Structure
Article

Increasing Sequence Diversity with Flexible Backbone Protein Design: The Complete Redesign of a Protein Hydrophobic Core

Grant S. Murphy,[1] Jeffrey L. Mills,[2,3] Michael J. Miley,[4,6] Mischa Machius,[4,5,6] Thomas Szyperski,[2,3] and Brian Kuhlman[5,7,*]
[1]Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-3290, USA
[2]Department of Chemistry, State University of New York at Buffalo, Buffalo, NY, 14260, USA
[3]Northeast Structural Genomics Consortium
[4]Center for Structural Biology
[5]Lineberger Comprehensive Cancer Center
[6]Department of Pharmacology
University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA
[7]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-7260, USA
*Correspondence: bkuhlman@email.unc.edu
DOI 10.1016/j.str.2012.03.026

SUMMARY

Protein design tests our understanding of protein stability and structure. Successful design methods should allow the exploration of sequence space not found in nature. However, when redesigning naturally occurring protein structures, most fixed backbone design algorithms return amino acid sequences that share strong sequence identity with wild-type sequences, especially in the protein core. This behavior places a restriction on functional space that can be explored and is not consistent with observations from nature, where sequences of low identity have similar structures. Here, we allow backbone flexibility during design to mutate every position in the core (38 residues) of a four-helix bundle protein. Only small perturbations to the backbone, 1–2 Å, were needed to entirely mutate the core. The redesigned protein, DRNN, is exceptionally stable (melting point >140°C). An NMR and X-ray crystal structure show that the side chains and backbone were accurately modeled (all-atom RMSD = 1.3 Å).

INTRODUCTION

A primary goal of protein design is to create proteins that have sequences, structures, and functions not found in nature. This goal can be reached by designing new protein structures from scratch or by modifying sequences and structures of proteins found in nature. The second approach is appealing, because in many cases it should be more likely to succeed, and it is the approach nature typically uses to evolve new functional proteins. There are many examples of naturally occurring protein pairs that are structurally homologous (have the same fold), but have different functions and low sequence identity (<15%). Recapitulating or expanding on this sequence diversity by design, however, is not straightforward. Most computational methods for protein design are built on side-chain optimization algorithms that work most efficiently with a fixed protein backbone (Gordon et al., 1999). When redesigning naturally occurring proteins with these methods, the computationally optimized sequences often closely resemble the native sequence, especially in the protein core, where >60% sequence identity is common (Desjarlais and Handel, 1999; Kuhlman and Baker, 2000; Pokala and Handel, 2001). It is clear from these studies and from the structural analysis of naturally occurring homologs that to expand sequence diversity it is necessary to allow perturbations to the protein backbone conformation. Even small changes to the backbone (<2 Å), can open large regions of sequence space (Yin et al., 2007). The challenge for protein designers is to identify backbone and sequence perturbations that are energetically favorable.

A variety of strategies have been developed for performing protein design with backbone flexibility (Apgar et al., 2009; Dantas et al., 2007; Davis et al., 2009; Desjarlais and Handel, 1999; Friedland et al., 2008; Fung et al., 2008; Georgiev and Donald, 2007; Grigoryan and Degrado, 2011; Havranek and Baker, 2009; Mandell and Kortemme, 2009; Su and Mayo, 1997), however, few have been experimentally validated with high-resolution structures of the designed protein (Correia et al., 2011; Harbury et al., 1995, 1998; Hu et al., 2007; Kuhlman et al., 2002, 2003; Murphy et al., 2009; Sammond et al., 2011). Perhaps the most tested approach has been iterative rounds of sequence optimization and backbone refinement with the molecular modeling program Rosetta. Sequence optimization is performed using a simulated annealing protocol that searches for low-energy combinations of side-chain rotamers. Structure refinement uses Monte Carlo sampling of small backbone torsion angle perturbations coupled with gradient-based minimization of dihedral angles. Both stages of optimization use an energy function that rewards tight packing, commonly observed side-chain and backbone torsion angles, favorable hydrogen-bond geometries and low energies of desolvation. This approach has been used to design a protein from scratch, design a protein-binding peptide and design new protein loop conformations

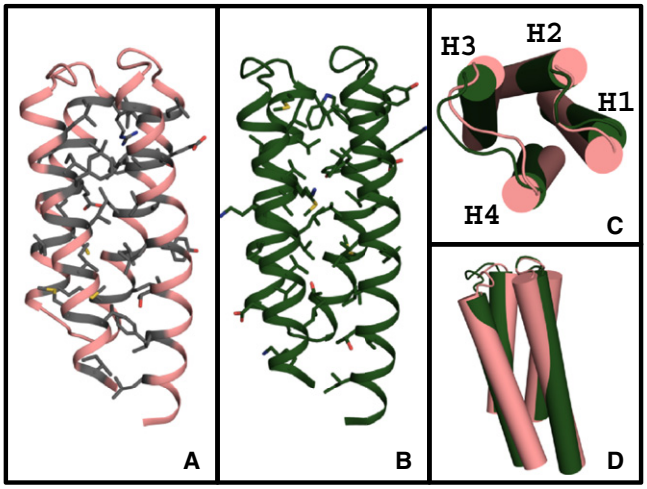


Figure 1. Global Comparison of the Wild-Type Template and DRNN Design Model

(A) Thirty-eight design positions shown as gray sticks were identified in the wild-type template.

