nature structural & molecular biology

Control of repeat-protein curvature by computational protein design

Keunwan Park[1,2], Betty W Shen[3], Fabio Parmeggiani[1,2], Po-Ssu Huang[1,2], Barry L Stoddard[3] & David Baker[1,2,4]

Shape complementarity is an important component of molecular recognition, and the ability to precisely adjust the shape of a binding scaffold to match a target of interest would greatly facilitate the creation of high-affinity protein reagents and therapeutics. Here we describe a general approach to control the shape of the binding surface on repeat-protein scaffolds and apply it to leucine-rich-repeat proteins. First, self-compatible building-block modules are designed that, when polymerized, generate surfaces with unique but constant curvatures. Second, a set of junction modules that connect the different building blocks are designed. Finally, new proteins with custom-designed shapes are generated by appropriately combining building-block and junction modules. Crystal structures of the designs illustrate the power of the approach in controlling repeat-protein curvature.

Repeat-protein scaffolds have attracted much attention as alternative binding scaffolds to antibodies[1–4] and also as building blocks of protein nanomaterials[5–7] because of their intrinsic modularity and high stability. The leucine-rich repeat (LRR) is a repeat-protein scaffold with a horseshoe-like global structure in which the concave surface is often a binding interface[8]. LRRs share a common structural motif (LxxLxLxxN/C), but different LRR modules generate proteins with distinct global curvatures when the repeat modules are packed on themselves[9]. Irregular LRR modules are frequently observed interspersed within arrays of canonical repeat modules; their presence contributes to the curvature diversity within the family. For example, Toll-like receptor 4 (TLR4) contains three distinct regions of LRR repeats, each having different curvatures that collectively generate a surface with high shape complementarity to the target surface of the MD2 protein[10]. Current engineering approaches have focused on changing residues at the binding surfaces of an already existing or consensus repeat protein[11–16], varying the numbers of repeat modules[17–19] and fusing naturally occurring repeat proteins[10,20,21]. Although powerful, these strategies do not allow for the customization of repeat-protein curvature for a specific application.

To create new repeat proteins with custom-specified curvature, we developed a general computational design approach. We demonstrate the power of the approach by designing 12 new proteins with different curvatures. Crystal structures show that the method allows control of repeat-protein curvature with atomic-level accuracy.

RESULTS

Strategy for curvature-tunable scaffold design

Our design strategy has three steps (**Fig. 1a**). The first step is the design of a set of idealized self-compatible building-block modules (BB_1 to BB_n) from which a series of proteins of variable length BB_i^n can be created directly by varying the number of building-block repeats without any further engineering. These 'homo-building-block' proteins will have a constant curvature defined by the base building-block module. The second step is the design of a set of junction modules ($JN_{BBi \rightarrow BBj}$) that connect building-block module *i* to building-block module *j*. A critical feature of the design at steps one and two is that the interfaces between individual building blocks, as well as those between building blocks in junction modules, have sufficiently low energy that the orientation between all units depends only on the identity of adjacent repeats and is independent of the longer-range context. This enables the third and final step, general module assembly, in which building-block and junction modules are combined to generate a protein with a desired overall curvature. Although the overall strategy is applicable to any repeat protein, in this paper we focus on LRRs. We describe the computational design and experimental characterization for each step in the following sections.

Step 1: building-block-module selection and design

Nature provides a diverse set of LRR modules, with lengths from 20 to 30 amino acids[8], but only a few possess high self-compatibility such that repeated stacking of the same module generates a well-folded protein structure. We generated a Markov transition model for naturally occurring LRR proteins to investigate the overall patterns of module organization in LRR structures. In the model, nodes correspond to individual modules (represented by the module length: L22 indicates an LRR module with 22 residues, etc.), and edges correspond to transitions between modules with strength proportional to the transition frequency observed between the modules in the Protein Data Bank (PDB) (Online Methods). The resulting transition network

[1]Department of Biochemistry, University of Washington, Seattle, Washington, USA. [2]Institute for Protein Design, University of Washington, Seattle, Washington, USA. [3]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [4]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to D.B. (dabaker@u.washington.edu).

Received 21 August 2014; accepted 25 November 2014; published online 12 January 2015; doi:10.1038/nsmb.2938

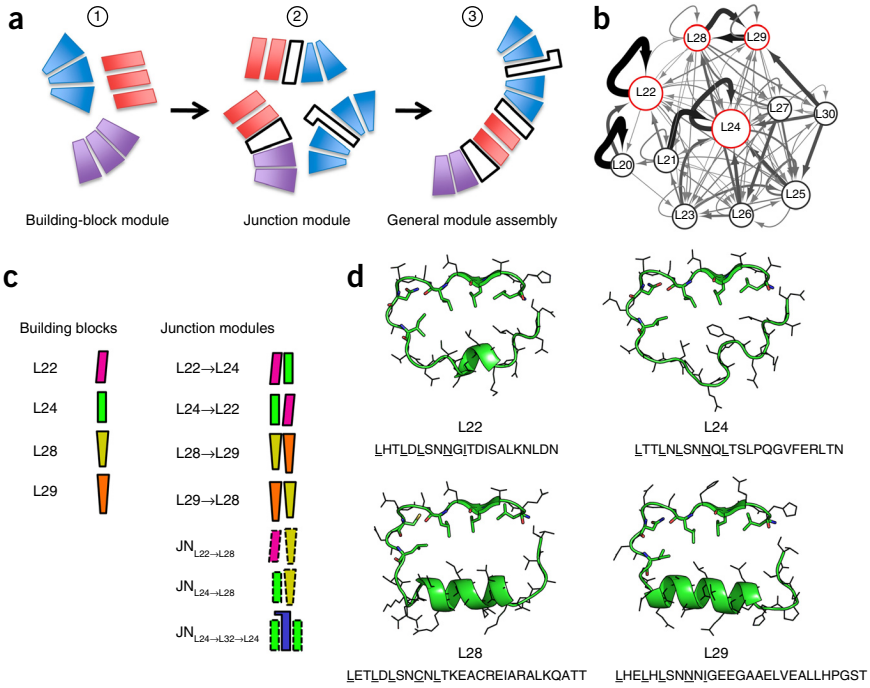

Figure 1 Assembly of LRRs from modules. (**a**) Overview of curvature-tunable scaffold design: idealized building-block-module design, junction-module design and general module assembly. (**b**) Module organization of natural LRRs. Nodes represent modules, and edges represent transitions between modules. The size of nodes and the thickness of edges are proportional to the frequencies observed in the PDB. (**c**) Graphical representation of designed building-block and junction modules. (**d**) Idealized designed building-block-module structures and sequences. The highly conserved residues are shown in sticks and underlines.

(**Fig. 1b**) has strong self-edges, corresponding to packing of identical modules for L22 and L24, and strong mutual transitions between L28 and L29. Accordingly, we selected these LRR types to design the idealized building blocks (**Fig. 1c**,**d**).

