Control over overall shape and size in de novo designed proteins

Yu-Ru Lin[a,1], Nobuyasu Koga[a,b,c,1,2], Rie Tatsumi-Koga[a,b], Gaohua Liu[d], Amanda F. Clouser[a], Gaetano T. Montelione[d,e], and David Baker[a,2]

[a]Department of Biochemistry, University of Washington and Howard Hughes Medical Institute, Seattle, WA 98195; [b]Research Center of Integrative Molecular Systems, Institute for Molecular Science, National Institutes of Natural Sciences, Okazaki, Aichi 444-8585, Japan; [c]Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan; [d]Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers, The State University of New Jersey, Piscataway, NJ 08854; and [e]Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, NJ 08854

Edited by William F. DeGrado, School of Pharmacy, University of California, San Francisco, CA, and approved August 25, 2015 (received for review May 14, 2015)

We recently described general principles for designing ideal protein structures stabilized by completely consistent local and nonlocal interactions. 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 codependencies 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 rule set by designing a series of proteins with the same fold but considerable variation in secondary structure length, loop geometry, β-strand registry, 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.

de novo design | protein design | ideal protein | control protein shape

Protein design holds promise for applications ranging from therapeutics to biomaterials, with recent progress in designing small molecule binding proteins (1, 2), inhibitors of protein–protein interactions (3, 4), and self-assembling nanomaterials (5–7). Most of these efforts have repurposed naturally occurring scaffolds, which are likely not optimal starting points for creating new functions because they generally contain sequence and structural idiosyncrasies that arose during evolutionary optimization for their natural functions (8). Robust design of new functional proteins would be considerably enabled by the capability of precisely designing from scratch arbitrary protein structures.

We previously described general principles that allowed the de novo design of ideal protein structures with five different folds (9). In this paper, 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 in our previous paper 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 specific loop conformations to allow more detailed control over local geometry and overall protein topology.

It is convenient to describe protein local geometry by using the ABEGO (10) alphabet illustrated in Fig. 1*A*. "A" indicates the alpha region of the Ramachandran plot (11); "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*A* throughout the paper. 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 *B* and *C*). 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. 2*A*). If the cross-product of a vector pointing in the direction of the first strand and a vector from the first strand to the second 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. 2*B* shows that in native protein structures (*SI Appendix*), two-residue loops always have L-chirality, and that the GG block is particularly common. As is evident in the schematic in Fig. 1*C*, 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. 2 *C* and *D*. For five-residue loops, the R-chirality is preferred over

Significance

We describe how protein size and shape can be sculpted by de novo protein design. Precise control over protein shape will be critical for completely de novo design of high-affinity binding proteins, enzymes, and protein-based nanomaterials. The systematic procedure for design of αβ-protein structures from scratch described in this paper should be broadly useful.

Author contributions: Y.-R.L., N.K., R.T.-K., G.L., G.T.M., and D.B. designed research; Y.-R.L., N.K., G.L., and A.F.C. performed research; N.K. contributed new analytic tools; Y.-R.L., N.K., R.T.-K., G.L., G.T.M., and D.B. analyzed data; and Y.-R.L., N.K., R.T.-K., G.L., G.T.M., and D.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The atomic coordinates have been deposited in the Protein Data Bank, www.pdb.org (PDB ID codes 2N2U, 2N2T, 2N76, and 2N3Z).

[1]Y.-R.L. and N.K. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: dabaker@u.washington.edu or nkoga@ims.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1509508112/-/DCSupplemental.

