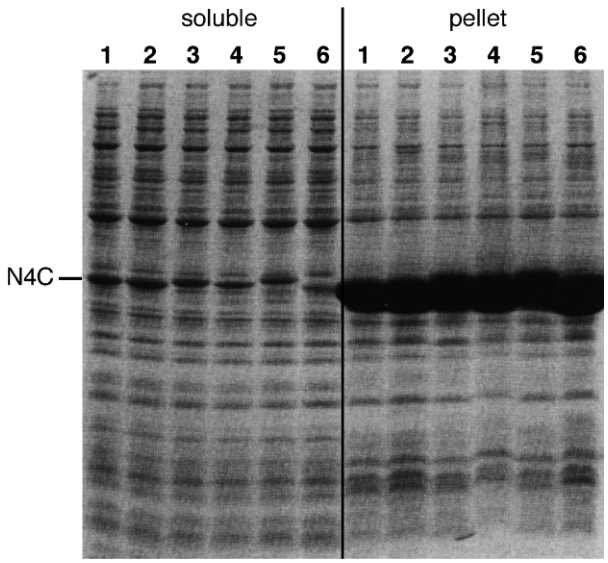

Figure 6. Solubility analysis of designed N4C LRR library members by SDS-PAGE stained with Coomassie brilliant blue. Six randomly chosen clones were analyzed after induction of expression and sonication. The position of the designed N4C LRR proteins is indicated.

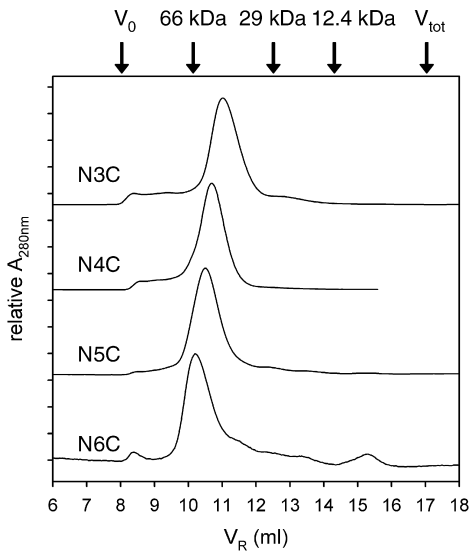

Figure 7. Size-exclusion chromatography of the designed N3C, N4C, N5C, and N6C LRR proteins. Proteins were analyzed after IMAC purification on a Superdex 75 column with TBS_{150} buffer. The void volume V_0 of the column is about 8 ml and the total column volumn V_t is about 17 ml as indicated. The retention volumes of the molecular mass standards bovine serum albumin (66 kDa), carbonic anhydrase (29 kDa), and cytochrome *c* (12.4 kDa) are indicated.

sonicated, and the soluble and insoluble fractions were analyzed by SDS-PAGE. Six independent clones of length N4C are shown (Figure 6). Other lengths look similarly (data not shown). No influence of the randomized framework positions on protein expression was obvious for any protein length. The soluble fraction always contained about 5–20 mg of designed LRR protein per 1 l culture, whereas the insoluble fraction contained about 200 mg/l (Figure 6). Purification of the designed repeat proteins to very high purity was achieved from the soluble fraction by using immobilized metal ion affinity chromatography (IMAC). Refolding of the insoluble LRR proteins was possible following IMAC purification under denaturing conditions. We routinely purified 5–10 mg of protein per 1 l culture from the soluble fraction or 30–50 mg from refolding of the insoluble fraction. A few of the purified LRR showed a tendency to aggregate; however, most of them could be handled at concentrations of above 200 μM (6–9 mg/ml) without any signs of precipitation. Storage at 4 °C was possible for at least one month without loss of material for all LRR proteins tested.

Biophysical characterization

For the biophysical characterization, four designed LRRs of lengths N3C, N4C, N5C, and N6C were chosen. Size-exclusion chromatography of designed LRRs in TBS_{150} (pH 7.4) was performed using a Superdex 75 column (Pharmacia) and showed single symmetric peaks (Figure 7). The apparent molecular mass values estimated from molecular mass standards were 1.4–1.5-fold higher than expected for globular proteins. CD spectrometry was used to measure the secondary structure content, showing α-helical secondary structure with minima of 222 nm and 206 nm (Figure 8(a)). The α-helical content was estimated to be around 35% for N3C, 36% for N4C, and 39% for N5C.[31] In the presence of 7 M urea, the same proteins did not show a minimum at 222 nm, consistent with the absence of α-helices. The tryptophan fluorescence emission spectra were recorded in the presence and in the absence of urea. Denaturation of the proteins was accompanied by a reduction of fluorescence intensity and by a shift of the emission spectra to higher wavelengths (Figure 8(b)). In these experiments, the protein of length N6C was not included, since it tended to aggregate during equilibration.

