A new class of synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension

James T. MacDonald [*] [†], **Burak V. Kabasakal** [‡], **David Godding** [‡] [§], **Sebastian Kraatz** [‡] [¶], **Louie Henderson** [‡], **James Barber** [‡], **Paul S. Freemont** [*] [†] **and James W. Murray** [‡]

[*]Centre for Synthetic Biology and Innovation, Imperial College London, London, SW7 2AZ, UK,[†]Department of Medicine, Imperial College London, London, SW7 2AZ, UK,[‡]Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK,[§]Department of Plant Sciences, University of Cambridge, Cambridge, CB2 3EA UK, and [¶]Laboratory of Biomolecular Research, Paul Scherrer Institut, CH-5232 Villigen PSI, Switzerland

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The ability to design and construct structures with atomic level precision is one of the key goals of nanotechnology. Proteins offer an attractive target for atomic design, as they can be synthesized chemically or biologically, and can self-assemble. However the generalized protein folding and design problem is unsolved. One approach to simplifying the problem is to use a repetitive protein as a scaffold. Repeat proteins are intrinsically modular, and their folding and structures are better understood than large globular domains. Here, we have developed a new class of synthetic repeat protein, based on the pentapeptide repeat family of beta-solenoid proteins. We have constructed length variants of the basic scaffold, and computationally designed *de novo* loops projecting from the scaffold core. The experimentally solved 3.56 Å resolution crystal structure of one designed loop matches closely the designed hairpin structure, showing the computational design of a backbone extension onto a synthetic protein core without the use of backbone fragments from known structures. Two other loop designs were not clearly resolved in the crystal structures and one loop appeared to be in an incorrect conformation. We have also shown that the repeat unit can accommodate whole domain insertions by inserting a domain into one of the designed loops.

computational protein design | synthetic repeat proteins | de novo backbone design

Abbreviations: RFR, repeat five residues; RMSD, root-mean-square deviation

During the course of evolution, natural proteins may be recruited to new unrelated functions conferring a selective advantage to the organism [1,2]. This accretion of new features and functions is likely to have left behind complex interlocking amino acid dependencies which can make reengineering natural proteins difficult and unpredictable [3]. For this reason, we and others hypothesize that it is more desirable to design *de novo* proteins as these provide a biologically-neutral platform onto which functional elements can be grafted [4]. Artificial proteins have been designed by decoding simple residue patterning rules that govern the packing of secondary structural elements and this has been particularly successful for α-helical bundle proteins [5–7]. An alternative approach is to assemble *de novo* folds from backbone fragments of known structures or idealized secondary structural elements and use computational protein design methods to design the sequence [4,8–10]. Both the computational and the simpler rules-based design approaches have concentrated on designing proteins consisting of canonical secondary structure linked with loops of minimal length.

A class of proteins that has attracted considerable interest is artificial proteins based on repeating structural motifs due to their intrinsic modularity and designability [11]. Repeat proteins have applications including their use as novel nanomaterials [12–14] and as scaffolds for molecular recognition [15, 16]. These proteins may be designed using both sequence consensus-based rules [17] or computational protein design methods [18, 19]. There are a number of families of beta-helical repeat proteins [20], from which we chose the pentapeptide repeat family, forming the RFR-fold (repeat five residues), which has a square cross-sectional profile, as the basis for the design of a new class of synthetic repeat protein (Fig. 1 A and B) [21].

The RFR-fold has a number of properties that make it attractive as a substrate for design. The structure is unusually regular, but is able to tolerate a wide range of residues on the outside of the solenoid barrel. The solenoids in natural RFR-fold proteins are nearly straight in contrast to several other forms of repeat protein such as the leucine rich repeat (LRR) which are highly curved. There are examples of natural RFR-fold proteins with loop extensions projecting from the barrel, making this class of protein particularly suitable for functionalization. The protein is similar in diameter to DNA, and some RFR-fold proteins are thought to play a role as DNA mimics [22]. Here, we have designed and solved the structures of a number of artificial RFR-fold proteins of different lengths.

Previously, computationally designed enzymes have reused backbone scaffolds from known natural proteins [23–25], although artificial helical bundle proteins have been functional-

Significance

The development of algorithms to design new proteins with backbone plasticity is a key challenge in computational protein design. In this paper, we describe a novel class of extensible synthetic repeat protein scaffolds with computationally designed variable loops projecting from the central core. We have developed new methods to computationally sample backbone conformations using a coarse-grained potential energy function without using backbone fragments from known protein structures. This was combined with existing methods for sequence design to successfully design a loop at atomic level precision. Given the inherent modular and composable nature of repeat proteins, this approach allows the iterative atomic-resolution design of complex structures with potential applications in novel nanomaterials and molecular recognition.

Reserved for Publication Footnotes

ized using an intuitive manual design process [26–28]. As the field of enzyme design becomes more ambitious it is likely that consideration of backbone plasticity will become increasingly important [29]. Backbone conformations from solved protein structures are guaranteed to be designable as there is at least one sequence known to fold into that structure. However, this is unlikely to be true for an arbitrary backbone conformation. The incorporation of backbone flexibility in protein design has been recognized as a key challenge in computational protein design [30] with current methods typically reusing backbone fragments from other known protein structures [31, 32]. Recently, we have developed algorithms to rapidly sample loop conformations using a coarse-grained C_α model [33] and to accurately reconstruct proteins backbones [34] as part of an approach that often gave sub-Å RMSD loop predictions [35]. In this paper, we have applied these techniques to *de novo* backbone design without using fragments from known protein structures while also explicitly considering alternative conformational states. We were able to solve the structures of four loop design proteins using X-ray crystallography and show that one of these structures matched the design at atomic level accuracy.