(B) The final design model for DRNN with the designed positions shown as green sticks.

(C and D) DRNN's backbone and helix crossing angles have been subtly changed by the flexible backbone design procedure. The helices are labeled H1–H4 in (C).

(A), (B), and (D) are in the same orientation, and (C) is a top-down view of the bundle.

(Dantas et al., 2003; Hu et al., 2007; Kuhlman et al., 2003). In this study, we explore whether iterative optimization of sequence and structure with Rosetta can be used to aggressively redesign an entire protein core.

Our specific goal was to mutate every residue in the core of the four-helix bundle protein, CheA phosphotransferase, while maintaining the overall fold and stability of the protein (Figure 1A). Several de novo design and redesign projects have focused on helix bundle proteins (Hecht et al., 1990). From these studies, it is evident that many sequences will adopt collapsed helical structures as long as the amphipathic nature of the helices is preserved and the sequence has significant helical propensity (DeGrado and Nilsson, 1997; Kamtekar et al., 1993). What is more challenging to design are sequences that adopt a specific pre-determined structure and show characteristics of natural helix bundle proteins, such as cooperative thermal unfolding. Many previously reported helical bundle designs formed a molten globule, i.e., an ensemble of collapsed structurally degenerate conformations. In cases where the structure for a design was experimentally determined, it often did not agree with the initial design model (Hill and DeGrado, 2000; Lovejoy et al., 1993; Willis et al., 2000). One striking success story is the accurate de novo design of a symmetric four-helix coiled-coil with a right-handed super-helical twist (Harbury et al., 1998). A key component of this work was optimization of packing energies via backbone refinement as well as sequence design with a reduced amino acid alphabet. Here, we show that flexible backbone design can be used to perturb the structure and sequence of a pre-existing protein with atomic-level accuracy.

RESULTS

Core Redesign of the CheA Four Helix Bundle

The four-helix bundle CheA phosphotransferase was chosen as the design template (Protein Data Bank [PDB] ID: 1TQG) because of its simple up-down helix bundle topology and its moderate size of 105 amino acids. Thirty-eight positions from the CheA X-ray crystal structure were identified as being completely or partially buried and were targeted for mutation (Figure 1A and Figure 2). Our initial hypothesis, based on previous protein redesign experiments, was that the protein backbone would need to be perturbed to completely redesign the protein core. To test this hypothesis, four different computational procedures were used to generate designed sequences: (1) fixed backbone design with all amino acid types allowed at each design position (FBAA), (2) fixed backbone design with the native amino acid disallowed at each design position (FBNN), (3) flexible backbone design with all amino acid types allowed at each design position (DRAA), and (4) flexible backbone design with the native amino acid disallowed at each design position (DRNN). In the naming scheme, FB stands for fixed backbone, DR stands for the design and backbone refinement strategy of flexible backbone design, AA denotes that all amino acids were allowed during design, and NN indicates that only non-native amino acids were allowed during design.

The fixed backbone design protocol used Rosetta's standard rotamer-optimization method, which uses Monte Carlo sampling of backbone-dependent side-chain rotamers to search for low-energy sequences. The flexible backbone protocol used the same sequence-optimization algorithm, but iterated sequence optimization with high-resolution backbone refinement using Monte Carlo sampling and gradient-based minimization of backbone torsion angles. Backbone perturbations with this protocol are generally modest, that is, 1–2 Å. Twenty-five thousand independent trajectories were generated for each protocol. As anticipated, in the two approaches where all amino acid types were allowed, FBAA and DRAA, the flexible backbone procedure DRAA generated sequences with lower sequence identity to the wild-type protein. The average sequence identity over the designed positions was 26% in the DRAA protocol and 65% with the fixed backbone protocol. To check if the fixed backbone protocol generated models with lower sequence identity, we searched for the best scoring fixed backbone models with less than 50% core identity to the wild-type sequence. Models with Rosetta energies within 6 Rosetta Energy Units of the lowest scoring fixed backbone model were identified that had sequence identities between 40% and 50%. The final FBAA sequence chosen for experimental characterization was selected from this filtered set. Sequence Logos of the 200 lowest energy sequences for each computational protocol illustrate the types of amino acids designed at each position (Figure S8).