The equilibrium denaturation of the designed LRR proteins was followed by both CD spectroscopy and tryptophan fluorescence using urea as denaturant (Figure 9). Both measurements show relatively broad transitions between 2 M and 4 M urea. The curves derived from CD spectroscopy and tryptophan fluorescence spectroscopy were not superimposable. It should be mentioned that the Trp residues are located in both terminal capping repeats but not in the hydrophobic core of the designed repeat domains. Since the curves do not coincide, and since the transitions are exceptionally broad, we think that unfolding does not follow a simple two-state mechanism. Tryptophan fluorescence may mostly

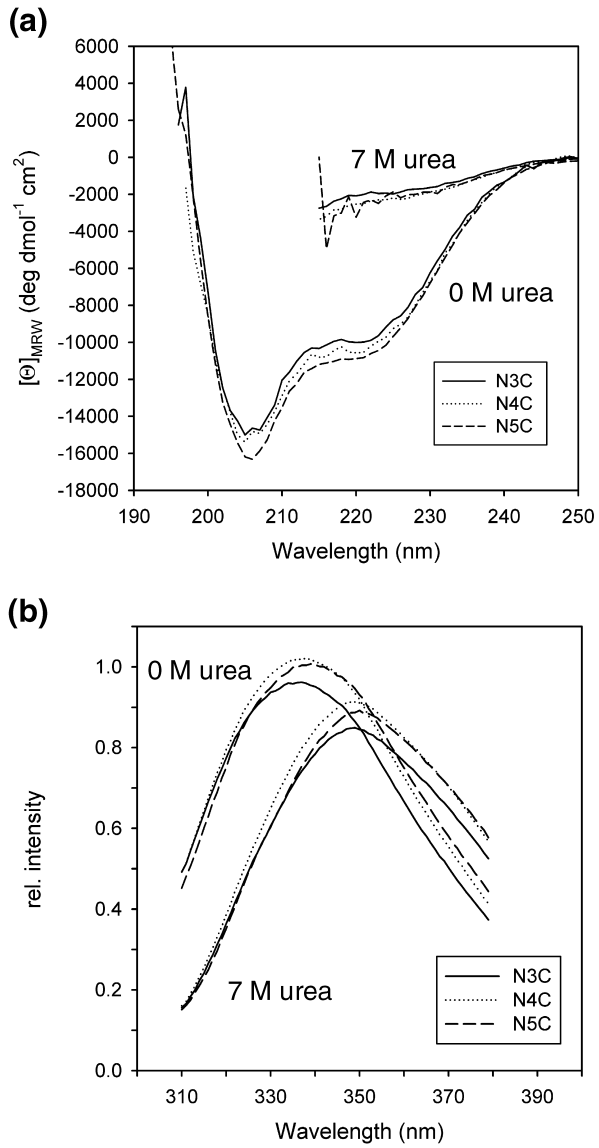

Figure 8. (a) CD spectra of the designed N3C, N4C, and N5C LRR proteins in 0 M and 7 M urea. Proteins were analyzed after IMAC purification and dialysis. (b) Fluorescence emission spectra of the designed N3C, N4C, and N5C LRR proteins in 0 M and 7 M urea. The excitation wavelength was 295 nm and the spectra were normalized to the maximum intensity observed. The fluorescence spectra of the proteins were measured in TBS_{150} buffer.