Results

Design of synthetic RFR-fold proteins of variable length. Residue frequency tables were derived from known RFR-fold proteins for each of the five positions in the repeat giving the consensus sequence ADLSG (Fig. 1C and SI Appendix, Table S2). A 120 residue stochastic repeat sequence (24 repeats or six superhelical turns) was drawn from the frequency tables and combined with N- and C-terminal capping sequences to protect the hydrophobic core from solvent exposure. The C-terminal cap also incorporated a dimer interface from the parent protein as a first step towards lattice and multimer design. The initial synthetic protein was named SynRFR24.1. Single turns were removed or added to create variant proteins of different lengths, SynRFR20.1 and SynRFR28.1. All three proteins were easily expressed and purified using standard techniques, and were found to crystallize in a variety of different crystal forms (Table 1 and SI Appendix, Table S1). A fourth variant protein, SynRFR24.2, was constructed with two amino acid changes (D196S and R198H). This crystallized in a new crystal form, not observed for the SynRFR24.1 protein, probably because the large arginine 198 side-chain blocked a crystal contact. All SynRFR proteins formed dimers in the crystal lattice (Fig. 1D) and in solution (SI Appendix, Fig. S2). Thermal stability measurements of these variable length proteins using a thermofluor assay showed melting temperatures of between 65 to 73 °C that did not appear to be correlated with repeat length (SI Appendix, Fig. S5).

Computational design of de novo loops. Given the inherent modularity and ease of expression, we decided to test whether these proteins could serve as an extended scaffold base for the design of *de novo* backbone embellishments as a step towards functionalization. Taking the 1.8 Å resolution SynRFR24.1 crystal structure (PDB: 4YC5) as the base scaffold, an eight residue insertion was created approximately midway along the stochastic repeat region of the protein (between residues 108-109). This loop length was chosen on the basis of the accuracy of previous loop structure prediction results [35]. 4000 backbone sequence-independent loop conformations were sampled using the PD2_loop_model software with no externally imposed restraints on secondary structure or any other feature (Fig. 2A). Briefly, the method samples plausible backbone loop conformations from a sequence-independent coarse-grained C_α potential energy function and then reconstructs other backbone atoms using a structural alphabet-based algorithm [34]. The C_α potential energy function includes pseudo-bond length, bond angle, and dihedral terms to ensure good local structure together with soft steric repulsive and pseudo-hydrogen bonding terms. Loop conformations were sampled by successive simulated annealing Monte Carlo runs followed by full backbone reconstruction. Previously, this method was successfully applied to loop prediction giving results that were comparable to fragment replacement-based methods despite the sequence-independence of the initial backbone conformational sampling [35]. Coarse-grained loop sampling was followed by sequence design using Rosetta [36] on each of the conformations to generate full-atom models.

Selection of designed loops. A significant proportion of the 4000 conformations were likely not designable so we developed an approach that explicitly considered alternative low energy conformational states in order to filter out bad designs. Each of the 4000 designed sequences was threaded onto each of the 4000 loop conformations then gradient minimized in the Rosetta force-field with the resulting energy and RMSD to the designed structure recorded (Fig. 2 B and C). With the assumption that we have sampled the important low energy states, we filtered the designs based on the probability that a design is in a folded state, $P_i > 0.9$ (equation [**1**]; Fig. 2D), calculated using the Boltzmann distribution, and other criteria (see Methods). The criterion that $P_i > 0.9$ removed 97.9 % of designs by itself.

Crystal structures of designed loop proteins. Of the ten loop extension designs selected for experimental characterization, five could be expressed and purified, and crystal structures were obtained for four (Table 1). Of these structures, SynRFR.t1428 was solved at 3.56 Å resolution and showed clear unbiased electron density that unambiguously matched the designed loop embellishment after molecular replacement using a model with the loop region excised (SI Appendix, Fig. S1). After refinement, the hairpin loop region residues (108 to 117) very closely matched the design with an all-atom RMSD value of 0.71 Å for the best chain (Fig. 3 A and B). The loop region forms a crystal contact with the non-crystallographic symmetry copy of itself leading to a higher-order assembly in the crystal lattice (Fig. 3C) but in solution, the protein was dimeric (SI Appendix, Fig. S3). The hairpin loop structure of SynRFR.t1428 forms a type I beta-turn with a proline, a tryptophan and a tyrosine forming a mini-hydrophobic core. A similar tyrosine and tryptophan stacking motif, albeit in different relative positions, can be seen in a designed beta-sheet protein with type I' beta-turns, Betanova, which was found to fold cooperatively in aqueous solution despite having no real hydrophobic core [37]. A previous study also engineered an extended beta-hairpin on an SH3 domain using sequences from a model peptide system in order to determine its effect on folding [38].

