Insight from designing ideal αβ monomers and homo-oligomers

Yu-Ru Lin

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

Year 2017

Reading Committee:

David Baker, Chair

Wilhelmus G.J. Hol

Ning Zheng

Program Authorized to Offer Degree:

Biochemistry

University of Washington

Abstract

Insight from designing ideal αβ monomers and homo-oligomers

Yu-Ru Lin

Chair of the Supervisory Committee:
Dr. David Baker
Biochemistry

Previously, general principles relating secondary structure patterns to tertiary packing motifs enable design of different protein topologies stabilized by consistent local and non-local interactions. With the goal to achieve fine control over protein shape and size within a particular topology, the first part of my thesis extended the design rules by systematically analyzing the co-dependencies between the lengths and packing geometry of successive secondary structure elements and the backbone torsion angles of the loop linking them. I then demonstrated the fine control by the applying the extended set of rules to design series of proteins with the same fold but considerable variation in secondary structure length, loop geometry, registry between β-strands and overall shape. Solution NMR structures of four designed proteins for two different folds showed that protein shape and size can be precisely controlled within a given protein fold.

These extended design principles can provide the foundation for custom design of protein structures performing desired functions.

The second part of my thesis focused on homo-oligomer design. Proteins in cells often form oligomers through non-covalent interaction to execute various biological properties. Natural oligomeric proteins involve scaffold support, enzymatic reactions, signaling transduction, etc. Moreover, the intrinsic protein flexibility often facilitates protein structural changes upon oligomerization. To understand how αβ- proteins interact to form oligomers and to enable future de novo protein applications in biomaterial or molecular machinery, the second part of my thesis investigated the design of cyclic homo-oligomers using de novo proteins as building blocks and experimented with introducing flexible backbone design strategies to oligomer design process. The results suggested that de novo proteins undergo structural change for oligomerization and for homo-oligomers formed from globular building blocks, pre-organized hydrogen bonds between amines and carbonyl groups of beta strands are more effective in achieving protein interaction and shape complementarity between subunits with higher specificity.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First, I would like to thank the Koga couple, Nobuyasu and Rie Tatsumi-Koga who have played very important roles in my graduate career. They introduced me to de novo protein design and guided me to set a solid foundation. It is an honor to have very knowledgeable de novo protein design pioneers as my mentors.

I would also like to thank all the developers of Rosetta macromolecule modeling package. Without their contribution to improving and perfecting Rosetta, I would not be able to perform protein design in my graduate work. For their computational advice and assistance, I would like to thank Darwin Alonso, Luki Goldschmidt, Patrick Vecchiato, Nobuyasu Koga, Javier Castellanos, Enrique Marcos, TJ Brunette, Fabio Parmeggiani, Tom Linsky, Scott Boyken, Robert A. Langan, Gustav Oberdorfer, Daniel Adriano Silva, Vikram Mulligan, Po-Ssu Huang, Lei Shi, Jason Klima, Brian Koepnick, Neil King, William Sheffler, Jorge Fallas, Yang Hsia, Hahnbeom Park, and Yu-Rei Wang.

For experimental work, I had a lot of help from Rie Tatsumi-Koga, Lauren Carter, Vanessa Nguyen, Alex Young-Seug Kang, Stephen Rettie, Clancy Wolf, Yang Hsia, Fabio Parmeggiani, Christopher D. Bahl, Benjamin Basanta.

Collaboration with Rutgers University and Columbia University of Northeast Structural Genomics Consortium was critical in structure determination of the designed proteins. I would

like to thank Gaetano T. Montelione, Gaohua Liu, Rongjin Guan, Rong Xiao, Gregory Kornhaber and Sergey M. Vorobiev.

I would also like to thank my Doctoral Advisory Committee, Rachel Klevit, Wim Hol, Ning Zheng and Gabriele Varani for making time to attend my committee meeting and provide advice and support. I want to give special thanks to Racel Klevit for providing me the opportunity to do undergraduate research in Klevit lab during my senior year at University of Washington. The experience and training obtained from Ying Liu in Klevit lab intrigues me to scientific research and lays solid foundation for my research skills. Certainly, I want to thank my thesis advisor, David Baker, for the fantastic opportunities and resources he has provided. His guidance has always been clear and yet inspiring. It has been a great honor to have him as my thesis advisor.

Lastly, I would like give thanks to my husband, Lawrence, and my family. I thank them for the love and support throughout the years.

INTRODUCTION

Proteins, biomolecules constructed with one or more amino acid polymers in various lengths, perform vast array of functions in biological world. Functions of proteins include but not limit to molecule transport, biochemical reactions catalysis, cell structure maintenance and signaling. How sequences made from 20 amino acids can achieve definite three-dimensional structures and perform biochemical reactions and interactions exquisitely always amaze and puzzle scientists.

With the molecular biology knowledge from years of research and advancement in techniques and equipment, scientist started to have the ability to rationally tinker proteins. The earliest protein design work can date back to 1975 when Gutte designed analogs of RNase S-protein(1–3). Protein design can be mainly categorized into structure study (how proteins fold) and function study (how proteins perform reactions) and my graduate research focus on the first category. Beginning with the early work of Christian Anfinsen in 1950s, scientists acknowledged that a chain of amino acids could have one or sometimes multiple definite low-energy three-dimensional structures and the conformational transition depends on the energy barrier between conformations and temperature applied(4,5). With the awareness of amphiphilic characteristic of α-helix, Kaiser and colleagues successfully built helical peptides(6). Later on, Richardson co-workers designed a β protein with no known sequence homologous protein(7). In order to design proteins in a robust way, with the help of computational resources, nowadays computer algorithms have been developed to calculate and predict the lowest energy structure of an amino acid sequence based on basic physical principles and Rosetta(8) developed in Baker Lab is one main software.

Previous success of building de novo ideal αβ protein topologies, including a novel topology(9,10) has shown the ability of Rosetta to predict and optimize protein backbone and sidechain interactions to achieve stable tertiary structure. Here, to have a robust stepwise protocol allowing precise control over shape of tertiary structure that can enable design of protein structures with desired functions, we focused on the problem of controlling proteins with structural variations within the same fold. One main determinant of protein structure variation is the loop connecting secondary structures, hence, we started by studying the common geometries of secondary structure connecting loops observed in nature. These commonly observed loop types then became the basic loop geometries we utilized throughout our designs. Another factor determining protein structures is the length of secondary structures which directly affects the size of a protein. With secondary structures interact to form tertiary structures, finding the optimal secondary structure lengths is crucial to achieve the lowest energy structures of proteins. By performing multiple Rosetta simulations, we found the dependency between secondary structures and how different connecting loop geometries affects secondary structure length dependency and packing geometries. We then used the results concluded from simulations to design a variety of proteins with ferredoxin-like fold and rossmann fold having different sizes and shapes. The close match between NMR structures and design models confirmed the principles we established allowed us to have precise control over protein size and shape with a robust stepwise protocol. Details of this work would be included in Section 1 and 2 below.

With the ability to control protein shape and size, we set out to explore the possible future application for de novo αβ proteins. One aspect would be exploring the assembly of protein

complexes since many biochemical reactions are performed by oligomeric protein molecules(11–16). In addition, even though there are many examples of computational designed oligomers, there has not been successful example of computationally design oligomers with de novo αβ proteins. We started with de novo αβ proteins characterized as building blocks for cyclic homo-oligomer design. In the first round of oligomer design, we used fixed-backbone design method where the backbone of subunit stayed as original while sampling for oligomer docking conformations. Experimental results from successful C2 design and C3 and C4 oligomer polydispersity plus comparisons done with naturally-occurring cyclic homo-oligomers led us to the modification of design method in the second round of design where we introduced building block backbone flexibility. However, although we could achieve interface shape complementarity comparable to that observed in naturally-occurring oligomers, we still could not have successful monodisperse oligomers with symmetry larger than 2-fold. The success in C2 homo-oligomers designed with both methods indicated the robustness of interaction facilitated by pre-organized backbone hydrogen bonds between β strands. In all, the challenges of building cyclic homo-oligomers with de novo αβ proteins remained in finding the balance between protein-protein interactions and building block stability. This work is described in Section 3 and 4 below.

SECTION 1. CONTROL OVER OVERALL SHAPE AND SIZE IN DE NOVO DESIGNED PROTEINS

ABSTRACT

General principles for designing ideal protein structures stabilized by completely consistent local and non-local interactions were previously described. The principles relate secondary structure patterns to tertiary packing motifs and enable design of different protein topologies. To achieve fine control over protein shape and size within a particular topology, we have extended the design rules by systematically analyzing the co-dependencies between the lengths and packing geometry of successive secondary structure elements and the backbone torsion angles of the loop linking them. We demonstrate the control afforded by the resulting extended set of rules by designing series of proteins with the same fold but considerable variation in secondary structure length, loop geometry, registry between β-strands and overall shape. Solution NMR structures of four designed proteins for two different folds show that protein shape and size can be precisely controlled within a given protein fold. These extended design principles provide the foundation for custom design of protein structures performing desired functions.

INTRODUCTION

Protein design holds promise for applications ranging from therapeutics to biomaterials, with recent progress in designing small molecule binding proteins(17,18), inhibitors of protein-protein interactions(19,20), and self-assembling nanomaterials(21–23) . Most of these efforts have repurposed naturally occurring scaffolds which are likely not optimal starting points for creating new functions since they generally contain sequence and structural idiosyncrasies that arose during evolutionary optimization for their natural functions(24). Robust design of new functional

proteins would be considerably enabled by the capability of precisely designing from scratch arbitrary protein structures.

Previously described general principles allowed the de novo design of ideal protein structures with five different folds(9). Here we focus on the "variations on a theme" problem of precisely controlling structural variation within the same fold. To achieve such control, we begin by characterizing the coupling between loop backbone geometry and the packing of the flanking secondary elements. We then use the resulting extended set of design principles to systematically vary structure for two different folds, and describe the experimental characterization of five of these de novo designed proteins.

RESULTS

Local structure building blocks

The design rules described previously relate the packing orientation of ββ-, βα- and αβ- units to the length of the loop connecting them(9). Here we begin by extending these rules to the level of particular loop types. This allows more detailed control over local geometry as well as overall protein topology.

It is convenient to describe protein local geometry using the ABEGO(25) alphabet illustrated in **Fig. 1.1A**. "A" indicates the alpha region of the Ramachandran plot(26), "B", the beta region, "G" and "E", the positive phi region, and "O", the cis peptide conformation. We color code the different ABEGO regions as shown in **Fig. 1.1A** throughout the section. For what follows, it is instructive to consider the change in chain orientation brought about by each of the 16 dipeptide combinations of the A, B, G and E backbone conformations (**Fig. 1.1B** and **1.1C**). These 16 two-

residue units can be viewed as "lego blocks" for assembling secondary structures in different orientations. For example, the AA block induces a 50° change in orientation of the polypeptide chain; the BB block, a 170° change; the BA block, a 140° change, and the EA block, a 30° change. Two-residue loops can be described by a single block, three-residue and longer loops by multiple blocks in series. In the following sections, we describe how these blocks determine the packing geometry of the flanking secondary structure elements.

ββ-connections

β-hairpins--two paired β-strands connected by a loop--have either R or L chirality (**Fig. 1.2A**). If the cross product of a vector pointing in the direction of the first strand and a vector from the first strand to the 2nd strand is parallel to the Cα-Cβ vector of the strand residue preceding or following the loop the chirality is R, otherwise it is L. **Fig. 1.2B** shows that in native protein structures (see **SECTION 2.**) two-residue loops always have L-chirality, and that the GG block is particularly common. As is evident in the schematic in **Fig. 1.1C**, the GG block is compatible with the twist of adjacent β-strands. The also observed EA and AA blocks similarly induce a twist in the ingoing and outgoing strands. Examples of L-hairpins with GG and EA loops are shown in **Fig. 1.2C** and **1.2D**. For five-residue loops, the R-chirality is preferred over the L-chirality. The most common five-residue loop, BAAGB, is shown in **Fig. 1.2E**.

In the standard β-turn type nomenclature(27), the AA and GG loops are the mirror-image turn types I and I' respectively, and the less common BG and EA loops are the turn types II and II'. We use the more general ABEGO torsion nomenclature to facilitate parallel analyses of loops connecting different secondary structure elements ($\beta\beta$, $\beta\alpha$, *and* $\alpha\beta$) and having different lengths.

βα-connections

The packing geometry of βα- and αβ- units can be described based on the orientation of the Cα-Cβ vector of the strand residue closest to the helix relative to the vector from the first secondary structure element to the second—if the vectors are parallel, the orientation is "Para", and if the two are antiparallel, it is "Anti" (see schematics in **Fig. 1.2F** and **1.2K**).

In βα-units, the Para orientation is favored for two-residue loops and the Anti orientation for three-residue loops(9). **Fig. 1.2G** shows the dependence of the orientation on the specific loop type in native structures. For two-residue loops, the Para orientation is almost always achieved with AB loop geometry, and for three-residue loops, the Anti orientation is achieved most often with BAB loop geometry. As illustrated in **Fig. 1.2H**, in βα-units with a AB loop, the consecutive B-residues in the β-strand follow a relatively straight trajectory, and then the A residue produces a direction change and together with the following B residues (See the AB block in **Fig. 1.1C**) produces a tight turn in backbone direction. The three-residue loop preferences inherit from the two-residue loop preferences: extending the strand by inserting one B residue before an AB loop to make a BAB loop flips the pleat at the end of the strand, switching the orientation from Para to Anti (**Fig. 1.2I**). The other common three-residue loop connecting a β- strand with a following helix is GBB, which leads to an Anti packing orientation with the G residue together with the preceding B-residue in the β-strand producing the change in chain direction (see the BG block in **Fig. 1.1C**). Although the A and G residues both change the direction of the polypeptide chain, because of the opposite sign of the φ angle, the change is in the opposite direction (compare the BA and BG images in **Fig. 1.1C**).

αβ-connections

For $\alpha\beta$- units, the preferred packing orientation is Para(9). As shown in **Fig. 1.2L**, the Para orientation is achieved by GB loop geometry, and the longer loops generated by inserting A residues at the beginning or B residues at the end (corresponding to changing the definition of the helix end and strand start) have the expected inherited orientation (AGB is "Para", GBB is "Anti", etc). The Para orientation is also achieved by the unrelated BA, GBA, and BAAB loops.

For tertiary structure design we select the most frequently observed loop geometries that favor interaction between the flanking secondary structure elements. For $\beta\beta$-connections, we selected the GG and EA loops for the L-chirality. For $\beta\alpha$-connections, we selected the AB loop for the Para orientation and the BAB and GBB loops for the Anti orientation. Although the BBB loop is also commonly observed, the loop geometry prevents close interaction between the flanking strand and helix (**Fig. 2.1**). For $\alpha\beta$-connections, we selected the GB, GBA and BAAB loops for the Para orientation. The BA loop is also frequently observed, but the loop geometry does not provide hydrogen-bonded helix capping (**Fig. 2.2**). The amino acid sequences in the loop regions were designed using Rosetta as described below with two exceptions where the local geometry strongly prefers a single amino acid: in GB, GBA and GBB loops, the G was set to glycine, and in BAAB loops, the first A was set to proline (**Fig. 2.3**).

Extended emergent rules

The different loop types have different geometries, and this changes the register of the attached secondary structure elements. The correlations between the lengths of the secondary structure

elements and the flanking loop types were determined through the secondary structure and ABEGO torsion constrained Rosetta folding simulations with a sequence-independent backbone model (9) (See **SECTION 2**.) for βαββ- (**Fig. 2.4**), βαβ- (**Fig. 2.5**) and βαβαβ- (**Fig. 2.6**) units; the most frequently observed helix length for each strand length and loop combination is indicated in **Tables 2.1-2.3**. For each choice of loop types, there is a distinct co-dependence of the secondary structure element lengths. For the βαββ-motif with the BAB loop preceding the helix and the BAAB loop following the helix as shown in **Fig. 1.3A**, the optimal helix length goes from 10 to 22 as the strand length increases (**Fig. 1.3B**). The change in motif size is illustrated in **Fig. 1.3C**. For a βαβ-motif with 5-residue strands and a GB loop connecting the helix to the second strand, the optimal helix length is ~14 if the loop preceding the helix is BAB, but 11 if this loop is GBB (**Fig. 1.3E**); this results from the different curvature of the two types of loops (**Fig. 1.3F** and **1.3G**).

Generation of structures with varying shape and size using extended rule set

The relationships between loop type and secondary structure packing geometry and length described in the previous sections allow the generation of structure diagrams of ideal αβ-proteins with different shapes and sizes. **Fig. 1.4** shows design backbone blueprints for a series of ferredoxin-like fold and Rossmann2x2 fold variants, referred to in the following as Fd and Rsmn2x2 respectively. Structures Fd_7A and the Rsmn2x2_6 were designed in the previous work(9). For the ferredoxin-like fold, strand lengths 5, 7, and 9 were used with or without a β-strand register shift between the 1st and 3rd strands. Suitable loop types for each secondary structure connection were selected based on the packing orientation as described above. Although there are three common αβ-loop types, using the BAAB loop for both αβ-connections

leads to the ferredoxin-like fold more frequently in de novo folding simulations (**Fig. 2.7**). Hence we used the BAAB loop for the *α*β-connections in the new ferredoxin-like fold designs. The helix lengths were then chosen based on the loop types and the strand length using **Tables 2.1, 2.2** and **2.3**; the lengths of the helices in the ferredoxin-like fold series are based only on the β*α*ββ- motif simulations (**Table 2.1**) while those of Rsmn2x2_5 are based on both the β*α*β- and β*α*β*α*β- motif simulations (**Table 2.2** and **2.3**).

For each blueprint, backbone structures were built up by carrying out multiple independent Rosetta folding simulations (See **SECTION 2**.). For each of the generated backbone structures, we designed amino-acid sequences by iterating between searching for the lowest energy combination of sidechain identities and conformations for fixed backbone structure (10) and searching for the lowest energy backbone structure for fixed amino acid sequence (28). Inward-pointing charged residues were introduced in edge β-strands and non-polar residues were disfavored at surface exposed positions to disfavor aggregation (more details on the sequence protocol we used are described in the Methods in (9)). The designed structures were then filtered based on the Rosetta full-atom energy, sidechain packing(29) and the local sequence-structure compatibility(9). For each designed sequence, we then carried out multiple independent Rosetta *ab initio* structure prediction simulations(30) starting from an extended conformation, and selected designed sequences with energy landscapes strongly funneled into the designed target structure for experimental characterization.

For the ferredoxin-like fold, we obtained synthetic genes (Genscript, Inc) encoding 6 designs for Fd_5S, 12 for Fd_5A, 10 for Fd_7S, and 12 for Fd_9A (sequences are provided in **Tables 2.10-**

2.13). All but one design (Fd_7S) are not homologous to any known proteins (Blast E-value < 0.02 against the non-redundant protein sequence database nr). For the Rossmann2x2 fold, 9 designs were selected for Rsmn2x2_5 for experimental characterization, only one of which has weak sequence similarity to a known protein (Blast E-value 0.019; the structures of this and the homologue of Fd_7S are not known). The proteins were expressed, purified and characterized by circular dichroism (CD) spectroscopy, size exclusion chromatography combined with multi-angle light scattering (SEC-MALS), and ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) nuclear magnetic resonance (NMR) spectroscopy.

For the ferredoxin-like fold, 37 of 40 designs (from Fd_5S, Fd_5A, Fd_7S and Fd_9A) are well expressed and highly soluble, although two of the soluble Fd_9A designs tend to aggregate after being stored at 4°C for two days perhaps due to the large hydrophobic core. The far-UV CD spectra show that 23 of the 37 soluble designs have the expected $\alpha\beta$- secondary structure content. In contrast, for the smallest variant—Fd_5S—none of the designs had CD spectra consistent with folded $\alpha\beta$-proteins. 26 of the 37 soluble designs were found to be monomeric by SEC-MALS. Two-dimensional ^{1}H-^{15}N HSQC spectra were measured for a total of 17 designs that were monomeric and had $\alpha\beta$ secondary structure content. Well-dispersed and sharp peaks indicate that these designed proteins fold into rigid tertiary structures, and not molten globule-like structures. The experimental results for the ferredoxin-like fold designs are summarized in **Table 1.1**, along with the designs of Fd_7A reported in the previous work (9).

For the Rossmann2x2 fold, 9 designs were tested for Rsmn2x2_5 (sequences are provided in **Table 2.14**). All the designs were expressed at high levels and all but two designs have high

solubility. 8 designs have the expected CD spectra for αβ-proteins, and of these, 6 designs were found to be monomeric by SEC-MALS. For the monomeric designs with the expected CD spectra, HSQC spectra were measured and 2 designs have well-dispersed and sharp ^{1}H-^{15}N HSQC peaks, suggesting well-packed tertiary structures. The properties of the Rsmn2x2_5 designs are summarized in **Table 1.1**, along with the previously described Rsmn2x2_6 (9).

For each target structure, we selected one design that was monomeric, had the expected secondary structure content, and well-dispersed NMR peaks for further thermodynamic characterization (**Fig. 1.5**). The free energy of unfolding of the ferredoxin-like fold designs ranges from 1.7 kcal/mol to 10.1 kcal/mol, with stability increasing with chain length: the 66 residue Fd_5A_3 design is marginally stable with a ΔG_{unfold} of 1.7 kcal/mol while the 98 residue Fd_9A_11 design has a ΔG_{unfold} of 10.1 kcal/mol. All designs of Fd_5S, which has 58 residues, did not fold; the hydrophobic core in such a structure may be too small to overcome the entropy loss in folding.

The solution NMR structures of the selected designs were determined using triple-resonance NMR with standard data collection and analysis protocols of the Northeast Structural Genomics (NESG) consortium (31) (**Table 2.9**). For Fd_5A_3, Fd_7S_6, and Rsmn2x2_5_6, the structures agree quite closely with the computational models for both the backbone and the core side chains (**Fig. 1.6A-D, G, H**). For Fd_9A_11, the design and NMR structure topologically are quite similar to one other, but the helices of the NMR structure are shifted and are more twisted than those of the design as shown in **Fig. 1.6E, F**.

We further compared the loop geometries at the ABEGO level (**Table 1.2 and 1.3**) in the design models and NMR structures. All but two of the 22 loops in the four NMR structures of the newly designed proteins have ABEGO patterns matching the design models. For L3 of Fd_5A_3 the design is GG, but the NMR structure is BG and for L2 of Fd_7S_6, the design is BAAB, but the NMR structure is BOBB, with a cis proline in the second position.

DISCUSSION

Classic early studies beginning nearly 40 years ago classified the loop types connecting regular secondary structure elements (β-strands and α-helices) observed in the native structures solved at that time (27,32–43). Chou & Fasman categorized β-turns into 11 types based on their backbone torsion angles (27) and Hutchinson & Thornton modified the classification after more protein structures were solved (27). An extensive study of short loops connecting regular secondary structures by Donate *et al.* in 1996 identified common groups of loop geometries connecting different secondary structure elements (33). The analysis of loop types in this section extends and updates this previous work, taking advantage of the much larger number of protein structures which have now been determined.