The RosettaHoles algorithm was used to evaluate packing density in the redesigned proteins compared to wild-type CheA and statistics from high-resolution X-ray crystal structures (Sheffler and Baker, 2009). RosettaHoles explicitly searches for small voids in the protein that are inaccessible to water, and assigns a score to each residue between 0 and 1 that reflects

RES#	8	11	12	15	18	19	22	25	26	29	38	39	41	42	43	45	46	49	52	Core ID	Total ID
%BRD	82	100	96	100	70	96	100	93	95	100	85	96	60	100	85	94	100	99	86		
WT00	L	F	V	T	Y	L	L	T	L	L	L	I	E	A	F	A	L	L	M	100%	100%
TRAD	L	F	T	L	K	L	L	D	L	L	L	I	R	A	F	D	L	I	Q	61%	86%
FBAA	L	F	A	A	L	L	I	F	L	L	M	I	K	V	L	A	F	L	L	34%	70%
FBNN	I	V	A	L	H	F	I	F	I	M	K	V	K	I	Q	E	F	A	I	0%	58%
DRAA	R	A	A	L	L	L	I	V	L	L	K	I	K	A	Q	L	F	I	K	29%	68%
DRNN	I	V	T	L	L	I	V	D	I	V	Y	W	K	I	Y	L	V	M	I	0%	58%

RES#	53	61	64	65	68	69	71	72	75	76	87	90	91	93	94	97	100	101	104	Core RH	Total RH
%BRD	100	100	78	100	93	91	69	100	93	70	95	100	60	100	100	100	70	100	73		
WT00	A	M	L	C	L	E	I	L	A	R	L	I	F	G	V	I	M	V	I	0.41	0.63
TRAD	A	I	L	A	A	E	I	L	A	R	L	I	K	L	V	I	E	M	I	0.28	0.47
FBAA	A	M	M	A	A	A	L	A	A	A	L	L	K	M	A	L	F	V	L	0.23	0.46
FBNN	F	A	I	A	A	H	L	A	S	S	I	L	K	Y	A	L	F	M	L	0.27	0.50
DRAA	A	A	Y	A	G	E	I	A	A	A	L	L	K	Y	A	I	E	L	Y	0.42	0.57
DRNN	T	V	V	L	I	M	L	V	M	L	I	V	K	K	L	V	E	L	K	0.50	0.61

Figure 2. Comparison of Wild-Type and Designed Sequences

The core sequences for wild-type(WT00), the traditional output from RosettaDesign (TRAD), and the four design experiments FBAA, FBNN, DRAA, and DRNN are shown. The core and total sequence identity and the core and total RosettaHoles scores are given for each sequence. The percent of burial for each core position is shown as %BRD. Residue number is listed as RES#. Gray boxes indicate that a position is conserved between the wild-type sequence and one or more of the designed sequences. The one letter amino-acid codes are colored red (E,D), orange (M,C), green (L,A), blue (K,R,H), black (I,V), pink (N,Q,S,T), plum (F,W,Y), and glycine is shown white on a black background.

See also Figure S8.

the quality of packing around that residue. RosettaHoles scores closer to 1 indicate fewer voids. Residues in high-resolution crystal structures generally have scores between 0.5 and 1.0 for the entire protein. Models generated with the FBAA and FBNN protocols had RosettaHoles scores between 0.2 and 0.3 for the core residues, while the DRNN and DRAA models had scores between 0.4 and 0.5.

For each of the four protocols, a single sequence was selected for experimental validation (Figure 2). Sequences were selected for experimental testing based on their total Rosetta energy, the quality of packing, correctly predicted secondary structure, performance in ab initio folding experiments and deviation from the wild-type sequence (see Experimental Procedures for more details). In choosing a sequence from the FBAA protocol, we also did not consider sequences that had >50% core sequence identity with the wild-type sequence. For comparison, Figure 2 also shows the lowest scoring sequence generated with the FBAA protocol, labeled as TRAD. The TRAD sequence has 61% identity with the wild-type sequence in the core of the protein.