We used a recently developed Rosetta repeat-protein-idealization method[22] to design ideal versions of each unit. Different instances of the naturally occurring repeat units have somewhat variable sequences; the idealization process generates a single low-energy repeat unit (both sequence and structure) guided by the available information for the family. Briefly, we generated an idealized polyvaline backbone structure with identical repeats by using RosettaRemodel[23] with LRR family–specific constraints. We then carried out Rosetta sequence design guided by a family-specific sequence profile while constraining the sequences to be identical for each repeat. The idealization of the L24 module (DLRR_B) was previously described in Parmeggiani *et al.*[22]. We applied the idealization procedure to the L22 module (DLRR_A) and the two-unit {L28→L29} module (DLRR_C) and obtained the sequences and models in **Figure 1d**.

We synthesized genes for proteins containing five to seven idealized building-block modules. We fused the N-terminal capping domain of internalin B (Ncap) to DLRR_A and DLRR_B to enhance protein solubility and expression[12,20], whereas we expressed DLRR_C without a capping motif, instead redesigning the sequences of the N- and C-terminal repeats to eliminate exposed hydrophobic residues. After expressing the idealized repeat designs in *Escherichia coli*, we found them to be soluble and to have high thermal stability (**Fig. 2c**).

We solved the crystal structures of DLRR_A ($L22^6$) and DLRR_B ($L24^7$; with superscript numbers indicating numbers of repeat units) (**Table 1**) and found that they closely match the design models (DLRR_A at Cα r.m.s. deviation 1.4 Å; DLRR_B at Cα r.m.s. deviation 1.7 Å; **Fig. 3a**,**b**). The crystal structures contain water-mediated networks localized to the convex side of the repeats; it may be possible to incorporate these in future design calculations (**Supplementary Fig. 1a**). Each of the idealized building-block repeats has the expected overall curvature: repeats of the L22 and L24 building blocks generate solenoid-like structures, whereas repeats of the {L28→L29} building block are almost circular and have a more curved concave surface (parametric descriptions of the global shapes generated by each building-block repeat in **Supplementary Fig. 1b** and **Supplementary Table 1**).

Step 2: design of junction modules

We devised a computational protocol for junction-module design that takes advantage of the conserved motif (LxxLxLxxN/C) in the idealized LRR building blocks: the core residues are kept constant to maintain a stable hydrophobic core, whereas the evolutionarily variable positions, primarily located on the convex side, are optimized to create a low-energy interface between adjacent modules. To generate a junction module $JN_{BBi\rightarrow BBj}$ connecting building block *i* and building block *j*, we start from a two-unit BB_i^2 module and a one-unit BB_j module (**Fig. 2a**). The second unit in BB_i^2 is superimposed on BB_j by aligning the core motif residues. RosettaCM[24] is then used to generate a hybrid structure $BB_i\rightarrow BB_j$ with coordinates based on those of the first unit in BB_i^2 before the core motif and those of BB_j after the motif. The residues at the fusion interface are optimized with RosettaDesign[25]. This redesigned hybrid two-unit structure $BB_i\rightarrow BB_j$ is the junction module $JN_{BBi\rightarrow BBj}$ between building block *i* and building block *j* (**Fig. 2b**).

A special case of a junction module is a three-unit module $JN_{BBi\rightarrow BBw\rightarrow BBi}$ that connects two identical copies of the same building block but has a structure different than that of the building block (**Fig. 2b**). We call such junction modules between two identical building blocks 'wedge' modules. Like other junction modules, wedge modules produce a local change in the protein curvature. We designed and characterized five junction modules connecting the building-block modules described in the previous section.

A junction module for L22→L24 has been generated previously without hydrophobic core design[12]; hence, we made direct fusion constructs between L22 and L24 in both directions (i.e., L22→L24 and L24→L22) to test compatibility between the two idealized modules. The hybrid model structures showed high structural compatibility without further design. Thus, the junction modules in these cases are simply the fusion of the two building blocks.

We expressed two fusion proteins for L22→L24 (DLRR_D) and L24→L22 (DLRR_E) in *E. coli* and found them to be soluble and monomeric in size-exclusion chromatography coupled to multiangle light scattering (SEC-MALS) experiments (**Supplementary Fig. 1e**). Far-UV CD spectra and thermal denaturation profiles suggested well-packed structures with the expected secondary-structure content (**Fig. 2d**). The fusion proteins had similar or higher stability than the