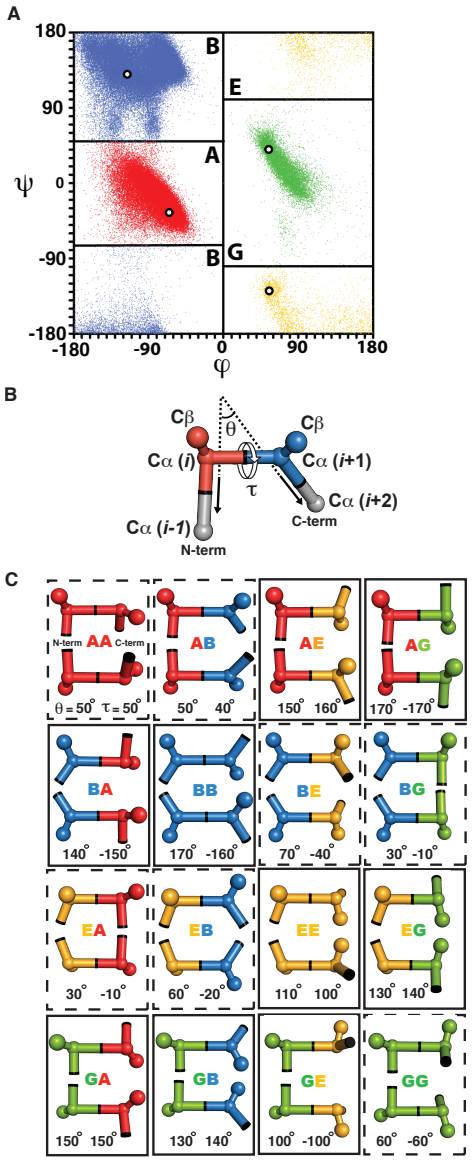
Common loop geometries such as type I, II, I', II' β-hairpins (27, 32, 33,40) and α-helical C-capping are re-identified as expected (33-38,41), and new loop geometries such as the GBB loop in βα-connections are identified. Most importantly, we uncover relationships between loop geometries and the packing orientations of the flanking secondary structures which to our knowledge have not been previously described. The analysis of the dependencies between loop types and secondary structure packing orientations enables the extension of our previous design rule set to more precisely control overall protein size and shape.

The framing of $\alpha\beta$-protein design principles in terms of specific loop types in this section makes possible a systematic building block based approach to designing new structures. The basic algorithm consists of 1) choosing a topology (placement of secondary structure elements with order along the sequence specified), 2) choosing the strand lengths and registers, 3) choosing from the loop types specified by the extended rules, and 4) choosing helix lengths compatible with the strand lengths and loop types. Complete information for steps 3 and 4 are provided in **Tables 2.1, 2.2 and 2.3**.

The very high similarity between the designed structures and the experimental NMR structures demonstrates the capability of this algorithmic approach to systematically and accurately vary protein shape and size. This capability will be invaluable in the creation of the next generation of designed functional proteins with backbones finely tuned to be optimal for their functions.

FIGURES

Figure 1.1

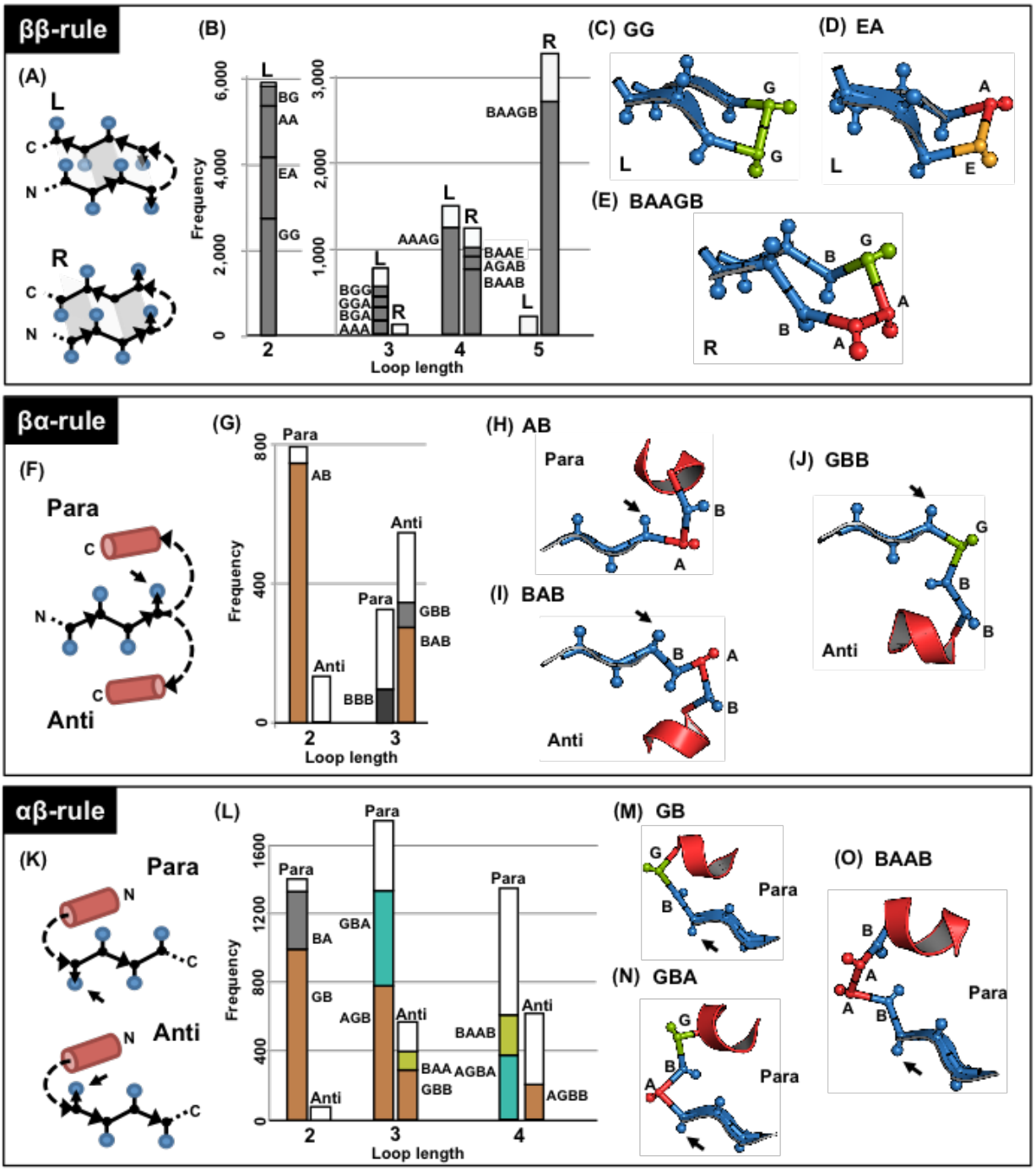


Discrete state model of protein local geometry.

(A) ABEGO representation of protein local structure shown on Ramachandran plot. A: alpha region; B: beta region; G, E: positive phi region. The most frequently observed torsion angles for each region are indicated by the white circle. (B) Two-residue "lego blocks" are represented by four consecutive C*α* atoms connected by virtual bonds. It is useful to consider the net change in

chain direction θ produced by each lego block and the net twist τ. θ is the angle between the vector from Cα(i) to Cα(i-*1*) and the vector from Cα(i+*1*) to Cα(i+*2*), and τ , the dihedral angle defined by Cα(i-*1*), Cα(i), Cα(i+*1*), and Cα(i+*2*). (C) Two views of each of the 16 lego blocks built from the A, B, G, and E geometries indicated by the white circles in (A). θ (left) and τ (right) are indicated at the bottom of the panels. For simplicity, the gray parts in (B) are omitted. While the E residues and most of the G residues are generally Gly, to make the structural feature of the blocks clear, Cβ atoms are shown.

Figure 1.2

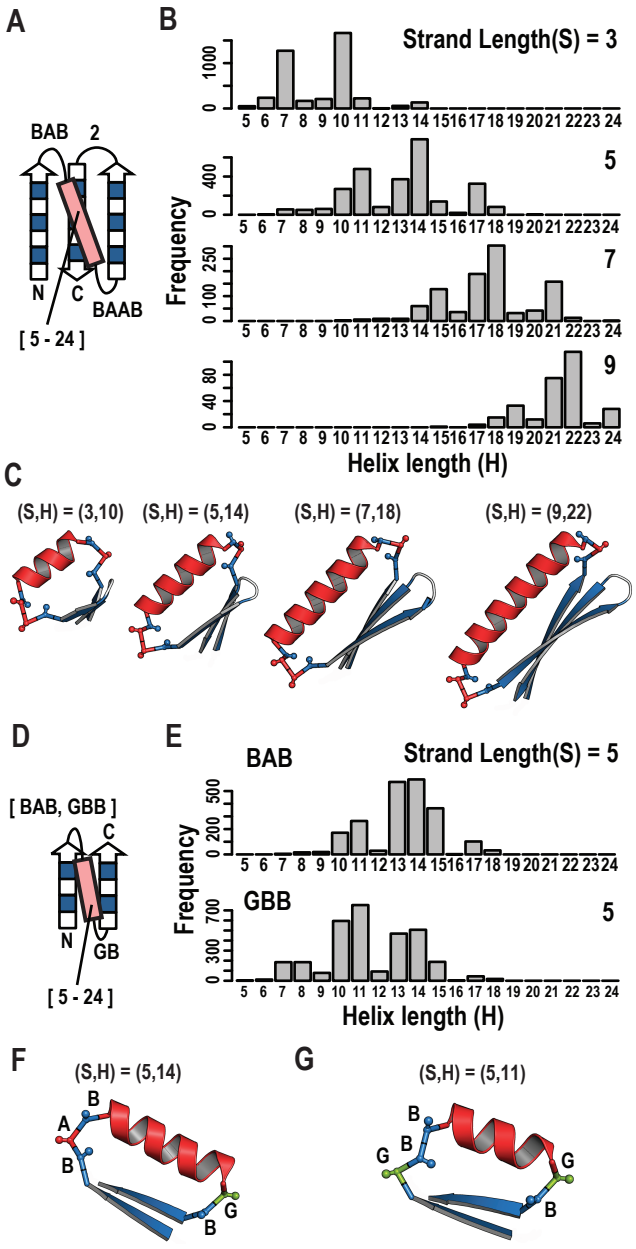


Common loop geometries for ββ-, β*α*-, and *α*β-**units in naturally occurring proteins.**

(A)(F)(K) Secondary structure packing orientation definitions of ββ-, β*α*-, and *α*β-units are illustrated. (B)(G)(L) Loop type distributions in naturally occurring protein structures for ββ-,

βα-, and αβ- units for different loop lengths. The white portions of the histograms indicate other loop types. (C-E, H-J, and M-O) Examples of the most frequently observed loop types. (B) The GG and EA loops are very frequent two-residue L-chirality loops and BAAGB is the only common R-chirality loop. (G) The AB loop is highly preferred for Para orientation. A "B" extension of an AB loop generates the most frequent Anti orientation loop type, BAB; the color coding in the histograms indicates such loop inheritance. GBB also has Anti orientation. (L) The two-residue loop used in designs is GB (see **Fig. 2.2**). Extension of the GB loop generates the AGB, GBB and AGBB loops. GBA is also a common three-residue loop and it gives rise to the four-residue AGBA loop. The less frequent three-residue loop, BAA, extends to the four-residue BAAB loop.

Figure 1.3

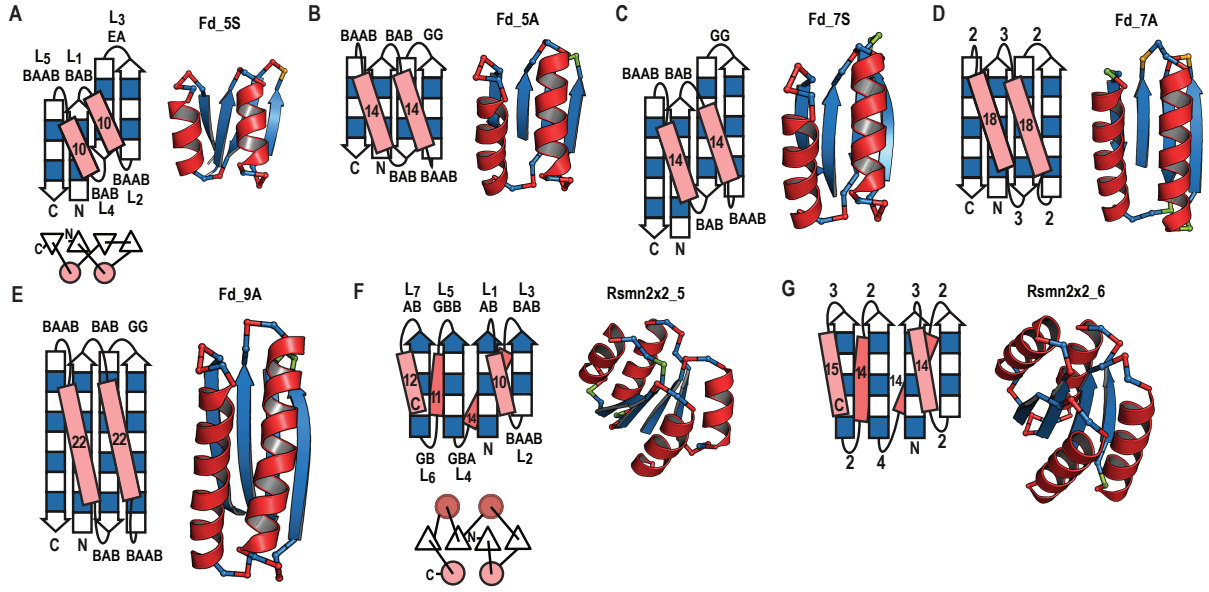


Loop types and strand length determine helix length.

(A) (D) Schematics of the βαββ-units and the βαβ-units found in the ferredoxin-like fold and the Rossmann fold respectively. (B) Helix length depends on the strand length. Multiple sequence-independent simulations of βαββ-unit folding were carried out with fixed loop types and different

strand and helix lengths, and the frequency of successful βαββ-unit folding was assessed. For different strand lengths, optimal folding of the structure occurs for different helix lengths. (C) Examples of four βαββ-units with the same loop types but different strand lengths and the corresponding optimal helix lengths. (E) Helix length depends on βα-loop type. Multiple sequence-independent simulations of βαβ-unit folding were carried out with a fixed αβ-loop type and strand lengths but different βα-loop types, and the frequencies of successful βαβ-unit folding with different helix lengths were determined. Different βα-loop types show different optimal helix lengths. (F) (G) The different geometries of BAB and GBB loops result in different optimal helix lengths.

Figure 1.4

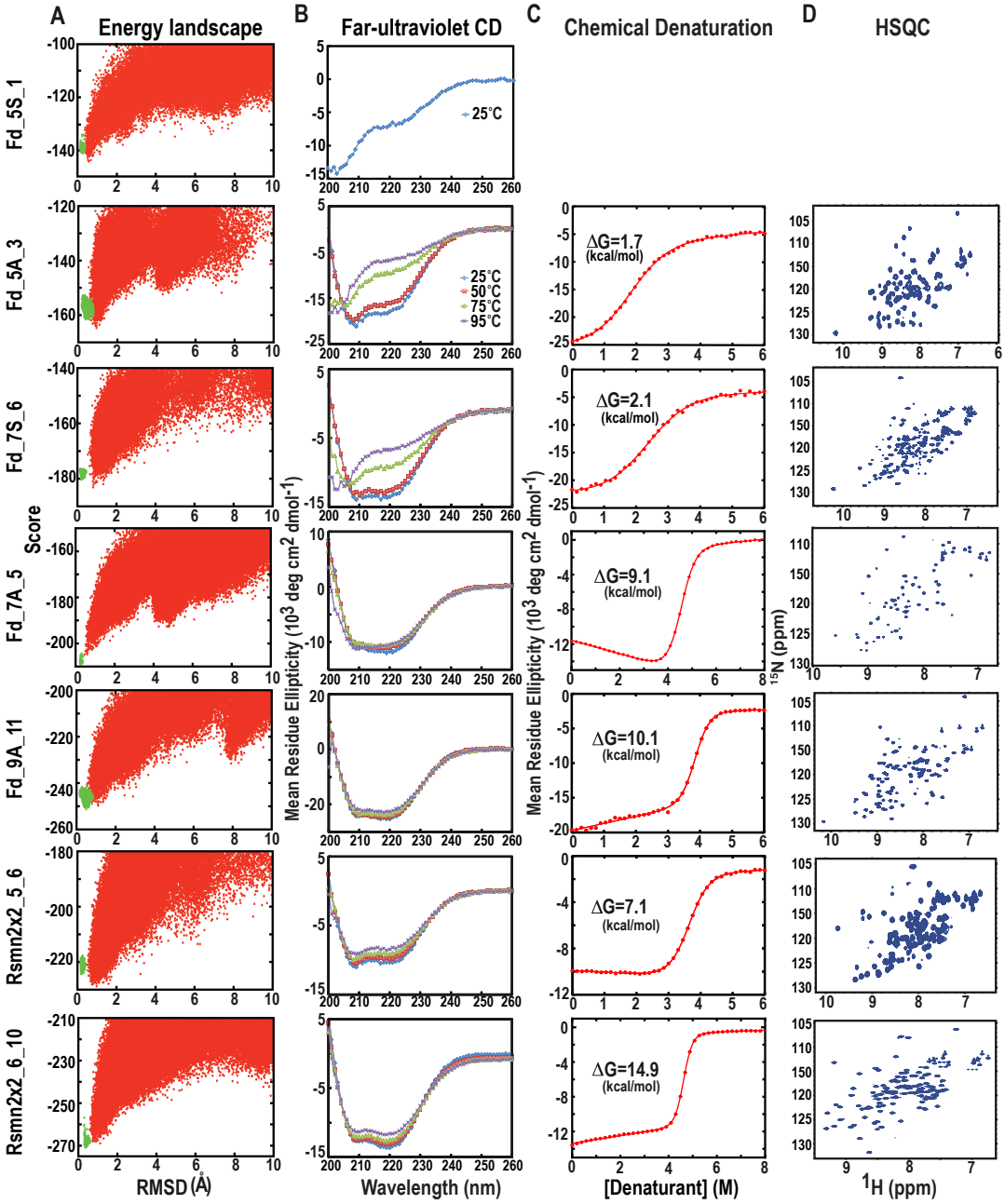


Backbone blueprints and design models for ferredoxin-like folds and Rossmann2x2 folds with different sizes and shapes.

(A-E) the ferredoxin-like fold, and (F-G) the Rossmann2X2 fold. Backbone blueprint for each topology (left) and a corresponding Rosetta generated backbone structure (right). (A) 5S: 58

residues, with register shift between the 1st and 3rd strands. (B) Fd_5A: 66 residues, without register shift. (C) Fd_7S: 74 residues, with register shift. (D) Fd_7A: 76 residues, without register shift. (E) Fd_9A: 98 residues, without register shift. (F) Rsmn2x2_5: 87 residues. (G) Rsmn2x2_6: 99 residues. Helices are represented by pink or red rectangles, and strands by arrows with individual positions indicated by filled and open boxes. The filled boxes represent pleats coming out of the page, and the open boxes, pleats going into the page. Designed loop types are indicated for Fd_5S, Fd_5A, Fd_7S, Fd_9A and Rsmn2x2_5. Fd_7A and Rsmn2x2_6 were designed by Koga *et al*. in 2012(9), where they were referred as Di-I_5 and Di-II_10 respectively, using loop length but not loop type based rules.

Figure 1.5

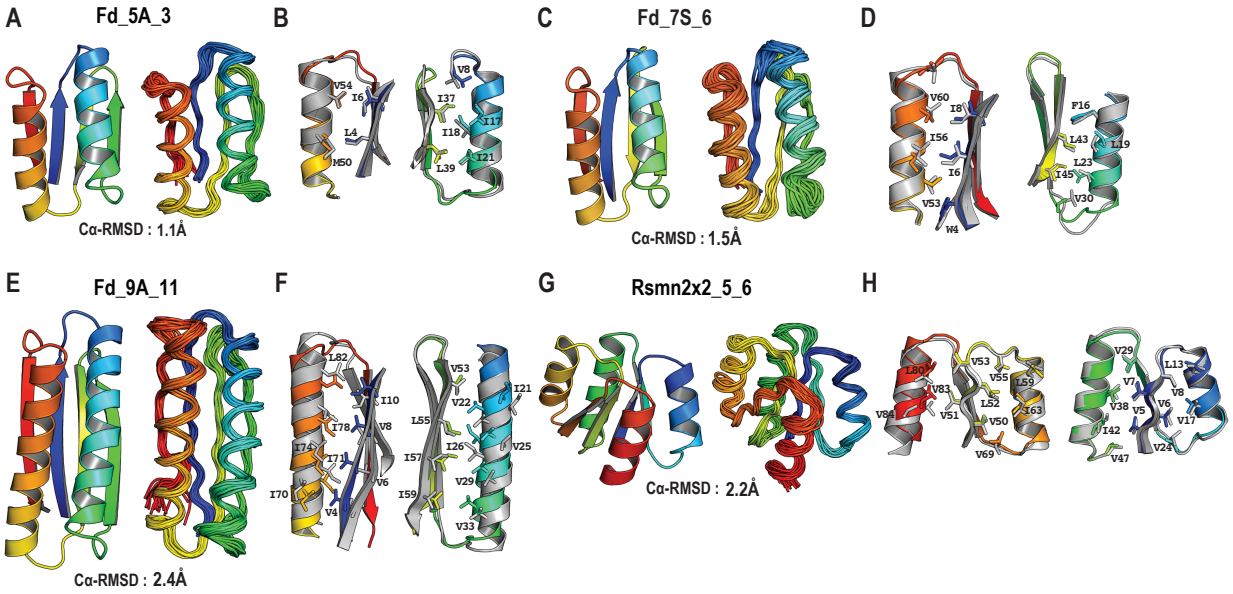


Experimental characterization of designed proteins.

(A) Energy landscapes obtained from Rosetta *ab initio* structure prediction simulations on Rosetta@home. Red points represent the lowest-energy structures obtained in independent

Monte Carlo structure prediction trajectories starting from an extended chain for each sequence; y axis, Rosetta all-atom energy; x axis, Cα root mean square deviation (RMSD) from the design model. Green points represent the lowest-energy structures obtained in trajectories starting from the design model. (B) The far-ultraviolet circular dichroism (CD) spectra at various temperatures. (C) Chemical denaturation with GuHCl or urea monitored by CD at 220 nm at 25°C. Urea was used for Fd_5A and Fd_7S denaturation and GuHCl for others. The data were fitted to a two-state model (red solid line) to obtain the free energy of unfolding ΔG. (D) Two-dimensional ^{1}H-^{15}N HSQC spectra at 25°C and 600 MHz. p.p.m., parts per million.