The computational experiments that incorporated flexible backbone design show subtle but important backbone movements (Figures 1B–1D and 3; Figures S3, S4, and S5). The backbone movements generated by this procedure are most often small local changes, with the most variation occurring at loops and termini. The designed sequence, DRNN, and the DRNN design model are the most varied from the native sequence and CheA crystal structure (Figures 1B–1D) and will be used to illustrate the types of backbone changes due to flexible backbone design. The final DRNN design model has a backbone RMSD of 1.6 Å compared to the CheA crystal structure. The largest backbone deviations between the design model and the crystal structure are seen in loop 3, helix 1, and helix 4. Although its sequence was not varied, loop 3 is pushed away from the center of the helix bundle because of the incorporation of a tryptophan at position 39, previously an isoleucine (Figure 3B). Using a global alignment, the backbone RMSD of loop 3 compared to the wild-type protein is 1.9 Å and the all-atom RMSD is 2.9 Å. Helix 1 is perturbed by 1.9 Å and helix 4 is perturbed by 2.1 Å (Figures 1C and 1D). The sequence identity of the 38 designed core residues is 0% compared to the native CheA Å, and the total sequence identity is 57%. A diverse set of mutations was predicted for the 38 core design positions; 27 mutations were hydrophobic/aromatic residues mutated to different hydrophobic/aromatic residues, 6 mutations were hydrophobic/aromatic residues mutated to polar residues, 3 mutations were polar residues mutated to hydrophobic/aromatic residues, and 2 mutations were polar amino acids mutated to polar amino acids. In this study, residue positions on the template CheA were classified as buried core positions if they were greater than 50% buried and made significant contacts with residues that were completely buried. This is an intentionally broad definition of the protein core and was intended to capture as much of the protein core as possible, without redesigning the entire protein.

Protein Expression and Behavior

Three of the designed proteins, FBAA, DRAA and DRNN expressed in *Escherichia coli* in soluble form at a variety of induction temperatures, 16°C–37°C, and produced greater than 33 mg/L of purified protein of culture. The proteins eluted as single peaks from size exclusion chromatography with apparent molecular weights consistent with the expected monomer weights, ~14 kD. In contrast, FBNN was found only in an

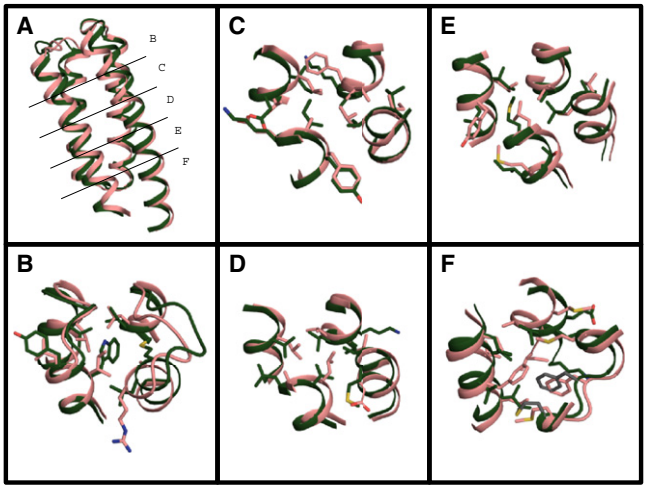


Figure 3. Comparison of Wild-Type Template and DRNN Design Model

The design and the wild-type bundle can be divided into five layers of interacting side chains.

(A) Shows the global view of the side-chain layers.

(B–F) Show the layers with wild-type in salmon and DRNN in green; positions that were not designed are shown in gray.

See also Figures S3, S4, and S5.

insoluble form. This behavior was seen at all tested temperatures and IPTG induction concentrations.

Biophysical Characterization of Redesigned CheA

Far-UV circular dichroism experiments confirmed that the designed proteins are primarily α-helical, with strong minima present at 208 nm and 220 nm (Figure 4A and Figures S1 and S2). Two of the designed proteins (FBAA, DRNN) did not unfold when subjected to temperatures of up to 97°C (Figure 4B and Figure S1). Chemical denaturation with guanidine hydrochloride (GdnHCl) shows that the designed proteins undergo highly cooperative unfolding events (Figure 4C and Figures S1 and S2). To determine accurate values for m, the temperature of the midpoint of unfolding (Tm), ΔH°, ΔCp°, and ΔG°, a Gibbs-Helmholtz surface was constructed by fitting several thermally induced denaturations in the presence of varying amounts of GdnHCl to the Gibbs-Helmholtz equation modified to take into account the effect of denaturant concentration (Table 1, Figures 4D and 4E, and Figures S1 and S2) (Kuhlman and Raleigh, 1998). The designed proteins are hyperthermostable with Tm values between 96 and 142°C and ΔG° values for unfolding between 5.5 and 16.2 kcal/mol. Remarkably, the computationally most ambitious design, DRNN, was the most stable. For comparison, the wild-type protein has a ΔG° of unfolding of 3.5 kcal/mol and a Tm of 91°C. The designed proteins have ΔCp° ranging from 0.83 to 1.1 kcal/mol*deg, which are typical values for proteins of this size (Myers et al., 1995). The ΔH° values range from 63 to 128 kcal/mol and the m values range from 1.9 to 3.4 kcal/(mol*M), the wild-type protein has values of 41 kcal/mol and 1.4 kcal/(mol*M) respectively.