Of the other loop designs, SynRFR.t1555, solved at 4.4 Å resolution, showed electron density consistent with the designed loop conformation but the resolution was too low to be conclusive (SI Appendix, Fig. S1). SynRFR.t801 had electron density over the entire loop that was clearly different to the design and had the same type III crystal form as one of the SynRFR24.1 structures (Table 1, Fig. 4). The density for the SynRFR.t3284 loop was not resolvable beyond the first few residues but had the same type IV crystal form observed for SynRFR24.2. Thermal stability assays of the loop variant proteins showed slighly lower melting temperatures compared

the length variant proteins ranging from 54 to 60 °C (SI Appendix, Fig. S6).

Loop energy landscape. In order to further characterize the loop energy landscape and understand our results, an extra 16,000 loop conformations were sampled with weak harmonic C_α coordinate restraints to the solved crystal structures using the PD2_loop_model software followed by gradient energy minimization using Rosetta for each of the loops with electron density in the loop region. These extra samples are shown as green points in Fig. 2C. The structure of SynRFR24.t801 was found to be in a completely different conformation to the designed structure, however a new potential energy minimum near to the experimentally solved structure was not observed. Although the general path of the SynRFR24.t801 loop backbone could be traced in the electron density, it was not well resolved so the restrained resampling procedure may not have sampled the correct region of conformational space. Another potential source of error is that the energy minimization protocol did not permit bond angle or bond length flexibility which could be important for the accurate modelling of the energy landscape [39]. Alternatively, this may indicate that the potential energy function can be further improved. The energy landscape for SynRFR24.t1428 supports a minimum around the designed structure but SynRFR24.t1555 appears to have a broad minimum 1-2 Å RMSD from the designed structure. This indicates conformational flexibility for this design and may explain why the loop was not well-ordered in the crystal structure.

Crystal lattice packing. Several of the synRFR proteins crystallized in the same crystal form. For example the SynRFR.t801 crystal form is the same as the type III $P3_221$ SynRFR24.1 structure, but the loop projects into the solvent voids (Fig. 4 A and B). The SynRFR24.2 I222 (form IV) structure has a packing motif of a bundle of three dimers with D3 symmetry, which is also found in the SynRFR.t3284 structure, in which the loops project into the solvent voids (Fig. 4 C and D). The C-terminal dimer axis exhibits flexibility, with the angle between the solenoid axes varying between 157° and 168° in the different structures, and there is also variation within a single crystal form. Such flexibility may assist in assembling future nanostructures. The large surface area and wide allowed variability within the RFR consensus repeat of the SynRFR solenoid should allow for fine control of lattice contact points, enabling precise lattice and multimer design in future constructs.

Whole domain insertions into the solenoid scaffold. To test whether the extended beta-solenoid structure can be decorated with whole domain embellishments, two variants with superfolder green fluorescent protein (sfGFP) domain insertions were created. Taking the SynRFR24.t1428 protein as the template, a sfGFP domain was inserted between the loop residues P112 and W113 with additional glycine/serine linkers to connect to the termini of the sfGFP domain. A second variant was simultaneously created with a W113A mutation in case the large hydrophobic tryptophan caused unwanted interactions. Both proteins were found to be well-expressed, soluble and fluorescent. The proteins were found to be dimeric in solution (SI Appendix, Fig. S4), suggesting that the C-terminal cap of the solenoid was still folded and able to form a dimer interface. It is unlikely that dimerization is mediated by the sfGFP domain as this is monomeric [40]. These data suggest that the solenoid is continuous, and has accomodated the large domain insertion.

Discussion

We have described the design and construction of a series of variable-length synthetic beta-solenoid RFR-fold proteins, which are capable of hosting computationally designed loops that decorate the structure. Initial results suggest entire protein domains can also be inserted. To our knowledge, these are the only artificial versions of this class of protein to have been created to date. The synthetic protein scaffolds crystallize in a variety of crystal forms, some identical between different protein designs. The regular extensible linear structure and DNA-like dimensions make synRFR proteins potential building blocks in the emerging field of protein origami as well as for co-assembling DNA-protein nanomaterials [41]. Proteins have several advantages over DNA in that they have many more functional groups for derivatization, are chemically richer and are capable of self-assembling *in vivo* without complex annealing protocols. The variety of crystal forms is also a first step towards crystal lattice design, enabling the construction of functional zeolite-like porous bioreactive materials. Multi-component designs of solenoids with ends capable of forming different multimers could also be used to construct closed cages [42] or extended complex lattices.

Here, we have been able to computationally design an automatically generated free-form *de novo* backbone embellishment on a *de novo* repeat scaffold without using backbone fragments from known protein structures. The ability to sample plausible and designable backbone conformations directly from a coarse-grained potential energy function rather than using fragment insertion permits the incorporation of functional geometric constraints and the use of sophisticated sampling techniques during the design process. In this work, we have developed a method to select promising loop designs by using the alternative sampled backbone conformations as decoys and the Boltzmann distribution to rank the designs. It is probable that very few short single loop projections into solvent from the solenoid core are designable and able to fold into well-defined rigid structures due to the lack of opportunity to form a well-packed core. Multiple surface loop projections are more likely to form stable well-defined structures and could be iteratively designed using successful single loop designs as starting points.

We have shown that the beta-solenoid scaffold may be capable of hosting whole domain insertions within the repeat units by inserting a sfGFP domain into the loop of SynRFR24.t1428. This could prove useful by providing a rigid scaffold as a basis for large artificial multi-enzyme complexes [43].