Figure 1.6



Comparison of computationally designed models with NMR structures. (A) (C) (E) (G) Overall comparison of design models (left) and NMR structures (right); the Cα root mean square deviation (RMSD) between the two is indicated. (B) (D) (F) (H) Comparison of core side-chain

packing in superpositions of design models (rainbow) and NMR structures(grey). (A)(B) Fd_5A_3; (C)(D) Fd_7S_6; (E)(F) Fd_9A_11; (G)(H) Rsmn2x2_5_6.

TABLES

Table 1.1 | Design Success Rate

	Designs Tested	Expressed[1]	Soluble[1]	Expected CD Spectrum[2]	Monomeric[3]	Well-resolved HSQC	Success[4]
Fd_5S	6	6	6	0	3	0	0 (0%)
Fd_5A	12	12	12	6	9	4	4 (33%)
Fd_7S	10	10	8	6	7	1	1 (10%)
Fd_7A	11	9	8	6	3	3	2 (18%)
Fd_9A	12	12	11	11	7	3	3 (25%)
Rsmn2x2_5	9	9	7	8	6	2	2(22%)
Rsmn2x2_6	12	12	12	10	4	4	4(33%)

The second column shows the number of designs experimentally tested for the backbone blueprint (**Fig 1.4**) indicated in the leftmost column. The subsequent columns give the number of designs that satisfy each criterion.

1. Expression and solubility were assessed by SDS-PAGE and mass spectrometry.

2. The expected CD spectrum is characteristic of $\alpha\beta$-proteins.

3. SEC-MALS was used to determine oligomerization state.

4. The successful designs are defined as those that satisfy all criteria. The details of the results are shown in **Tables 2.4-2.8**.

Table 1.2 | ABEGO-based loop comparison between design models and NMR structures for the five loops in the three ferredoxin-like folds.

In each cell, the loop type for the design model (left) and the most frequent loop type in the NMR ensemble (right) are shown.

	L1	L2	L3	L4	L5
Fd_5A_3	BAB/BAB	BAAB/BAAB	GG/BG[1]	BAB/BAB	BAAB/BAAB
Fd_7S_6	BAB/BAB	BAAB/BOBB[2]	GG/GG	BAB/BAB	BAAB/BAAB
Fd_9A_11	BAB/BAB	BAAB/BAAB	GG/GG	BAB/BAB	BAAB/BAAB

1. The B conformation at the position 1 was confirmed by chemical shift data(44).

2. The cis proline conformation at position 2 was confirmed by both proline Cβ/Cγ chemical shifts and a characteristic strong sequential Hα-Hα NOE.

Table 1.3 | ABEGO-based loop comparison between design model and NMR structures for the seven loops in the Rossmann2x2 fold.

	L1	L2	L3	L4
Rsmn2x2_5_6	AB/AB	BAAB/BAAB	BAB/BAB	GBA/GBA

	L5	L6	L7
Rsmn2x2_5_6	GBB/GBB	GB/GB	AB/AB

SECTION 2. SUPPLEMENTAL INFORMATION AND METHOD FOR CONTROL OVER OVERALL SHAPE AND SIZE IN DE NOVO DESIGNED PROTEINS

SI AND METHODS

Database for naturally occurring protein structures.

The PISCES server(45) was used to collect 6875 X-ray structures from the PDB with resolution ≤ 2.5Å, R-factor ≤ 0.3, sequence lengths from 40 to 10000, and ≤ 25% sequence identity, and their secondary structures were assigned using DSSP.

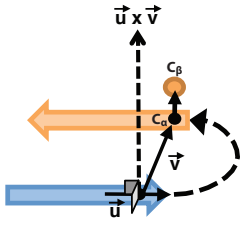
Definition for ββ-units.

ββ-units were identified as adjacent secondary structure elements (no intervening secondary structure element other than a loop) in which the strands were at least 2 residues.

Definition for the chirality of ββ-unit.

The chirality of ββ-unit is defined as follows:

$$\begin{cases} \text{R when } (\vec{u}\times\vec{v})\cdot\overrightarrow{c_\alpha c_\beta} > 0 \\ \text{L when } (\vec{u}\times\vec{v})\cdot\overrightarrow{c_\alpha c_\beta} < 0 \end{cases} \quad (1)$$



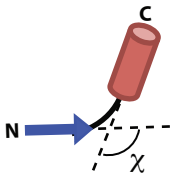
$\vec{u}$: a vector along the first strand. The vector from the N (backbone amide nitrogen) to the C (backbone carbonyl carbon) atoms of the strand residue preceding the connecting loop was used.

$\vec{v}$: a vector from the center of the first strand to the center of the second strand. The vector from the Cα atom of the strand residue preceding the loop to the Cα atom of the strand residue following the loop was used.

$\overrightarrow{c_\alpha c_\beta}$: a vector from the Cα to the Cβ atoms of the strand residue closest to the loop.

Definition for βα-units.

βα-units were identified as adjacent secondary structure elements (no intervening secondary structure element other than a loop) in which the strands were at least 2 residues and the helices were at least 5 residues in length. For plotting the histograms in **Fig. 1.2** in the main text, we considered βα-units in which the angle χ between the β-strand and the α-helix was ≤ 60°.

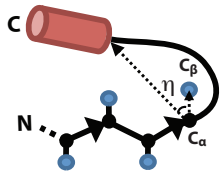


To compute χ, a strand vector was defined as the vector from the N to C atoms of the last residue in the strand immediately preceding the helix and a helix vector as the vector from the N (backbone amide nitrogen) atom of the first helix residue to the C (backbone carbonyl carbon) atom of the fourth helix residue. χ is then the angle between the strand vector and the helix vector.

Definition for the orientation of βα-unit.

The orientation of βα-unit is defined as follows:

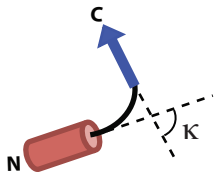
$$\begin{cases} \text{parallel (Para) when the } \eta \text{ angle} \le 80° \\ \text{antiparallel (Anti) when the } \eta \text{ angle} \ge 100° \end{cases} \quad (3)$$



where η is the angle between the $\overrightarrow{c_\alpha c_\beta}$ vector of the last residue in the strand and the vector from the Cα atom of the last residue in the strand to the average of the coordinates of the first 11 backbone atoms (N, C, and Cα) in the helix.

Definition for αβ-units.

αβ-units were identified as adjacent secondary structure elements (no intervening secondary structure element other than a loop) in which the strands were at least 2 residues and the helices were at least 5 residues in length. For plotting the histograms in **Fig. 1.2** in the main text, we considered αβ-units in which the angle κ between the helix vector and the strand vector was ≤ 60°.

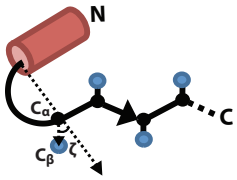


To compute κ, a helix vector was defined as the vector from the N atom of the fourth helix residue from the last to the C atom of the last helix residue and a strand vector as the vector from the N to C atoms of the first residue in the strand immediately following the helix.

Definition for the orientation of αβ-unit.

The orientation of αβ-unit is defined as follows:

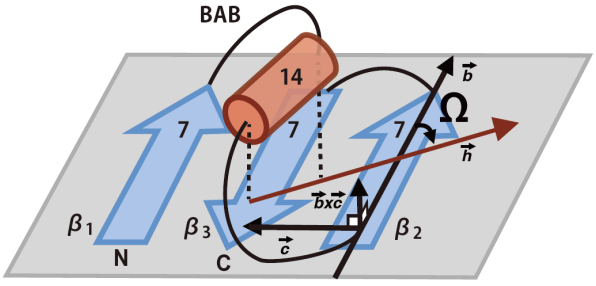
$$\begin{cases} \text{parallel (Para) when the } \zeta \text{ angle} \le 80° \\ \text{antiparallel (Anti) when the } \zeta \text{ angle} \ge 100° \end{cases} \quad (5)$$



where ζ is the angle between the $\overrightarrow{c_\alpha c_\beta}$ vector of the first residue in the strand and the vector from the average of the coordinates of the last 11 backbone atoms (N, C and Cα) in the helix to the Cα atom of the first strand residue.

Definition for the tilt angle of the helix relative to β-sheet.

The tilt angle of the helix relative to the β-sheet, Ω, was calculated from $\beta_1\alpha\beta_2\beta_3$-unit folding simulations with the BAB loop for the βα-connection, with strand lengths 7 and helix



length 14 (**Fig. 1.3H, I**). The Ω angle is defined as the angle between a strand vector along the β_2 strand, $\vec{b}$, and a helix vector projected on a β-sheet plane, $\vec{h}$. The $\vec{b}$ strand vector was defined as the vector from the N to C atoms of the first residue in the β_2 strand and the helix vector as the vector from the C atom of the last helix residue to the N atom of the fourth helix residue from the last. The β-sheet plane is specified by the $\vec{b}$ vector and a vector from the Cα atom of the first residue of the β_2 strand to the Cα atom of the last strand residue of the β_3 strand, $\vec{c}$.

Rosetta folding simulations.

For building backbone structures based on the blueprints with various secondary structure lengths, loop geometries and β-strand register shifts, secondary-structure-and-ABEGO constrained Monte-Carlo Rosetta folding simulations were carried out with a sequence-independent backbone model consisting of the N, NH, Cα, C, CO and Cβ atoms, with a pseudo-atom representing a generic side chain (the centroid model of Rosetta), using the Rosetta potential function(9), in which steric repulsion (vdw=1.0), overall compaction (rg=1.0), secondary structure pairings (ss_pair = 1.0, rsigma = 1.0, and hs_pair = 1.0) and hydrogen bonds (hbond_sr_bb = 1.0, hbond_lr_bb = 1.0) are used. Note that no amino-acid-sequence dependent score terms are included. For the steric radius of the side-chain pseudo-atom, the radius of Val was used. For enhancing sampling efficiency for the conformations described in the backbone blueprints, we modified the ss_pair and rsigma score functions so that only the strand residue pairs that are formed in the backbone blueprints are favored.

For sampling backbone structures, the fragment assembly method was employed(46). Backbone fragment sets consisting of 1, 3 or 9 consecutive residue fragments were prepared in advance

from a non-redundant set of X-ray structures; the fragments have information only on the phi, psi and omega torsion angles. We performed Monte Carlo simulations in which in each attempted Monte Carlo trial, a new conformation is generated by replacing the torsion angles (phi, psi and omega) of a randomly selected frame consisting of 1, 3 or 9 consecutive residues with the torsion angles of a randomly selected fragment compatible with the secondary structure and ABEGO type assigned in the backbone blueprints. The total number of Monte-Carlo steps in one trajectory is 300×(length of simulated chain).

Protein expression and purification.

For all designed sequences, a spacer was added at the C-terminus in order to separate the designed region and the C-terminal 6xHis-tag. (GS for Fd_5S, Fd_5A; S for Fd_7S; SSWS for Fd_9A; SG for Rsmn2x2_5). The synthetic genes encoding Fd_5S, Fd_5A and Fd_9A designed sequences, obtained from GenScript, were cloned into the pET21 expression vector. Double stranded DNA fragments encoding Fd_7S and Rsmn2x2_5 designed sequences were obtained from Integrated DNA Technologies. DNA fragments of Fd_7S designed sequences and pET21 vectors were digested with NdeI and XhoI restriction enzymes at 37 °C for 1 hour. Digested designed sequences and vectors were purified with QIAquick PCR purification kit. Vectors and designed sequences were ligated with T4 ligase at room temperature for 1 hour. DNA fragments of Rsmn2x2_5 designed sequences were cloned into pET21 vectors digested with NdeI and XhoI restriction enzymes by Gibson Assembly Cloning method(47). Cloned-pET21 plasmids were then transformed into XL1-Blue cells and sequences were verified by GENEWIZ.

The designed proteins were expressed in E. coli BL21 Star (DE3) cells as uniformly (U-) ^{15}N-labeled proteins for all designs. The U-^{15}N-labeled proteins were expressed using MJ9 minimal media(48), which contain ^{15}N ammonium sulfate as sole nitrogen source and ^{12}C glucose as sole carbon source. The expressed proteins with a 6xHis-tag at the C-terminus were purified through a nickel affinity column. The purified proteins were then dialyzed against typical PBS buffer, 137 mM NaCl, 2.7mM KCl, 10mM Na_2HPO_4, 1.76 mM KH_2PO_4, at pH 7.4; this buffer was used for all the experiments except NMR structure determination. The expression, solubility, and purity of the designed proteins were assessed by SDS-PAGE and mass spectrometry (TSQ LC/MS, Thermo Scientific).

Circular dichroism (CD).

All CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD spectra of designed proteins were measured from 260 to 200 nm in PBS buffer (pH 7.4) at protein concentrations of 14-28 uM and temperatures of 25, 50, 75, and 95 °C, using a 1 mm path length cuvette. The protein concentrations were determined from the absorbance at 280 nm(49) using UV spectrophotometer (NanoDrop, Thermo Scientific). T_m is the melting temperature where the fraction of folded proteins is equal to the fraction of unfolded proteins during temperature denaturation. Chemical denaturations with GuHCl (for Fd_9A and Rsmn2x2_5) and urea (for Fd_5A and Fd_7S) were monitored at 220 nm for 3-7 uM protein samples in PBS buffer at 25 °C in a 1 cm path length cuvette. The GuHCl/urea concentration was automatically controlled by a Microlab titrator (Hamilton). The chemical denaturation curves were fit using a two-state unfolding and linear extrapolation model(50). The free energy change, ΔG, for the unfolding transition and its dependency on the denaturant, m-value, were obtained from the fitting.

Size exclusion chromatography combined with multi-angle light scattering (SEC-MALS).

SEC-MALS experiments were performed using a miniDAWN TREOS static light scattering detector (Wyatt Technology) combined with a HPLC system (LC 1200 Series, Agilent Technologies). The volume 100 ul of 400-700 uM protein samples in PBS buffer (pH 7.4) was injected into a Superdex 75 10/300 GL column (GE Healthcare) equilibrated with PBS buffer at a flow rate of 0.5 ml/min. The protein concentrations were calculated from the absorbance at 280 nm detected by the HPLC system. Static light scattering data were collected at three different angles, 41.4°, 90.0°, and 138.6°, at 658 nm. These data were analyzed by the ASTRA software (version 5.3.4, Wyatt Technology) with a change in the refractive index with concentration, a dn/dc value, 0.185 ml/g.

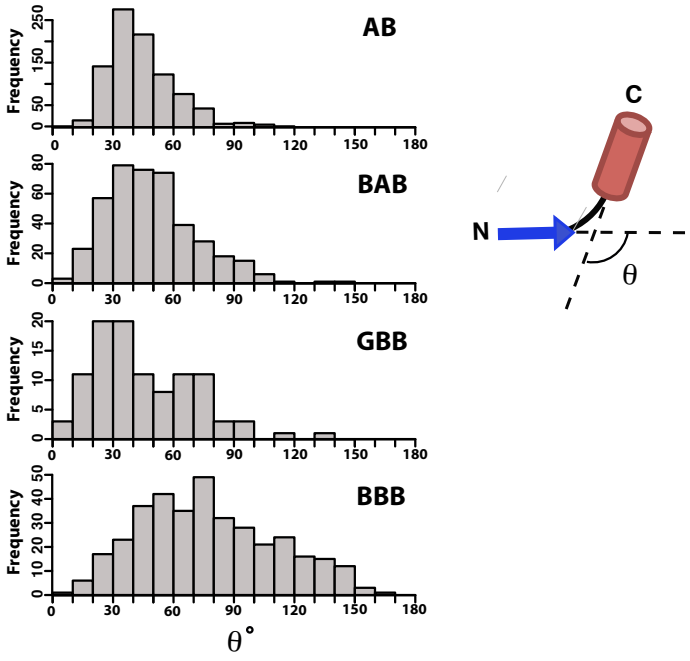
NMR structure determination.

The selected designs were expressed and purified using standard protocols established by the Northeast Structural Genomics (NESG) consortium (51). The designs were expressed in *E. coli* BL21 (DE3) pMGK cells as *U*-^{15}N, 5%^{13}C-enriched proteins, and *U*-^{15}N, *U*-^{13}C-enriched proteins using MJ9 minimal media(48). The *U*-^{15}N, 5%^{13}C-labeled proteins were generated for stereo-specific assignments of isopropyl methyl groups of valines and leucines(52) and for residual dipolar couplings (RDC) measurement(53). For NMR structure determination, all NMR spectra were recorded at 25 °C using cryogenic NMR probes. The NMR structures were determined using standard triple-resonance NMR data collection and analysis protocols as previously described(31). Dihedral angle constraints were obtained from TALOSN. RDC from one or two alignment media were used for the structure determination of Fd_5A_3, Fd_9A_11

and Rsmn2x2_5_6. Both RDC and TALOSN restraints were carefully checked and only applied to residues at regular secondary structure elements or at well-defined loop regions. The predicted order parameter (S2 ≥ 0.68) is used as guidance to select TALOSN/RDC restraints for loop residues. Consensus hydrogen bonds identified from the 20 best CYANA conformers were applied as additional distance restraints in the final structure refinement in explicit water using CNS programs.

FIGURES

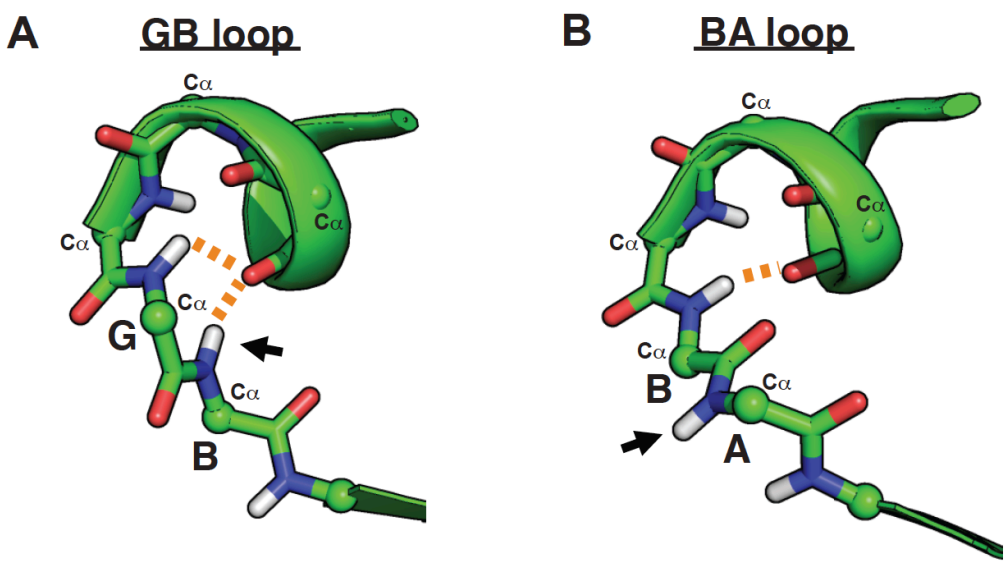
Figure 2.1



Packing geometry of βα-units in naturally occurring proteins.

Consider the cross angle between the vectors of strand and helix, θ (Detail definition for θ is described in **SI and Methods**: Database for naturally occurring protein structures). Despite of common observation of BBB loop in native βα-units, further analysis of native "all" βα-units (The histograms in **Fig. 1.2** were plotted with the dataset of the θ angle < 60°, but these histograms were plotted without the angle restriction.) indicates that compared to AB, BAB and GBB loops which commonly lead to small θ angle, BBB loop has a wide range of θ angle and gives around 80° θ angle the most. Large θ angle indicates poor packing between strand and helix and therefore BBB loop is not used in design process.

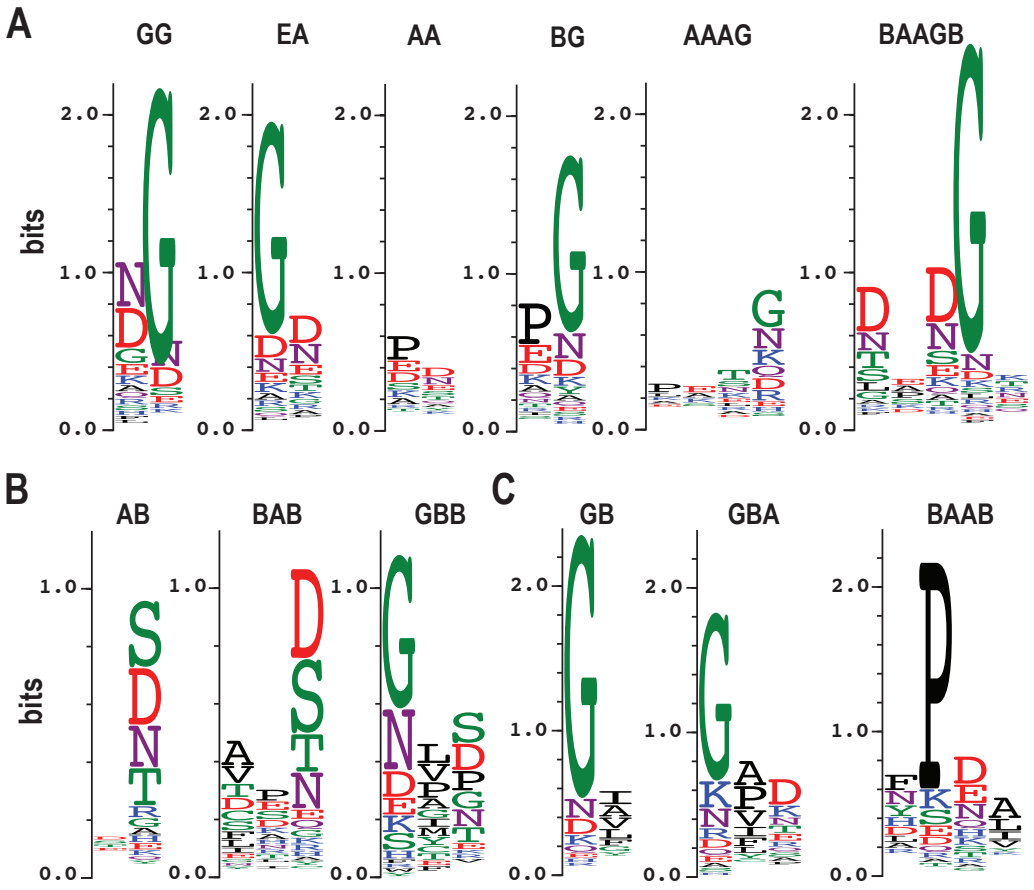
Figure 2.2



GB loop provides C-terminal helix capping.

Examples of naturally occurring protein structures for (A) GB and (B) BA loops. Even though both GB and BA loops are highly observed in αβ-units, the mainchain amide of "B" geometry in GB loop, which is indicated by an arrow in (A), provides a hydrogen-bonded C-terminal helix capping. On the other hand, in BA loop, the corresponding mainchain amide indicated by an arrow flips to the solvent without making the helix capping.