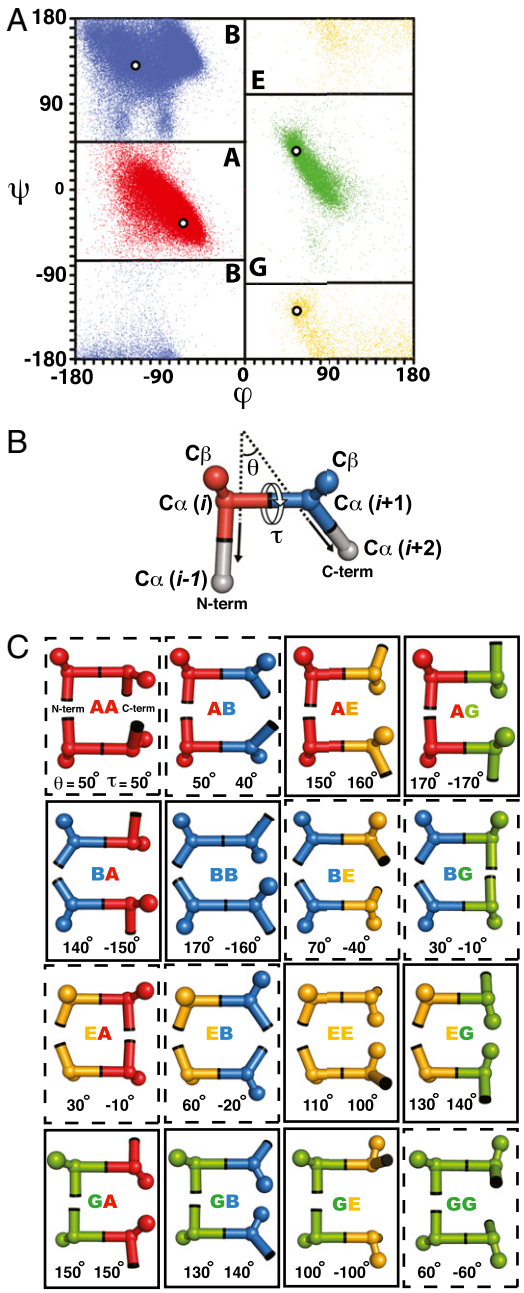

Fig. 1. Discrete state model of protein local geometry. (*A*) ABEGO representation of protein local structure shown on Ramachandran plot. A, alpha region; B, beta region; G and E, positive phi region. The most frequently observed torsion angles for each region are indicated by the white circle. (*B*) Two-residue "lego blocks" are represented by four consecutive Cα atoms connected by virtual bonds. It is useful to consider the net change in chain direction θ and the net twist τ produced by each lego block. 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 images. 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.

the L-chirality. The most common five-residue loop, BAAGB, is shown in Fig. 2*E*.

In the standard β-turn type nomenclature (12), 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 (ββ, βα, and αβ) 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. 2 *F* and *K*).

In βα-units, the Para orientation is favored for two-residue loops and the Anti orientation for three-residue loops (9). Fig. 2*G* 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. 2*H*, 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 (see the BA block in Fig. 1*C*) and together with the following B residue 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. 2*I*). 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*C*). 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*C*).

αβ-connections. For αβ-units, the preferred packing orientation is Para (9). As shown in Fig. 2*L*, 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"). 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 ββ-connections, we selected the GG and EA loops for the L-chirality. For βα-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 (*SI Appendix*, Fig. S1). For αβ-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 (*SI Appendix*, Fig. S2). The amino acid sequences in the loop regions were designed by 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 (*SI Appendix*, Fig. S3).

Extended Emergent Rules. The different loop types have different geometries, which change 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 secondary structure and ABEGO torsion constrained Rosetta folding simulations with a sequence-independent backbone model (9) (*SI Appendix*) for βαββ- (*SI Appendix*, Fig. S4), βαβ- (*SI Appendix*, Fig. S5) and βαβαβ- (*SI Appendix*, Fig. S6) units; the most frequently observed helix length for each strand length and loop combination is indicated in *SI Appendix*, Tables S1–S3. For each choice of loop types,