Because DRNN was the most aggressive redesign of CheA and the most stable redesign, we choose it for high-resolution structure determination by NMR and X-ray crystallography.

X-Ray Crystal Structure of DRNN

The structure of the designed protein DRNN was determined by X-ray crystallography using diffraction data to a resolution of 1.85 Å. The structure was determined by molecular replacement using the design model with all side-chain atoms removed (to test for potential model bias). In the resulting $2F_o$-F_c electron density map, almost all of the side chains of the designed residues were clearly defined (Figure 5A). The final model has excellent stereochemical parameters (as determined by Molprobity [Davis et al., 2004]) and also ranks in the ~95th percentile for RosettaHoles packing score, 0.64, in the 1.0–2.0 Å resolution range (Figures 5B–5F) (Sheffler and Baker, 2009).

There is strong agreement between the DRNN design model and the experimentally determined structure (Figure 6 and Figure S9). The all-atom RMSD between the design model and both chains A and B in the asymmetric unit of the experimental structure are 1.5 Å and 1.3 Å, respectively. The 38 core design positions were predicted with good accuracy, 34 positions were observed in the correct rotamer state. Three design positions (Y37, K90, K92) were observed in different rotamer states due to the presence of crystal contacts (K90), or hydrogen bonding with nearby waters (Y37 and K92) that were not included in the design model. Valine 29 was observed in a rotamer different from that in the design model for unknown reasons. The prediction of the backbone of loop 3, which was extensively remodeled, is also highly accurate, with RMSD values of 0.32 Å and 0.38 Å, respectively, over backbone atoms for both chains A and B. Additionally, a hydrogen bond between the side chain of W39 and the backbone carbonyl oxygen of P33 in loop 1 is present in the crystal structure as designed.

We also compared the DRNN X-ray crystal structure to the 1TQG X-ray crystal structure, the starting template for the flexible backbone design procedure. The DRNN X-ray crystal structure is more similar to the DRNN design model than the starting template (Figure S9). The Cα RMSD between the DRNN crystal structure and the DRNN model is 0.8 Å, while the Cα RMSD between the DRNN crystal structure and the 1TQG starting template is 1.7 Å. The structures were further compared by making a histogram of distances between equivalent Cα atoms in the DRNN design model or 1TQG template and the DRNN crystal structure. While 48% of the equivalent Cα atoms were within 0.5 Å of each other when comparing the DRNN model to the DRNN crystal structure, only 29% were within 0.5 Å when comparing the 1TQG template to the DRNN crystal structure. Visually, the most striking comparison is for loop 3, where the DRNN design model is similar to the DRNN crystal structure while loop 3 from the template is more tightly packed against loop 1 (Figure 9).

NMR Structure of DRNN

To also obtain an NMR solution structure, DRNN was nominated as a PSI:Biology community outreach target assigned to the Northeast Structural Genomics Consortium (http://www.nesg.org; NESG target ID OR38). The 2D [^{15}N, ^{1}H]-HSQC spectrum of DRNN (Figure 7A) shows that a homogeneous NMR sample containing well-folded DRNN was obtained. Furthermore, the estimated correlation time for isotropic reorientation (τ_c = 5 ns) confirms that DRNN is monomeric in solution.