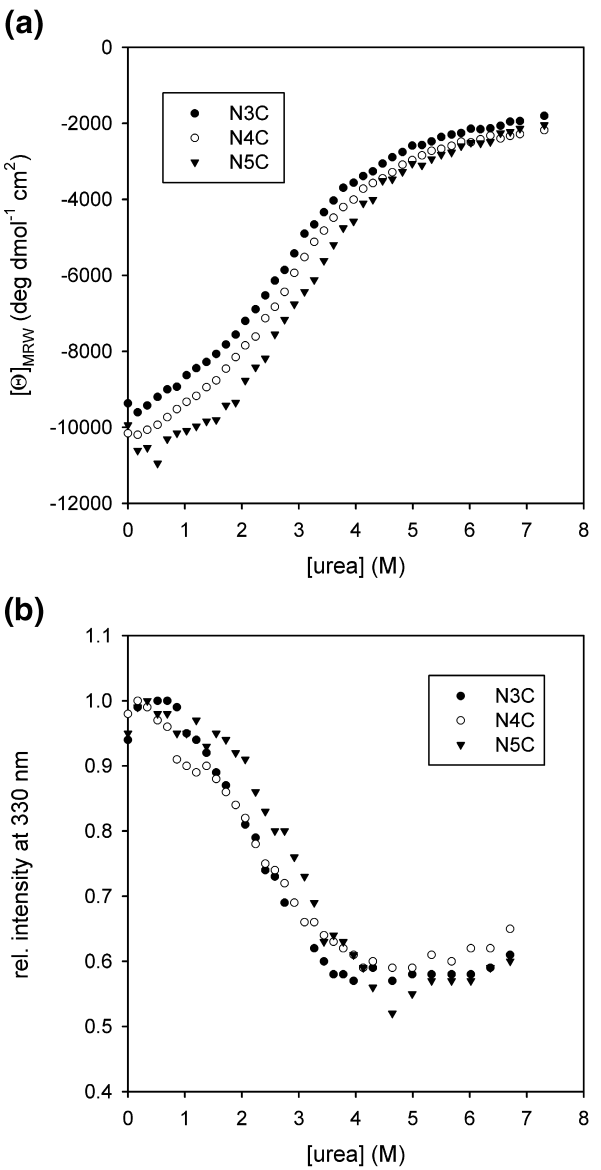

Figure 9. Equilibrium denaturation of the designed N3C, N4C, and N5C LRR proteins followed by CD spectroscopy and fluorescence measurement. Proteins were analyzed after IMAC purification and dialysis. (a) The mean residue ellipticity measured at 222 nm is plotted against increasing concentrations of urea. (b) The normalized fluorescence intensity measured at 330 nm after excitation at 295 nm is plotted against increasing concentrations of urea.

monitor the unfolding of the capping repeats, which appear to unfold at lower concentrations of denaturant than the internal repeats. In contrast, CD spectroscopy monitors contributions from all parts of the LRR domain. If the CD data of protein N5C were fit to a two-state model, a midpoint of denaturation of about 2.9 M urea would be obtained, but the cooperativity of this protein would then be only about 3 kJ/mol M, which is clearly less than expected for highly cooperatively unfolding proteins of this size.[32] Consequently, a two-state behavior of the designed LRR proteins is unlikely and further quantitative analysis of the equilibrium denaturation data was not undertaken. An apparently greater stability of protein N5C compared to proteins N3C and N4C is visible in both the CD and the tryptophan fluorescence measurements. However, it needs to be pointed out that the proteins tested differ in eight positions per repeat. Therefore, it is too early to conclude that longer proteins are more stable. Equilibrium

denaturation measurements of designed LRR proteins with identical repeats are underway and should reveal the contribution of repeat number to stability and cooperativity.

Discussion

Natural repeat proteins

Repeat proteins are involved in many important protein–protein interactions in most organisms. Recently, the abundance of repeat proteins has been analyzed and it has become clear that the complexity of higher organisms is accompanied by a relative increase in the proportion of repeat proteins in the respective genomes.[33] One attractive hypothesis is that repeat proteins, in general, can evolve more quickly than non-repeat proteins, since the modular architecture is less susceptible to detrimental sequence variations.[33] This apparent robustness is another advantage we sought to exploit when establishing our design strategy to generate novel binding molecules based on the modular nature of repeat proteins.[10] For the implementation of our strategy, we scrutinized the four major repeat protein classes of elongated shape:[34] the tetratricopeptide repeat, the HEAT/armadillo repeat, the AR and the LRR. LRR and AR[11] proteins were most attractive to us due to their abundance and the wealth of published data describing their structures and binding properties. The attraction of LRR proteins, in particular, stems from the extensive binding surface areas. On the other hand, relatively little was known about the biophysical properties of LRR proteins. In order to implement our strategy for LRR proteins, we first had to derive self-compatible LRR modules. Self-compatibility should result directly from the repeat consensus.[10] Furthermore, highly conserved framework positions as well as poorly conserved positions involved in target binding should become apparent from the repeat consensus.

Designed LRR proteins

We chose to analyze the mammalian RIs because of their intracellular location and their femtomolar affinities to RNases. One important consequence from the initial alignments was the definition of the A-type/B-type double repeat module. If an LRR consensus had been derived averaging over both A and B-type RI-like LRR, many positions would have appeared unclear. This choice was confirmed recently by the finding that the gene of another RI-like LRR, MATER, has LRR exons encoding the A-type/B-type double LRR with 57 amino acid residues.[8] The same exon arrangement is found in human RI (M. T. S., unpublished results), suggesting exon duplication as one likely evolutionary scenario.[12] Another important result from the sequence alignments was the adoption of capping repeats, reflecting a natural design principle of repeat proteins.[10] It should be mentioned that our design has not yet explored the full potential of engineering the capping repeats. In the case of human RI, many functionally important residues have been found in the region around Tyr434 of the C-terminal capping repeat (Figure 2(d)). These positions can, in principle, be used for randomization to extend the putative interaction area even across both terminal repeats (cf. Figure 1(c)). On the other hand, rational design might further improve the capping function of the terminal repeats by incorporating more hydrophilic amino acids to further improve folding yields.