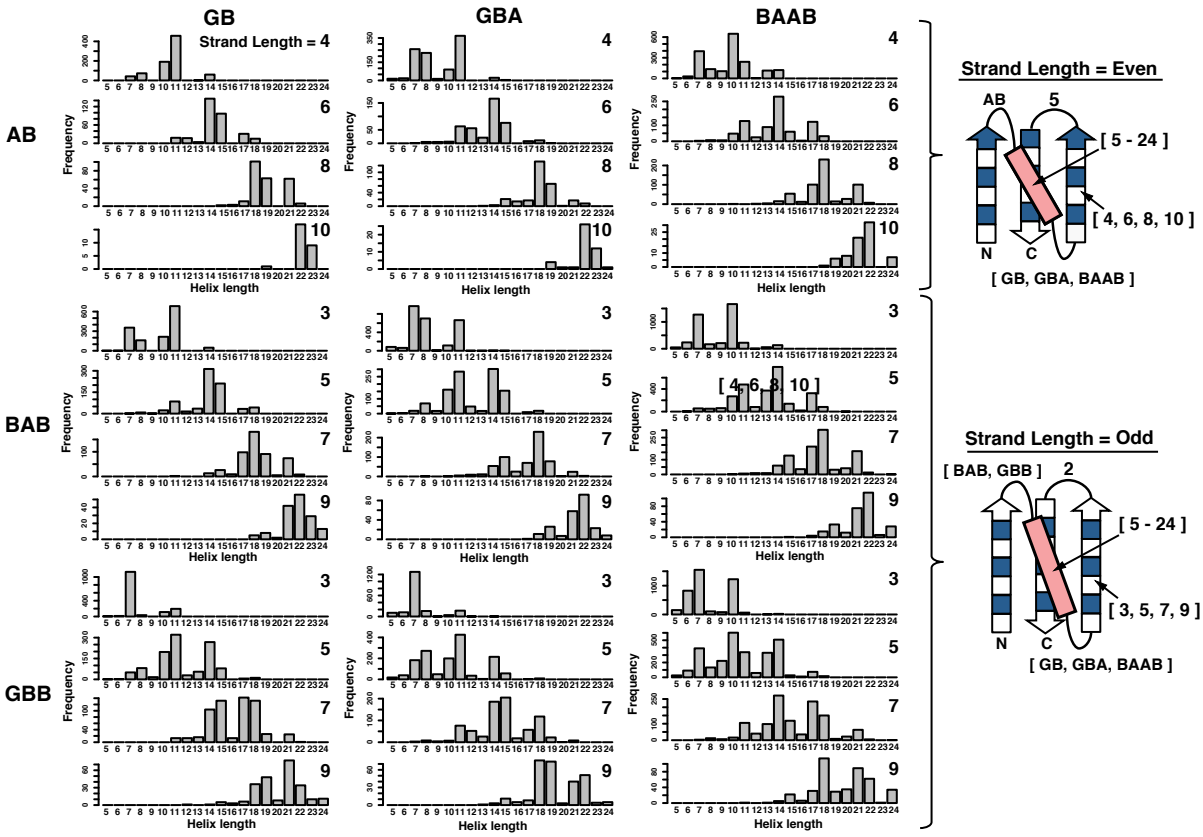
Figure 2.3



Sequence profiles for loop types for naturally occurring protein structures.

Sequence profiles for naturally occurring protein structures, created by WebLogo(54), are shown for ββ-connection (A), βα-connection (B), and αβ-connection (C).

Figure 2.4

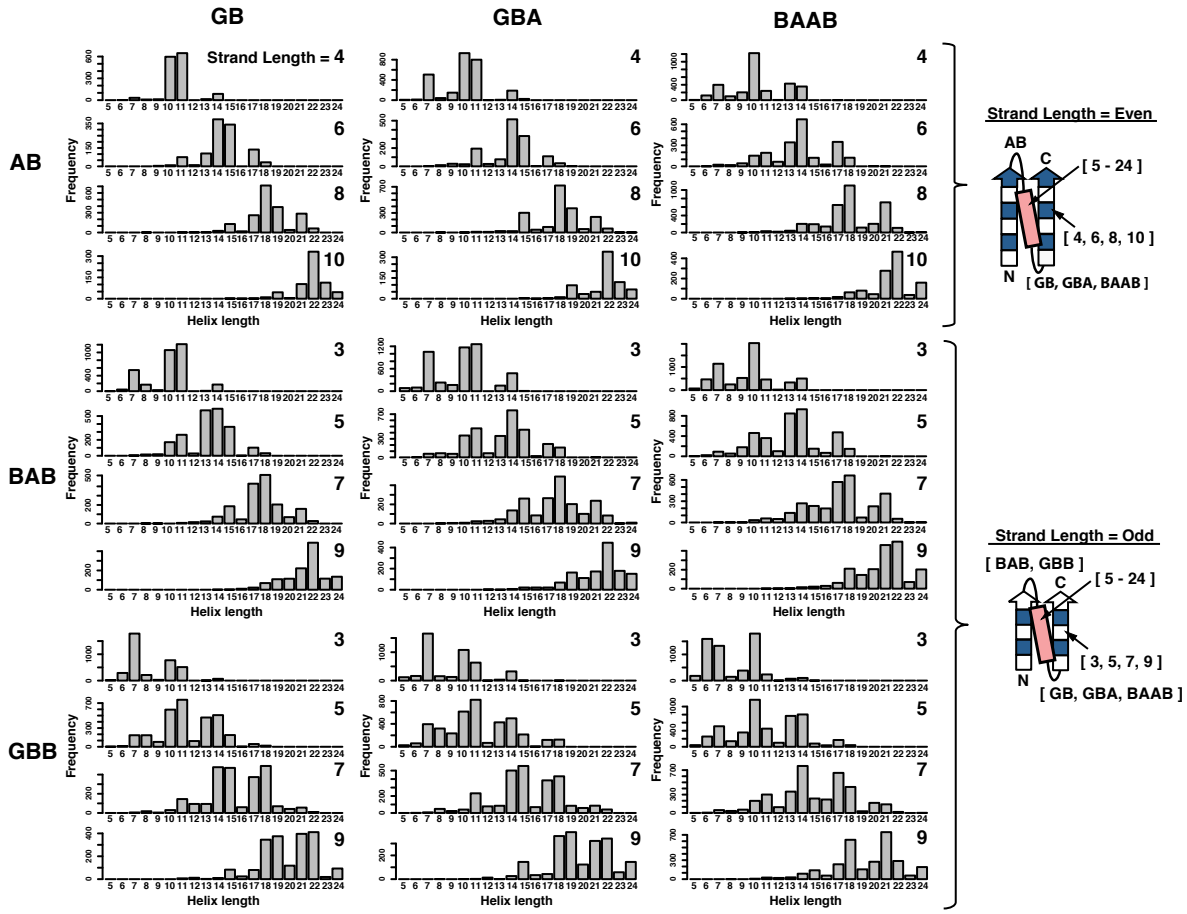


Codependency among optimal helix length, strand length, and loop geometries for βαββ-unit.

For detecting optimal helix length in βαββ-unit with various strand lengths and loop geometries, the secondary-structure-and-ABEGO constrained Rosetta folding simulations (**SI and Methods**) were carried out. Simulation inputs for secondary-structure lengths and loop types are described in the schematic views (the three strands have same lengths). For each combination of secondary structure lengths and loop geometries, we performed 3000 (S=3,4,5,6), 5000(S=7), and 10000(S=8,9,10) independent Monte Carlo simulation trajectories at temperatures 0.5 (S=3) and 1.0 (S=4,5,6,7,8,9,10), followed by secondary structures assignments for end point structures using DSSP, and then counted the number of end point structures, in which 1) the secondary-

structure lengths and loop geometries agreed with the input assignment, and 2) the three strands made antiparallel pairings as illustrated in the schematic views. In general, optimal helix length increases with strand at a ratio of four (helix residues) to two (strand residues). Note that optimal helix lengths are about four-residues longer in units using BAB or AB loops in βα-connection than those using GBB loops with same strand lengths.

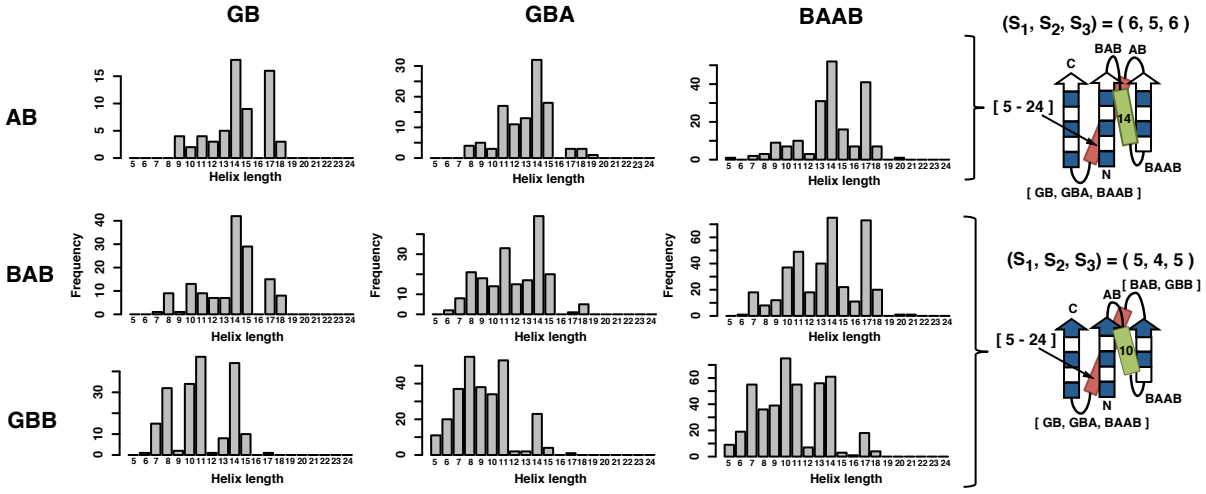
Figure 2.5



Codependency among optimal helix length, strand length, and loop geometries for βαβ-unit.

For detecting optimal helix length in βαβ-unit with various strand lengths and loop geometries, the secondary-structure-and-ABEGO constrained Rosetta folding simulations (**SI and Methods**) were carried out. Simulation inputs for secondary-structure lengths and loop types are described in the schematic views (the two strands have same lengths). For each combination of secondary structure lengths and loop geometries, we conducted 3000 (S=3,4,5,6), 5000(S=7), and 10000(S=8,9,10) independent Monte Carlo simulation trajectories at temperatures 0.5 (S=3) and 1.0 (S=4,5,6,7,8,9,10), followed by secondary structures assignments for end point structures using DSSP, and then counted the number of end point structures, in which 1) the secondary-structure lengths and loop geometries agreed with the input assignment, and 2) the two strands made a parallel pairing as illustrated in the schematic views. Similar to βαββ-unit, optimal helix length increases with strand at a ratio of four (helix residues) to two (strand residues) and BAB and AB loops in βα-connection allow longer optimal helix lengths.

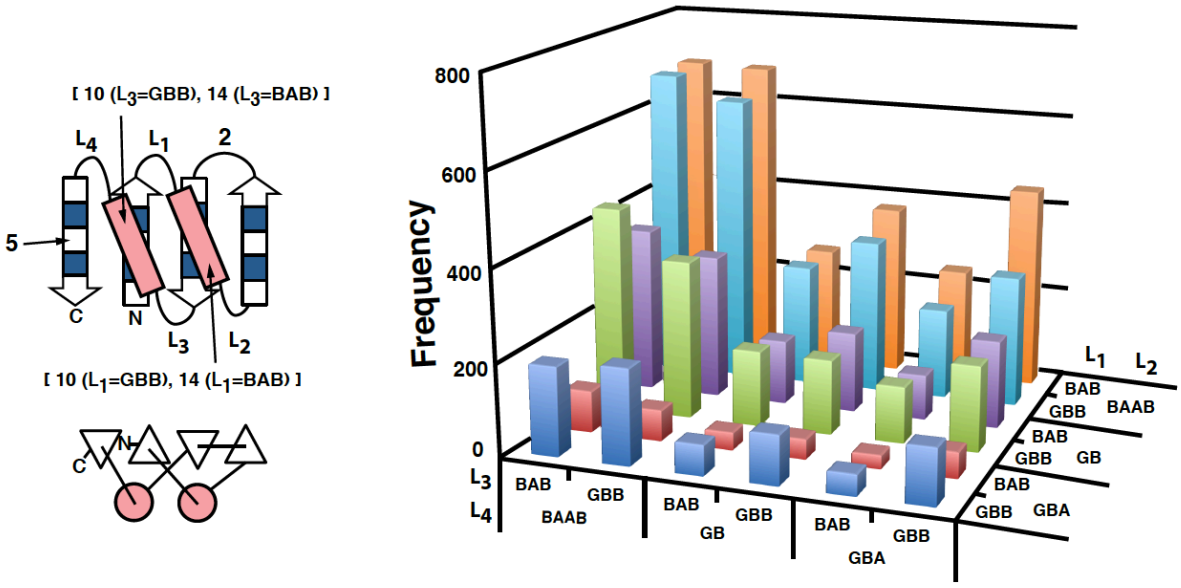
Figure 2.6



Codependency among optimal helix length, strand length, and loop geometries for βαβαβ-unit.

For detecting optimal helix lengths in βαβαβ-unit (red helix in the schematic view) with various loop geometries and strand lengths, the secondary-structure-and-ABEGO constrained Rosetta folding simulations (**SI and Methods**) were carried out. Simulation inputs for secondary-structure lengths and loop geometries are described in the schematic views. For each combination of secondary structure lengths and loop geometries, we conducted 3000 independent Monte Carlo simulation trajectories at a temperature 1.0, followed by secondary structures assignments for end point structures using DSSP, and then counted the number of end point structures, in which 1) the secondary-structure lengths and loop geometries agreed with the input assignment, and 2) the three strands made parallel pairings as illustrated in the schematic views.

Figure 2.7



Combined βαββ-units preferred certain loop geometry.

Ferredoxin-like fold can be seen as a combination of two βαββ-units. First unit involves strand 1, 2 and 3 and helix 1 and second unit includes strand 3, 4 and helix 2. We conducted 10000 independent Monte Carlo simulation trajectories at a temperature 1.0, followed by secondary

structures assignments for end point structures using DSSP, and then counted the number of end point structures, in which 1) the secondary-structure lengths and loop geometries agreed with the input assignment, and 2) the four strands made antiparallel pairings as illustrated in the schematic views. With fixed secondary structure length, simulation shows BAAB loop is highly preferred for L_2 and L_4.

TABLES

Table 2.1 | Summary of Fig. 2.3: Codependency among optimal helix length, strand length, and loop geometries for βαββ-unit.

βαββ	βα-loop				αβ-loop
Strand Length	AB	BAB	GBB		
3	NA	11	7	GB	
		7	7	GBA	
		10	7	BAAB	
4	11	NA	NA	GB	
	11			GBA	
	10			BAAB	
5	NA	14	11	GB	
		14	11	GBA	
		14	10	BAAB	
6	14	NA	NA	GB	
	14			GBA	
	14			BAAB	
7	NA	18	17	GB	
		18	15	GBA	
		18	14	BAAB	
8	18	NA	NA	GB	
	18			GBA	
	18			BAAB	
9	NA	22	21	GB	
		22	18	GBA	
		22	18	BAAB	
10	22	NA	NA	GB	
	22			GBA	
	22			BAAB	

Table 2.2 | Summary of Fig. *2*.4: Codependency among optimal helix length, strand length, and loop geometries for βαβ-unit.

βαβ	βα-loop				αβ-loop
Strand Length	AB	BAB	GBB		
3	NA	11	7	GB	
		11	7	GBA	
		10	10	BAAB	
4	11	NA	NA	GB	
	10			GBA	
	10			BAAB	
5	NA	14	11	GB	
		14	11	GBA	
		14	10	BAAB	
6	14	NA	NA	GB	
	14			GBA	
	14			BAAB	
7	NA	18	18	GB	
		18	15	GBA	
		18	14	BAAB	
8	18	NA	NA	GB	
	18			GBA	
	18			BAAB	
9	NA	22	22	GB	
		22	19	GBA	
		22	21	BAAB	
10	22	NA	NA	GB	
	22			GBA	
	22			BAAB	

Table 2.3 | Summary of Fig. 2.5: Codependency among optimal helix length, strand length, and loop geometries for βαβαβ-unit.

βαβαβ	βα-loop				αβ-loop
Strand Length (E_1,E_2,E_3)	AB	BAB	GBB		
5, 4, 5	NA	14	11	GB	
		14	8	GBA	
		14	10	BAAB	
6, 5, 6	14	NA	NA	GB	
	14			GBA	
	14			BAAB	

Table 2.4 | Summary of experimental results of 6 designs for Fd_5S.

	Expressed	Soluble	Expected CD spectrum	T_m (°C)	Monomeric	Well-resolved HSQC
1	✔	✔				
2	✔	✔				
3	✔	✔			✔	
4	✔	✔			✔	
5	✔	✔			✔	
6	✔	✔				

Each row corresponds to the results for each design. The columns give the results for each criterion, of which the details are described in **Table 1.1**. T_m is the melting temperature. The design that satisfies a criterion is shown with a check mark and the design that does not satisfy a criterion is shown in white blank. The case that the experiment was not conducted is shown in gray blank.

Table 2.5 | Summary of experimental results of 12 designs for Fd_5A.

	Expressed	Soluble	Expected CD spectrum	T_m (°C)	Monomeric	Well-resolved HSQC
1	✔	✔				
2	✔	✔			✔	
3	✔	✔	✔	55	✔	✔
4	✔	✔				
5	✔	✔	✔	55	✔	
6	✔	✔	✔	55	✔	
7	✔	✔	✔	40	✔	✔
8	✔	✔	✔	60	✔	✔
9	✔	✔			✔	
10	✔	✔			✔	
11	✔	✔				
12	✔	✔	✔	60	✔	✔

The summary was given in the same way as **Table 2.4**.

Table 2.6 | Summary of experimental results of 10 designs for Fd_7S.

	Expressed	Soluble	Expected CD spectrum	T_m (°C)	Monomeric	Well-resolved HSQC
1	✔	✔	✔	85	✔	
2	✔	◆	✔	80	✔	❖
3	✔	✔			✔	
4	✔					
5	✔	◆	✔	60	❖	
6	✔	✔	✔	70	✔	✔
7	✔	✔			✔	
8	✔	✔	✔	>95	✔	
9	✔					
10	✔	✔	✔	90	✔	

The summary was given in the same way as **Table 2.4**. ◆represents low concentration. ❖ indicates that experiment was not conducted due to low concentration.

Table 2.7 | Summary of experimental results of 12 designs for Fd_9A.

	Expressed	Soluble	Expected CD spectrum	Tm(°C)	Monomeric	Well-resolved HSQC
1	✔	✔	✔	>>95	✔	
2	✔	✔	✔	>95		
3	✔	✔	✔	>95	✔	
4	✔					
5	✔	✔	✔	>95	✔	
6	✔	✔	✔	>95	✔	
7	✔	✔	✔	>95		
8	✔	✔	✔	>>95		
9	✔	✔	✔	>95	✔	✔
10	✔	✔	✔	>95	✔	✔
11	✔	✔	✔	>95	✔	✔
12	✔	✔	✔	>95		

The summary was given in the same way as **Table 2.4**.

Table 2.8 | Summary of experimental results of 9 designs for Rsmn2x2_5.

	Expressed	Soluble	Expected CD spectrum	Tm(°C)	Monomeric	Well-resolved HSQC
1	✔	≈	✔	>>95	✔	
2	✔	≈	✔	>95	✔	
3	✔	✔	✔	>>95		
4	✔	✔	✔	>95	✔	✔
5	✔	✔	✔	>95	✔	
6	✔	✔	✔	>>95	✔	✔
7	✔	✔			✔	
8	✔	✔	✔	85		
9	✔	✔	✔	>>95		

The summary was given in the same way as **Table 2.4**. ≈ represents Mild precipitation.

Table 2.9 | NMR and refinement statistics for designed protein structures.

Design ID	**Fd_5A_3**	**Fd_7S_6**	**Fd_9A_11**	**Rsmn2x2_5_6**
NESG ID	**OR358**	**OR303**	**OR414**	**OR446**
PDB ID	**XXXX**	**XXXX**	**2MQ8**	**XXXX**
NMR distance and dihedral constraints				
Distance constraints				
Total NOE	1489	1863	2762	2505
Intra-residue	412	472	580	565
Inter-residue				
Sequential (|i-j| = 1)	375	504	683	655
Medium-range (|i-j| ≤ 4)	313	360	584	614
Long-range (|i-j| ≥ 5)	389	527	915	671
Intermolecular				
Hydrogen bonds	10	26	38	28
Total dihedral angle restraints				
phi	46	54	79	65
psi	46	54	79	65
Total RDCs	37		72	104
Q(%, alignment media 1[§])		-		25.2
Q(%, alignment media 2[§])	6.5	-	28.2	36.4
Structure statistics				
Violations				
RMS of distance violation/constraint[¶] (Å)	0.01	0.01	0.01	0.01
RMS of dihedral angle violation/constraint (°)	0.87	1.21	0.96	0.55
Max distance constraint violation (Å)	0.33	0.37	0.36	0.30
Max dihedral angle violation (°)	6.9	7.9	7.7	7.0
Average pairwise r.m.s.d.** (Å)				
Heavy	1.61±0.13	1.59±0.12	1.33±0.15	1.45±0.16
Backbone	0.62±0.06	0.74±0.14	0.60±0.09	0.74±0.12
RPF Scores				
Recall	0.983	0.975	0.984	0.989
Precision	0.95	0.952	0.98	0.973
F-measure	0.966	0.963	0.982	0.981
DP-scores	0.84	0.83	0.922	0.905
Structure Quality Factors - overall statistics scores (raw/Z-scores[⌘])				
Procheck G-factor (phi / psi only)**	-0.26/-0.71	-0.26/-0.71	-0.02/0.24	0.12/0.79
Procheck G-factor (all dihedral angles)**	-0.18/-1.06	0.19/-1.12	-0.06/-0.35	0.11/0.65
Verify3D	0.43/-0.48	0.42/-0.64	0.43/0.48	0.52/0.02
ProsaII (-ve)	1.00/1.45	0.89/0.99	-/-	1.32/2.77
MolProbity clashscore	14.82/-1.02	14.36/-0.94	13.69/-0.82	17.87/-1.54
Ramachandran Plot Summary from Richardson Lab's Molprobity**				
Most favored regions (%)	94.4	98.6	99	98.9
Allowed regions (%)	5.1	1.1	1.0	1.1
Disallowed regions (%)	0.6	0.2	0.0	0.0

***** Analyzed for the 20 lowest energy refined structures of design for each of the four folds by using PDBSTAT and PSVS 1.4(55,56).

[§] PEG and phage were used as alignment media 1 and 2.