These advances provide a solid basis for the design of functionalised extensions, of single and multiple loops, to be incorporated into new crystal lattices and oligomers. The ability of the SynRFR proteins to act as stable platforms for variable loops may also prove useful for molecular recognition applications [15]. In the future we can envisage more complex multiple loop decorations, including co-factor binding sites, enzyme active sites, and complete protein domains.

Materials and Methods

Design of beta-solenoid repeats. Residue frequency tables were derived from known Repeat Five Residue (RFR) proteins then manually edited to remove cysteine and proline residues, and to ensure alanine at position 1 and leucine at position 3. These were found to have a consensus repeat sequence of ADLSG. A stochastic repeat sequence was created by drawing residues from the residue frequency table. The N-terminal cap, including a cleavable hexahistidine tag, (sequence: MGSSHHHHHH SSGLVPRGSHMNVGEILRHYAAGKRNFQHINLQEIELTNASLTGADLSY) was taken from the HetL protein from Nostoc sp. Strain PCC7120 (PDB: 3DU1) and the C-terminal cap (sequence: ADLSGARTTGARLDDADLRGATVDPVLWRTASLVGARV

DVDQAVAFAAAHGLCLAGGSGC) was taken from the MfpA protein from $M.\ tuberculosis$ (PDB: 2BM4). The C-terminal cap forms a homodimeric interface in all the crystal structures. A one-turn solenoid extension variant, synRFR28.1, was created by modelling an extra 20 residue turn into the solenoid structure after residue 139. The model was created by superimposing the highest resolution SynRFR24.1 crystal structure (PDB: 4YC5) onto itself with a one-turn shift. Two halves from each structure were then recombined to create the final extended solenoid model. Sequences for the one turn insertion were designed using RosettaDesign permitting only residues that appear in the residue frequency table. A one-turn solenoid deletion, synRFR20.1, was created by deleting 20 residues from SynRFR24.1 (Δ120-139).

Computational loop design. Using the highest resolution type I SynRFR24.1 crystal structure as the base scaffold (PDB: 4YC5), an 8 residue insert was created between residues 108 and 109 at a "corner" of the square solenoid repeat. 4000 backbone loop conformations were sampled using algorithms we have previously developed and implemented in the PD2 software package [34, 35]. Using the Rosetta3 software package [36], a sequence was designed for each of the loop structures with a protocol that cycles through rounds of sequence design and gradient energy minimisation (FlxbbDesign). The amino acid identites of residues immediately adjacent to the loop were also allowed to vary in addition to the loop region itself. Each of the 4000 sequences was threaded onto all 4000 structures and gradient energy minimised using the FastRelax protocol. For both the design and relaxation protocols, the Talaris2013 scoring function was used. Good sequence designs were expected to have the lowest potential energies close to the desired loop conformation. Assuming the loops follow the Boltzmann distribution, the sequence threading calculations permitted the ranking of each design by explicitly considering alternative low energy conformational states using equation **[1]**.

$$P_i = \frac{\sum\limits_{j \in F} e^{-\frac{E_i(j)}{k_B T}}}{\sum\limits_{j=1}^{N} e^{-\frac{E_i(j)}{k_B T}}} \qquad \textbf{[1]}$$

where P_i was the probability of the designed loop, i, being in the desired folded conformation, $E_i(j)$ was the gradient minimised energy of sequence i on structure j, F was the set of correctly folded loops (defined as less than 1 Å RMSD from the designed structure), N was 4000 (i.e. all sampled conformations). A list of 10 designs for experimental characterisation was selected by picking structures with $P_i > 0.9$, the lowest folded loop energy less than the mean lowest folded loop energy (< -324.8) Rosetta Energy Units (REU), the energy gap between the lowest energy structure < 1 Å RMSD and the lowest energy structure > 1 Å RMSD being < -2 REU, RosettaHoles score < 2.3, and with no residues in forbidden regions of the Ramachandran plot.

Cloning, expression and purification. The SynRFR24.1 gene sequence was codon optimised, synthesised and cloned into the pET11a expression plasmid by GeneArt. Turns were added and deleted using PCR followed by recircularisation using Gibson assembly or restriction enzyme digestion and ligation. Variable loop regions were supplied as linear DNA gBlocks from IDT, the original SynRFR24.1 plasmid (including the non-variable parts of the coding sequence) was linearised by PCR and the final construct formed using In-Fusion HD (Clontech Laboratories, Inc.). 100 μg/ml ampicillin was used for selection in all media. All ligation and assembly reactions were transformed into the $Escherichia\ coli$ strain NEB10β (New England Biolabs) and grown overnight on LB Agar medium. Colonies were picked and grown overnight in 5 ml Lysogeny Broth (LB) medium, plasmid miniprepped (QIAprep Spin Miniprep Kit, Qiagen) and sequence verified (Eurofins Genomics) using the standard T7 and T7 terminator primers. Verified plasmids were transformed into chemically competent BL21-Gold DE3 (Agilent Technologies) or KRX cells (Promega). For each SynRFR variant, 1 l LB or Terrific Broth (TB) medium was inoculated with 1 ml from 5 ml overnight cultures. The cultures were grown until an OD600 reading of 0.6 whereupon expression was induced with 1 mM IPTG or 0.1% rhamnose for KRX cells. After 4 hours of induction, the cells were harvested and resuspended in lysis buffer (100 mM bicine and 150 mM NaCl buffer titrated to pH 9.0 with NaOH) with EDTA-free SIGMAFAST protease inhibitor cocktail tablets (Sigma). The cells were sonicated and clarified by spinning at 40,000 RCF for 40 minutes. The proteins were purified with a Ni-NTA column, washed with 100 mM bicine, 150 mM NaCl, 25mM imidazole at pH 9.0 and eluted in 100 mM bicine, 150 mM NaCl, 250mM imidazole at pH 9.0. The proteins were further purified by gel filtration using Superdex75 HiLoad 16/60 (GE Healthcare) or Superdex200 HiLoad 16/60 (GE Healthcare) columns. Proteins were concentrated in bicine buffer using 10 kDa cutoff centrifugal concentrators (Millipore).