The success of our design strategy became apparent when protein expression was analyzed. For all investigated repeat lengths, soluble LRR proteins were amongst the most abundant proteins of *E. coli* extracts. This is in stark contrast to the situation of recombinant RI, where expression has proven difficult, probably due to the problematic oxidation of the 30–32 reduced cysteine residues.[35] The comparison of designed LRR proteins of different lengths did not indicate a significant influence of the protein size on soluble expression yields, nor was a significant influence of the amino acids at randomized positions obvious. It should be emphasized that, since we are aiming at molecules useful for intracellular applications, the performance inside the cell is of special interest. The high expression level and solubility of the designed LRR proteins in the bacterial cytosol underlines the folding efficiency and stability against proteolysis, both indicators of well-folded proteins. In addition, the observed high tolerance to randomization underlines the fact that designed LRRs are good scaffolds for the selection of new binding molecules. This demonstrates the success of our design and shows that modularity can be achieved with self-compatible repeat modules derived by consensus analysis. The analysis of purified designed LRR proteins partly rationalizes the observed *in vivo* behavior. Size-exclusion chromatography suggested that the designed LRR proteins are monomers in solution (Figure 7). The observed greater apparent molecular mass in size-exclusion chromatography reflects a larger hydrodynamic radius of the proteins. This has been observed before in the case of rna1p, an 11 repeat LRR protein, where the observed molecular mass was 1.35-fold larger than expected.[36]

Comparison of designed repeat proteins

LRR and AR proteins, designed according to the same principles,[11] have a different equilibrium denaturation behavior. Designed AR proteins show high cooperativity of 12–21 kJ/mol M, when analyzed with a two-state model.[37] Similarly, natural AR proteins exhibit cooperativities of unfolding like globular proteins of the same size.[38] At first glance this may be surprising, since the absence of long-range interactions and the simple

topology of repeat proteins, where only neighboring segments interact, would rather suggest a modular folding model (i.e. low unfolding cooperativity).[38] However, the interrepeat contacts in the AR proteins are substantial, while hydrophobic contacts within one repeat are limited, which may rationalize the cooperative nature of unfolding. In contrast, the equilibrium denaturation behavior of designed LRR proteins confirms the expectation of modular folding. Apparent *m*-values of unfolding derived from a two-state model appear to be roughly two to four times lower than those determined for designed AR proteins of similar size[37] and five times lower than typical for globular proteins of the same size.[32] This difference indicates a smaller mutual stabilization of the repeats in the designed LRR proteins and a more modular unfolding reaction than in the AR proteins. To date, no unfolding studies of LRR proteins have been published and the question, whether the low cooperativity of the designed LRR proteins is unusual or a hallmark of this protein family, has to await further experiments.

When comparing the design of the LRR proteins presented here to the design of the AR proteins in the accompanying paper,[11] it appears that the biggest difference was the initial data set. Whereas the AR consensus was calculated from several hundred repeat sequences from very divergent proteins, the LRR consensus was based on the 28 repeat sequences of four homologous proteins. Therefore, in the case of the designed LRR proteins we cannot exclude the possibility that residues conserved for functional reasons have been misclassified as residues conserved for structural reasons. Nonetheless, this concern is small because the consensus was calculated from seven functionally different repeats in each mammalian RI. Furthermore, following the publication of more genomes, a new consensus for all detectable RI-like LRR protein sequences was derived, which confirms the consensus presented here (M. T. S., unpublished results).

Conclusion

The LRR has evolved as a versatile building block of protein-binding domains, which rely on repeat shuffling, elongation or deletion to alter their binding specificities. By consensus design, we have derived self-compatible LRR modules enabling us to exploit this modularity for the generation of binding protein libraries. Together with adapted natural capping repeats, these partly randomized LRR modules could be assembled successfully into repeat proteins ranging from 130 to 415 amino acid residues. All tested proteins are well expressed in *E. coli*, can be purified easily in large amounts and are monomeric in solution. Importantly, our design has yielded molecules of adjustable size with overall comparable biophysical properties. Thus, consensus design has again proven very useful to obtain proteins with properties superior to the natural counterparts. Furthermore, previous problems of other designed binding molecules, such as low solubility at physiological pH,[39] low production yields in bacteria[40] and proteolytic instability could be overcome with our designed repeat proteins. Together with the successful design of AR protein libraries, we were able to implement our strategy for the design of binding molecules in two independent repeat protein classes. Therefore, we believe that this strategy could be applied to other repeat proteins.

Material and Methods

Materials

Chemicals were purchased from Fluka (Switzerland). Oligonucleotides containing trinucleotide mixtures[27] were from MorphoSys AG (Germany), standard oligonucleotides were from Microsynth (Switzerland). Vent DNA polymerase and restriction endonucleases were from New England Biolabs (USA) or Fermentas (Lithuania). All cloning and protein expression was performed in *E. coli* XL1-Blue (Stratagene, USA) using plasmid pQE30 (QIAgen, Switzerland) derivatives.