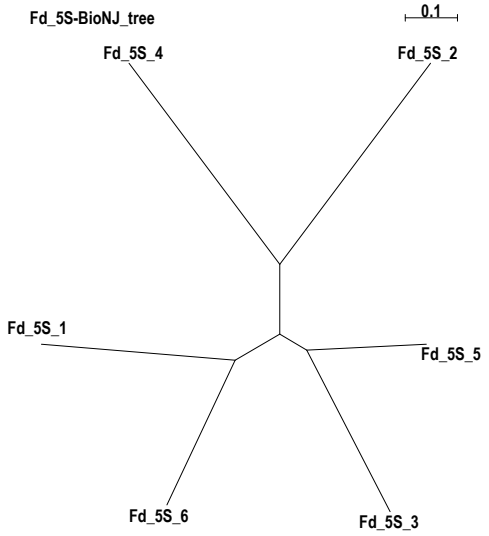
¶ Calculated by using sum over r^{-6}.

** Calculated among 20 refined structures for ordered residues that have sum of phi and psi order parameters(57) S(phi)+S(psi)>1.8(42). The ordered residues of Fd_5A_3: 2-7, 11-66; Fd_7S_6: 4-9, 12-25,28-74; Fd_9A_11: 2-48, 52-95; Rsmn2x2_5_6: 4-66, 69-87.

⌘ With respect to mean and standard deviation for a set of 252 X-ray structures with sequence lengths < 500, resolution ≤ 1.80 Å, R-factor ≤ 0.25 and R-free ≤ 0.28; a positive value indicates a 'better' score.

Table 2.10 | Sequences and phylogenetic tree of 6 designs for Fd_ 5S.

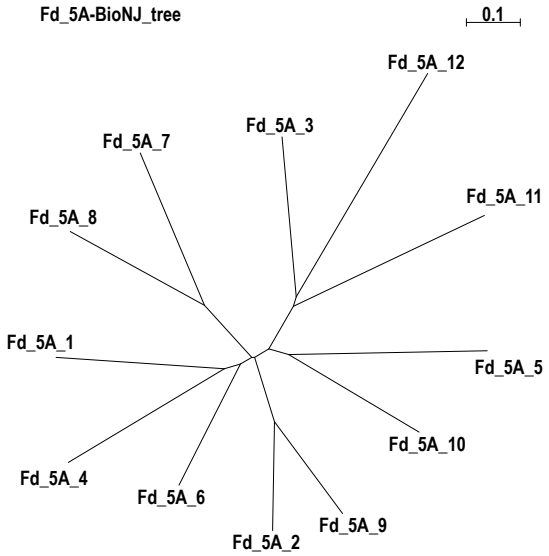
Fd_5S_1	mLTWEIRVDDEELAEEIERDDPQATVTRKGNTVEVRVTSEDVVKRAR ERDPEATITRTGgslehhhhhh
Fd_5S_2	mVTYKITITDKERMEELKKREPSATITRRNGEYEIELTDKDLMEEFK KEDPEVTITQTGgslehhhhhh
Fd_5S_3	mGKYEVIVQDEELAKRMEKRKPNATVTRQGNDYKVDLNSEKIMREIL KEKPNATVTTRGgslehhhhhh
Fd_5S_4	mVKYDIKLDDENLVRKLKEKRPNATITTRGNDYKIDLQSKEAVEEMR RERPNATIRTKGgslehhhhhh
Fd_5S_5	mGRYNVRVDDKELAERLREELPNATVQTQGNKYEVDLESEEQVKEIR KRKPEATITTQGgslehhhhhh
Fd_5S_6	mGQYRIRVDSRELAEDVRKERPNATVTQDNGTYEVRATDEDLRKEIE KRDPNATITQTGgslehhhhhh



Computationally designed sequences are shown in uppercase and residues added to allow expression, purification and the spacer between the designed sequences and the C-terminal His-tag are shown in lowercase.

Table 2.11 | Sequences and phylogenetic tree of 12 designs for Fd_5A.

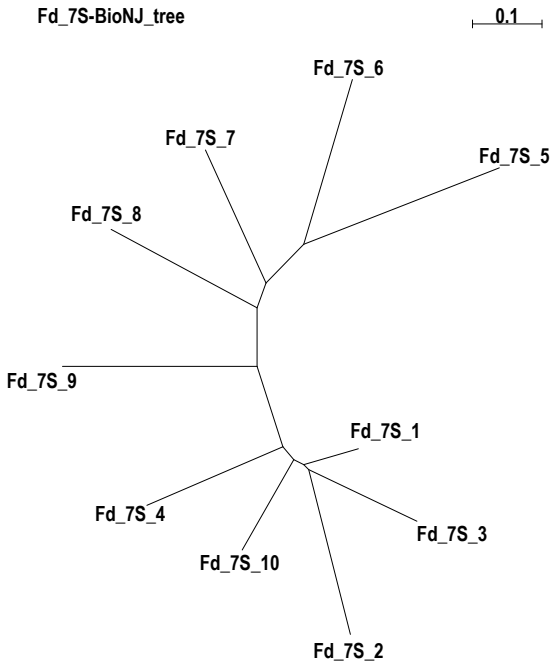
Fd_5A_1	mVTVKLDVNDDDLAERVLEDIRKKWPSATVTRTGGDIEVTSNTDEEA EKVEKIMREQAPQATVRTTGgslehhhhhh
Fd_5A_2	mVKVRLTSSDEDRAEEIARRIREKWPNATVQRTNGDIQVESQTDETA KKLAELMKKEKPEATVTRTGgslehhhhhh
Fd_5A_3	mVDLKIDVSDDEEAEKIIREIREQWPKATVTRTNGDIKLDAQTEKEA EKMEKAVKKVKPNATIRKTGgslehhhhhh
Fd_5A_4	mVTVKLDLSDERKAEEILKKIREEYPGATVRTTNGDWTVKANDEEKA KNVAKIMKEKAPDATVTRTGgslehhhhhh
Fd_5A_5	mVEIKLKATDDNKAEKILDRMKKRWPNATVTRTNGEVTVRADTQEKA QKMKQDMEELLPEATVETTGgslehhhhhh
Fd_5A_6	mIEVKLRTNDEERAEKILQKIREKWPNATVTRTNGDIQVKATTEEEI KRIEDNMKETAPSATITTTGgslehhhhhh
Fd_5A_7	mVDVKLKLNDEKQAEELAKKIRDKWPDATVTRTNGEVRVDVTEEQEL EEIEDQMKREYPDGTIRTTGgslehhhhhh
Fd_5A_8	mVTIRLESSDEREAERLARRIKDEWPSATVRKTNGDVTIDVQSQDEL KEIEDKMKEEYPSATVRTTGgslehhhhhh
Fd_5A_9	mVRVRLTSTDQEKAEKIARDIRDKWPNATVTRTNGEIDVESQSDDDA KRIEEEMEKQQPEATVTTTGgslehhhhhh
Fd_5A_10	mVEVELRVQDEDKAEKIARKIQNRWPNVTVTRTNGDVRLRAQTEDKA KRLKEQMREEDPSATLTTTGgslehhhhhh
Fd_5A_11	mVTLEIRLQDEQEAERLLQRIKQEWPNATITRTNGTLKIDSDTEQKA ERMEKQIRKQDPDATITRTGgslehhhhhh
Fd_5A_12	mVRITVRSSDKERLDKIRDDIERRWPKATVTKTNGDLKIQAQTQEDA ERIQKQIRRDDPNATVRTTQgslehhhhhh



This table was given in the same way as **Table 2.10**.

Table 2.12 | Sequences and phylogenetic tree of 10 designs for Fd_7S.

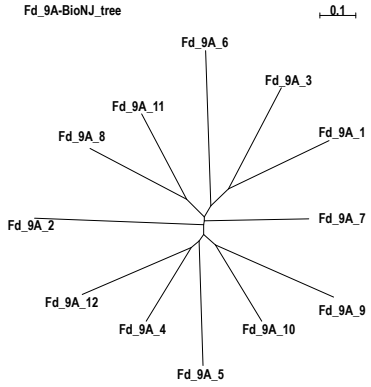
Fd_7S_1	mGTLQIQLTSSSEEEIRRILEKIRKEYPSAQTTETTTNGKWRLRIRSSD EEEVRRVLERLKKEVPSATVRETTTGslehhhhhh
Fd_7S_2	mGTVQIQLKTSSEQEIRRILEHIRKEFPSAQTTETTYNGKWQLRIQTSD KQQIEEVLERLKKEKPSAQVQRTTQGslehhhhhh
Fd_7S_3	mGTYQIQLTSSSEREIRRLLEKIRKEYPSAQSTETTTNGKWKLRIRTSD EKKIREVLDKLRNDVPSAQTRETTTGslehhhhhh
Fd_7S_4	mGKYQIQLTSSSENLIRRILQKIKQERPSVQSTETTQNGKWQLQLRSSD DERVQEMLERLKKEVPSAQVRETTRGslehhhhhh
Fd_7S_5	mGQLTIKIRAENEELFRKLIERLKEERPSSQYTRTDQNGKRQLQITSRS EREVREILDRMRKEVPQAQVQETTQGslehhhhhh
Fd_7S_6	mGQWQIKIYSENEREFRELIERLEEERPSVQYTETTRNGRRQLTIRSND KNEVDRILEEVRRKVPNARVRETETGslehhhhhh
Fd_7S_7	mGQWRLQITSENEEEVRRLIKEIKKERPSVQVTETTQNGRRQLRITSNS EEKFERILDELRREVPSAQTRRTTTGslehhhhhh
Fd_7S_8	mGQWQITIQSSDEELFRRLVERIRRERPEIQVTETTQNGRRQLRIKSRD KNKVEELLKRLREEVPKATVRETETGslehhhhhh
Fd_7S_9	mGQTQLQIYAEDEEKFRRLVEEIRREVPSVQVTETTYNGKWRLQIRSSS DEEIKRVVERVKEEVPSAQTRRTTTGslehhhhhh
Fd_7S_10	mGTYQIQLTSSSEEEIRRILEDIRKEYPNAQSTETTQNGKWTLKIRSSN KEIVERVLQRLREEVPSAQVRETETGslehhhhhh



This table was given in the same way as **Table 2.10**.

Table 2.13 | Sequences and phylogenetic tree of 12 designs for Fd_9A.

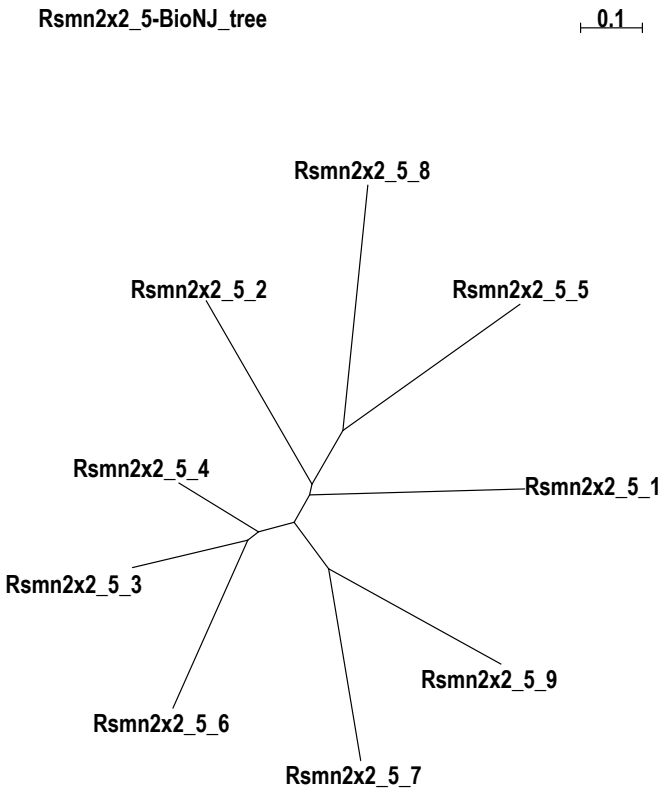
	mVTIEVQVKVKADDRNEAKKIIDEIIKEIEEELRKQRPNVRVTRTVRT
Fd_9A_1	TDGTVQLEIRVKANDKEEAEKLVKQVEEAIERVLKEQKPNATITRTIR
	RTVgswslehhhhhh
	mLTLKLEIEIRADDPDEAKRLVERVAEEVEKQIEKERPNVTVTRQITT
Fd_9A_2	RDGKIKLEVKITAETEDDAKQLVDQIKDEIERRIRKDRPEVRLTRTVK
	KTVgswslehhhhhh
	mLEIEVKLRVKSSDKDEAKNIIENIKEELEKKIRKQRPNARITRTITQ
Fd_9A_3	TDGEVELTIKVKAETEEKVKKLVEELIKEMERKLKEQRPNARITRTIR
	TKVgswslehhhhhh
	mLTVKITIKVTADDKERAEKIVKEIEQELERQVREEFPNARITRTITT
Fd_9A_4	RDGTVELEIKVKAESLDKLKRLIREIEREIERRLKEVDPNARITRTVT
	TEVgswslehhhhhh
	mVKLELEVEVTADSEDDAKRLVEEIEEEIERRVKERYPNARVTRTITQ
Fd_9A_5	EDGRVTLTVKVEAESQEKARELLEEITREIERKLKEQDPNYTVTRTIR
	REVgswslehhhhhh
	mLELEVEIRIRLDNTDEAEEIVKEIAQEIEEEIRKKWPSATVTRTVKT
Fd_9A_6	QDGEVRLTVKIKASSREDVERLREVIEKKVEDVARKRQPNATVTRTIR
	ETVgswslehhhhhh
	mLTIKVQVKIQAEDEDEAKKIVKKISQEVKRRIEDKRPNATITRTIRT
Fd_9A_7	RDGKIELEIKVKAESEQKIEELIKEIEKEIERVAKEEKPNATITRTVT
	REVgswslehhhhhh
	mLQVKIEIRIQADDEREAEKIVEQILKEVEKRVEDNYPNATVTRQITR
Fd_9A_8	TDGEVQLRIKVKAETTEKARKIVEDIEKNIEEVIKKERPNATVTRRVQ
	TEVgswslehhhhhh
	mLKVKLEVRITSDSEEDARKIVKQITDEIDKKLKEKRPNVTITRKIRT
Fd_9A_9	RQGTVELELEIRAESRDDVKELVEELAKEIERVVREQKPNATVTRTIK
	RTVgswslehhhhhh
	mVTVKVEVRITADDENNAEDIIKDVSEEIERRVKEQYPNATITRRITR
Fd_9A_10	RDGTLELEVKVKAESKDKVERLVEELAREIERRARERDPNVTITRTKR
	KTVgswslehhhhhh
	mLTVEVEVKITADDENKAEEIVKRVIDEVEREVQKQYPNATITRTLTR
Fd_9A_11	DDGTVELRIKVKADTEEKAKSIIKLIEERIEEELRKRDPNATITRTVR
	TEVgswslehhhhhh
	mVTVQLKIEIRADSEEKAEKIAKEIEKEIERVVRDQLPNARVTRTITR
Fd_9A_12	RDGTVQLEVRITANDEETVEELIKRIARDIERVLKKVDPNVTITRTVT
	RRVgswslehhhhhh



This table was given in the same way as **Table 2.10**.

Table 2.14 | Sequences and phylogenetic tree of 9 designs for Rsmn2x2_5.

RSMN2x2_5_1	mgIVVVIVTTEEEERRVKERVPKADVLRVTTKEEAEKVAEKLKRKGVQYVVFVGV DDNIIDEIKQRANVQVRRVDDENKLKDVVEKLLGsglehhhhhh
RSMN2x2_5_2	mgQVVVIVGSDEYKRKVEEEYPNVEVRQVTTREKASRVAEEIKKRGITKVVVVGVS EDIIEEIRERANVQVYRVRTKDELKDVLNRLQGsglehhhhhh
RSMN2x2_5_3	mgKLVVVVSSQEEAKKVQEKVPNAQVRLVTTEEDAEKVAEEIRKRGIQNVVFVGV DEDLIKKIKQEANVQVYRVTSEDDLEKIVKDLQGsglehhhhhh
RSMN2x2_5_4	mgRVLVIVGSEELKKKVEEKVPDVEVRRVTTEEDAKKVAKELRERGVQYVVFVGV DEKIIKKIEQEANVQVRRVTSEDDLEKIVKKLNGsglehhhhhh
RSMN2x2_5_5	mgQLVVLVSNEDYKRVAEEVDPNVQVRDVTSKEQAKQVAEELEKRGVQYVIVVNV DDEIVREIEQRANVRVVQVDDEEKLREKIEKLQGsglehhhhhh
RSMN2x2_5_6	mgRLVVVVTSEQLKEEVRKKFPQVEVRLVTTEEDAKQVIKEIQKKGVQKVVLVGVS EKLLQKIKQEANVQVYRVTSNDELEQVVKDVKGsglehhhhhh
RSMN2x2_5_7	mgTVVVLVSTEELERRVKKKVPNVKVRRFETEEDIKKIIKELKEEGVQRVVVVGLD EERVQRIRQQANVDVYEVRSEDRLEDHLKNVQGsglehhhhhh
RSMN2x2_5_8	mgQLLVVVSSDDYKEAVERVDPNIRVLTATTSEDIKEIAKRVQKEGVQRVVVVGVD KNRIEKLREEANVRVIDVDSRDKLREKVEELRGsglehhhhhh
RSMN2x2_5_9	mgRVLIVVKTEDLEKKVKEEVPEIEVRRITTEEDLKKIVKEIEEKGIQVVVVVGVDE EKIKRIRQQANVRVTQVNDDDELKEVVRKVRGsglehhhhhh



This table was given in the same way as **Table 2.10**.

SECTION 3. CYCLIC OLIGOMER DESIGN WITH BACKBONE REMODELED DE NOVO αβ-PROTEINS

ABSTRACT

We have previously shown that monomeric globular αβ- proteins can be designed de novo with considerable control over topology, size and shape. In this section, we investigate the design of cyclic homo-oligomers from these starting points. We experimented with both keeping the original monomer backbones fixed during the cyclic docking and design process, and allowing the backbone of the monomer to conform to that of adjacent subunits in the homo-oligomer. The latter flexible backbone protocol generated designs with shape complementarity approaching that of native homo-oligomers, but experimental characterization showed that the fixed backbone designs were more stable and less aggregation prone. C2 homo-oligomers with β- strand backbone interactions were designed using both fixed and flexible backbone protocols, and found experimentally to adopt homodimeric structures close to the design models. In contrast, C3-C5 designed homo-oligomers with primarily nonpolar residues at interfaces all formed a range of oligomeric states. Taken together, our results suggest that for homo-oligomers formed from globular building blocks, improved structural specificity will be better achieved using monomers pre-organized for shape complementary subunit-subunit interactions and more polar interfaces.

INTRODUCTION

Globular αβ- protein homo-oligomers play important roles in nature, including molecular machines(11,12) , catalysis(13,14) , and regulation(15,16). Considerable control over αβ- protein

monomer topology and shape has been achieved with de novo protein design, but incorporating sequence features that specify a particular oligomerization state is a further challenge. Previous homo-oligomer design efforts have focused primarily on all α- proteins with non-globular structures such as coiled coils(21,58–63) and repeat proteins(64). Compared to oligomers made from elongated helical bundles, homo-oligomeric structures made from globular building blocks have the advantage of multiple reconfigurable interfaces associated with subunit rotations along multiple axes.

The design of homo-oligomers using αβ- proteins has been challenging, likely at least in part because a significant fraction of the interface will generally be involved. For example, Huang et al starting with protein G obtained a mixture of species(65). The best results have been obtained with strand-strand interfaces as in a computationally designed β- sandwich homodimer(66). Here we explore the design of a wide range of homo-oligomeric geometries starting from de novo αβ- proteins.

RESULTS

Starting with previously described de novo designed αβ- proteins(9,67), we experimented with both fixed and flexible backbone methods for designing cyclic homo-oligomers.

Fixed-backbone oligomer design

Each of the de novo designed αβ- proteins were docked into C_n oligomer conformations by repeatedly (1) applying a random rotation to the monomer, (2) sliding the monomers together until they come into contact and (3) optimizing the identities and conformation of residues within

10Å of newly formed interfaces to minimize the binding energy using Rosetta Monte Carlo sequence design calculations (see **SECTION 4.)**(9). During the sequence design calculations, the energy is minimized with respect to the backbone and rigid body degrees of freedom but the changes in both are generally quite small The thousands of alternative dock+designs generated using this procedure were ranked based shape complementarity(68), Rosetta binding energy(69), and the requirement that buried polar groups make hydrogen bonds(59) (**Fig. 3.1**). The structural specificity of the top ranked designs was evaluated by carrying out multiple independent Rosetta symmetric docking(70) calculations, and those with energy landscapes strongly funneling into the designed homo-oligomer conformation were selected for experimental characterization.

We used this fixed backbone protocol to design cyclic oligomers from de novo designed ferredoxin folds with a range of size and shapes(67). We obtained synthetic genes for 8 designed homodimers built from Fd_7A, 6 designed trimers built from Fd_7A_v1, and 8 designed tetramers built Fd_9A (**Table 4.8-4.10**). The different oligomer designs built from the same monomer differ in sequence and in rigid body orientation. Synthetic genes were cloned into pET21 or pET29 E. coli expression vectors, and the designs were expressed, purified by immobilized metal affinity column (Ni-NTA) and characterized by circular dichroism (CD) spectroscopy and size exclusion chromatography combined with multi-angle light scattering (SEC-MALS.). We use the following naming convention: the oligomerization state (C2, C3, C4) is followed by the name of the monomer design and then by the number of the design in the series: for example, C2_Fd_7A_8 is the 8^{th} dimer design built from the Fd_7A monomer.

Of the 22 designs, all but 2 (both from the C4_Fd_9A set) had CD spectra expected for αβ-proteins, suggesting that the many (20-26) amino acid residue changes made to create the designed interface did not disrupt the monomer fold. However, the SEC-MALS results indicated that only 6 of 8 C2_Fd_7A designs had the molecular weights expected for the designed homo-oligomers, and all of the C3_Fd_7A_v1 and C4_Fd_9A designs were polydisperse with multiple alternative oligomer conformations. **Fig. 3.2B-3.2E** shows experimental data for one of the best-behaved designs from each of the three groups (C2, C3, C4); the data for the remaining designs are in **Table 4.1-4.3**.

We succeeded in solving the crystal structure of design C2_Fd_7A_8 (PDB:4PWW) at 1.47Å resolution. The crystal structure reveals a dimer very similar to the design model (backbone RMSD 1.2Å; **Fig. 3.3, Table 4.15**). The helices at the interface twist around each other in both design and crystal structure, but the extent of supercoiling is more significant in the crystal structure resulting in a more shape complementary interface. The extent of twisting of the helix brought about by the backbone minimization step during the design calculations is indicated in the comparison to the original design model on the right panel in **Fig. 3.3**.

Naturally-occurring αβ-protein oligomers have higher interface shape complementarity and area

An obvious limitation of fixed backbone approaches is that the shape complementarity between subunits is limited by the fixed backbone of the monomer. The increase in shape complementarity observed in the Fd_7A_8 crystal structure suggested that the alternative oligomeric states observed in the C3 and C4 designs possibly resulted from insufficiently shape

complementary interfaces. To determine whether the shape complementarity of the designs could be a contributor to the lack of success of the larger homo-oligomers, we compared them to naturally-occurring αβ- homo-oligomers.