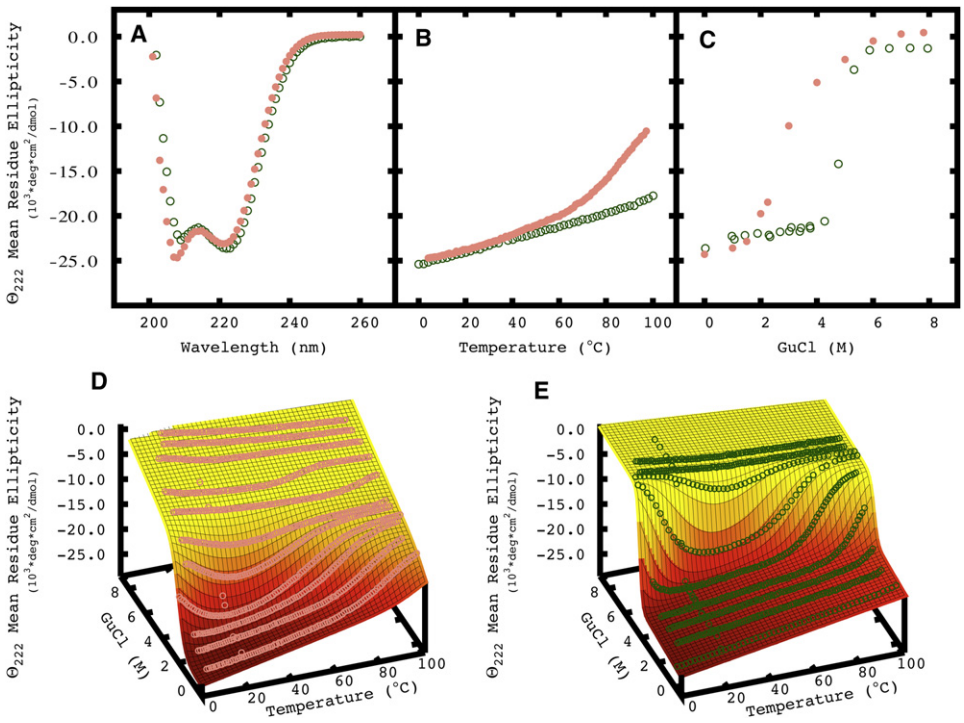


Figure 4. Biophysical Characterization of DRNN and Wild-Type Template

(A) Far-UV circular dichroism.

(B and C) Thermal denaturation (B) and chemical denaturation (C) of DRNN (green) and wild-type (salmon).

(D and E) Global fits (mesh) of thermal and chemical denaturation data for wild-type (D) and DRNN (E) obtained by fitting the data to the Gibbs-Helmholtz equation. All experiments were carried out at 10–20 μM protein concentration in 50 μM sodium phosphate at pH 7.4 and 20°C.

A high-quality NMR solution structure was obtained (Table S3), which like the crystal structure is similar to the design model: the RMSD calculated for the backbone heavy atoms N, C, and C′ between the DRNN design model and the mean coordinates of the 20 conformers representing the solution is 2 Å. Deviations between the design model and NMR structure are, however, primarily observed for the poorly defined conformations of the N-terminus of helix 1 and C terminus of helix 4 (Figure 7B). Hence, the corresponding RMSD calculated for residues 15–105 only is 1.2 Å. Both the DRNN design model and the DRNN X-ray crystal structure are in excellent agreement with the NMR derived conformational constraints, i.e., only 11 out of 1,406 distance constraints are violated by more than 0.5 Å in the crystal structure or the design model.

Comparison of χ1-angles in the NMR structure (Figure 8) and the design model reveals that 35 of the 38 designed core residues are in the expected (i.e., designed) rotameric state, and that significantly different rotamer states are observed only for L15, L18, and T53. Notably, the closest agreement between the NMR structure and the design model is observed in the region surrounding W39, with the all heavy atom RMSD calculated for the 19 closest neighbors of W39 being only 1.35 Å (Figure 7C and Figure S6).

Table 1. Thermodynamic Parameters for Wild-Type and Designed Sequences

Parameter	ΔG° (Kcal/mol)	Tm (°C)	ΔCp° (Kcal/mol*K)	ΔH° (Kcal/mol)	m (Kcal/mol*M)
WT	3.5	91	0.61	41	1.4
FBAA	14.9	144	0.83	107	2.3
DRAA	5.5	96	0.90	63	1.9
DRNN	16.2	142	1.08	128	3.4

Values for ΔG°, Tm, ΔCp°, ΔH°, and m were calculated by globally fitting a surface of chemical and thermal melts using the Gibbs-Helmholtz equations.

See also Figures S1 and S2.

DISCUSSION

The experimentally determined X-ray and NMR structures of DRNN show that it is possible to use flexible backbone design to aggressively sample sequence space compatible with a naturally occurring protein fold. The redesigned protein DRNN has zero core sequence identity with the parent CheA, but adopts a structure that is similar to that of CheA with distinct conformational perturbations that were predicted by the design protocol (Figure 9). The remodeling of protein sequences and conformations is a common path used by nature to evolve new functional proteins. Our results suggest that it should be possible to use computational protein design to achieve precise placements of backbone and side-chain atoms as a critical step in building novel binding and active sites. Of the four proteins that were

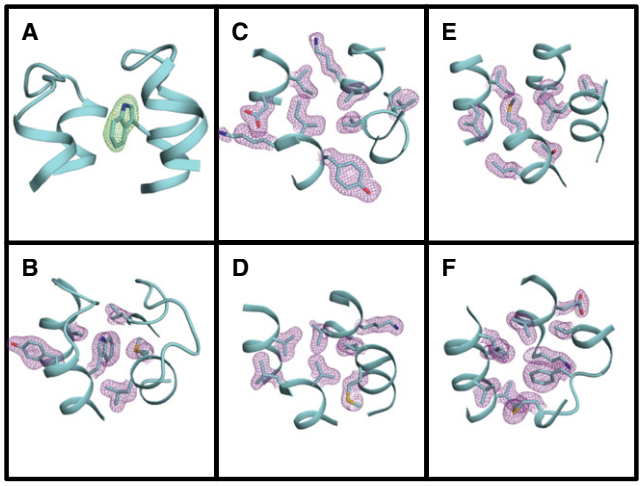


Figure 5. X-Ray Crystal Structure of DRNN

(A) Fo-Fc electron density (green) for residue W39 after molecular replacement using DRNN without side-chain atoms as the search model.