X-ray crystallography. The proteins were concentrated to ∼10 mg/ml and used to set up vapour diffusion sparse-matrix crystallization trials with a TTP mosquito robot. Crystals were optimised in manually set up trays where necessary. Crystals were cryoprotected in the mother liquor and 30% volume added of glycerol or PEG400, were flash-cooled in liquid nitrogen and stored for data collection. Diffraction data were collected at Diamond Light Source synchrotron with the exception of SynRFR24.2 which was collected with an in-house rotating-anode source.

ACKNOWLEDGMENTS. The authors would like to thank Diamond Light Source for beamtime (proposals mx7299, mx1227, mx9424). BBSRC through a Eurocores (funding JTM) grant number: BB/J010294/1. EPSRC through the Syntegron project (funding JTM) grant numer: EP/K034359/1. JWM was funded by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/F023308/1). The Imperial College High Performance Computing cluster was used for computational protein design. Ciaran McKeown and Kirsten Jensen are thanked for help and support.

1. Aharoni A et al. (2005) The 'evolvability' of promiscuous protein functions. Nature Genetics 37(1):73–76.
2. Toscano M, Woycechowsky K, Hilvert D (2007) Minimalist Active-Site Redesign: Teaching Old Enzymes New Tricks. Angewandte Chemie International Edition 46(18):3212–3236.
3. Dutton PL, Moser CC (2011) Engineering enzymes. Faraday Discussions 148:443–448.
4. Lin YR et al. (2015) Control over overall shape and size in de novo designed proteins. Proc Natl Acad Sci U S A pp. E5478–E5485.
5. Regan L, DeGrado WF (1988) Characterization of a helical protein designed from first principles. Science 241(4868):976–978.
6. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262(5140):1680–1685.
7. Woolfson DN (2005) The design of coiled-coil structures and assemblies. Advances in Protein Chemistry 70(04):79–112.
8. Kuhlman B et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302(5649):1364–8.
9. Koga N et al. (2012) Principles for designing ideal protein structures. Nature 491(7423):222–227.
10. Thomson AR et al. (2014) Computational design of water-soluble alpha-helical barrels. Science 346(6208):485–488.
11. Javadi Y, Itzhaki LS (2013) Tandem-repeat proteins: Regularity plus modularity equals design-ability. Current Opinion in Structural Biology 23(4):622–631.
12. Lee G et al. (2006) Nanospring behaviour of ankyrin repeats. Nature 440(7081):246–249.
13. Tsai Cj et al. (2007) Principles of nanostructure design with protein building blocks. Proteins 68(1):1–12.
14. Phillips JJ, Millership C, Main ERG (2012) Fibrous nanostructures from the self-assembly of designed repeat protein modules. Angewandte Chemie - International Edition 51:13132–13135.
15. Binz HK, Amstutz P, Plückthun A (2005) Engineering novel binding proteins from nonimmunoglobulin domains. Nature Biotechnology 23(10):1257–1268.
16. Karanicolas J et al. (2011) A de novo protein binding pair by computational design and directed evolution. Molecular Cell 42(2):250–60.
17. Boersma YL, Plückthun A (2011) DARPins and other repeat protein scaffolds: Advances in engineering and applications. Current Opinion in Biotechnology 22(6):849–857.
18. Parmeggiani F et al. (2014) A general computational approach for repeat protein design. Journal of Molecular Biology 427(2):563–575.
19. Park K et al. (2015) Control of repeat-protein curvature by computational protein design. Nature Structural & Molecular Biology 22(2):167–174.
20. Kajava AV, Steven AC (2006) Beta-rolls, beta-helices, and other beta-solenoid proteins. Advances in Protein Chemistry 73:55–96.
21. Bateman A, Murzin AG, Teichmann SA (1998) Structure and distribution of pentapeptide repeats in bacteria. Protein Science 7(6):1477–1480.
22. Hegde SS et al. (2005) A Fluoroquinolone Resistance Protein from Mycobacterium tuberculosis That Mimics DNA. Science 308(5727):1480–1483.
23. Bolon DN, Mayo SL (2001) Enzyme-like proteins by computational design. Proc Natl Acad Sci U S A 98(25):14274–14279.
24. Rothlisberger D et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453(7192):190–195.
25. Jiang L et al. (2008) De novo computational design of retro-aldol enzymes. Science 319(5868):1387–91.
26. Kaplan J, DeGrado WF (2004) De novo design of catalytic proteins. Proc Natl Acad Sci U S A 101(32):11566–11570.
27. Lichtenstein BR, Cerda JF, Koder RL, Dutton PL (2009) Reversible proton coupled electron transfer in a peptide-incorporated naphthoquinone amino acid. Chemical Communications (2):168–170.