We found that naturally-occurring αβ- cyclic homo-oligomers from the PDB(71) (see **SECTION 4.**) generally have higher backbone (sequence independent) shape complementarity than the fixed backbone C3 and C4 homo-oligomer docked configurations that the experimentally characterized designs were based on. In addition, the naturally-occurring oligomers have larger interface areas due to the larger size of the subunits and high fraction of monomer surface area involving in the oligomer interfaces (**Fig. 3.4A-3.4C**).

With the backbone movement resulting in super-helix-like helical interface observed with crystal structure of C2_Fd_7A_8, we hypothesized that improved oligomeric interaction specificity could be achieved using backbone (N, NH, C, Cα, CO and Cβ) remodeling to increase the surface area and shape complementarity(68) of the designed interface.

Flexible-backbone oligomer design

To remodel the backbone geometry of the monomers to increase shape complementarity between subunits, we used a flexible-backbone design method combining Rosetta folding simulations(9) with oligomer rigid body sampling. In the first step, fixed backbone C_N docking calculations were carried out to identify potential oligomer interfaces. Segments of the monomer at the oligomer interface were then subjected to Rosetta remodeling in a Monte Carlo flexible docking trajectory in which small rigid body moves are alternated with broken chain remodeling of a

randomly selected interface segment followed by loop closure. For example, the helices and the flanking loops at the interface of both C2 and C5 oligomers were selected for further backbone remodel in **Fig. 3.1**(72–74). In the backbone remodeling step, the backbone torsion angles of randomly selected interface segments were replaced by fragments of the same length from the PDB (See **SECTION 4.**). The flexible-backbone protocol increased the sequence-independent backbone shape complementarity and interface area significantly beyond the fixed backbone designs to close to that observed in the native complexes (**Fig. 3.4A-3.4C**).

We used the flexible-backbone method to design C3 trimers based on Fd_7A (C3_Fd_7A_v2, building block: 2KL8) with β-strands and alpha helices at the interface, C4 tetramers of Rsmn2x2_6 (C4_Rsmn2x2_6, building block: 2KPO), C5 pentamer of Rsmn2x2_6 (C5_Rsmn2x2_6, building block: 2KPO) and a de novo designed C2 dimer with an extended sheet interface as in the C2_Fd_7A_8 crystal structure (we call this design CFR because it resembles the structure of a C terminal fragment of Top7(75)). We chose C4_Rsmn2x2_6 over C4_Fd_9A for flexible-backbone design because it has a larger core and is more stable and hence it can likely better maintain the overall fold even with substantial backbone remodeling. To generate C5 homo-oligomers with high shape complementarity, we truncated α4 of 2KPO and remodeled α1 to interact with α2 and α3 of a neighboring subunit (C5_Rsmn2x2_6; **Fig. 3.2A**).

Oligomer conformations with backbone interface area larger then 240Å^2 were selected for sequence design. Residues at the interface and within 8Å of remodeled segments were redesigned to optimize both monomer stability and interactions across the oligomer interface(10,

65). Overall, designs made with the flexible backbone protocol had higher backbone and sidechain shape complementarity and interface surface than those made with the fixed backbone protocol starting from the same building blocks (**Fig. 3.4D-3.4F**).

For each design, to assess in silico the folding of the monomeric building block (perturbed more than in the fixed backbone case as both the sequence and the structure differ from the starting design), we carried out multiple independent Rosetta *ab initio* structure prediction calculations(30) starting from an extended chain. Designs with energy landscapes funneled into the remodeled monomer structure were then subjected to Rosetta symmetric docking calculations to assess the designed homo-oligomeric interface(70).

Genes encoding designs with docking energy landscapes strongly funneled into the design target conformation were obtained for experimental characterization; these include 10 designs for CFR, 18 for C3_Fd_7A_v2, 12 for C4_Rsmn2x2_6 and 8 for C5_Rsmn2x2_6 (**Table 4.11-4.14**). Solubly expressed designs were, as in the fixed backbone experiments, characterized with CD spectroscopy and SEC-MALS after Ni-NTA purification. The computational model and experimental results of the design with the highest thermal stability for each target oligomer conformation are shown in **Fig. 3.2**. Most of the C3_Fd_7A_v2 and C4_Rsmn2x2_6 designs were soluble and had the expected far-UV CD spectra, but α4 truncation of 2KPO appeared to decrease tertiary structure stability as only one C5_Rsmn2x2_6 design had an αβ-protein CD spectrum at 25°C (**Table 4.5-4.7**). Unfortunately, the flexible backbone design protocol did not solve the polydispersity problem; multiple species were again observed for all the C3-C5 solubly expressed designs.

More success was observed with the C2 flexible backbone design. Design 10 of CFR (CFR_10) had the CD spectrum of an αβ-protein and had the expected molecular weight by SEC-MALS (**Fig. 3.2**). We were unable to crystalize CFR_10 to compare with the design model, but SEC-MALS indicated dimerization of CFR_10 and the solution small-angle X-ray scattering (SAXS)(76,77) profile was consistent with the design model (χ=1.64, **Fig. 4.1**)(78,79).

DISCUSSION

Taken together, we can draw several conclusions from the successes and failures described in this section in designing assemblies of ~100-residue de novo αβ- protein with fixed- and flexible- backbone design methods and using αα-, αβ-, and ββ- interfaces.

First, success with C2_Fd_7A and CFR suggested the robustness of designed interfaces with extended strand-strand interactions forming an extended beta sheet, as found by Stranges and Kuhlman(66). Our recently described homo-tetrameric TIM barrel also involves a beta sheet extending across the interface(80). Pre-organized β- strands with exposed backbone amine and carbonyl groups allow strong association between subunits without introduction of large hydrophobic patches and hence partially circumvent the tradeoff between subunit solubility and interface stability. For interface geometries where backbone beta strand pairing is not possible, extensive designed sidechain polar hydrogen bonding networks could increase structural specificity as observed for two ring helical bundles(59).

Second, flexible backbone methods can generate assemblies with subunit-subunit interfaces having shape complementarity in the range of native complexes, and with the monomers predicted to fold to the intended subunit structures. In contrast, the shape complementarity of assemblies for C3-C5 generated using fixed backbone methods is generally quite a bit lower than that of native complexes.

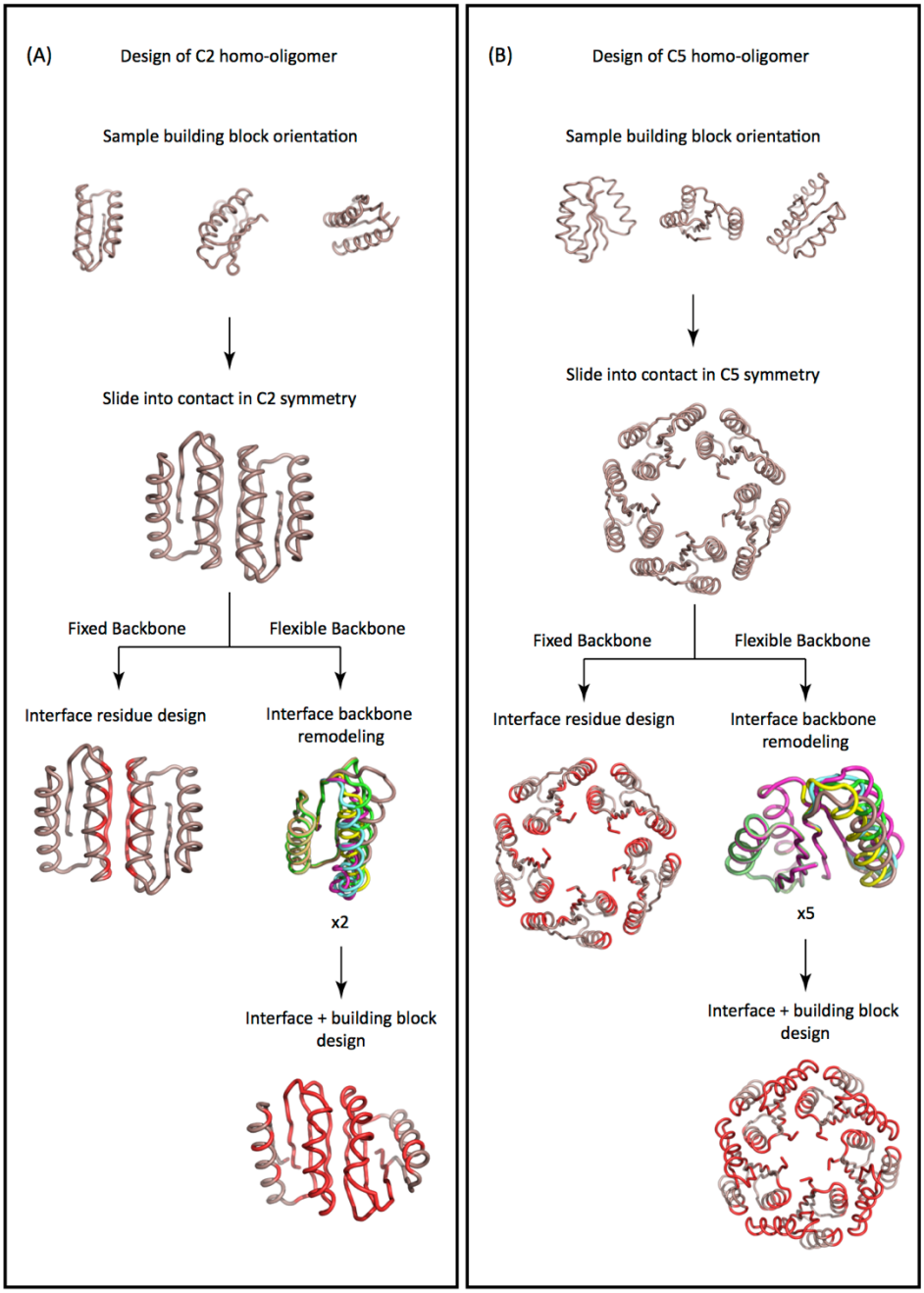
Third, despite the good in silico metrics of the flexible backbone C3-C5 designs, it is difficult to control the precise oligomerization state. The flexible backbone designs were readily expressed and purified, but were not monodisperse. To gain insight into the origins of these shortcomings, we compared our designs with naturally-occurring homo-oligomers and found that the latter generally have more polar interactions across the interface (**Fig. 3.4G-3.4I**). More extensive negative design could also improve success rates, backbone remodeling may result in flexibility that is consistent with oligomerization states beyond the design target.

The described de novo designed αβ- proteins are small globular proteins, and redesigning a large fraction of the surface residue for non-polar subunit-subunit interactions could impact monomer solubility. Loss of characteristic beta strand and alpha helical hydrophobic-polar patterning upon introduction of the subunit-subunit interface could disrupt monomer folding (and perhaps lead to the observed aggregates and multiple oligomeric states). One approach to maintain building block stability while designing protein-protein interaction could be addition of small extra elements for interaction as frequently observed in native homo-oligomeric protein structures(81). Such inserted or appended structural elements can likely be optimized for interface shape complementarity without disturbing the stability of original building block.

In summary, for flexible backbone homo-oligomer design, how to balance interactions across the homo-oligomer interface with monomer foldability and stability remains an outstanding challenge. On the computational side, the flexible backbone interface design problem is closely related to long studied and similarly challenging problem-flexible backbone protein-protein docking. The very large size of the joint search space (rigid body degrees of freedom X internal monomer degrees of freedom) make comprehensive sampling and robust identification of the global energy minimum quite challenging.

FIGURES

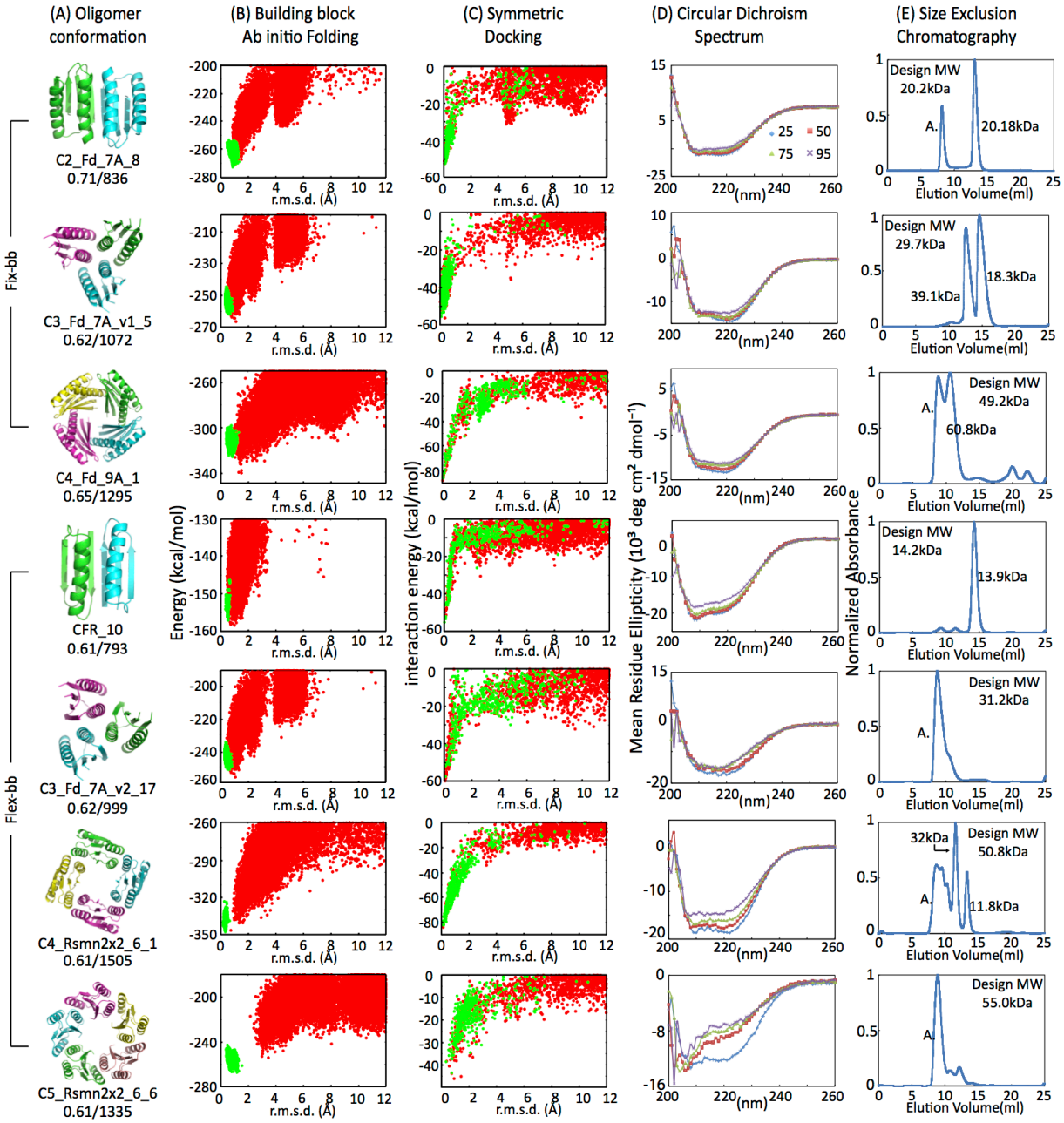
Figure 3.1



Design protocol.

Schematic of fixed and flexible backbone design protocols for (**Left panel**) C2 homooligomers based on design Fd_7A and (**Right panel**) C5 homooligomers based on design Rsm2x2_6.

Figure 3.2

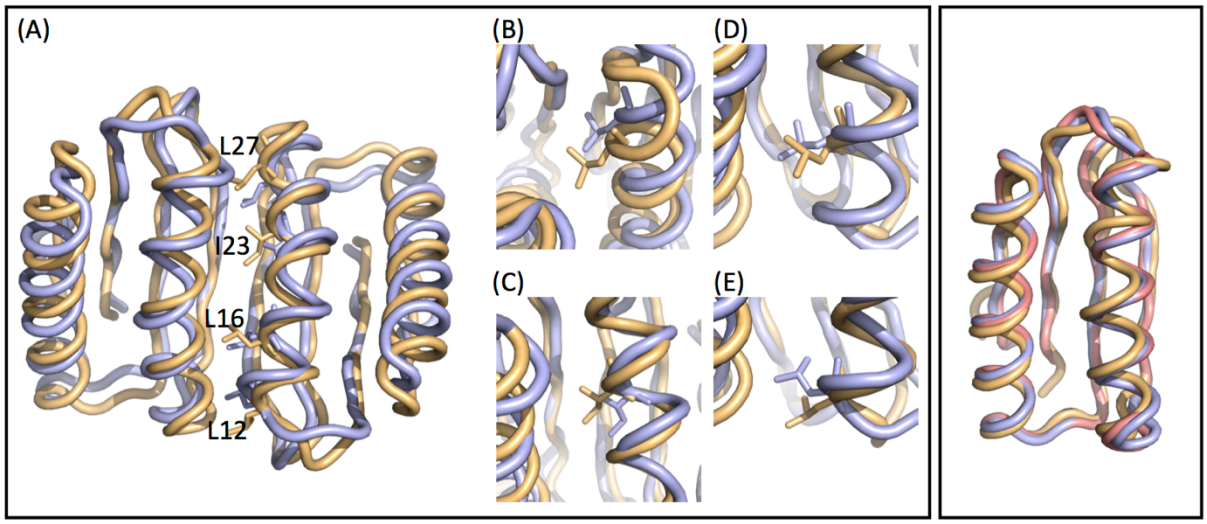


Characterization of computational designed oligomers

(A) Design models of the best design of each oligomer conformation grouped by design methods, fixed-backbone (fix-bb) and flexible-backbone (flex-bb). Numbers below names are shape complementarity/interface area calculated with RosettaScript. (B) Energy landscapes

obtained from Rosetta *ab initio* structure prediction simulations on Rosetta@home. Red points represent the lowest-energy structures obtained in independent Monte Carlo structure prediction trajectories starting from an extended chain for each sequence; y axis, Rosetta all-atom energy; x axis, Cα root mean square deviation (RMSD) from the design model. Green points represent the lowest-energy structures obtained in trajectories starting from the design model. (C) Energy landscapes obtained from Rosetta symmetric docking. Red points represent the lowest-energy docking conformations result from independent global sampling docking trajectories. X-axis: Rosetta interaction energy. Y-axis: Cα root mean square deviation (RMSD) from the design oligomer conformation. Green points represent the lowest-energy structures obtained from local sampling docking trajectories. (D) The far-ultraviolet circular dichroism (CD) spectra at various temperatures. (E) Size-exclusion chromatography spectra with molecular weight determined through multi-angle light scattering (MALS).

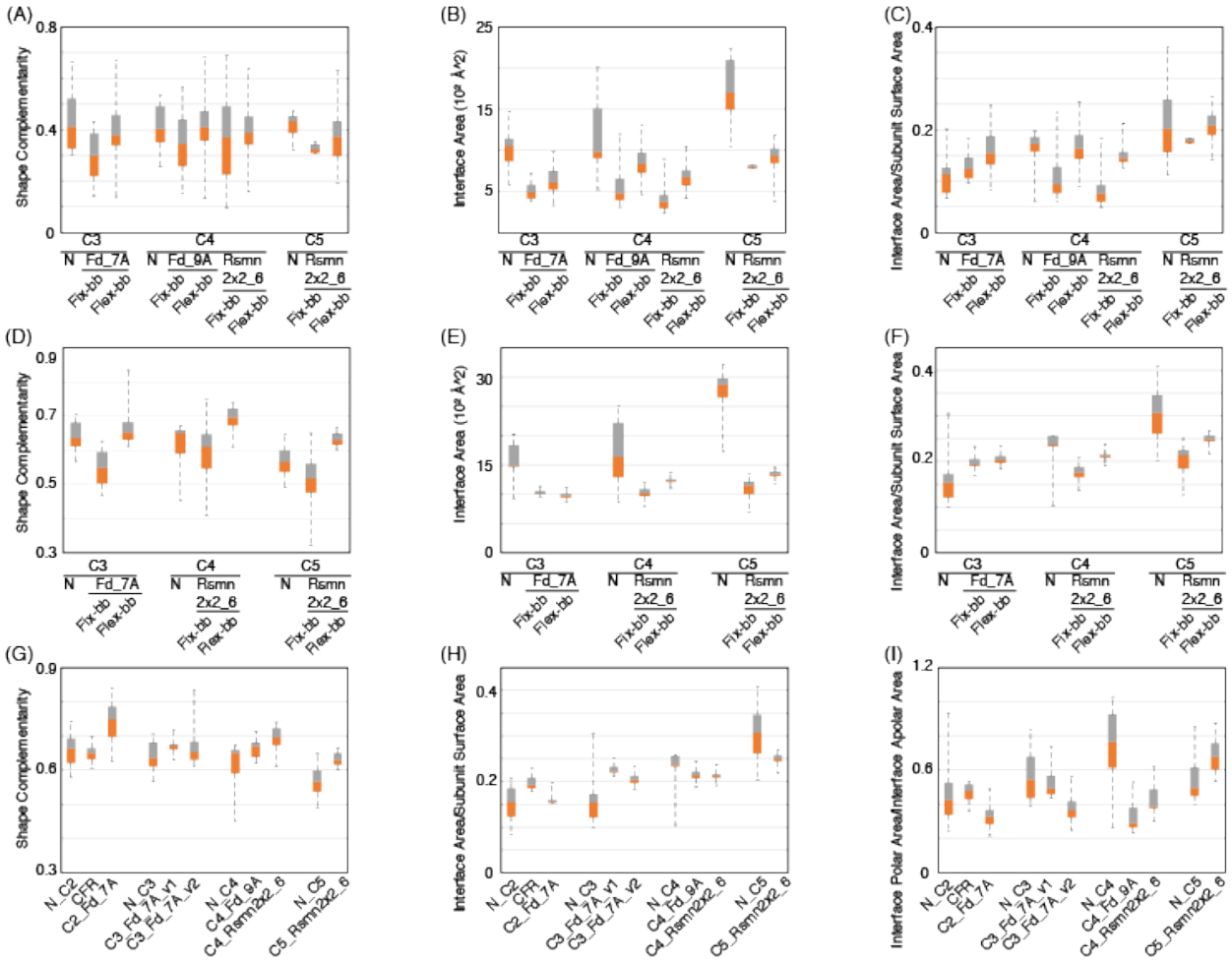
Figure 3.3



Crystal structure of design C2_Fd_7A_8.

(**Left panel**) (A) Computational design model (yellow) aligned with crystal structure (light blue) with aliphatic residues at interface from one monomer shown. Note the helix at interface was designed to bend toward the interacting helix in design model and curvature of bending is even more significant in crystal structure. (B) (C) (D) (E) Close-up view of L27, I23, L16 and L12 respectively. Due to curvature of helix bending, side chains are packed deeper into interface core in crystal structure. (**Right panel**) Backbone superimpose of Fd_7A (pink), chain A of computational design model C2_Fd_7A_8 (yellow) and crystal structure (light blue).

Figure 3.4



Comparison of interfaces in naturally occurring and fixed and flexible backbone designed homo-oligomers.