(B–F) Ribbon-presentation of the DRNN backbone in cyan.

The final 2Fo-Fc density (purple) for molecule A of the DRNN X-ray crystal structure in the five layers used to describe the wild-type and design model; sticks are shown for all design positions and residues 56M and 58F in (F).

See also Table S2.

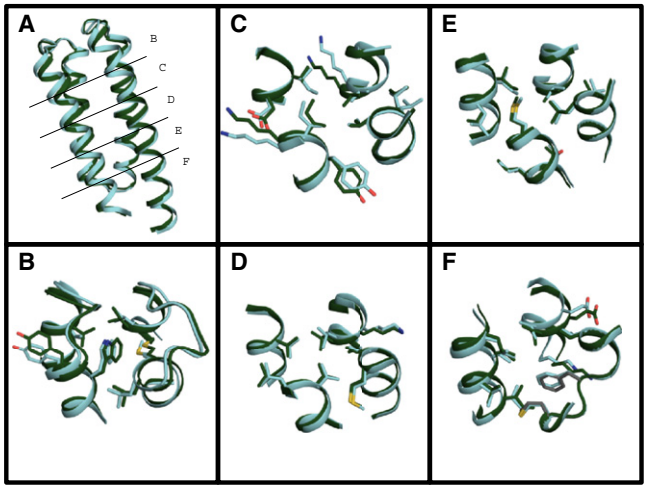


Figure 6. Comparison of DRNN Design Model and DRNN X-Ray Crystal Structure

The DRNN design model (green) and chain B of the X-ray crystal structure (cyan) shown in a global view (A) and as the five layers that make the bundle core (B–F); positions that were not designed are shown in gray in (F).

experimentally characterized, only FBNN failed to express in a soluble form in bacteria. This result suggests that there may be a limit to the degree that a sequence can be redesigned without explicit modeling of backbone relaxation, although additional experiments of this type are needed before more general conclusions can be made. Also, if constrained to using a fixed backbone during the design process, protocols that adjust the energy function to soften repulsive forces may be better suited for dramatically redesigning protein cores (Dahiyat and Mayo, 1997; Grigoryan et al., 2007).

The DRNN sequence has exceptional thermostability with a T_m > 140°C and a free energy of folding of −16 kcal/mol at 25°C. High stability has also been observed in previous computational redesigns of naturally occurring proteins (Dantas et al., 2003, 2007; Malakauskas and Mayo, 1998; Schweiker and Makhatadze, 2009). In many of these studies the whole protein was redesigned or mutations were dispersed between buried and exposed residues. Our results confirm that high thermostability can be achieved by computational remodeling of just the hydrophobic core. This was also demonstrated in a recent study from Borgo and Havranek (Borgo and Havranek, 2012). Iterative computational cycles of point mutations and backbone relaxation were used to identify small sets of mutations that fill voids in protein cores. The redesigns were stabilized by several kilocalories per mole.

Why is DRNN more stable than the wild-type protein? Possible sources of stability include the incorporation of amino acids with higher intrinsic propensity to form a helix, a burial of more hydrophobic surface area, and a preference for lower energy side-chain rotamers. One of the Rosetta scoring terms used during sequence optimization is based on the probability of observing an amino acid with a particular ϕ and ψ angle in naturally occurring protein structures. This scoring term accounts for the intrinsic preferences of the amino acids to be in α helices and β strands. Interestingly, the value for this score term on average is only slightly more favorable, 1%–2%, for DRNN than the wild-type protein. In fact, eighteen of the designed residues in DRNN are β-branched amino acids (valine, threonine or isoleucine), which are typically enriched in β strand structure (Minor and Kim, 1994). In contrast, ten of the designed positions are β-branched amino acids in the wild-type sequence.

Each amino-acid side chain has intrinsic preferences for the various rotamers that it can adopt. These preferences are highly dependent on the backbone ϕ and ψ angles of the residue. These preferences are incorporated in the Rosetta scoring function by evaluating the log odds of observing a particular rotamer in the protein database, conditioned on ϕ and ψ angle. Rosetta uses backbone-dependent rotamer statistics compiled by Dunbrack (Shapovalov and Dunbrack, 2011). On average, the rotamers used in DRNN (both in the model and in the crystal structure) are only slightly more favorable, 2%–3%, than the rotamers adopted in the wild-type structure (Figure S7).