28. Koder RL et al. (2009) Design and engineering of an O(2) transport protein. Nature 458(7236):305–309.
29. Eiben CB et al. (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nature Biotechnology 30(2):190–2.
30. Ollikainen N, Smith Ca, Fraser JS, Kortemme T (2013) Flexible Backbone Sampling Methods to Model and Design Protein Alternative Conformations. Methods in Enzymology 523:61–85.
31. Hu X, Wang H, Ke H, Kuhlman B (2007) High-resolution design of a protein loop. Proc Natl Acad Sci U S A 104(45):17668–17673.
32. Murphy PM, Bolduc JM, Gallaher JL, Stoddard BL, Baker D (2009) Alteration of enzyme specificity by computational loop remodeling and design. Proc Natl Acad Sci U S A.
33. MacDonald JT, Maksimiak K, Sadowski MI, Taylor WR (2010) De novo backbone scaffolds for protein design. Proteins 78(5):1311–1325.
34. Moore BL, Kelley LA, Barber J, Murray JW, MacDonald JT (2013) High quality protein backbone reconstruction from alpha carbons using Gaussian mixture models. Journal of Computational Chemistry 34(22):1881–1889.
35. MacDonald JT, Kelley LA, Freemont PS (2013) Validating a Coarse-Grained Potential Energy Function through Protein Loop Modelling. PLoS ONE 8(6):e65770.
36. Leaver-Fay A et al. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in Enzymology 487(11):545–574.
37. Kortemme T (1998) Design of a 20 – Amino Acid , Three-Stranded beta-Sheet Protein. Science 281(5374):253–256.
38. Viguera AR, Serrano L (2001) Bergerac-SH3: "frustation" induced by stabilizing the folding nucleus. Journal of Molecular Biology 311(2):357–371.
39. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Science 23(1):47–55.
40. Pédelacq JD, Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein. Nature Biotechnology 24(1):79–88.
41. Mou Y, Yu JY, Wannier TM, Guo CL, Mayo SL (2015) Computational design of co-assembling protein–DNA nanowires. Nature 525(7568):230–233.
42. King NP et al. (2014) Accurate design of co-assembling multi-component protein nanomaterials. Nature 510(7503):103–108.
43. Dueber JE et al. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. Nature Biotechnology 27(8):753–9.

Table 1. Summary of crystal structures of SynRFR proteins solved in this study.

Protein	Crystal form	Dimer pairs *	Dimer angle/° †	Bundle motif‡	Resolution/Å	Space group	PDB
SynRFR24.1	I	A:A'	157	NO	1.76	$P4_122$	4YC5
SynRFR24.1	II	A:B' B:A'	166, 166	YES	3.31	H32	4YDT
SynRFR24.1	III	A:A'	164	NO	2.41	$P3_221$	4YCQ
SynRFR24.2	IV	A:A' B:C' C:B'	165, 162, 162	YES	3.55	I222	4YEI
SynRFR28.1	V	A:A'	168	YES	3.39	H32	4YFO
SynRFR28.1	VI	A:B C:D E:F	160, 164, 167	NO	3.33	$P2_12_12_1$	5DZB
SynRFR20.1	VII	A:A'	156	NO	2.99	$P4_32_12$	5DRA
SynRFR24.t1555	VIII	A:A'	161	NO	4.40	$I4_122$	5DN0
SynRFR24.t1428	IX	A:B' B:A'	160, 160	NO	3.56	$P3_221$	5DNS
SynRFR24.t3284	IV	A:A' B:C' C:B'	164, 158, 158	YES	3.27	I222	5DQA
SynRFR24.t801	III	A:A'	168	NO	2.28	$P3_221$	5DI5

*Dimers are shown between chains related by a two-fold axis at the C-termini, a prime indicates a symmetry-related partner.

†The C-terminal dimer axis exhibits flexibility, with the angle between the solenoid axes of the dimer varying between 157° and 168° in the different structures.

‡The "dimer bundle" packing motif is a bundle of 3 dimer pairs with D3 symmetry, seen in Fig. 4 C and D.