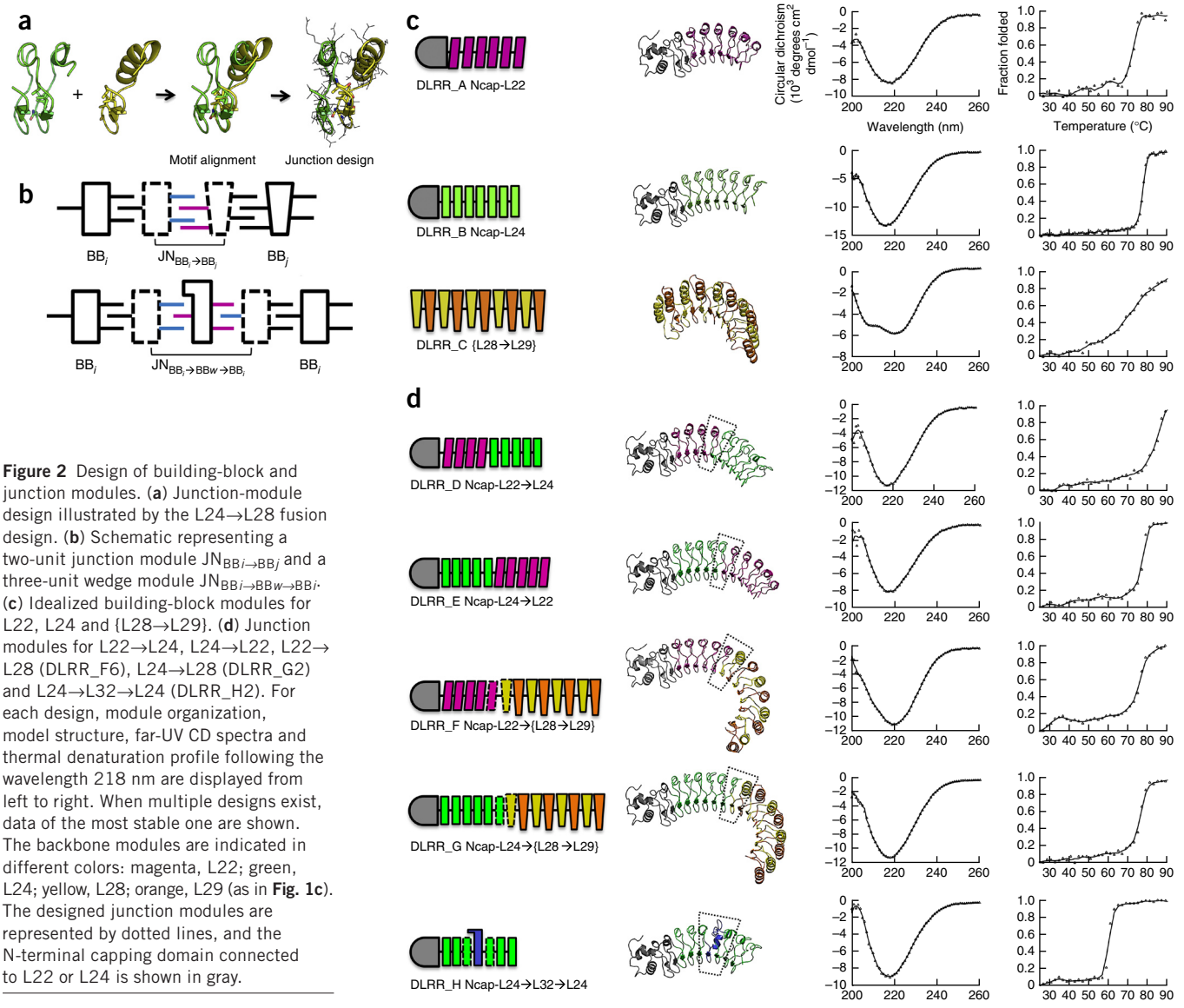

Figure 2 Design of building-block and junction modules. (**a**) Junction-module design illustrated by the L24→L28 fusion design. (**b**) Schematic representing a two-unit junction module $JN_{BBi \rightarrow BBj}$ and a three-unit wedge module $JN_{BBi \rightarrow BBw \rightarrow BBi}$. (**c**) Idealized building-block modules for L22, L24 and {L28→L29}. (**d**) Junction modules for L22→L24, L24→L22, L22→L28 (DLRR_F6), L24→L28 (DLRR_G2) and L24→L32→L24 (DLRR_H2). For each design, module organization, model structure, far-UV CD spectra and thermal denaturation profile following the wavelength 218 nm are displayed from left to right. When multiple designs exist, data of the most stable one are shown. The backbone modules are indicated in different colors: magenta, L22; green, L24; yellow, L28; orange, L29 (as in **Fig. 1c**). The designed junction modules are represented by dotted lines, and the N-terminal capping domain connected to L22 or L24 is shown in gray.

original L22 (DLRR_A) or L24 (DLRR_B) designs (**Table 2**; L22 and L24 evidently have high compatibility despite the rare occurrence of fusions between them in nature (**Fig. 1b**). The crystal structure for L24→L22 (DLRR_E) at 1.9-Å resolution showed high consistency with the design model (**Fig. 3c**) and the original L22 and L24 structures (**Supplementary Fig. 1c**).

Designs of junction modules for L22→L28 and L24→L28 are challenging because of substantial differences in the module length (22 or 24 versus 28), secondary structure in the variable region (3^{10} helix or loop versus α-helix), curvature on the concave surface (moderately curved versus highly curved) and global shape (superhelical versus circular). The initial fusion models generated by RosettaCM[24] (before redesign) contained side chain clashes and cavities at the interface between the modules (**Supplementary Fig. 1d**). We therefore redesigned residues at the fusion interface to improve the all-atom Rosetta energy and packing as assessed by RosettaHoles[26]. We based the junction designs solely on building-block models generated by Rosetta[25] because the crystal structures of the building blocks were not determined. We experimentally characterized six designs for L22→L28 (DLRR_F) and six designs for L24→L28 (DLRR_G) (**Table 2**).

All designs, when expressed in *E. coli*, were highly soluble and monomeric in SEC-MALS experiments (**Supplementary Figs. 2** and **3**). They displayed well-defined far-UV CD spectra with minima near 218 nm, similar to those of previously characterized LRRs with primarily β-sheet secondary structure. Thermal denaturation experiments showed cooperative unfolding for all fusion designs (**Fig. 2d**), suggesting a well-packed hydrophobic core. Fusion of more-stable LRR modules to less-stable LRR modules via a well-designed junction appeared to increase overall stability: the stability of all the junction module–containing designs was greater than that of the original {L28→L29}[5] design (DLRR_C).

We determined the crystal structure of the L24→L28 fusion (DLRR_G3) to evaluate the accuracy of the design. The crystal structure, determined at 2.5-Å resolution, shows the atomic details of the junction module as well as the structures of L24 and {L28→L29} modules (**Fig. 3d**). The assumption underlying our approach that curvature can be locally controlled is supported by the similarity of the L24 modules (Cα r.m.s. deviation 0.3 Å) in the DLRR_G3 structure to those in the all L24 DLRR_B structure and by the similarity of the {L28→L29} modules (Cα r.m.s. deviation 1.3 Å) to the {L28→L29} modules in the DLRR_C model. The key core side chain interactions

Table 1 Data collection and refinement statistics

Crystal	DLRR_A	DLRR_E	DLRR_G3	DLRR_H2	DLRR_I	DLRR_K
Data collection						
Space group	$P2_1$	$P2_12_12_1$	$F222$	$P2_12_12_1$	$C2$	$P22_12_1$
Cell dimensions						
a, *b*, *c* (Å)	57.66, 245.07, 57.73	32.12, 77.71, 101.89	91.13, 136.38, 161.74	89.78, 96.50, 136.36	109.49, 42.71, 67.82	36.87, 93.37, 126.24
α, β, γ (°)	90, 115.36, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 102.4, 90	90, 90, 90
Resolution (Å)	50 (2.36)[a]	42.6 (1.93)	23.5 (2.53)	50 (2.9)	50 (1.73)	50 (2.8)
R_{sym}	0.081 (0.183)	0.063 (0.171)	0.067 (0.153)	0.092 (0.529)	0.076 (0.252)	0.192 (0.742)
$I / \sigma I$	24.0 (6.4)	17.7 (8.0)	15.5 (4.1)	17.2 (3.85)	33.7 (4.5)	8.9 (2.3)
Completeness (%)	96.7 (83.7)	99.8 (96.0)	98.1 (85.5)	99.8 (99.6)	96.3 (83.7)	99.7 (99.2)
Redundancy	5.7 (3.0)	6.4 (5.0)	4.5 (1.9)	7.2 (7.1)	10.3 (2.3)	6.2 (5.8)
Refinement						
Resolution (Å)	50 (2.36)	42.6 (1.93)	23.5 (2.53)	50 (2.9)	50 (1.73)	50 (2.8)
No. reflections	34,180	19,993	17,061	25,484	31,150	10,729
R_{work}(%)	18.9 (22.3)	15.86 (17.50)	18.47 (23.4)	21.16 (32.8)	17.07 (21.50)	20.75 (28.4)
R_{free} (%)	24.2 (27.7)	22.38 (23.7)	24.65 (36.1)	25.15 (48.5)	21.99 (31.70)	28.53 (36.0)
No. atoms						
Protein	6,771	2,388	3,456	7,841	2,577	3,582
Ligand/ion	8	12	29	20	–	1
Water	230	106	96	1	199	18
B factors						
Protein	12.13	14.65	11.73	69.14	10.84	16.96
Ligand/ion	35.53	39.67	54.64	85.26	–	42.89
Water	18	35.39	21.91	50.44	25.43	18.6
r.m.s. deviations						
Bond length (Å)	0.0137	0.0181	0.0138	0.0126	0.0194	0.0136
Bond angles (°)	1.661	1.81	1.629	1.651	2.052	1.475

[a]Values in parentheses are for highest-resolution shell.

in the junction module are very similar in the design model and crystal structure (Cα r.m.s. deviation 0.9 Å; **Fig. 3d**).