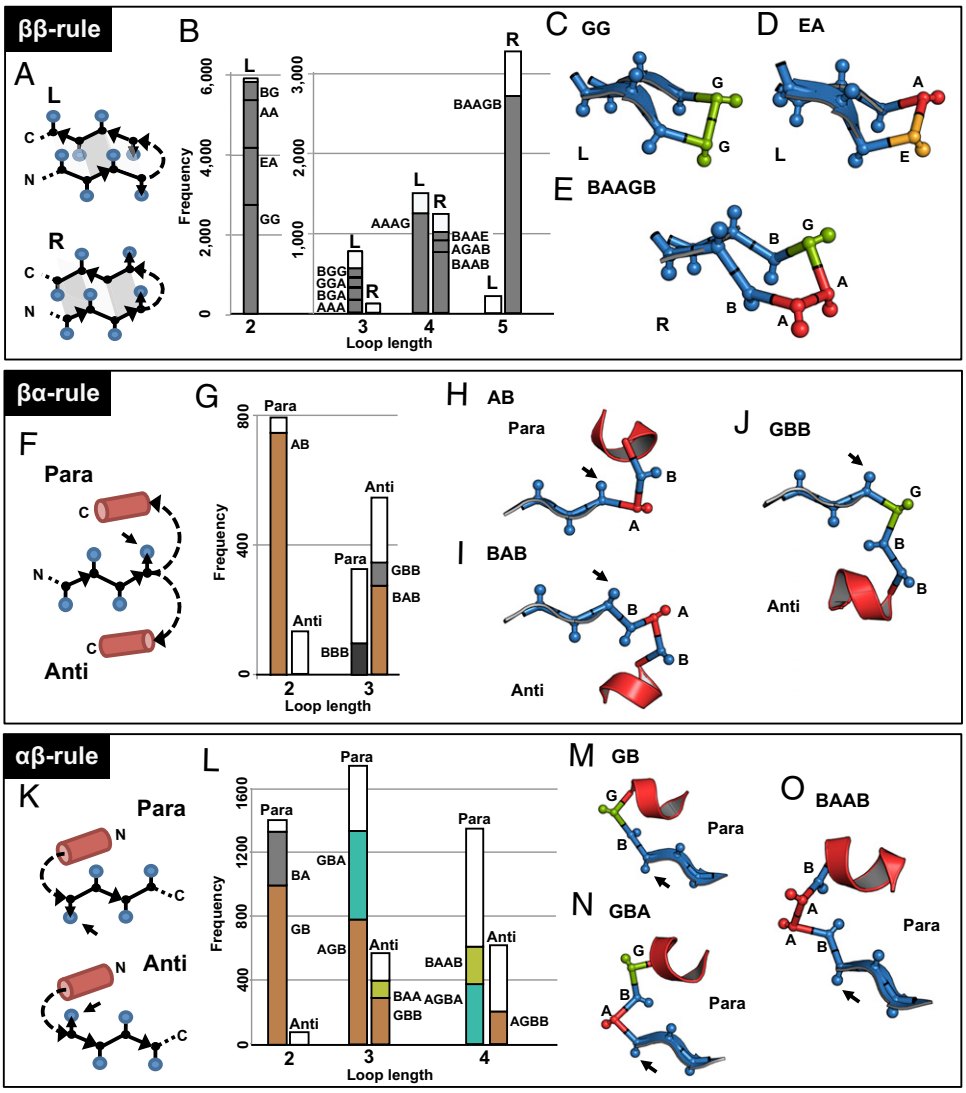

Fig. 2. Common loop geometries for ββ-, βα-, and αβ-units in naturally occurring proteins. (*A*, *F*, and *K*) Secondary structure packing orientation definitions of ββ-, βα-, and αβ-units are illustrated. (*B*, *G*, and *L*) Loop-type distributions in naturally occurring protein structures for ββ-, 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 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 (*SI Appendix*, Fig. S2). 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.

there is a distinct codependence 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. 3*A*, the optimal helix length goes from 10 to 22 as the strand length increases (Fig. 3*B*; the change in overall size and shape is illustrated in Fig. 3*C*). For a βαβ-motif with five-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. 3*E*); these differences result from the different curvature of the two types of loops (Fig. 3 *F* and *G*). For αβ-units, the tilt angle of the helix relative to the β-sheet (Fig. 3*H* and *SI Appendix*) is determined by the loop type: With the BAAB loop, the helix is parallel to the β-strands, whereas with the GBA loop the helix runs diagonally to the β-strands (Fig. 3*I*).

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. 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 paper (9). For the ferredoxin-like fold, strand lengths 5, 7, and 9 were used with or without a β-strand register shift between the first and third strands. Suitable loop types for each secondary structure connection were selected based on the packing orientation as described above, and for αβ-connections, the helix-sheet tilt angle (Fig. 3 *H* and *I*). In the αβ-connections, the GBA loop that leads to helices which run diagonally on the β-sheet is less compatible with the ferredoxin-like fold than the BAAB loop that leads to helices parallel to the β-sheet: The two helices in the ferrredoxin-like fold can readily pack together on the β-sheet in the latter case but not the former (*SI Appendix*, Fig. S7). Hence, we used the BAAB loop for the αβ-connections in the new ferredoxin-like fold designs. The helix lengths were then chosen based on the