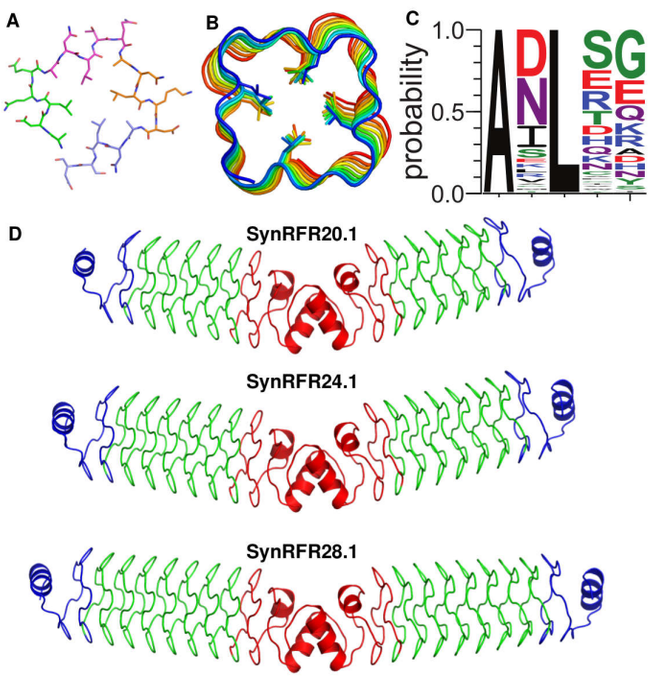

Fig. 1. The repeat five residue (RFR) beta-solenoid proteins. (a) a single superhelical turn composed of four repeats with square cross-sectional profile. Each five residue repeat forms one face of the square and twenty residues forms a helical turn with a ∼5 Å rise. (b) view down the beta-solenoid, showing leucine residues from position 3 in the repeat motif forming the hydrophobic core. (c) Logo plot of residue frequencies used to produce stochastic sequence region of the synRFR24.1 protein at each position in the RFR repeat and (d) the solved crystal structures of the three synthetic length variants SynRFR20.1, SynRFR24.1, SynRFR28.1.

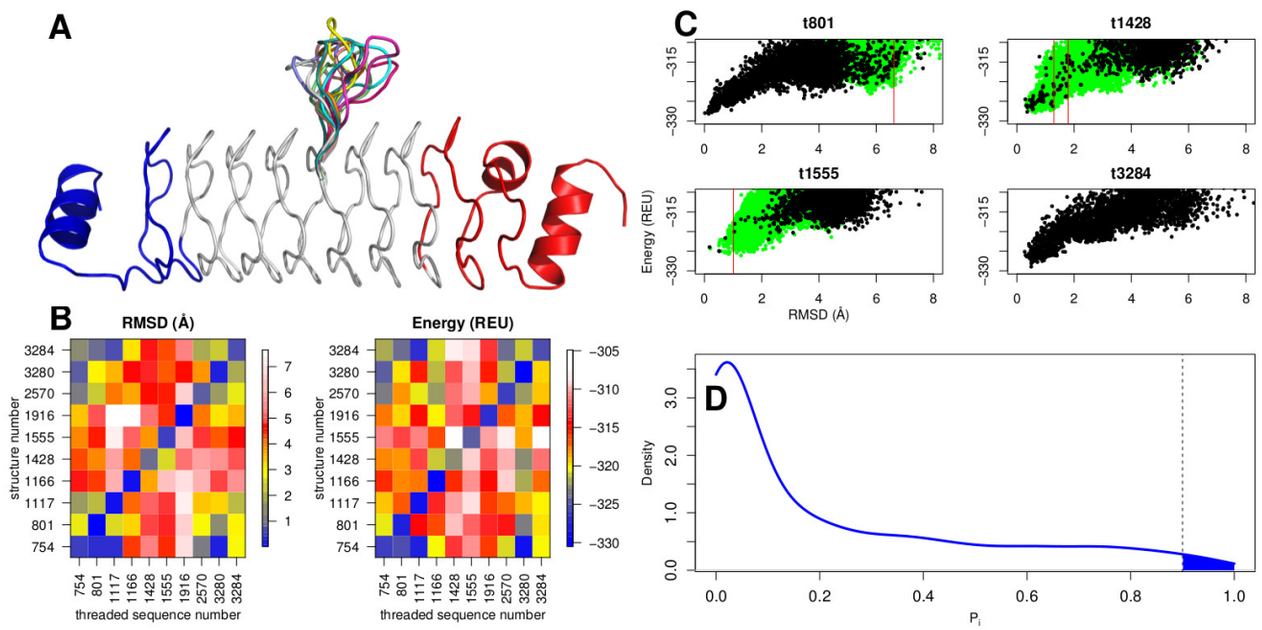

Fig. 2. Automated computational design and selection of *de novo* loops. (a) Superposition of the ten designed loop variants based on the SynRFR24.1 scaffold selected for experimental characterisation. (b) Sequence vs Structure energy/RMSD matrices for the ten selected loops. The designed sequence for each of the 4000 sampled loop structures was threaded onto all 4000 sampled loop structures and energy minimised. The resulting energies and all-atom loop RMSD (to the designed loop structure) values were recorded giving two 4000×4000 matrices (SI Appendix, Fig. S7). Rows and columns corresponding the selected ten structures and sequences are shown here. (c) Loop RMSD vs minimised Rosetta energy plots for the experimentally solved loop structures. The vertical red lines correspond to RMSD values of the solved crystal loop structure to the designed structure. The black points represent the 4000 originally sampled loop conformations after sequence threading and energy minimisation. The green points represent an additional 16000 conformations sampled with additional harmonic restraints to sample the region around the solved crystal structure conformation for each loop. (d) Histogram of P_i values for all 4000 designs. P_i is the probability that sequence, i, is in a folded conformation assuming the loop conformations follow a Boltzmann distribution. The vertical line at $P_i = 0.9$ and blue shaded region under the curve represent the selected designs. All RMSD values in this figure were calculated by superposing the C_α atoms of the non-loop regions of the solenoid scaffold and calculating the all-atom RMSD of the region around the loop compared to the designed structure (residues 105-120) without further superposition.