Bioinformatics

All database analyses were performed using GCG (Wisconsin Package Version 10.2, Genetics Computer Group, USA). The numbering of the LRR follows the most common scheme based on the first crystal structure of pig RI.[17] Thus, the conserved leucine residues are found in positions 2, 5, 7, 12, 20, and 24 (Figure 3(a)). PDB[41] entries 1DFJ[20] and 1A4Y[18] were used for structural analyses. Alignments were performed using the program PILEUP of the GCG software with standard settings, i.e. a gap creation penalty of 8 and a gap extension penalty of 2.

Cloning of the designed LRR protein libraries

All oligonucleotides are listed in Table 3. To obtain the N-terminal capping repeat flanked by appropriate restriction endonuclease sites, the DNA of pTRP-PRI[19] was amplified with oligonucleotides MTS37 and MTS4 giving PCR-fragment "N". Primer MTS37 introduced a *Bam*HI site at the 5′-end of the coding strand, replaced the N-terminal first direct repeat SLDIQ by only Q, and cysteine 12 was changed to serine. Primer MTS4 introduced a *Bss*HII, an *Xho*I, and a *Hin*dIII site at the 3′-end of the coding strand. The PCR-fragment "N" was digested and ligated into the *Bam*HI and *Hin*dIII sites of pQE30 yielding plasmid pQE_N. DNA sequencing on a LI-COR 4000 system (LI-COR, USA) using an EXCELII kit (Epicentre, USA) confirmed the correct DNA sequence. The C-terminal capping repeat flanked by appropriate restriction endonuclease sites was amplified from annealed oligonucleotides MTS29 and MTS30 by MTS5b and MTS3pQE giving PCR-fragment "C". Primer MTS5 introduced an *Xho*I site at the 5′-end of the coding strand; primer MTS3 introduced a *Hin*dIII site at the 3′-end of the coding strand. The PCR fragment "C" was digested and ligated into the *Xho*I and *Hin*dIII sites of

Table 3. Oligonucleotides used for the cloning of the designed LRR protein library based on mammalian RIs

Name	Sequence in 5′–3′ direction (restriction sites are underlined)[a]	Description[b]
MTS37	CGGGATCCCAGagcctggacatccagTCTgaggagctg (*Bam* HI)	fwd PCR primer to obtain N-terminal capping repeat and to introduce C → S mutation (human RI position 12)
MTS4	GCATAAGCTTATCACTCGAGGCGCGCGTAGGGctgctggagcagagg (*Hin*dIII, *Xho* I, *Bss* HII)	rev PCR primer to obtain N-terminal capping repeat
MTS29	CTCGAGCAGCTGGTCCTGTACGACATTTACTGGTCTGAGGAGATGgag gaccggctccaggc (*Xho* I)	fwd assembly primer to obtain C-terminal capping repeat
MTS30	ATGACCCTCAGGGATGGCTTGTCCTTCTCCAGGgcctggagccggtcctc	rev assembly primer to obtain C-terminal capping repeat
MTS3pQE	GCATAAGCTTATCAagagatgaccctcagggatgg (*Hin*dIII)	rev PCR primer to amplify MTS30
MTS5b	CATGCCATGGACGCGTCTCGAGcagctggtcc (*Nco* I, *Mlu* I, *Xho* I)	fwd PCR primer to amplify MTS29
MTS7	TTGGCGCGCCTGGAG111CTG111CTG111111222gacctcaccgagg ccggc (*Bss* HII)	fwd assembly primer, five randomized positions
MTS8	ccgcaggctcgggttggaGCGGAGCACGCTGGCCAGGTCCTTCANgcc ggcctcggtgaggtc	rev assembly primer, N = A, C, G, or T
MTS9	TccaacccgagcctgcggGAGCTG444CTGAGC444aacaagctcgg cgatgca	fwd assembly primer, two randomized positions
MTS10	CCGCTCGAGACGCGTGCCGGGGTCCAGCAGCCCCTGCAAGAGCAGCC GCACGCCtgcatcgccgagcttgtt (*Xho* I, *Mlu* I)	rev assembly primer
MTS11b	TAATACGACTCACTATAGGGttggcgcgcctggag (*Bss* HII)	fwd PCR primer to amplify the assembly MTS7-10
MTS14b	GGCTTTGTTAGCAGCCGGATCctcgagacgcgtgccggggtc (*Bam* HI, *Xho* I, *Mlu* I)	rev PCR primer to amplify the assembly MTS7-10

[a] Lower-case letters indicate regions designed for annealing. Digits abbreviate the trinucleotide mixtures[27] used for oligonucleotide synthesis: 111 encodes amino acids D, E, N, Q, R, K, S, Y, W; 222 encodes amino acids N, S, T; 444 encodes amino acids G, D, N, S, T, H.
[b] fwd, forward; rev, reverse.

pQE_N yielding plasmid pQE_NC. DNA sequencing confirmed the correct DNA sequence.