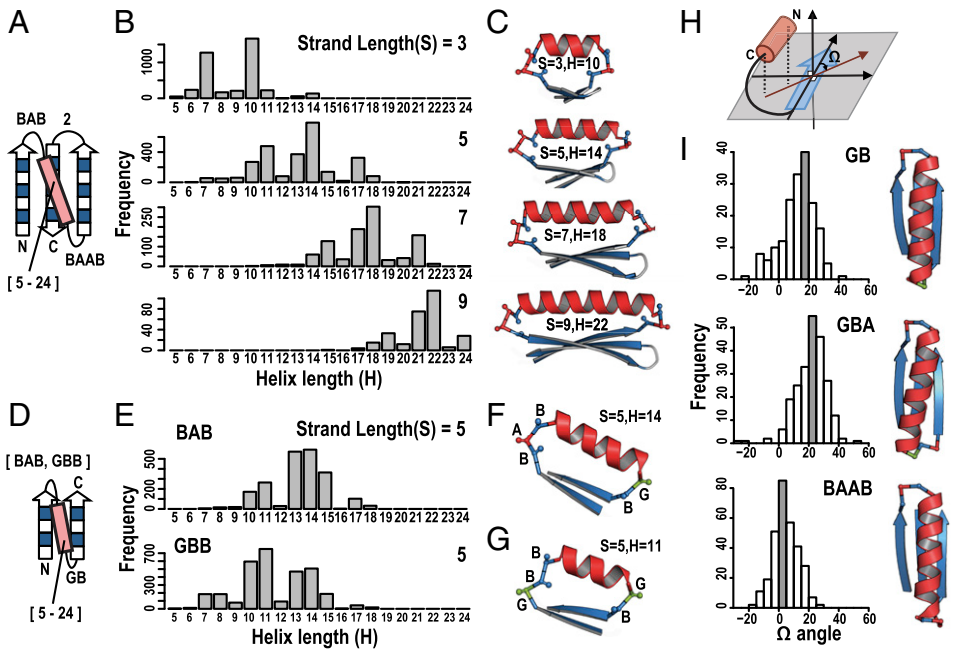

Fig. 3. Loop geometry controls secondary structure lengths and helix-sheet tilt angles in alpha-beta super secondary structure elements. (*A* and *D*) Schematics of the βαββ-units and the βαβ-units found in the ferredoxin-like fold and the Rossmann fold, respectively. (*B*) Helix length depends on the 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 frequencies of successful βαββ-unit folding were 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 yield different optimal helix lengths. (*F* and *G*) BAB and GBB loops result in different optimal helix lengths. (*H*) The tilt angle Ω of the α-helix relative to the β-sheet for αβ-units. (*I*) The Ω angle depends on αβ-loop type. The angle distribution was calculated from βαββ-unit folding simulations with the BAB loop for the βα-unit, with strand lengths 7 and helix length 14.

loop types and the strand length using *SI Appendix*, Tables S1–S3; the lengths of the helices in the ferredoxin-like fold series are based only on the βαββ-motif simulations (*SI Appendix*, Table S1), whereas those of Rsmn2x2_5 are based on both the βαβ- and βαβαβ-motif simulations (*SI Appendix*, Tables S2 and S3).

For each blueprint, backbone structures were built up by carrying out multiple independent Rosetta folding simulations

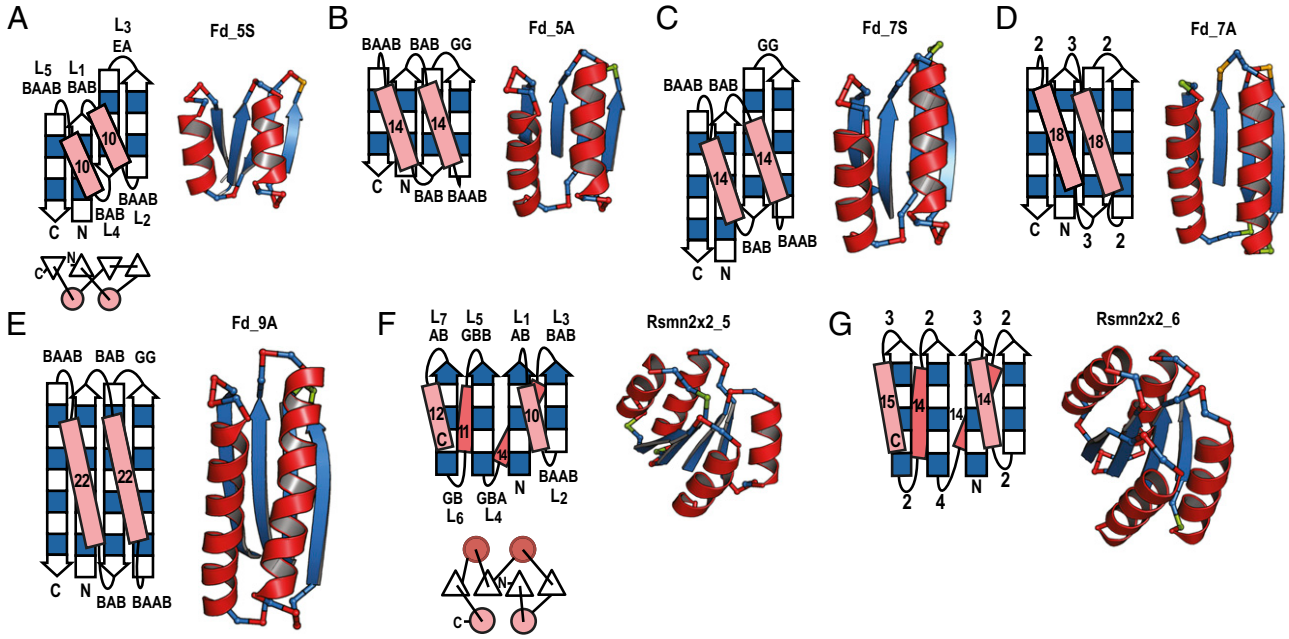

Fig. 4. Backbone blueprints and design models for ferredoxin-like folds and Rossmann2×2 folds with different sizes and shapes. The ferredoxin-like fold (*A*–*E*) and the Rossmann2×2 fold (*F* and *G*). Backbone blueprint for each topology (*Left*) and a corresponding Rosetta generated backbone structure (*Right*). (*A*) Fd_5S: 58 residues, with register shift between the first and third 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*) Rsmn2×2_5: 87 residues. (*G*) Rsmn2×2_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 Rsmn2×2_5. Fd_7A and Rsmn2×2_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.