In addition to the junction modules linking the different building-block modules, we designed a wedge module inserted between L24 modules. In native LRR proteins, inserting an 'irregular' module between the regular modules is a common way to generate structural diversity by altering the overall curvature or forming a binding interface other than the concave surface (for example, the diverse LRR-module organization and irregular binding surfaces in TLR family[27] and plant LRR proteins[28]). We chose the idealized L24 repeat structure (DLRR_B) as a base scaffold because it had the highest stability among the three idealized LRRs.

For the wedge-module design, we retrieved L24→L*x*→L24 triples (in which *x* denotes any length of LRR) from the LRRML database[29] to identify irregular modules flanked by the L24 modules, finding a total of 21 unique irregular modules. We selected the 32-length LRR unit (L32) found in the Toll-like receptor 3 structure[30] (PDB 2A0Z[30], 532–563) as a starting point. L32 has a relatively rigid and structured loop located on its convex surface, which could be useful in future binding-pocket designs. We applied the junction-module design process to the two fusion interfaces (L24→L32 and L32→L24), which resulted in the wedge module $JN_{L24\rightarrow L32\rightarrow L24}$ (DLRR_H). We then selected and experimentally characterized four designs for L24→L32→L24 (**Table 2**).

All designs, when expressed in *E. coli*, were soluble. Two designs were monomeric in a SEC-MALS experiment. Thermal

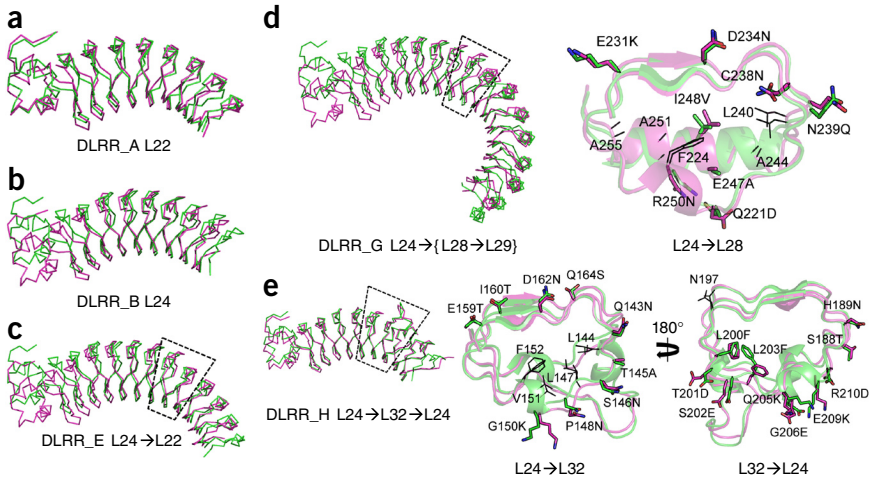

Figure 3 Crystal structures of the building-block-module and junction-module designs. (**a**–**e**) Design models (green) for L22 (**a**), L24 (**b**), L24→L22 (**c**), L24→L28 (DLRR_G3) (**d**) and L24→L32→L24 (DLRR_H2) (**e**) superimposed on the crystal structures (magenta). Close-up views for the designed junction modules (dashed region) are shown for **d** and **e**. Residues mutated from the original building-block sequences are annotated and shown as sticks. Additional residues that vary within the designs are shown as lines. Except for that of DLRR_A, the crystal structures have missing electron density at the C terminus (10–20 amino acids). PyMOL (http://www.pymol.org/) is used in all structural visualizations.

Table 2 Summary of fusion designs and experimental results

Design name	Module organization[a]	Modules (repeat units)[b]	Designs tested	Soluble	Folded (CD)	Monomeric	X-ray	T_m (°C)[e]	r.m.s. deviation[f]	r.m.s. deviation[g]
DLRR_A	Ncap-L22^6	6 (6)	1	1	1	1	1	73	1.4 (0.8, 2.0)	0.4 (0.5, 1.0)
DLRR_B[c]	Ncap-L24^7	7 (7)	1	1	1	1	1	78	1.7 (1.5, 2.9)	0.3 (0.6, 0.4)
DLRR_C[d]	{L28→L29}^{5}	5 (10)	5	5	1	–		71		
DLRR_D	Ncap-L22^4→L24^5	9 (9)	1	1	1	1		87		
DLRR_E	Ncap-L24^5→L22^5	10 (10)	1	1	1	1	1	77	2.1 (1.4, 2.0)	0.4 (0.7, 0.7)
DLRR_F	Ncap-L22^4-JN$_{L22\rightarrow L28}$→L29→{L28→L29}^{3}	9 (13)	6	6	6	6		77		
DLRR_G	Ncap-L24^5-JN$_{L24\rightarrow L28}$→L29→{L28→L29}^{3}	10 (14)	6	6	6	6	1	81	2.6 (3.1, 3.8)	0.8 (0.8, 2.2)
DLRR_H	Ncap-L24^2-JN$_{L24\rightarrow L32\rightarrow L24}$-L24^2	5 (7)	4	4	4	2	1	65	0.9 (0.5, 1.0)	0.8 (0.7, 1.2)
DLRR_I	Ncap-L24^2-JN$_{L24\rightarrow L32\rightarrow L24}$-JN$_{L24\rightarrow L32\rightarrow L24}$-L24^2	6 (10)	1	1	1	1	1	53	1.7 (1.2, 2.3)	0.5 (0.5, 0.7)
DLRR_J	Ncap-L22^4→L24^2-JN$_{L24\rightarrow L28}$→L29→{L28→L29}^{2}	10 (13)	1	1	1	1		82		
DLRR_K[h]	Ncap-L24^2-JN$_{L24\rightarrow L32\rightarrow L24}$-L24^3-JN$_{L24\rightarrow L28}$→L29→{L28→L29}^{2}	10 (15)	1	1	1	1	1	75		1.1 (1.2, 3.9)
DLRR_L	Ncap-L22^3→L24^3-JN$_{L24\rightarrow L32\rightarrow L24}$-L24^3-JN$_{L24\rightarrow L28}$→L29→{L28→L29}^{2}	14 (19)	1	1	1	1		83		