Oligonucleotides MTS7 and MTS9 were synthesized partly from trinucleotides to encode predetermined mixtures of amino acids. Oligonucleotides were designed to overlap with each other over 18 or 19 bases and to have melting temperature of at least 56 °C. To obtain a library encoding one repeat module, the partly randomized oligonucleotides MTS7, MTS8, MTS9, and MTS10 were assembled by PCR and were amplified with a tenfold molar excess of MTS11b and MTS14b during 30 PCR cycles (denaturation at 95 °C for one minute, annealing at 50 °C for 30 seconds, primer extension at 72 °C for 30 seconds) on a MJ Research PTC-200 instrument (MJ Research, USA). In the case of the LRR library described here, this initial PCR assembles the A-type/B-type RI double repeat module.

The single repeat module library was cloned with *Bss* HII and *Xho* I into similarly treated plasmid pQE_NC yielding plasmids pQE_N1C (Figure 5(a)). The assembly of several repeat modules was carried out by stepwise random ligation as follows. The PCR assembled library was divided in two aliquots and digested either with *Bss* HII or *Mlu* I. These restriction endonucleases create compatible overhangs that do not regenerate a restriction endonuclease site after ligation, similar to a previously described repetitive cloning strategy.[42] After purifying the DNA using QIAquick spin columns (QIAgen, Switzerland), ligation yielded a library composed of two repeat modules (Figure 5(b)). Again, the DNA was purified and divided into two aliquots and digested with either *Bss* HII or *Mlu* I. Isolation of the desired band by agarose gel separation was again followed by ligation to yield a library composed of four repeat modules. Libraries of three, five and six repeat modules were obtained by ligating one or two repeat modules to two or four repeat modules, respectively. The multiple repeat module libraries were cloned *via* *Bss* HII and *Xho* I into similarly treated plasmid pQE_NC. The resulting ligation mixes were QIAquick purified and used for electroporation of XL1-Blue cells prepared as described,[43] except that cells were grown at 27 °C. DNA sequencing was carried out to determine the composition of the final libraries. The strategy presented below describes a way to obtain multimers of DNA fragments in a defined direction using compatible restriction endonucleases and ligation. One such possibility is to use the restriction endonucleases *Bss* HII and *Mlu* I, which were compatible with the designed protein sequence. It should be noted that the ligation of identical ends reconstitutes the original recognition sites and these DNA molecules can therefore be distinguished by restriction digestion.

Protein production and purification

Expression was performed essentially as described (QIAgen "QIAexpressionist", Switzerland). Briefly, 50 ml overnight cultures of XL1-Blue grown in LB medium supplemented with ampicillin (50 μg/ml) and 1% (w/v) glucose were used to inoculate 1 l of LB medium. Cultures were grown at 37 °C to $A_{600} = 0.6$ and induced by 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Expression was continued for three to four hours at 37 °C. Bacteria were collected by centrifugation and stored at −20 °C. Cells were resuspended in TBS_{500} (50 mM Tris–HCl (pH 8.0), 500 mM NaCl) before lysis by sonication or French press. Lysed cells were centrifuged at 20,000 rpm in a Sorvall SS-34 rotor for 15 minutes. The composition of the soluble and insoluble fractions of single clones was analyzed by SDS-PAGE stained with Coomassie brilliant blue R-250 (Figure 6). Visual comparison using purified LRR proteins of known concentrations allowed an estimate of soluble

and insoluble expression yields. For protein purification, the supernatant was adjusted to 10% (v/v) glycerol and 20 mM imidazole and applied onto Ni-NTA superflow (QIAgen, Switzerland) columns equilibrated in the same buffer. Purification from the insoluble fraction was done after solubilization of the inclusion bodies in lysis buffer containing 8 M urea. Refolding of the immobilized protein was achieved on the Ni-NTA column by dilution of urea to TBS_{500} buffer with 20 mM imidazole and 10% (v/v) glycerol. Column washing and elution was carried out according to the manufacturer's instructions. Proteins were dialyzed after purification against 50 mM Tris–HCl (pH 8.0), 150 mM NaCl, 10% glycerol.