Table 1. Design success rate

Structure	Designs tested	Expressed*	Soluble*	αβ-protein CD spectrum	Monomeric†	Well-resolved HSQC	Success (%)‡
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 blueprints (Fig. 4) indicated in the leftmost column. The subsequent columns give the number of designs that satisfy each criterion.
*Expression and solubility were assessed by SDS/PAGE and mass spectrometry.
†SEC-MALS was used to determine oligomerization state.
‡The successful designs are defined as those that satisfy all criteria. The details of the results are shown in *SI Appendix*, Tables S4–S8.

(*SI Appendix*). 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 (13) and searching for the lowest energy backbone structure for fixed amino acid sequence (14). Inward-pointing charged residues were introduced in edge β-strands and nonpolar residues were penalized at surface exposed positions to disfavor aggregation (the sequence design

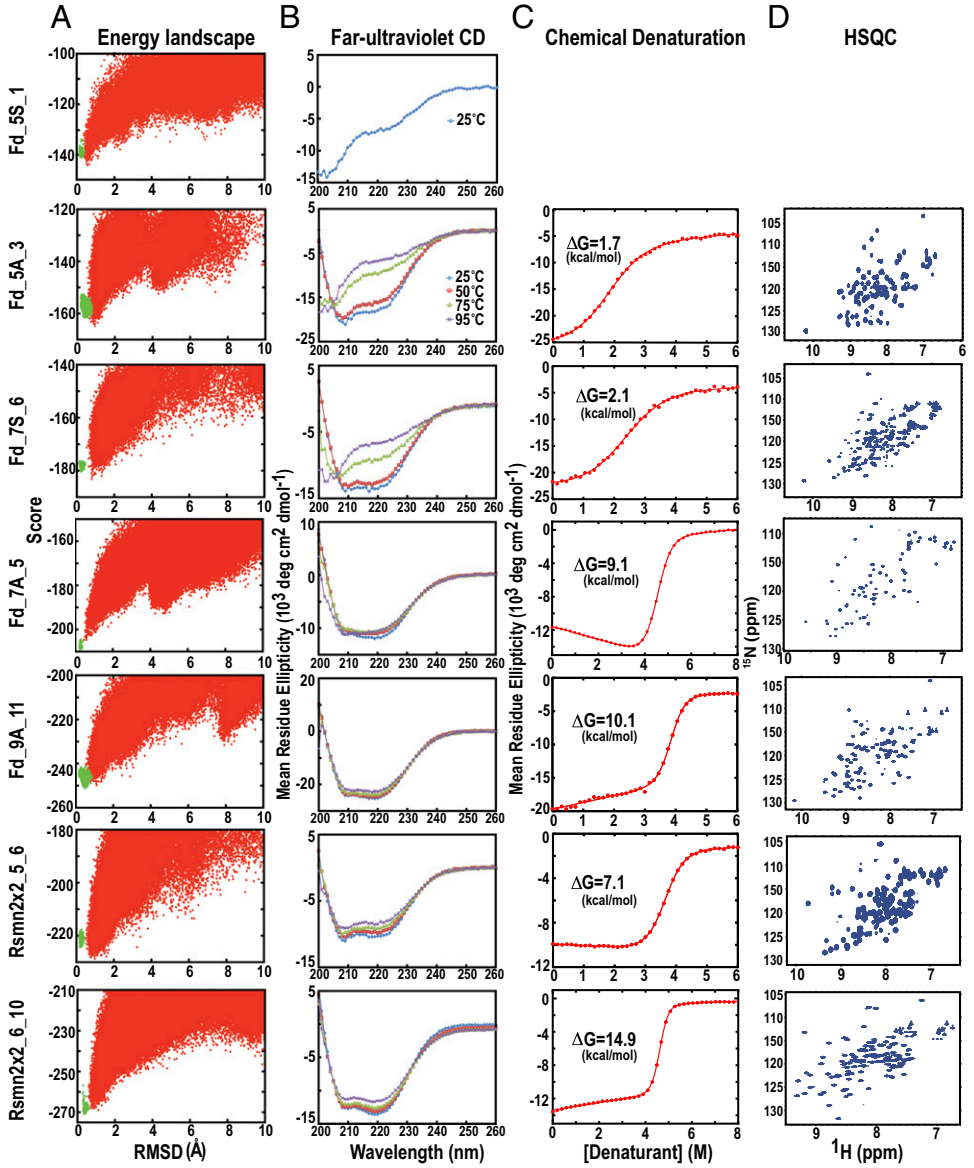