(A-C) Evaluation of backbone complementarity, interface area and the ratio of interface area between neighboring subunits and total surface area per subunit between naturally-occurring oligomers (N) and cyclic conformations sampled utilizing different de novo building blocks. Cyclic homo-oligomer conformations were generated with either fixed-backbone (fix-bb) or flexible-bb (flex-bb) design method. Only backbone atoms were included in all calculations. (D-F) Same evaluation as (A-C) but including all atoms rather than just the backbone. (G-H) Same analysis as D and F for all of the experimentally characterized designs. (I) Evaluation of interface polarity for all the experimentally characterized designs.

SECTION 4. SUPPLEMENTAL INFORMATION AND METHOD FOR CYCLIC OLIGOMER DESIGN WITH BACKBONE REMODELED DE NOVO αβ-PROTEINS

SI AND METHODS

Computational design

For C2_Fd_7A, C3_Fd_7A_v1, C4_Fd_9A, C3_Fd_7A_v2, C4_Rsmn2x2_6 and C5_Rsmn2x2_6, de novo monomeric proteins were used as initial building blocks for symmetrical cyclic oligomer design. In order to model cyclic symmetry within Rosetta framework, symmetric definition files are required to generate models in desired(82). In each sampling trajectory of fixed-backbone design method for C_N symmetry, the subunit was initially having the cyclic axis aligned to the vector [0,0,1] and perturbed by a translation perpendicular to the axis of symmetry and a random rotation in three-dimensional space. The applied perturbation was selected from a Gaussian distribution bounded by user-defined distances and angles. N copies of subunit then slide into contact based on the information provided by symmetry definition file provided.

For flexible-backbone design, after potential oligomer interface was identified with fixed-backbone C_N docking calculations, segments at the interface were then subjected to Rosetta remodeling in with flexible-backbone design method. Each homo-oligomer sampling trajectory required a monomeric building block aligned to the oligomer symmetry axis that aligned to vector [0,0,1] and a blueprint file specifying segments for backbone remodel. In these Monte Carlo calculations, backbone fragment assembly at the specified segments followed by loop

closure alternated with small rigid body moves that symmetrize oligomer assembly with slide-into-contact method according to symmetry defined. Monte Carlo simulation was performed at temperature 1.0 and the rigid body perturbation was selected from a Gaussian distribution bounded by 0.7Å in translation and 2 degrees in rotation. Backbone fragment set consisting of 1, 3 or 9 consecutive residue fragments were prepared in advance from a non-redundant set of X-ray structures; the fragments have information only on the phi, psi and omega torsion angles. In each attempted Monte Carlo trial, a new backbone conformation was generated by replacing the torsion angles of a randomly selected interface segment consisting of 1, 3 or 9 consecutive residues with the torsion angles of a randomly selected fragment compatible with the assigned secondary structure(9,73-75). CFR had the complete backbone structure assembled with fragment-based structural building based on blueprints with secondary structure lengths, loop geometries and β-strand register shifts specified.

Various docking conformations were scored based on number of Cβ in contact between subunits, where a contact is defined as two Cβ are within 10Å. Top 30 conformations sampled for each building block had residues at interface designed and optimized using Monte Carlo simulations. Design protocol initially designed with soft repulsive score term and then the positions designed were allowed to minimize side-chain torsion angles. This design+minimize cycle was repeated four times with increasing repulsive score term and the standard score term was used in final cycle to obtain a sequence that correspond to the local minimum of the energy function. The designs were then filtered through Rosetta energy score terms, including binding energy(69) where the difference between the Rosetta energy of the bound (oligomeric) and unbound (monomeric) states less than −20.0, shape complementarity(68) greater than 0.6 and no buried

unsatisfied charged residues(59).

Comparison of shape complementarity and interface area between naturally-occurring and computational generated αβ- cyclic homo-oligomers

Shape complementarity and interface area (solvent accessible surface area) were computed by RosettaScript. The regions to calculate solvent accessible surface area can be assigned as interface, polar residues, non-polar residues or the subunit of oligomer. Shape complementarity calculates Lawrence & Coleman shape complementarity(68).

PDB ID of naturally-occurring αβ- cyclic homo-oligomers selected for analyses(71)

C2 symmetry: 1ID1, 1LSS, 1NN5, 1PQW, 2AK4, 2APX, 2BDT, 2JJ8, 2OP5, 2ZDP, 3CZR and 3DHX

C3 symmetry: 2CZ4, 2ZFH,1C9K,1KHT,1VIY,1VL0 and 2P2L

C4 symmetry: 1DSX, 1J8D, 1L3A, 1S1G, and 6Q21

C5 symmetry: 1JG5, 1N3R, 1T0T, 1VR4, 1W8S, 1Y5Y, and 1ZIS

Dataset for sequence-independent (backbone only) analyses

Cyclic homo-oligomer conformations generated with fixed- or flexible- backbone design method were turned into poly-valine model and conformations with no heavy atom clashing (distance below 3.5Å was considered clashing and 2.6Å for backbone carbonyl oxygens and nitrogens) and interface size larger than 240 $Å^2$ were selected for analyses. Poly-valine interface of selected naturally-occurring oligomers were also calculated for sequence independent shape complementarity and interface area.

Dataset for backbone and sidechains analyses

Designs in flexible-backbone group were the exact designs ordered for experimental characterizations (C3_Fd_7A_v2, C4_Rsmn2x2_6 and C5_Rsmn2x2_6). Oligomer docking conformations in fixed-backbone group were generated with PDB 2KL8, 2KPO, and α4-truncated 2KPO aligned to backbone of C3_Fd_7A_v2, C4_Rsmn2x2_6 and C5_Rsmn2x2_6. Sequences at interface were then designed as described in the previous section.

Protein expression and purification

For each CFR, C2_Fd_7A, C3_Fd_7A_v1, C3_Fd_7A_v2, C4_Rsmn2x2_6 and C5_Rsmn2x2_6 designed sequence, a spacer was added at the C-terminus in order to separate the designed region and the C-terminal 6xHis-tag (SWG for CFR, GSWS for C2_Fd_7A, C3_Fd_7A_v2 and C4_Rsmn2x2_6, GWS for C5_Rsmn2x2_6, GSSWS for C3_Fd_7A_v1.) For each C4_Fd_9A designed sequence, a SWSG spacer was added at the N-terminus in order to separate the designed region and the N-terminal 6xHis-tag. The genes encoded designed sequences of CFR, C2_Fd_7A, C3_Fd_7A_v2 (cloned into plasmid pET21), RSMN4 and RSMN5 (cloned into pET29) were obtained from GenScript. The double stranded DNA fragments encoding and C3_Fd_7A_v2 and C4_Fd_9A designed sequences were obtained from IDT and cloned into pET29 vector digested with NdeI and XhoI restriction enzymes at 37 °C for 4 hours by Gibson Assembly Cloning method. Cloned-pET29 plasmids were then transformed into XL1-Blue cells and sequences were verified by GENEWIZ.

The designed proteins were expressed in *E. coli* BL21 Star (DE3) cells induced with IPTG for 18 hours at 18°C. Cell lysis was performed in TBS (25 mM Tris-HCl, 150mM NaCl, pH8.0) with sonication at 20W for 3 minutes total on time, using 10/10 seconds on/off rounds. After lysis and

centrifugation at 20,000xG for 30 minutes, the expressed proteins with a 6xHis-tag were purified through a nickel affinity column (Ni-NTA, Qiagen.) The purified proteins were then dialyzed against typical TBS buffer at pH 8.0; this buffer was used for all the experiments except crystal structure determination. The expression, solubility, and purity of the designed proteins were assessed by SDS-PAGE and mass spectrometry (TSQ LC/MS, Thermo Scientific).

Circular dichroism (CD)

All CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD spectra of designed proteins were measured from 260 to 200 nm for 14-28 uM protein samples in TBS buffer (pH8.0) at various temperatures of 25, 50, 75, and 95 °C in a 1 mm path length cuvette. The protein concentrations were determined from the absorbance at 280 nm3 using UV spectrophotometer (NanoDrop, Thermo Scientific).

Size exclusion chromatography combined with multi-angle light scattering (SEC-MALS)

SEC-MALS experiments were performed using a miniDAWN TREOS static light scattering detector (Wyatt Technology) combined with a HPLC system (LC 1200 Series, Agilent Technologies). The volume 100 ul of 400-700 uM protein samples in TBS buffer (pH 8.0) was injected into a Superdex 75 Increase 10/300 GL column (GE Healthcare) equilibrated with TBS buffer at a flow rate of 0.5 ml/min. The protein concentrations were calculated from the absorbance at 280 nm detected by the HPLC system. Static light scattering data were collected at three different angles, 41.4°, 90.0°, and 138.6°, at 658 nm. These data were analyzed by the ASTRA software (version 5.3.4, Wyatt Technology) with a change in the refractive index with concentration, a *dn/dc* value, 0.185 ml/g.

Small-angle X-ray Scattering (SAXS)

CFR_10 was re-expressed and purified for low-resolution structure determination while in solution by small-angle X-ray scattering (SAXS). Purified size-exclusion chromatography eluted sample and concentrated sample of CFR_10 were sent for data collection at the SIBYLS High Throughput SAXS Advanced Light Source in Berkeley, California(77). Experimental diffraction data was then analyzed with java-based application, ScÅtter. Minimum q value (q_{min}) was determined by Guinier analysis. Data resolution, reflected by maximum q value (q_{max}), was determined by characteristic asymptote in signal intensity described by Porod's Law(83). Refined data set and corresponding computational design model were input to the FoXS server to compute the agreement (evaluated as χ) between the experimental and model-computed profile(79).

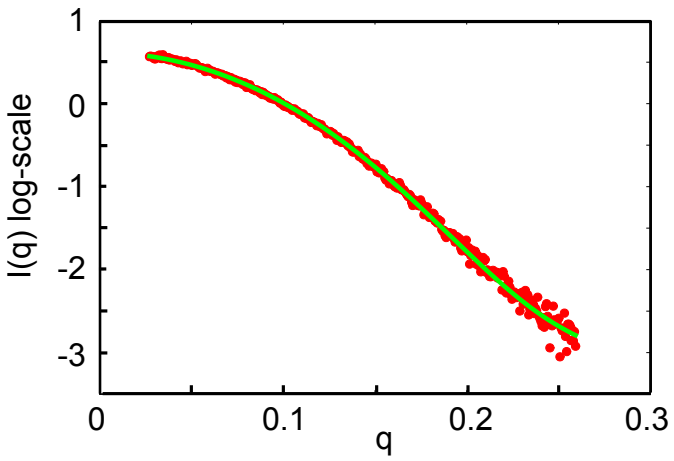
Crystallization, Data Collection, and Structure Refinement

Crystallization screening was performed using a microbatch-under-oil crystallization method at 18°C(84). After optimization, OR494 crystals useful for structure determination were grown in drops composed of 1.0 µL of protein and 1.0 µL of precipitant solution [2.0 M ammonium sulfate, 0.1 M citric acid, pH 3.5] under paraffin oil (Hampton Research). The crystals were cryoprotected with 15% ethylene glycol prior to flash-freezing in liquid nitrogen for data collection. A data set was data collected at beamline X4C at the National Synchrotron Light Source (λ = 0.97907 Å). The diffraction data from single crystal was processed with the HKL2000 package(85).

The structure was solved by molecular replacement using program BALBES(86). A structure of computational design protein (PDB ID 4KY3) was used as a searching model and the OR494 model was completed using iterative cycles of manual rebuilding in Coot(87). The OR494 structure was refined against 1.47 Å data with the program PHENIX(88). The data processing and refinement statistics for the crystal structure determination are summarized in Table S15. The quality of the final structure was assessed using PROCHECK(89). The atomic coordinates and structure factors are available in the Protein Data Bank under accession code 4PWW.

FIGURES

Figure 4.1



Comparison between CFR SAXS profiles.

Experimentally measured SAXS data are in red and SAXS profile computed from CFR design model with FoXS server (see **SI and Methods**) is in green.

TABLES

Table 4.1 | Summary of C2_Fd_7A designs

C2_Fd_7A Design number	Soluble	αβ protein	SEC-MALS MW matches expected MW
Design 1	❖	✓	✚
Design 2	❖	✓	✓
Design 3	❖	✓	✓
Design 4	✓	✓	✓
Design 5	✓	✓	✓
Design 6	✓	✓	✓
Design 7	◆	✓	✚
Design 8	✓	✓	✓

◆ Low concentration

❖ Gradually become insoluble at 4ºC storage

✚ Did not conduct the experiment due to low concentration

Table 4.2 | Summary of C3_Fd_7A_v1 designs

C3_Fd_7A_v1 Design number	Soluble	αβ protein	SEC-MALS MW matches expected MW
Design 1	✓	✓	
Design 2	✓	✓	
Design 3	❖	✓	
Design 4	❖	✓	
Design 5	✓	✓	
Design 6	✓	✓	

❖ Gradually become insoluble at 4ºC storage

Table 4.3 | Summary of C4_Fd_9A designs

C4_Fd_9A Design number	Soluble	αβ protein	SEC-MALS MW matches expected MW
Design 1	❖	✓	
Design 2	❖	✓	
Design 3	✓	✓	
Design 4	✓	✓	
Design 5	✓		
Design 6	✓	✓	
Design 7	◆	✚	
Design 8	✓	✓	

◆ Low concentration

❖ Gradually become insoluble at 4ºC storage

✚ Did not conduct the experiment due to low concentration

Table 4.4 | Summary of CFR designs

CFR Design number	Soluble	αβ protein	SEC-MALS MW matches expected MW
Design 1	✓		
Design 2	✓	✓	
Design 3	✓		
Design 4	❖	✓	
Design 5	❖		
Design 6	✓	✓	
Design 7	✓	✓	
Design 8	❖	✓	✚
Design 9	✓	✓	
Design 10	✓	✓	✓

❖ Gradually become insoluble at 4ºC storage

✚ Did not conduct the experiment due to low concentration

Table 4.5 | Summary of C3_Fd_7A_v2 designs

C3_Fd_7A_v2 Design number	Soluble	αβ protein	SEC-MALS MW matches expected MW
Design 1	◆	✚	
Design 2	◆	✚	
Design 3	◆	✚	
Design 4	◆	✚	
Design 5	✓	✓	
Design 6	✓	✓	
Design 7	❖	✚	
Design 8	✓	✓	
Design 9	✓	✓	
Design 10	❖	✓	
Design 11	✓	✓	
Design 12	✓	✓	
Design 13			
Design 14	❖	✚	
Design 15			
Design 16	❖	✚	
Design 17	❖	✓	
Design 18	❖	✓	

◆ Low concentration

❖ Gradually become insoluble at 4ºC storage

✚ Did not conduct the experiment due to low concentration

Table 4.6 | Summary of C4_Rsmn2x2_6 designs

C4_Rsmn2x2_6 Design number	Soluble	αβ protein	SEC-MALS MW matches expected MW
Design 1	✓	✓	
Design 2	✓	✓	
Design 3	❖	✚	
Design 4	✓	✓	
Design 5	❖	✚	
Design 6	❖	✓	
Design 7	✓	✓	
Design 8	✓	✓	
Design 9	✓	✓	
Design 10	✓	✓	
Design 11	✓	✓	
Design 12	✓	✓	

◆ Low concentration

❖ Gradually become insoluble at 4ºC storage

✚ Did not conduct the experiment due to low concentration

Table 4.7 | Summary of C5_Rsmn2x2_6 designs

C5_Rsmn2x2_6 Design number	Soluble	αβ protein	SEC-MALS MW matches expected MW
Design 1	✓		
Design 2	✓		
Design 3	❖	✚	
Design 4	❖	✚	
Design 5	✓		
Design 6	✓	✓	
Design 7	❖		
Design 8	✓		

◆ Low concentration

❖ Gradually become insoluble at 4ºC storage

✚ Did not conduct the experiment due to low concentration

Table 4.8 | Sequences of 8 designs of C2_Fd_7A

Fd_7A	mEMDIRFRGDDLEAFEKALKEMIRQARKFAGTVTYTLDGNDLEIRIT GVPEQVRKELAKEAERLAKEFNITVTYTIRlehhhhhh
Design 1	mEMDIRFRGDDLEALMKALQEMFRQAAKFGATITAKLDGNDLEIRIT GVPEQVRKELAKEAERLAKEFGITVTRTIRgswslehhhhhh
Design 2	mEMDIRFRGDDAEALLKAAAEMIKQAAKFGATIELRWDGNDLEIRIT GVPEQVRKELAKEAERLAKEFGITVTRTIRgswslehhhhhh
Design 3	mEMDIRFRGDDLEALWKALWEMAKQAAKFGATIEARLDGNDLEIRIT GVPEQVRKELAKEAERLAKEFGITVTRTIRgswslehhhhhh
Design 4	mEMDIRFRGDDFEALAKAALEMAKQALKFGATITLSLHGNDLEIRIT GVPEQVRKELAKEAERLAKEFGITVTRTIRgswslehhhhhh
Design 5	mEMDIRFRGDDLDALLKAAAEMIKQALKFGATIELRIEGNDLEIRIT GVPEQVRKELAKEAERLAKEFGITVTRTIRgswslehhhhhh
Design 6	mEMDIRFRGDDLEALMKALQEMAKQAAKFGATIEARLDGNDLEIRIT GVPEQVRKELAKEAERLAKEFGITVTRTIRgswslehhhhhh
Design 7	mEMDIRFRGDDVEALAKALAEMARQAAKFGATIKLELRGNDLEIRIT GVPEQVRKELAKEAERLAKEFGITVTRTIRgswslehhhhhh
Design 8	mEMDIRFRGDDLEALLKAAIEMIKQALKFGATITLSLDGNDLEIRIT GVPEQVRKELAKEAERLAKEFGITVTRTIRgswslehhhhhh

Sequence of Fd_7A is included for comparison. Residues mutated are highlighted in red.

Table 4.9 | Sequences of 6 designs of C3_Fd_7A_v1

Fd_7A	mEMDIRFRGDDLEAFEKALKEMIRQARKFAGTVTYTLDGNDLEIRIT GVPEQVRKELAKEAERLAKEFNITVTYTIRlehhhhhh
Design 1	mEMDIRFRGDDNSALAVAAAAMSLAARNFGATVTRTVDGNDLEIRIT GVPEQVLKELAKEAEKIAKAAGITVTRTIRgsswslehhhhhh
Design 2	mEMDIRFRGDDQSALALAAAVMRLAAENFGATVTETVDGNDLEIRIT GVPEQVLKELAKKAEKLAKIAGITVTRTIRgsswslehhhhhh
Design 3	mEMDIRFRGDDRSALAVAAAAMSILARNFGATVTRTFDGNDLEIRIT GVPEQVLKELAKEAEKLAKAAGITVTRTIRgsswslehhhhhh
Design 4	mEMDIRFRGDDNNALALAALIMSLAARNFGATVTRTDDGNDLEIRIT GVPEQVLKELAKLAEEAAKLAGITVTRTIRgsswslehhhhhh
Design 5	mEMDIRFRGDDTSALALAAAAMDIEARRFGATVTRTVDGNDLEIRIT GVPEQVLKELAKKAEELAKAAGITVTRTIRgsswslehhhhhh
Design 6	mEMDIRFRGDDNQALALAAAAMSAAARDFGATVTRTLDGNDLEIRIT GVPEQVLKELAKEAEKLAKAAGITVTRTIRgsswslehhhhhh

Sequence of Fd_7A is included for comparison. Residues mutated are highlighted in red.

Table 4.10 | Sequences of 8 designs of C4_Fd_9A

Fd_9A	mLTVEVEVKITADDENKAEEIVKRVIDEVEREVQKQYPNATITRTLTRDDGTV ELRIKVKADTEEKAKSIIKLIEERIEEELRKRDPNATITRTVRTEVgsswslehhhhhh
Design 1	mhhhhhhswsgLTVLVVVIITADDENKAEEIVKRVIDEVEREVQKEYPNATI TRTLTRVNGLVVLVIVVKADTRVKALVIMALIVVRIEEELRKRDPNAEIRRVTLTEV
Design 2	mhhhhhhswsgLTVIVIVIIRADDENKAEEIVKRVIDEVEREVQRDYPNATI TRRLTRVNGLVVLVIVVKADTRVKAAIIMVAIVLRIEEELRKRDPNAEITRVTITEV
Design 3	mhhhhhhswsgLTVIVIVVIKADDENKAEEIVKRVIDEVEREVQKNYPNATI TRTLTRVNGVVVLVIKVKADTEEKAAIIAVAIVVRIEEELRKRDPNATIRRTTETKV
Design 4	mhhhhhhswsgLTVIVIVIIRADDENKAEEIVKRVIDEVEREVQRDYPNATI TRRLTRVNGLVVLVIVVKADTRVKAAIIMVAIVLRIEEELRKRDPNAEITRVVITEV
Design 5	mhhhhhhswsgLTVIVVVVITADDENKAEEIVKRVIDEVEREVQKNYPNATI TRTLTRVNGVVVLVIKVKADTEEKAIAIMVAIVVLIEEELRKRDPNATIRRTVSTKV
Design 6	mhhhhhhswsgLTVIVVVIIRADDENKAEEIVKRVIDEVEREVQREYPNATI TRRLTRVNGLVVLVIVVKADTTEKALIIMVAIVLRIEEELRKRDPNAEITRVVLTEV
Design 7	mhhhhhhswsgLTVIVVVVIRADDENKAEEIVKRVIDEVEREVQKNLPNATI TRTLTRVNGMVVLVILVKADTRAKAAAIIVLIVVIIEEELRKRDPNAEIRRVVLVEV
Design 8	mhhhhhhswsgLTVIVVVIITADDENKAEEIVKRVIDEVEREVQRDYPNATI TRRLTRVNGLVVLVIVVKADTTVKAAIIMVAIVVRIEEELRKRDPNAEIRRVVITEV

Sequence of Fd_9A is included for comparison. Residues mutated are highlighted in red.