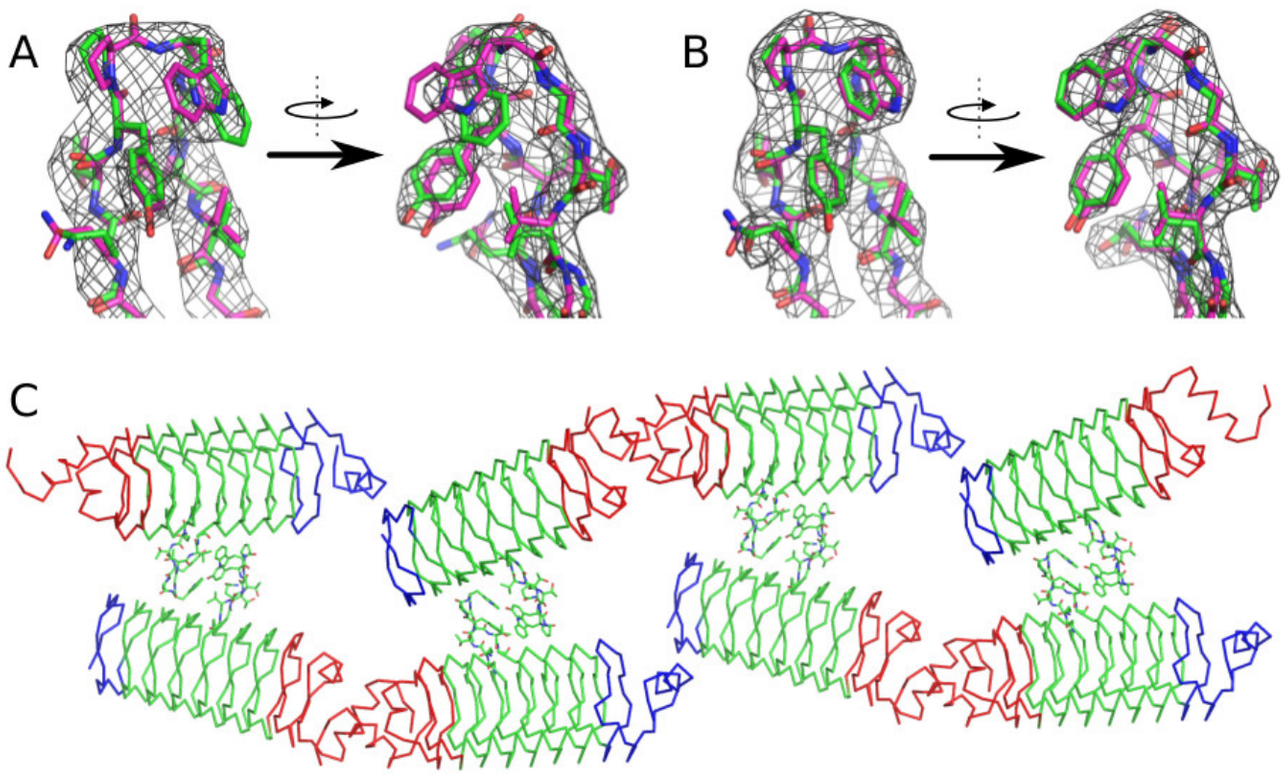

Fig. 3. Crystal structure of SynRFR24.t1428 loop design. The protein was found to crystallise with two chains in the asymmetric unit. The designed loop structure (shown in magenta) is shown superposed on the experimentally solved structures (shown in green) of (a) chain A and (b) chain B together with 2Fo-Fc electron density map contoured at 1σ. The loops very closely matched the model structure with all atom RMSD values of 1.47 Å for chain A and 0.71 Å for chain B after superposition. If the non-loop region C_α atoms of the scaffold were superposed, the all atom RMSDs compared to the design were determined to be 2.15 Å (chain A) and 1.52 Å (chain B) for the loop region residues. Chain A has a flipped tryptophan side-chain compared to the design. (c) The loop embellishment mediates a higher order assembly of SynRFR24.t1428 in the crystal lattice.

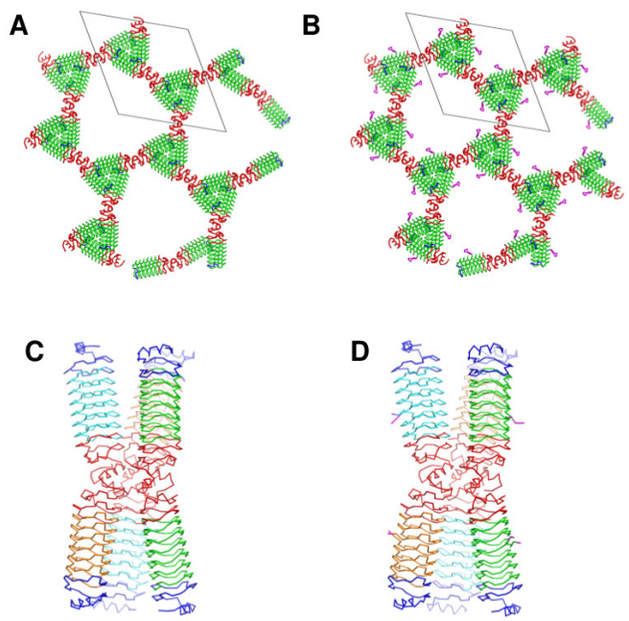

Fig. 4. The lattice of the $P3_221$ synRFR24.1 protein (a), and the synRFR24.t801 protein (b). The dimer bundle packing motif of synRFR24.2 (c), also seen in (d) the structure of synRFR24.t3284.