Fig. 5. In silico energy landscapes and experimental characterization of designed proteins. (*A*) Energy landscapes obtained from Rosetta ab initio structure prediction simulations on Rosetta@home. Red points represent the lowest-energy structures obtained in independent Monte Carlo structure prediction 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-UV 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 Δ*G*. (*D*) Two-dimensional ^{1}H-^{15}N HSQC spectra at 25 °C and 600 MHz. p.p.m., parts per million.

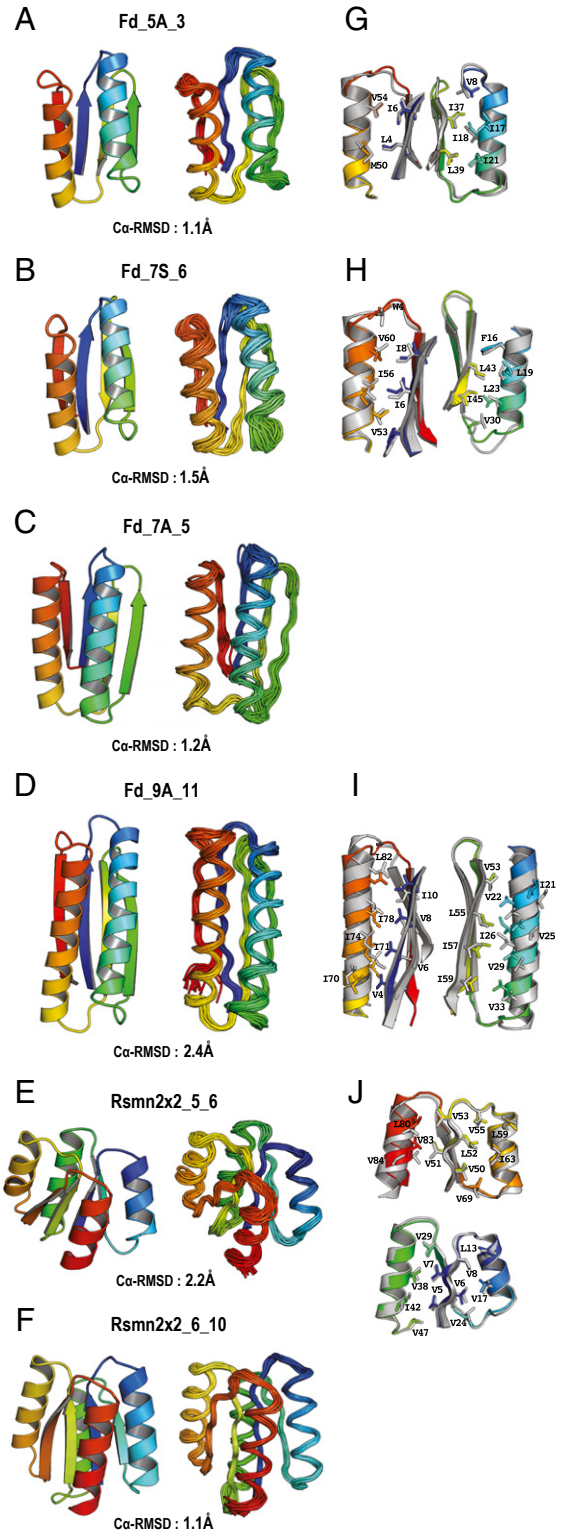

Fig. 6. Comparison of computational design models with experimentally determined NMR structures. (*A*–*F*) Comparison of protein backbones of design models (*Left*) and NMR structures (*Right*); the Cα root mean square deviation (RMSD) between the two is indicated. (*G*–*J*) Comparison of core side-chain packing in superpositions of design models (rainbow) and NMR structures (gray). (*A* and *G*) Fd_5A_3 (2N2U). (*B* and *H*) Fd_7S_6 (2N2T). (*D* and *I*) Fd_9A_11 (2N76). (*E* and *J*) Rsmn2×2_5_6 (2N3Z). (*C*) Fd_7A_5 and (*F*) Rsmn2×2_6_10 designed by Koga et al. in 2012 (9) are included here for shape and size comparison.