Biophysical characterization

Size-exclusion chromatography was carried out on a Superdex 75 HR10/30 analytical gel-filtration column using an ÄKTAexplorer chromatography system (Amersham Biosciences, Switzerland) thermostaticaly controled at 15 °C. The buffer was TBS_{150} (10 mM Tris–HCl (pH 7.4), 150 mM NaCl). Designed LRR proteins were analyzed at loading concentrations from 10 μM to 200 μM. The void volume of the column is about 8 ml and the total column volumn is about 17 ml. Protein standards (Sigma, USA) for the calculation of the apparent molecular mass were bovine serum albumin (66 kDa), carbonic anhydrase (29 kDa), and cytochrome *c* (12.4 kDa). Circular dichroism spectra were recorded using 3–10 μM protein in 10 mM sodium phosphate (pH 6.5), 125 mM NaCl on a Jasco J-715 instrument (Jasco, Japan) with a 1 mm circular cuvette and three spectra (2 nm band width, four seconds response time, and 1 nm data pitch) were recorded and averaged. All spectra were corrected for buffer absorption. The CD signal was converted to the mean residue ellipticity using the protein concentration determined by absorbance measurements and the calculated molar extinction coefficient[44] at 280 nm (Table 1). The α-helix percentage was estimated using the relation $f_\alpha = 100(-[\Theta]_{208\ nm} - 4000)/(33{,}000 - 4000)$.[31] The temperature in all CD measurements was 22 °C. Intrinsic fluorescence measurements were performed on a PTI QM-2000-7 instrument (Photon Technology International, USA) with an excitation wavelength of 295 nm (4 nm bandwidth) and an emission scan from 310 nm to 380 nm (4 nm bandwidth). A cuvette with 1 cm pathlength was thermostatically controlled using a water-bath. The temperature in all fluorescence measurements was 22 °C. Equilibrium denaturation was allowed for at least 18 hours at 22 °C, which was sufficient to reach equilibrium. The CD signal was recorded for 120 seconds at 222 nm and averaged. The fluorescence spectra were averaged over three recordings and the intensity was normalized to the maximum intensity observed. The same samples were used for CD and fluorescence measurements. The reversibility of the unfolding reaction was verified by dilution of native sample into 3 M urea buffer and by dilution of denatured sample to 0.2 M urea buffer.

GenBank accession numbers

The sequences of the characterized designed LRRs named N3C, N4C, N5C and N6C have been deposited in the GenBank database with the accession numbers AY266453, AY266454, AY266455, and AY266456, respectively.

Acknowledgements

The authors thank Dr David L. Zechel for critical reading of the manuscript and Dr Annemarie Honegger for EXCEL® macros. The generous gift of the plasmid pTRP-PRI from Dr Bert L. Vallee is gratefully acknowledged. We thank MorphoSys AG for the trinucleotide containing oligonucleotides. M.T.S. was the recipient of a Kekulé Stipendium of the Fonds der Chemischen Industrie, with financial participation of the Bundesministerium für Bildung und Forschung. H.K.B. was the recipient of a predoctoral scholarship from the Roche Research Foundation. This project was supported by the Swiss Cancer Research grant KFS 1055-09-2000 and the National Center of Competence in Research Structural Biology.

References

1. Nygren, P.-Å. & Uhlén, M. (1997). Scaffolds for engineering novel binding sites in proteins. *Curr. Opin. Struct. Biol.* **7**, 463–469.
2. Skerra, A. (2000). Engineered protein scaffolds for molecular recognition. *J. Mol. Recognit.* **13**, 167–187.
3. Cattaneo, A. & Biocca, S. (1999). The selection of intracellular antibodies. *Trends Biotechnol.* **17**, 115–121.
4. Kobe, B. & Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **11**, 725–732.
5. Janeway, C. A., Jr & Medzhitov, R. (2002). Innate immune recognition. *Annu. Rev. Immunol.* **20**, 197–216.
6. 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.
7. Koo, Y. B., Ji, I., Slaughter, R. G. & Ji, T. H. (1991). Structure of the luteinizing hormone receptor gene and multiple exons of the coding sequence. *Endocrinology*, **128**, 2297–2308.
8. Tong, Z.-B., Nelson, L. M. & Dean, J. (2000). Mater encodes a maternal protein in mice with a leucine-rich repeat domain homologous to porcine ribonuclease inhibitor. *Mamm. Genome*, **11**, 281–287.
9. Ellis, J., Dodds, P. & Pryor, T. (2000). The generation of plant disease resistance gene specificities. *Trends Plant Sci.* **5**, 373–379.
10. 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.
11. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A. (2003). *J. Mol. Biol.* **332**, 489–503.
12. Kobe, B. & Deisenhofer, J. (1994). The leucine-rich repeat: a versatile binding motif. *Trends Biochem. Sci.* **19**, 415–421.
13. Buchanan, S. G. S. C. & Gay, N. J. (1996). Structural and functional diversity in the leucine-rich repeat family of proteins. *Prog. Biophys. Mol. Biol.* **65**, 1–44.
14. Shapiro, R. (2001). Cytoplasmic ribonuclease inhibitor. *Methods Enzymol.* **341**, 611–628.
15. Marino, M., Braun, L., Cossart, P. & Ghosh, P. (2000). A framework for interpreting the leucine-rich repeats