[a]The superscripts represent the number of repeat units. [b]The alternating two-unit {L28→L29} is considered one module. [c]Experimental data of DLRR_B are from Parmeggiani *et al.*[22]. [d]DLRR_C forms a dimer. [e]T_m is estimated by calculating the infection point of the melting curve at 218 nm, and the highest T_m value is represented when multiple designs exist. [f]r.m.s. deviation is Cα r.m.s. deviation (Å) between crystal structure and model generated from design models of building blocks and junction modules. [g]r.m.s. deviation is Cα r.m.s. deviation (Å) between crystal structure and model generated from crystal structures of building-block and junction modules (**Supplementary Fig. 5c**). r.m.s. deviations for the first and the last unit in global structure alignment are provided in parentheses. [h]Model of DLRR_K is generated by module assembly without an initial design model.

denaturation experiments showed that insertion of the wedge module generally decreased stability of the base scaffold, but unfolding was still cooperative (**Fig. 2d** and **Supplementary Fig. 4**). The crystal structure of DLRR_H2 determined at 2.9-Å resolution was consistent with the design model (Cα r.m.s. deviation 0.9 Å), thus confirming the accuracy of the junction-module design protocol (**Fig. 3e**).

Step 3: Curvature specification by general module assembly

The crystal structures described thus far demonstrate that the building-block modules (L22, L24, L28 and L29), junction modules (L22→L24, L24→L22, L24→L28, L28→L29 and L29→L28) and wedge modules (L24→L32→L24) all have structures that are very similar to the design models regardless of overall protein context. In principle, this enables the design of combinations of modules to achieve a desired curvature. We represent the space of possible LRR structures as a network consisting of building-block modules (nodes) connected by junction modules (edges) as in **Figure 1b** (**Fig. 4a**). Any sequence of modules generated by following the edges in the network corresponds to an LRR structure with unique curvature. For example, all the 18,786 possible fusion structures consisting of 12 modules are depicted as lines connecting the center of masses for each repeat module in the structure (**Fig. 4b**). The curvature diversity is orders of magnitude greater than that of the original LRRs containing the same number of building-block modules.

We chose to use models of the individual building-block and junction modules extracted from the crystal structures described thus far in the general module assembly process rather than the original design models of these units. Although the building-block modules are similar to previously described structures, the designed junction modules have sequences (**Supplementary Fig. 5a**) and structures (**Supplementary Fig. 5b**) quite different from those of previously described LRRs. Because of the imperfect state of computational protein design, we consider the crystal structures (which differ from the design models by Cα r.m.s. deviation 0.2–1.0 Å) to be more accurate representations of the structures that these modules are likely to adopt in new designs (**Supplementary Fig. 5c**).

General module assembly and experimental characterization

As a proof of concept for general module assembly, we designed four multiple-fusion constructs (DLRR_I, DLRR_J, DLRR_K and DLRR_L; **Fig. 4c**). The designs contain more than two fusion interfaces, resulting in large superhelical structures comparable in size to TLR4 (ref. 31; PDB 3FXI, 626 residues) and plant steroid receptor BRI1 (ref. 32; PDB 3RIZ, 743 residues) (module organization and module origins for each design in **Table 2**, **Fig. 4c** and **Supplementary Table 2**).

Experimental characterization showed that the general module assembly protocol is quite robust. All of the multiple-fusion designs, when expressed in *E. coli*, were soluble and monomeric with well-defined CD spectra, cooperative unfolding transitions and high thermal stability (**Fig. 4c** and **Supplementary Fig. 1e**). This is notable because all are quite large and complex proteins. We succeeded in solving the crystal structures of two of the designs.

The structure of design DLRR_I, containing two successive L32 wedge modules with multiple flanking N- and C-terminal L24 modules, was solved at 1.7-Å resolution (**Fig. 4d**). In agreement with the assumption of context-independent structure of the individual modules, the two L32 wedge modules in DLRR_I and the single L32 wedge module in DLRR_H2 are nearly identical over the backbone and core side chains (Cα r.m.s. deviation 0.3–0.5 Å). Over the full ten-repeat-unit structure, the crystal structure is closer to the model (Cα r.m.s. deviation 0.5 Å; **Supplementary Fig. 5c**) assembled from the crystal structures of the individual building-block and junction modules extracted from DLRR_B and DLRR_H than to the model (Cα r.m.s. deviation 1.7 Å; **Fig. 4d**) assembled from the design models of the individual modules (**Table 2**), thus supporting our decision to use the crystal structures of the building blocks rather than the original design models in the general module assembly calculations.

Design DLRR_K consists of two L24 modules followed by the L32 module, three additional L24 modules, the L24→L28 junction module and three {L28→L29} modules—a total of 15 repeat units. Such complexity of module organization is rarely if ever observed in naturally occurring LRRs. The protein is monomeric and stable, with a melting temperature (T_m) of 75 °C. The crystal structure of