Table 2. ABEGO-based comparison between design model and NMR structures for the five loops in the three ferredoxin-like folds

	L1	L2	L3	L4	L5
Fd_5A_3	BAB/BAB	BAAB/BAAB	GG/BG*	BAB/BAB	BAAB/BAAB
Fd_7S_6	BAB/BAB	BAAB/BOBB†	GG/GG	BAB/BAB	BAAB/BAAB
Fd_9A_11	BAB/BAB	BAAB/BAAB	GG/GG	BAB/BAB	BAAB/BAAB

The columns L1–L5 correspond to the five loops shown in Fig. 4. In each cell, the loop type for the design model (left) and the most frequent loop type in the NMR ensemble (right) are shown.

*The B conformation at the position 1 was confirmed by chemical shift data (30).

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

protocol is described in detail in the methods and figure S12 in ref. 9). The designed structures were then filtered based on the Rosetta full-atom energy, sidechain packing (15), and the local sequence-structure compatibility (9). For each designed sequence, we then carried out multiple independent Rosetta ab initio structure prediction simulations (16) starting from an extended conformation, and selected designed sequences with energy landscapes strongly funneled into the designed target structure (Fig. 5*A*) for experimental characterization.

For the ferredoxin-like fold, we obtained synthetic genes (Genscript) encoding six designs for Fd_5S, 12 for Fd_5A, 10 for Fd_7S, and 12 for Fd_9A (sequences are provided in *SI Appendix*, Tables S10–S13). All but one design (Fd_7S) are not homologous to any known proteins (Blast *E* value <0.02 against the nonredundant protein sequence database nr). For the Rossmann2x2 fold, nine 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 Rsmn2x2_5 design sequence similar protein and the homolog 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) 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 2 days perhaps due to the large hydrophobic core. The far-UV CD spectra show that 23 of the 31 soluble designs for Fd_5A, Fd_7S, and Fd_9A have the expected αβ-secondary structure content. In contrast, for the smallest variant—Fd_5S—none of the designs had CD spectra consistent with folded αβ-proteins. Twenty-six 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 αβ-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, along with the designs of Fd_7A reported in the previous paper (9).

For the Rossmann2x2 fold, nine designs were tested for Rsmn2 × 2_5 (sequences are provided in *SI Appendix*, Table S14). All of the designs were expressed at high levels, and all but two designs have high solubility. Eight designs have the expected CD spectra for αβ-proteins, and of these, six designs were found to be monomeric by SEC-MALS. For the monomeric designs with the expected CD spectra, HSQC spectra were measured, and two designs have well-dispersed and sharp ^{1}H-^{15}N HSQC peaks, suggesting well-packed tertiary structures. The properties of the

Table 3. ABEGO-based comparison between design model and NMR structures for the seven loops in the Rossmann2x2 fold

	L1	L2	L3	L4	L5	L6	L7
Rsmn2x2_5_6	AB/AB	BAAB/BAAB	BAB/BAB	GBA/GBA	GBB/GBB	GB/GB	AB/AB

The columns L1–L7 correspond to the seven loops shown in Fig. 4. In each cell, the loop type for the design model (left) and the most frequent loop type in the NMR ensemble (right) are shown.

Rsmn2x2_5 designs are summarized in Table 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. 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, whereas 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 consortium (17) (*SI Appendix*, Table S9). In addition to distance restraints derived from NOESY data, dihedral angle restraints were derived for each design from backbone chemical shift data by using TALOSN, and used for structure calculations. Residual dipolar coupling (RDC) restraints from at least one alignment media were also obtained for three designs, Fd_5A_3, Fd_9A_11, and Rsmn2x2_5_6, and used in these structure calculations. RDC and chemical shift-based restraints were included only for residues in regular secondary structures and ordered regions of surface loops. 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. 6 *A*, *B*, *E*, *G*, *H*, and *J*). 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. 6 *D* and *I*.

We further compared the loop geometries at the ABEGO level (Tables 2 and 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 (12, 18–29). Chou and Fasman categorized β-turns into 11 types based on their backbone torsion angles (18) and Hutchinson and Thornton modified the classification after more protein structures were solved (12). An extensive study of short loops connecting regular secondary structures by Donate et al. identified common groups of loop geometries connecting different secondary structure elements (19). The analysis of loop types in this paper extends and updates this previous work, taking advantage of the much larger number of protein structures that have now been determined. Common loop geometries such as type I, II, I′, II′ β-hairpins (12, 18, 19, 26) and α-helical C-capping (19–24, 27) are reidentified as expected, and previously unidentified 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 αβ-protein design principles in terms of specific loop types in this paper makes possible a systematic building block-based approach to designing new structures. The basic algorithm consists of (*i*) choosing a topology (placement of secondary structure elements with order along the sequence specified), (*ii*) choosing the strand lengths and registers, (*iii*) choosing from the loop types specified by the extended rules, and (*iv*) choosing helix lengths compatible with the strand lengths and loop types. Complete information for steps 3 and 4 are provided in *SI Appendix*, Tables S1–S3.