of the Listeria internalins. *Proc. Natl Acad. Sci. USA*, **97**, 8784–8788.
16. Dangl, J. L. & Jones, J. D. G. (2001). Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.
17. Kobe, B. & Deisenhofer, J. (1993). Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature*, **366**, 751–756.
18. Papageorgiou, A. C., Shapiro, R. & Acharya, K. R. (1997). Molecular recognition of human angiogenin by placental ribonuclease inhibitor—an X-ray crystallographic study at 2.0 Å resolution. *EMBO J.* **16**, 5162–5177.
19. Lee, F. S. & Vallee, B. L. (1989). Expression of human placental ribonuclease inhibitor in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* **160**, 115–120.
20. Kobe, B. & Deisenhofer, J. (1995). A structural basis of the interactions between leucine-rich repeats and protein ligands. *Nature*, **374**, 183–186.
21. Lee, F. S., Fox, E. A., Zhou, H.-M., Strydom, D. J. & Vallee, B. L. (1988). Primary structure of human placental ribonuclease inhibitor. *Biochemistry*, **27**, 8545–8553.
22. Blackburn, P., Wilson, G. & Moore, S. (1977). Ribonuclease inhibitor from human placenta. Purification and properties. *J. Biol. Chem.* **252**, 5904–5910.
23. Kobe, B. & Deisenhofer, J. (1995). Proteins with leucine-rich repeats. *Curr. Opin. Struct. Biol.* **5**, 409–416.
24. 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.
25. Lehmann, M. & Wyss, M. (2001). Engineering proteins for thermostability: the use of sequence alignments *versus* rational design and directed evolution. *Curr. Opin. Biotechnol.* **12**, 371–375.
26. 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.
27. 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.
28. Blázquez, M., Fominaya, J. M. & Hofsteenge, J. (1996). Oxidation of sulfhydryl groups of ribonuclease inhibitor in epithelial cells is sufficient for its intracellular degradation. *J. Biol. Chem.* **271**, 18638–18642.
29. Hofsteenge, J., Kieffer, B., Matthies, R., Hemmings, B. A. & Stone, S. R. (1988). Amino acid sequence of the ribonuclease inhibitor from porcine liver reveals the presence of leucine-rich repeats. *Biochemistry*, **27**, 8537–8544.
30. Hénaut, A. & Danchin, A. (1996). Analysis and predictions from *Escherichia coli* sequences, or *E. coli in silico*. In *Escherichia coli and Salmonella: Cellular and Molecular Biology* (Neidhardt, F. C., Curtiss, R. III, Ingraham, J. L., Lin, E. C. C., Brooks Low, B. & Magasanik, B., eds), 2nd edit., pp. 2047–2066, American Society for Microbiology, Washington, DC.
31. Greenfield, N. & Fasman, G. D. (1969). Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, **8**, 4108–4116.
32. Myers, J. K., Pace, C. N. & Scholtz, J. M. (1995). Denaturant *m* values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138–2148.
33. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* **293**, 151–160.
34. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131.
35. Klink, T. A., Vicentini, A. M., Hofsteenge, J. & Raines, R. T. (2001). High-level soluble production and characterization of porcine ribonuclease inhibitor. *Protein Expr. Purif.* **22**, 174–179.
36. Hillig, R. C., Renault, L., Vetter, I. R., Drell, IV, T., Wittinghofer, A. & Becker, J. (1999). The crystal structure of rna1p: a new fold for a GTPase-activating protein. *Mol. Cell*, **3**, 781–791.
37. 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.
38. Marchetti Bradley, C. & Barrick, D. (2002). Limits of cooperativity in a structurally modular protein: response of the Notch ankyrin domain to analogous alanine substitutions in each repeat. *J. Mol. Biol.* **324**, 373–386.
39. Tramontano, A., Bianchi, E., Venturini, S., Martin, F., Pessi, A. & Sollazzo, M. (1994). The making of the minibody: an engineered β-protein for the display of conformationally constrained peptides. *J. Mol. Recognit.* **7**, 9–24.
40. Schlehuber, S. & Skerra, A. (2002). Tuning ligand affinity, specificity, and folding stability of an engineered lipocalin variant—a so-called 'anticalin'—using a molecular random approach. *Biophys. Chem.* **96**, 213–228.
41. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
42. Forrer, P. & Jaussi, R. (1998). High-level expression of soluble heterologous proteins in the cytoplasm of *Escherichia coli* by fusion to the bacteriophage Lambda head protein D. *Gene*, **224**, 45–52.
43. Tung, W. L. & Chow, K.-C. (1995). A modified medium for efficient electrotransformation of *E. coli*. *Trends Genet.* **11**, 128–129.
44. Gill, S. C. & von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **182**, 319–326.
45. Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.

Edited by J. Thornton

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