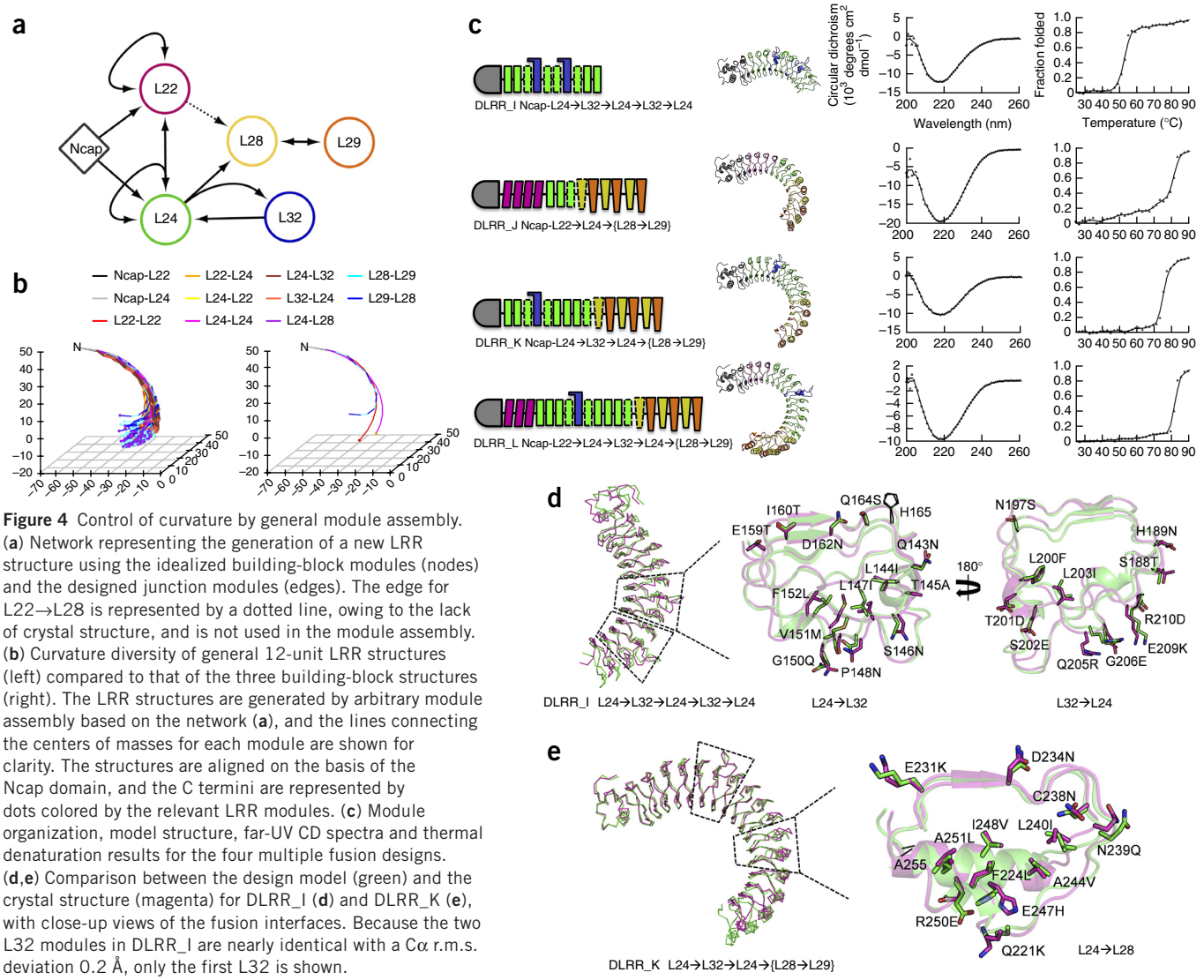

Figure 4 Control of curvature by general module assembly. (**a**) Network representing the generation of a new LRR structure using the idealized building-block modules (nodes) and the designed junction modules (edges). The edge for L22→L28 is represented by a dotted line, owing to the lack of crystal structure, and is not used in the module assembly. (**b**) Curvature diversity of general 12-unit LRR structures (left) compared to that of the three building-block structures (right). The LRR structures are generated by arbitrary module assembly based on the network (**a**), and the lines connecting the centers of masses for each module are shown for clarity. The structures are aligned on the basis of the Ncap domain, and the C termini are represented by dots colored by the relevant LRR modules. (**c**) Module organization, model structure, far-UV CD spectra and thermal denaturation results for the four multiple fusion designs. (**d**,**e**) Comparison between the design model (green) and the crystal structure (magenta) for DLRR_I (**d**) and DLRR_K (**e**), with close-up views of the fusion interfaces. Because the two L32 modules in DLRR_I are nearly identical with a Cα r.m.s. deviation 0.2 Å, only the first L32 is shown.

DLRR_K at 2.8-Å resolution is very close to the general module assembly model (built from crystal structures of the individual modules from previous structures), with a Cα r.m.s. deviation of 1.1 Å (**Fig. 4e**).

The structures of DLRR_I and DLRR_K demonstrate that assembly of designed building-block and junction modules can produce new structures with predefined shapes quite accurately.

DISCUSSION

We have described a general approach to creating repeat proteins with custom-designed shapes through combination of designed building-block and junction modules. The generation of scaffolds with defined curvatures with our computational approach is very likely to be simpler than that which occurred during the complex evolution of naturally occurring LRRs and is considerably more controlled than what can be achieved in library selection approaches. The strategy allows the ready programming of a rich diversity of scaffolds with distinct curvatures: over 18,000 distinct 12-repeat-unit structures can, in principle, be generated with our current set of building-block and junction modules (**Fig. 4b** and **Supplementary Table 3**). The stable and well-expressed DLRR_L design (**Fig. 4c**) has a complex organization with five different types of modules (19 repeat units) in total; for this length there are over 5,000,000 distinct possibilities with our current module set, and increasing the repertoire of idealized designs of building-block and junction modules would enrich the curvature diversity still further.

Our approach integrates protein structural analysis with energy-driven design calculations to arrive at the idealized building-block and junction modules and further uses computation and experiment to achieve high-accuracy models of the complex repeat proteins generated by the module assembly process. Although a completely energy-driven approach would be preferable on aesthetic grounds, making use of information extracted from naturally occurring LRRs and from the crystal structures of idealized LRRs described in this study allows the generation of large families of LRR proteins with tunable curvatures to address current challenges. The critical role of computation in the overall process is illustrated by the junction modules: both the sequences (**Supplementary Fig. 5a**) and the structures (**Supplementary Fig. 5b**) of the designed junction modules differ considerably from those of their closest counterparts in naturally occurring LRRs and hence could not have been obtained without energy-driven design calculations. These calculations are not perfect, however, and because the small differences between the design models and the corresponding crystal structures are amplified through lever-arm effects when many modules are combined, we use crystal structures of the designed building-block and junction modules in

the general module assembly calculations rather than the original design models.

The ability to custom design repeat proteins with well-defined shapes and curvatures has immediate application to the design of a next generation of high-affinity binding proteins. Studies of native protein-protein interactions have shown that shape complementarity is a major determinant of protein binding affinity[33–36]. In particular, naturally occurring LRR-based binding proteins often achieve high affinity and specificity by having shapes closely conforming to the surfaces of the target proteins. The importance of this shape tuning for LRR-protein molecular recognition is illustrated with the naturally occurring LRR proteins internalin A (InlA) and RNase inhibitor (RI) (**Supplementary Fig. 5d**). Each protein has a curvature adapted to its target (E-cadherin and RNase A, respectively), thus resulting in well-packed complementary protein-protein interfaces with hotspot clusters at both the N and C termini. Swapping the targets for each of these LRR proteins results in substantial clashes and large gaps (**Supplementary Fig. 5d**).