The 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.

Materials and Methods Summary

Rosetta folding simulations for building backbone structures were performed on a sequence-independent backbone model with a pseudoatom representing a generic side chain, using a potential function that favors compact structure and hydrogen bonds between amide hydrogen and carbonyl oxygen and disfavors overly close atom pairs. Each Monte Carlo simulation attempt replaces the torsion angles of a randomly selected residue with torsion angles randomly selected from the region of the Ramachandran plot compatible with the assigned secondary structure and ABEGO type.

ACKNOWLEDGMENTS. We thank the hundreds of thousands of Rosetta@home volunteers for making this work possible. This work was supported by grants from Japan Science and Technology Agency, Precursory Research for Embryonic Science and Technology (N.K.), the National Institutes of General Medical Science Protein Structure Initiative program Grant U54 GM094597 (to G.T.M.), Defense Threat Reduction Agency, and Howard Hughes Medical Institute (D.B.). N.K. acknowledges a grant of computer time from the Research Center for Computational Science.

1. Tinberg CE, et al. (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* 501(7466):212–216.
2. Chan WL, Zhou A, Read RJ (2014) Towards engineering hormone-binding globulins as drug delivery agents. *PLoS One* 9(11):e113402.
3. Root MJ, Kay MS, Kim PS (2001) Protein design of an HIV-1 entry inhibitor. *Science* 291(5505):884–888.
4. Fleishman SJ, et al. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332(6031):816–821.
5. Hume J, et al. (2014) Engineered coiled-coil protein microfibers. *Biomacromolecules* 15(10):3503–3510.
6. Patterson DP, et al. (2014) Characterization of a highly flexible self-assembling protein system designed to form nanocages. *Protein Sci* 23(2):190–199.
7. King NP, et al. (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336(6085):1171–1174.
8. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247(4):536–540.
9. Koga N, et al. (2012) Principles for designing ideal protein structures. *Nature* 491(7423):222–227.
10. Wintjens RT, Rooman MJ, Wodak SJ (1996) Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J Mol Biol* 255(1):235–253.

11. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–99.
12. Hutchinson EG, Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. *Protein Sci* 3(12):2207–2216.
13. Kuhlman B, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302(5649):1364–1368.
14. Tyka MD, et al. (2011) Alternate states of proteins revealed by detailed energy landscape mapping. *J Mol Biol* 405(2):607–618.
15. Sheffler W, Baker D (2009) RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci* 18(1):229–239.
16. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. *Methods Enzymol* 383:66–93.
17. Liu G, et al. (2005) NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc Natl Acad Sci USA* 102(30): 10487–10492.
18. Chou PY, Fasman GD (1977) Beta-turns in proteins. *J Mol Biol* 115(2):135–175.
19. Donate LE, Rufino SD, Canard LH, Blundell TL (1996) Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction. *Protein Sci* 5(12):2600–2616.
20. Aurora R, Rose GD (1998) Helix capping. *Protein Sci* 7(1):21–38.
21. Richardson JS, Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science* 240(4859):1648–1652.
22. Scheerlinck JP, et al. (1992) Recurrent alpha beta loop structures in TIM barrel motifs show a distinct pattern of conserved structural features. *Proteins* 12(4):299–313.
23. Pavone V, et al. (1996) Discovering protein secondary structures: Classification and description of isolated alpha-turns. *Biopolymers* 38(6):705–721.
24. Wintjens R, Wodak SJ, Rooman M (1998) Typical interaction patterns in alphabeta and betaalpha turn motifs. *Protein Eng* 11(7):505–522.
25. Kuhn M, Meiler J, Baker D (2004) Strand-loop-strand motifs: Prediction of hairpins and diverging turns in proteins. *Proteins* 54(2):282–288.
26. Mattos C, Petsko GA, Karplus M (1994) Analysis of two-residue turns in proteins. *J Mol Biol* 238(5):733–747.
27. Schellman C (1980) The αL conformation at the ends of helices. *Protein Folding* (Elsevier, New York), pp 53–61.
28. Efimov AV (1991) Long and mediumsize irregular regions in proteins as combinations of small standard structures. *Molecular Conformation and Biological Interactions*, eds Balaram P, Ramaseshan S (Indian Acad Sci, Bangalore, India), pp 19–29.
29. Srinivasan N, Sowdhamini R, Ramakrishnan C, Balaram P (1991) Analysis of short loops connecting secondary structural elements in proteins. *Molecular Conformation and Biological Interactions*, eds Balaram P, Ramaseshan S (Indian Acad Sci, Bangalore, India), pp 59–73.
30. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR* 44(4):213–223.