Table 4.11 | Sequences of 10 designs of CFR

Design 1	mTQVNVKIGMSDEKTANKVAQAVADEVTKDNPNSEVRNRVDGNTVEVEVQG swglehhhhhh
Design 2	mGRVQLTLDITSEELRTKAAKNAVKAAKESFPNLEVTNTVDGTKDTVEVQG swglehhhhhh
Design 3	mNEVEVEVRLSDEETAKQLAREITKKLKEQVPNSEVTNTVDGNKVQVEVQG swglehhhhhh
Design 4	mGQVTVVLDFTSEEEATKSVKEAVKRLKEAWPNLDVENRVDGTKVEVRVQG swglehhhhhh
Design 5	mGRRQVEVNVTSDDTSRKAIKLAKQLNKDEGPNNDVTSTVDGTKVKVQVDG swglehhhhhh
Design 6	mGELEITVNMTSEDAAKNAVKSIVENAKRAWPNLEVTNRVDGDKVEVRVQG swglehhhhhh
Design 7	mSRVTVKFAMSDQKSADRAAKRASESAKEANPNKEVTNTVDGNTVKVEVRG swglehhhhhh
Design 8	mTRIEVRIEVTNEDEARKWAKELAKVVTKLFPESEVTRRVDGNTVEVKVQG swglehhhhhh
Design 9	mKSVNIEVEITSEDKAQEAVDKIVELLKKLFPNLDVENRVDGTKVEVRVQG swglehhhhhh
Design 10	mSTVIVEIRVDDEEQAKQIAKKVEELLKKERPNSEVTNTVDGNTVKVKVQG swglehhhhhh

Table 4.12 | Sequences of 18 designs of C3_Fd_7A_v2

Fd_7A	mEMDIRFRGDDLEAFEKALKEMIRQARKFAGTVTYTLDGNDLEIRITGVPE QVRKELAKEAERLAKEFNITVTYTIRlehhhhhh
Design 1	mEMDIRFRVDDESAFMQAAFIMLAQVVNFAGKFHVERDGNDLEIRITGVPE QVRKELAKEAEYLAKQFGITVTYTIRgswslehhhhhh
Design 2	mEMDIRFRVDDDKAFIKAMILMLAQAVTFAGNFEFTQDGNDLEIRIEGVPE QVRKELAKEAERLAKEFGITVTYTIRgswslehhhhhh
Design 3	mEMDIRFRVDDQEAFIRAMIHMAVQAMNFAGKFEFTSDGNDLEIRITGVPE QVRKELAKEAEREAKRFGITVTYTIRgswslehhhhhh
Design 4	mEMDIRFRVDDENAFIEAMFVMLAQAVSFAGQFEVTRDGNDLEIRINGVPE QVRKELAKEAERLAKRFGITVTYTIRgswslehhhhhh
Design 5	mEMDIRFRVDDSEAFLSAAVMMVVQAFQFAGRFEMTRDGNDLEIRITGVPE QVRKELAKEAERLAKQFGITVTYTIRgswslehhhhhh
Design 6	mEMDIRFRVDDDSAFLEAAILMVVQAMTFAGNFEFTKDGNDLEIRIEGVPE QVRKELAKEAERLAKRFGITVTYTIRgswslehhhhhh
Design 7	mEMDIRFRVDDDRREMQAAFSMLVQAVNFAGKFEMTQDGNDLEIRITGVPE QVRKELAKEAERLAKRFGITVTYTIRgswslehhhhhh
Design 8	mEMDIRFRVDDEDAKSRAAFMMLVQAMNFAGKLEFTSDGNDLEIRITGVPE QVRKELAKEAEREAKEFGITVTYTIRgswslehhhhhh
Design 9	mEMDIRFRVDDDSAFIKAMASMIVQAMNFAGQFEFQQDGNDLEIRITGVPE QVRKELAKEAEKLAKEFGITVTYTIRgswslehhhhhh
Design 10	mEMDIRFRVDDDRREIQAAVSMLAQVMQFAGKFEFTQDGNDLEIRITGVPE QVRKELAKEAERLAKEFGITVTYTIRgswslehhhhhh
Design 11	mEMDIRFRVDDENAFISAAVVMVQQAIEFAGTFKMTRDGNDLEIRITGVPE QVRKELAKEAERLAKEFGITVTYTIRgswslehhhhhh
Design 12	mEMDIRFRVDDDKREMQAAFDMLAQVLKFAGTLEFTQDGNDLEIRITGVPE QVRKELAKEAERLAKEFGITVTYTIRgswslehhhhhh
Design 13	mEMDIRFRVDDENAFLSAAIVMVVQALSFAGQFEMTQDGNDLEIRINGVPE QVRKELAKEAERLAKEFGITVTYTIRgswslehhhhhh
Design 14	mEMDIRFRVDDQSAFLSAAVDMLAQVMSFAGQFEFTQDGNDLEIRINGVPE QVRKELAKEAERLAKEFGITVTYTIRgswslehhhhhh
Design 15	mEMDIRFRVDDKEAFIDAQLIMVAQAMNFAGQVEFTQDGNDHEIRITGVPE QVRKELAKEAERLAKKFGITVTYTIRgswslehhhhhh
Design 16	mEMDIRFRVDDEKAFIEAAIMMVVQAIQFAGKFEFQKDGNDLEIRITGVPE QVRKELAKEAEKLAKEFGITVTYTIRgswslehhhhhh
Design 17	mEMDIRFRVDDENAFINAAMMMLVQAVTFAGNLEFTQDGNDLEIRIDGVPE QVRKELAKEAERLAKEFGITVTYTIRgswslehhhhhh
Design 18	mEMDIRFRVDDKEAFSDAQLSMVAQAIKFAGQVEMTQDGNDHEIRITGVPE QVRKELAKEAERLAKEFGITVTYTIRgswslehhhhhh

Sequence of Fd_7A is included for comparison. Residues mutated are highlighted in red.

Table 4.13 | Sequences of 12 designs of C4_Rsmn2x2_6

Rsmn2x2_6	mLLYVLIISNDKKLIEEARKMAEKANLELRTVKTEDELKKYLEEFRKESQNIKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSPDEAKRWIKEFSEEggslehhhhh h
Design 1	mLLIVLIISNDEALIKAARSVANQANLQLVTVDDEEVLEALLFAARENSQNIKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSFKEAIKWIKEFSEEgswslehhhhh h
Design 2	mLLVVLIISNDKVLILAARLLANQANLELYTVDTEEILKALLFAVRSSDQNIKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSAEEALKWIKEFSEEgswslehhhhh h
Design 3	mLYVLIISNDENLIIAARVAARNNNLDLRTVKTEEFLEVVLLALAIASQNVKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSNDEAAKWVREFSEEgswslehhhh hh
Design 4	mLYVLIISNDEHLIIAVQILAESQNLKLRTVKTEEVLDAALLAARAASQNAKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSTEEALKWLQEFSEEgswslehhhhh h
Design 5	mLLYVLIISNDEDLFIAARVLASDRNLELRTVKTESVLKAILIAVRFADQNAKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSTEEAAKWIKEFSEEgswslehhhhh h
Design 6	mLLIVLIISNDHSLILAARILAEQQNLKLYTVETEKNLEAALLAARQSDQNIKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSFKEAAEWIREFSEEgswslehhhhh h
Design 7	mQLYVLIISNDKHLIVAAQILAKSSNLDLKTVKTEEFLHAALLAIRFADQNAKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSTEEALKWVLEFSEEgswslehhhhh h
Design 8	mLLYVLIISNDEDLILAARLLAKSQNLELRTVKTESALKVLLIAARFADQNTKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSTEEAAKWIKEFSEEgswslehhhhh h
Design 9	mLLYVLIISNDRHLLIAARILAKSQNLDLRTVKTEEVLSAALLAARFADQNAKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSTEEALKWIKEFSEEgswslehhhhh h
Design 10	mLLYVLIISNDEQILVAARLLAESLNLKLRTVKTEEVLKALLQAARWSNQNAKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSAEEAARWLKEFSEEgswslehhhh hh
Design 11	mKLYVLIISNDENLFIAARLLAKSTNLDLRTVKTESVLKALLLAARFADQNAKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSNEEALKWIKEFSEEgswslehhhhh h
Design 12	mRLYVLIISNDKSLIIAARVAAKESNLELETVKTEEILKAVLVALRFVTQNVKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSIEEAAKWVVEFSEEgswslehhhhh h

Sequence of Rsmn2x2_6 is included for comparison. Residues mutated are highlighted in red.

Table 4.14 | Sequences of 8 designs of C5_Rsmn2x2_6

Rsmn2x2_6	mLLYVLIISNDKKLIEEARKMAEKANLELRTVKTEDELKKYLEEFRKESQNIKV LILVSNDEELDKAKELAQKMEIDVRTRKVTSPDEAKRWIKEFSEEggslehhhhh h
Design 1	mLSYVLAIANDKKLEERLREVSRKANTEQRTVKEDEELRKYLEEFRKESQNIKV KILVRDDEKKAAAELLRDSSEIDVETRK~~STSPDEAKRWIKEFSEE~~gwslehhhhhh
Design 2	mLSYVLAIANDKEAEEKLREAARKANTEQRTVKSTEEAKKYLEEFRKESQNIKV EILVRDDKDDAEAKILSESSEIDVRTSK~~STSPDEAKRWIKEFSEE~~gwslehhhhhh
Design 3	mLKYVLLIANEKEAREELKKAAKEANTESRTVKSEEELKKYLEEFRKESQNIEV TILVREDRKDASAKILSQSSEIDVKTSK~~STSPDEAKRWIKEFSEE~~gwslehhhhhh
Design 4	mLKYVLIISNRKKAIEKARSLAEENNLESRTVKTKKELKKYLDEFRKESQNIKV DILVSNEEERARAEMAAKESEIDVEVRKE~~TSPDEAKRWIKEFSEE~~gwslehhhhhh
Design 5	mLSYVLLIANDKKLDEELKEAARKANTESRTVKEDDELKKYLEEFRKESQNIKV KILVRKDDKKASAEILRDSSEIDVETRK~~STSPDEAKRWIKEFSEE~~gwslehhhhhh
Design 6	mLKYVLLIANDKEAEDELRRASSKANAESRTVKEEEELKKYLEEFRKESQNIDV KILVRRDKDEAVAKVLSDSSEIDVKVDK~~STSPDEAKRWIKEFSEE~~gwslehhhhhh
Design 7	mLSYVLLIANDKEADRRLKDAARKANTESRTVKRSDELDKYLKEFRKESQNIK VRILVRESDKSAAAKILSESSEIDVKEERE~~TSPDEAKRWIKEFSEE~~gwslehhhhhh
Design 8	mLKYVLAIANDKEANEELRRAAEKANTESRTVKDEDELKKYLDEFRKESQNIK VDILVREDKKDAQAKILSKSSEIDVETRK~~STSPDEAKRWIKEFSEE~~gwslehhhhh h

Sequence of Rsmn2x2_6 is included for comparison. Residues mutated are highlighted in red.

Table 4.15 | X-ray data and refinement statistics for OR494[a]

Crystal Parameters	
Space group	$I2_12_12_1$
Cell dimensions:	
a, b, c (Å)	33.55, 69.65, 73.14
α, β, γ, (º)	90, 90, 90
Matthews coefficient (Å^3/Da)	2.1
Solvent Content (%)	41.3
Data Collection[b]	
Wavelength (Å)	0.9791
Resolution (Å)	36.6-1.47 (1.52-1.47)
R_{merge} (%)	5.5 (40.3)
No. of unique reflections	27452
No. of reflections in R_{free} set	1390
Mean Redundancy	4.5 (3.7)
Overall completeness (%)	97.5 (91.2)
Mean I/σ	33.3 (4.2)
Refinement Residuals[c]	
R_{free} (%)	21.6 (32.5)
R_{work} (%)	17.9 (27.2)
Completeness (%)	99.1 (97.5)
Model Quality[d]	
RMSD bond lengths (Å)	0.006
RMSD bond angles (º)	0.936
MolProbity Ramachandran statistics	
Most favored (%)	98.8
Allowed (%)	1.2
Disallowed (%)	0.0
Mean main chain B-factor (Å^2)	21.6
Mean overall B-factor (protein + solvent) (Å^2)	25.3
Mean solvent B-factor (Å^2)	34.8
Model Contents	
Protomers in ASU	1
Protein residues	1-89
No. of protein non-H atoms	674
No. of heterogen. atoms	17
No. of water molecules	96

[a] Entries in parentheses report data from the highest resolution shell. Entries in parentheses report data from the limiting resolution shell. Data collection and refinement statistics come from HKL2000 (84) and PHENIX (87), respectively.
[b] All observations with $I \geq -3\sigma_I$ were merged and included in calculating data quality statistics.
[c] Reflections with $F \geq 1.91\ \sigma_F$ were included in calculating *R*-factors.

REFERENCES

1. Gutte, B. A synthetic 70-amino acid residue analog of ribonuclease S-protein with enzymic activity. *J. Biol. Chem.* **250,** 889–904 (1975).
2. Gutte, B. Study of RNase A mechanism and folding by means of synthetic 63-residue analogs. *J. Biol. Chem.* **252,** 663–670 (1977).
3. Gutte, B., Däumigen, M. & Wittschieber, E. Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids. *Nature* **281,** 650–655 (1979).
4. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181,** 223–230 (1973).
5. Okazaki, K., Koga, N., Takada, S., Onuchic, J. N. & Wolynes, P. G. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci.* 103, 11844–11849 (2006).
6. Kaiser, E. T. & Kézdy, F. J. Secondary structures of proteins and peptides in amphiphilic environments. (A review). *Proc. Natl. Acad. Sci. U. S. A.* **80,** 1137–1143 (1983).
7. Richardson, J. S. & Richardson, D. C. The de novo design of protein structures. *Trends Biochem. Sci.* **14,** 304–309 (1989).
8. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487,** 545–574 (2011).
9. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491,** 222–227 (2012).
10. Kuhlman, B. *et al.* Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **302,** 1364–1368 (2003).
11. Turner, J. *et al.* Crystal structure of the mitotic spindle kinesin Eg5 reveals a novel conformation of the neck-linker. *J. Biol. Chem.* **276,** 25496–25502 (2001).

12. Smith, P. C. *et al.* ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. *Mol. Cell* **10,** 139–149 (2002).
13. Griffin, M. D. W. *et al.* Evolution of Quaternary Structure in a Homotetrameric Enzyme. *J. Mol. Biol.* **380,** 691–703 (2008).
14. Partanen, S. T. *et al.* The 1.3 A crystal structure of human mitochondrial Delta3-Delta2-enoyl-CoA isomerase shows a novel mode of binding for the fatty acyl group. *J. Mol. Biol.* **342,** 1197–1208 (2004).
15. Segura-Peña, D. *et al.* Quaternary Structure Change as a Mechanism for the Regulation of Thymidine Kinase 1-Like Enzymes. *Structure* **15,** 1555–1566 (2007).
16. Stieglitz, K., Stec, B., Baker, D. P. & Kantrowitz, E. R. Monitoring the Transition from the T to the R State in E. coli Aspartate Transcarbamoylase by X-ray Crystallography: Crystal Structures of the E50A Mutant Enzyme in Four Distinct Allosteric States. *J. Mol. Biol.* **341,** 853–868 (2004).
17. Tinberg, C. E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501,** 212–216 (2013).
18. Chan, W. L., Zhou, A. & Read, R. J. Towards Engineering Hormone-Binding Globulins as Drug Delivery Agents. *PLoS ONE* **9,** e113402 (2014).
19. Root, M. J., Kay, M. S. & Kim, P. S. Protein Design of an HIV-1 Entry Inhibitor. *Science* **291,** 884–888 (2001).
20. Fleishman, S. J. *et al.* Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science* **332,** 816–821 (2011).
21. Hume, J. *et al.* Engineered Coiled-Coil Protein Microfibers. *Biomacromolecules* **15,** 3503–3510 (2014).

22. Patterson, D. P. *et al.* Characterization of a highly flexible self-assembling protein system designed to form nanocages. *Protein Sci.* **23,** 190–199 (2014).

23. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510,** 103–108 (2014).

24. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).

25. Wintjens, R. T., Rooman, M. J. & Wodak, S. J. Automatic Classification and Analysis of αα-Turn Motifs in Proteins. *J. Mol. Biol.* **255,** 235–253 (1996).

26. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7,** 95–99 (1963).

27. Hutchinson, E. G. & Thornton, J. M. A revised set of potentials for beta-turn formation in proteins. *Protein Sci. Publ. Protein Soc.* **3,** 2207–2216 (1994).

28. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405,** 607–618 (2011).

29. Sheffler, W. & Baker, D. RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci. Publ. Protein Soc.* **18,** 229–239 (2009).

30. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. in *Methods in Enzymology* (ed. Johnson, L. B. and M. L.) **383,** 66–93 (Academic Press, 2004).

31. Liu, G. *et al.* NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 10487–10492 (2005).

32. Chou, P. Y. & Fasman, G. D. Beta-turns in proteins. *J. Mol. Biol.* **115,** 135–175 (1977).

33. Donate, L. E., Rufino, S. D., Canard, L. H. J. & Blundell, T. L. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction. *Protein Sci.* **5,** 2600–2616 (1996).
34. Aurora, R. & Rose, G. D. Helix capping. *Protein Sci. Publ. Protein Soc.* **7,** 21–38 (1998).
35. Richardson, J. S. & Richardson, D. C. Amino acid preferences for specific locations at the ends of alpha helices. *Science* **240,** 1648–1652 (1988).
36. Scheerlinck, J. P. *et al.* Recurrent alpha beta loop structures in TIM barrel motifs show a distinct pattern of conserved structural features. *Proteins* **12,** 299–313 (1992).
37. Pavone, V. *et al.* Discovering protein secondary structures: classification and description of isolated alpha-turns. *Biopolymers* **38,** 705–721 (1996).
38. Wintjens, R., Wodak, S. J. & Rooman, M. Typical interaction patterns in alphabeta and betaalpha turn motifs. *Protein Eng.* **11,** 505–522 (1998).
39. Kuhn, M., Meiler, J. & Baker, D. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *Proteins* **54,** 282–288 (2004).
40. Mattos, C., Petsko, G. A. & Karplus, M. Analysis of two-residue turns in proteins. *J. Mol. Biol.* **238,** 733–747 (1994).
41. Schellman, C. The αL conformation at the ends of helices, in: Protein Folding (R. Jaenicke, ed.). *Elsevier Amst.* 53–61. (1980).
42. Efimov, A. V. [Long irregular regions in proteins as combinations of small standard structures]. *Mol. Biol. (Mosk.)* **24,** 851–858 (1990).
43. Srinivasan, N., Sowdhamini, R., Ramakrishnan, C. & Balaram, P. Analysis of short loops connecting secondary structural elements in proteins. *Mol. Conform. Biol. Interact. Indian Acad. Sci.* pp 59-73. (1991).

44. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44,** 213–223 (2009).

45. Wang, G. & Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinforma. Oxf. Engl.* **19,** 1589–1591 (2003).

46. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268,** 209–225 (1997).

47. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).

48. Jansson, M. *et al.* High-level production of uniformly ^{15}N- and ^{13}C-enriched fusion proteins in Escherichia coli. *J. Biomol. NMR* **7,** 131–141 (1996).

49. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4,** 2411–2423 (1995).

50. Santoro, M. M. & Bolen, D. W. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl .alpha.-chymotrypsin using different denaturants. *Biochemistry (Mosc.)* **27,** 8063–8068 (1988).

51. Acton, T. B. *et al.* Preparation of protein samples for NMR structure, function, and small-molecule screening studies. *Methods Enzymol.* **493,** 21–60 (2011).

52. Neri, D., Szyperski, T., Otting, G., Senn, H. & Wüthrich, K. Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional 13C labeling. *Biochemistry (Mosc.)* **28,** 7510–7516 (1989).

53. Tjandra, N., Grzesiek, K. & Bax, A. Publication: Magnetic Field Dependence of Nitrogen−Proton J Splittings in 15 N-Enriched Human Ubiquitin Resulting from Relaxation Interference and Residual Dipolar Coupling. *J. Am. Chem. Soc.* **118,** 6264–6272 (1996).

54. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).

55. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66,** 778–795 (2007).

56. Huang, Y. J., Powers, R. & Montelione, G. T. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J. Am. Chem. Soc.* **127,** 1665–1674 (2005).

57. Hyberts, S. G., Goldberg, M. S., Havel, T. F. & Wagner, G. The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. *Protein Sci. Publ. Protein Soc.* **1,** 736–751 (1992).

58. Mou, Y., Huang, P.-S., Hsu, F.-C., Huang, S.-J. & Mayo, S. L. Computational design and experimental verification of a symmetric protein homodimer. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 10714–10719 (2015).

59. Boyken, S. E. *et al.* De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352,** 680–687 (2016).

60. Fletcher, J. M. *et al.* A Basis Set of de Novo Coiled-Coil Peptide Oligomers for Rational Protein Design and Synthetic Biology. *ACS Synth. Biol.* **1,** 240–250 (2012).

61. Dolphin, G. T. A designed branched three-helix bundle protein dimer. *J. Am. Chem. Soc.* **128,** 7287–7290 (2006).

62. Egelman, E. H. *et al.* Structural plasticity of helical nanotubes based on coiled-coil assemblies. *Struct. Lond. Engl. 1993* **23,** 280–289 (2015).

63. Barth, P. & Senes, A. Toward high-resolution computational design of the structure and function of helical membrane proteins. *Nat. Struct. Mol. Biol.* **23,** 475–480 (2016).

64. Fallas, J. A. *et al.* Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem.* **advance online publication,** (2016).

65. Huang, P.-S., Love, J. J. & Mayo, S. L. A de novo designed protein–protein interface. *Protein Sci. Publ. Protein Soc.* **16,** 2770–2774 (2007).

66. Stranges, P. B., Machius, M., Miley, M. J., Tripathy, A. & Kuhlman, B. Computational design of a symmetric homodimer using β-strand assembly. *Proc. Natl. Acad. Sci.* **108,** 20562–20567 (2011).

67. Lin, Y.-R. *et al.* Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5478-5485 (2015).

68. Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234,** 946–950 (1993).

69. Karanicolas, J. & Kuhlman, B. Computational design of affinity and specificity at protein–protein interfaces. *Curr. Opin. Struct. Biol.* **19,** 458–463 (2009).

70. André, I., Bradley, P., Wang, C. & Baker, D. Prediction of the structure of symmetrical protein assemblies. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 17656–17661 (2007).

71. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

72. Qian, B. *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450,** 259–264 (2007).

73. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77,** 363–382 (2008).

74. Huang, P.-S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PloS One* **6,** e24109 (2011).

75. Dantas, G. *et al.* Mis-Translation of a Computationally Designed Protein Yields an Exceptionally Stable Homodimer: Implications for Protein Engineering and Evolution. *J. Mol. Biol. 36251004-1024* (2006). doi:10.1016/j.jmb.2006.07.092

76. Classen, S. *et al.* Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the Advanced Light Source. *J. Appl. Crystallogr.* **46,** 1–13 (2013).

77. Dyer, K. N. *et al.* High-throughput SAXS for the characterization of biomolecules in solution: a practical approach. *Methods Mol. Biol. Clifton NJ* **1091,** 245–258 (2014).

78. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105,** 962–974 (2013).

79. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* **44,** W424-429 (2016).

80. Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12,** 29–34 (2016).

81. Hashimoto, K. & Panchenko, A. R. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl. Acad. Sci.* **107,** 20352–20357 (2010).

82. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling Symmetric Macromolecular Structures in Rosetta3. *PLOS ONE* **6,** e20450 (2011).

83. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40,** 191–285 (2007).

84. Chayen, N. E., Shaw Stewart, P. D., Maeder, D. L. & Blow, D. M. An automated system for micro-batch protein crystallization and screening. *J. Appl. Crystallogr.* **23,** 297–302 (1990).

85. Otwinowski, Z. & Minor, W. [20] Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).

86. Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. BALBES: a molecular-replacement pipeline. *Acta Crystallogr. D Biol. Crystallogr.* **64,** 125–132 (2008).

87. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132 (2004).

88. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 1948–1954 (2002).

89. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).