With the capability provided by the approach described in this paper, it is now possible to design new proteins with high backbone-shape complementarity to essentially any macromolecular target of interest. Coupled with protein-interface design methodology previously used to create new binding proteins based on already existing scaffolds[37,38], this should allow the design of high-affinity, high-specificity binding proteins. Such an approach complements directed evolution methods[13,39,40] for obtaining high-affinity binding proteins on the basis of a single stable protein backbone, which, although powerful, still require considerable effort. For creating a high-affinity binding protein to a target of interest in the near future, a combination of our shape-complementary-scaffold design approach, protein-protein-interface design for chemical complementarity and limited directed evolution to optimize interactions not accurately described by computational design may prove particularly effective.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Coordinates and structure factors have been deposited in the Protein Data Bank under accession codes 4R58 (DLRR_A), 4R5C (DLRR_E), 4R5D (DLRR_G3), 4R6J (DLRR_H2), 4R6F (DLRR_I) and 4R6G (DLRR_K).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Bolduc for data collection for DLRR_A and the members of the protein production facility at the Institute for Protein Design for protein production. This work was supported by grants from the Howard Hughes Medical Institute (to D.B.), the Defense Threat Reduction Agency (HDTRA1-11-1-0041 to D.B.) and US National Institutes of Health (R01 GM49857 to B.L.S.). F.P. was supported as the recipient of a Swiss National Science Foundation Postdoctoral Fellowship (PBZHP3-125470) and a Human Frontier Science Program Long-Term Fellowship (LT000070/2009-L). This work was facilitated though the use of advanced computational, storage and networking infrastructure provided by the Hyak supercomputer system at the University of Washington.

AUTHOR CONTRIBUTIONS

K.P. and D.B. conceived the project; K.P. performed the computational design with assistance from F.P. and P.-S.H.; K.P. expressed, purified and characterized the designs with assistance from F.P.; B.W.S. crystallized the designs and determined the structures; K.P. and D.B. drafted the manuscript with input from all authors; B.L.S. and D.B. supervised research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Binz, H.K., Amstutz, P. & Pluckthun, A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* **23**, 1257–1268 (2005).
2. Skerra, A. Alternative non-antibody scaffolds for molecular recognition. *Curr. Opin. Biotechnol.* **18**, 295–304 (2007).
3. Gebauer, M. & Skerra, A. Engineered protein scaffolds as next-generation antibody therapeutics. *Curr. Opin. Chem. Biol.* **13**, 245–255 (2009).
4. Javadi, Y. & Itzhaki, L.S. Tandem-repeat proteins: regularity plus modularity equals design-ability. *Curr. Opin. Struct. Biol.* **23**, 622–631 (2013).
5. Grove, T.Z., Regan, L. & Cortajarena, A.L. Nanostructured functional films from engineered repeat proteins. *J. R. Soc. Interface* **10**, 20130051 (2013).
6. Phillips, J.J., Millership, C. & Main, E.R. Fibrous nanostructures from the self-assembly of designed repeat protein modules. *Angew. Chem. Int. Edn Engl.* **51**, 13132–13135 (2012).
7. Han, S.H., Lee, M.K. & Lim, Y.B. Bioinspired self-assembled peptide nanofibers with thermostable multivalent α-helices. *Biomacromolecules* **14**, 1594–1599 (2013).
8. Kobe, B. & Kajava, A.V. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **11**, 725–732 (2001).
9. Enkhbayar, P., Kamiya, M., Osaki, M., Matsumoto, T. & Matsushima, N. Structural principles of leucine-rich repeat (LRR) proteins. *Proteins* **54**, 394–403 (2004).
10. Kim, H.M. *et al.* Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. *Cell* **130**, 906–917 (2007).
11. Parker, R., Mercedes-Camacho, A. & Grove, T.Z. Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Sci.* **23**, 790–800 (2014).
12. Lee, S.C. *et al.* Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proc. Natl. Acad. Sci. USA* **109**, 3299–3304 (2012).
13. Binz, H.K. *et al.* High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.* **22**, 575–582 (2004).
14. Forrer, P., Stumpp, M.T., Binz, H.K. & Pluckthun, A. A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett.* **539**, 2–6 (2003).
15. Grove, T.Z., Cortajarena, A.L. & Regan, L. Ligand binding by repeat proteins: natural and designed. *Curr. Opin. Struct. Biol.* **18**, 507–515 (2008).
16. Main, E.R., Xiong, Y., Cocco, M.J., D'Andrea, L. & Regan, L. Design of stable alpha-helical arrays from an idealized TPR motif. *Structure* **11**, 497–508 (2003).
17. Filipovska, A. & Rackham, O. Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Mol. Biosyst.* **8**, 699–708 (2012).
18. Reichen, C., Hansen, S. & Pluckthun, A. Modular peptide binding: from a comparison of natural binders to designed armadillo repeat proteins. *J. Struct. Biol.* **185**, 147–162 (2014).
19. Mak, A.N., Bradley, P., Cernadas, R.A., Bogdanove, A.J. & Stoddard, B.L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335**, 716–719 (2012).
20. Ryou, J.H., Park, K., Lee, J.J., Kim, D. & Kim, H.S. Soluble expression of human glycoprotein Ibα in *Escherichia coli* through replacement of the N-terminal capping domain. *Protein Expr. Purif.* **101**, 21–27 (2014).
21. Jung, K. *et al.* Toll-like receptor 4 decoy, TOY, attenuates gram-negative bacterial sepsis. *PLoS ONE* **4**, e7403 (2009).
22. Parmeggiani, F. *et al.* A general computational approach for repeat protein design. *J. Mol. Biol.* doi:10.1016/j.jmb.2014.11.005.
23. Huang, P.S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS ONE* **6**, e24109 (2011).
24. Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).
25. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
26. Sheffler, W. & Baker, D. RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci.* **18**, 229–239 (2009).
27. Park, B.S. & Lee, J.O. Recognition of lipopolysaccharide pattern by TLR4 complexes. *Exp. Mol. Med.* **45**, e66 (2013).
28. Matsushima, N. & Miyashita, H. Leucine-rich repeat (LRR) domains containing intervening motifs in plants. *Biomolecules* **2**, 288–311 (2012).
29. Wei, T. *et al.* LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs). *BMC Struct. Biol.* **8**, 47 (2008).
30. Bell, J.K. *et al.* The molecular structure of the Toll-like receptor 3 ligand-binding domain. *Proc. Natl. Acad. Sci. USA* **102**, 10976–10980 (2005).
31. Park, B.S. *et al.* The structural basis of lipopolysaccharide recognition by the TLR4–MD-2 complex. *Nature* **458**, 1191–1195 (2009).
32. Hothorn, M. *et al.* Structural basis of steroid hormone perception by the receptor kinase BRI1. *Nature* **474**, 467–471 (2011).

33. Chen, R. & Weng, Z. A novel shape complementarity scoring function for protein-protein docking. *Proteins* **51**, 397–408 (2003).
34. Gabb, H.A., Jackson, R.M. & Sternberg, M.J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272**, 106–120 (1997).
35. Lawrence, M.C. & Colman, P.M. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234**, 946–950 (1993).
36. Jones, S. & Thornton, J.M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 13–20 (1996).
37. Fleishman, S.J. *et al.* Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**, 816–821 (2011).
38. Procko, E. *et al.* Computational design of a protein-based enzyme inhibitor. *J. Mol. Biol.* **425**, 3563–3575 (2013).
39. Epa, V.C. *et al.* Structural model for the interaction of a designed Ankyrin Repeat Protein with the human epidermal growth factor receptor 2. *PLoS ONE* **8**, e59163 (2013).
40. Lee, J.J. *et al.* A high-affinity protein binder that blocks the IL-6/STAT3 signaling pathway effectively suppresses non-small cell lung cancer. *Mol. Ther.* **22**, 1254–1265 (2014).

ONLINE METHODS

Markov transition model for natural LRR modules. To construct a Markov transition model for natural LRR modules, all sets of two consecutive LRR modules were collected from the LRRML database[29] and labeled on the basis of module length. From these data, we computed the transition probability $P_{a\to b} = N_{a\to b}/\Sigma_i N_{a\to i}$, where $N_{a\to b}$ represents the frequency of transitions from module length a to b in the PDB. In the network model in **Figure 1b**, the size of a node was scaled by the frequency of a module length in the PDB, and the thickness of an edge was scaled by the transition probability.

Computational design of junction modules. The initial fusion models were generated by RosettaCM[24] from the motif-aligned scaffolds as described in the main text (**Fig. 2a**) and were refined with the Rosetta relax protocol with coordinate constraints[41] to reduce perturbation of the structure. The fusion interface between the two heterogeneous building blocks was redesigned to improve structural compatibility with the Rosetta FastRelax protocol. The protocol runs four cycles of repack, design and minimization, and during each cycle the weight for the repulsive energy term gradually increases to obtain a well-packed and low-energy structure. During the design procedure, residue type constraints were added in order to favor original residue identities. After generating 1,000 design sequences, the top 10% of design sequences by both Rosetta energy and packing were retrieved and manually inspected to select the final sequences.

LRR structure modeling by iterative module assembly. Structures of building-block and junction modules were extracted from the crystal structures of the designed LRR proteins containing one or two building block–module types. Specifically, two-unit or three-unit module structures of Ncap-L22 (DLRR_A), L22-L22 (DLRR_A), Ncap-L24 (DLRR_B), L24-L24 (DLRR_B), L22→L24 (DLRR_B), L24→L22 (DLRR_E), L24→L28→L29 (DLRR_G3), L28→L29 (DLRR_G3), L29→L28 (DLRR_G3) and L24→L32→L24 (DLRR_H2) were used to elongate a LRR structure through module assembly mediated by the common flanking module. For example, module assembly of L22→L24 and L24→L22 though the common L24 unit generates the three-unit structure L22→L24→L22. The module assembly was then iteratively applied to elongate the overall structure module by module this resulted in the mature form of a general LRR structure. Finally, energy minimization with Rosetta was performed to eliminate potential structural defects. The crystal structure of L22→L24 was obtained from DLRR_B, which has the L22-containing N-terminal capping domain (of internalin B) fused to L24. The L22→L28 was not used in the general module assembly, owing to the lack of the crystal structure.

Gene cloning, protein expression and purification. Genes encoding building-block LRRs were synthesized and cloned into pET21_NESG (DLRR_A) or pET15_NESG (DLRR_B and DLRR_C) expression vectors by GeneScript. The gene fragments for each junction module were separately prepared by PCR assembly of six to eight 50- to 60-nucleotide oligos or by gene synthesis from Integrated DNA Technologies. Another gene fragment for the building-block module to be fused to was also obtained by PCR. The two gene fragments were then inserted into the plasmid of the appropriate building-block protein by Gibson cloning[42]. The C-terminal His_6 tag was added to all design sequences with Gly-Ser or Gly-Ser-Trp linkers, in which tryptophan was for measuring protein concentration easily.

The proteins were expressed in *E. coli* BL21 Star (DE3) cells at 37 °C for 4 h after induction with 0.1 mM IPTG. The cell pellets were resuspended in 20 ml of lysis buffer containing 20 mM Tris, pH 8.0, 500 mM NaCl, 30 mM imidazole and 5% v/v glycerol. Roche complete EDTA-free protease inhibitor tablet, lysozyme (1 mg/ml), and DNase (1 mg/ml) were also added to the lysis buffer. After sonication, the proteins were purified with a Ni-NTA column and eluted with 20 mM Tris, pH 8.0, 500 mM NaCl and 250 mM imidazole. The proteins were further purified with a Superdex 200 column (GE Healthcare) equilibrated in 20 mM Tris and 50 mM NaCl at pH 8.0. The soluble expression and purity were also tested with SDS-PAGE and MS (LCQ Fleet Ion Trap mass spectrometer, Thermo Scientific).

Biophysical characterization. CD with an AVIV 62S DA spectrometer was used to investigate secondary-structure contents and thermal stability. Far-UV CD spectra from 200 nm to 260 nm were measured for the protein samples in 20 mM Tris, pH 8.0, and 50 mM NaCl. Thermal denaturation experiments were also performed by following the minimum at 218 nm and increasing the temperature from 25 °C to 90 °C. Size-exclusion chromatography coupled to multiangle light scattering (SEC-MALS) was performed to assess the oligomeric state of protein samples. A Superdex 200 10/300 GL column (GE Healthcare) was equilibrated in PBS buffer and used on an HPLC system (LC 1200 Series, Agilent Technologies) connected to miniDAWN TREOS static light-scattering detector (Wyatt Technologies). The collected data were analyzed by ASTRA software (Wyatt Technology).

X-ray crystallography. Crystals of designed LRR-repeat proteins were grown by standard vapor-phase diffusion methods with a TTP labtech 'Mosquito' crystallization robot with 50-nanoliter drops of protein at concentrations ranging from 15 mg/mL to 40 mg/mL, equilibrated against 100 volumes of microliter individual reservoir solutions. The reservoir compositions that produced each crystal are provided in **Supplementary Table 4**. Crystals were then flash-cooled by rapid emersion into artificial mother liquors corresponding to the crystallization reservoir solutions supplemented with either ethylene glycol (to 25% v/v) or with PEG 3350 (to 35% w/v). Diffraction data were collected on cryocooled crystals with either an in-house CCD area detector with a rotating anode X-ray generator (DLRR_A, DLRR_G3, DLRR_H2 and DLRR_K) or with a CCD area detector at the Advanced Light Source X-ray synchrotron facility (DLRR_E and DLRR_I). All data were processed and scaled with HKL2000 (ref. 43). Molecular replacement was performed with PHASER[44] with computational coordinates of the individual designs produced by Rosetta as search models. Model building was performed with COOT[45], and refinement was performed with REFMAC[46]. Protein sequences for all designs are available in **Supplementary Table 5**.

41. Nivón, L.G., Moretti, R. & Baker, D. A Pareto-optimal refinement method for protein design scaffolds. *PLoS ONE* **8**, e59004 (2013).
42. Gibson, D.G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
43. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
44. McCoy, A.J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
45. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
46. Winn, M.D., Murshudov, G.N. & Papiz, M.Z. Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods Enzymol.* **374**, 300–321 (2003).