The hydrophobic effect is the primary driving force for protein folding (Dill, 1990) and the burial of more hydrophobic atoms can increase protein stability (Lim et al., 1994; Munson et al., 1996). To evaluate the number of hydrophobic atoms buried in DRNN and wild-type CheA, the solvent accessible surface area of each atom was calculated using a 1.4 Å probe, representative of water solvent. Fourteen additional non-hydrogen hydrophobic atoms were completely buried in DRNN, versus the wild-type CheA and an additional 16 hydrophobic atoms are greater than 50% buried (Table S1). This suggests that the extreme thermostability of DRNN may be partially due to the burial of an additional 27 hydrophobic atoms. However, a similar analysis of the FBAA, FBNN, and DRAA design models indicates that there is not a simple correlation between the number of buried hydrophobic atoms and the observed changes in protein stability

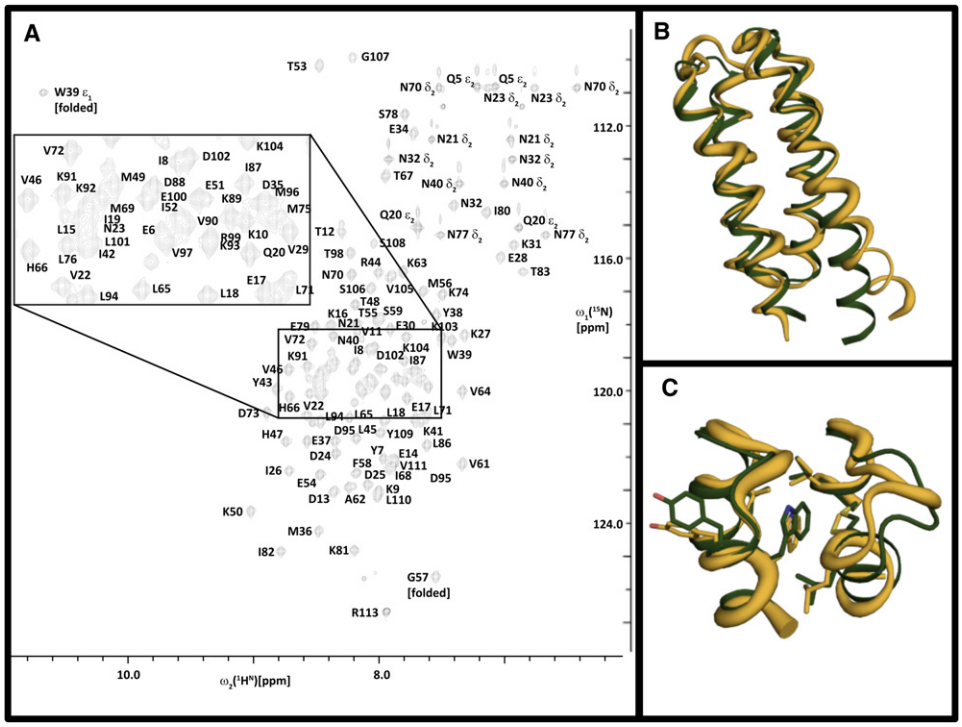


Figure 7. 2D [^{15}N,^{1}H] HSQC and NMR Solution Structure of DRNN

(A) 2D [^{15}N,^{1}H]-HSQC spectrum (~1 mM protein concentration, 20 mM sodium phosphate, pH 6.5) recorded at 750 MHz ^{1}H resonance frequency. Resonance assignments are indicated using the one-letter code for amino acids.

(B) Global comparison of the DRNN model (green) and the DRNN solution structure (orange).

(C) The region around W39 of the DRNN model and the solution structure (corresponds to layer B in Figures 5 and 6).

See also Figure S7.

(Table S1). While the FBAA design was nearly as stable as DRNN, in the FBAA design model there is one less buried hydrophobic atom than in the wild-type protein. In summary, we have not identified a single metric or characteristic that explains why FBAA and DRNN are more stable than DRAA and the wild-type protein. Like DRNN, the FBAA, FBNN, and DRAA models all have favorable Ramachandran dihedral angles and the side chains are modeled using favorable side chain torsion angles.

In this study we characterized DRNN using both X-ray crystallography and NMR spectroscopy. The X-ray structure is valuable for validating the details of side-chain packing in the protein core, while the NMR structure allows one to detect internal dynamics in solution. The NMR spectra obtained for DRNN show that the protein's global conformation is not affected at room temperature by chemical exchange on the chemical shift timescale (milli- to micro-seconds). In future work, it will be interesting to explore the backbone and side-chain dynamics of DRNN at faster timescales (nanoseconds) and compare results with the wild-type protein and other computationally designed proteins: in a previous study of a designed three-helix bundle, DeGrado and co-workers demonstrated, by measuring NMR spin relaxation parameters, that the side chains in the core of a designed protein were more dynamic on average than is commonly observed for natural proteins (Walsh et al., 2001).

In conclusion, the redesign strategy applied here promises to be valuable for the stabilization of enzymes, ligand-binding proteins, and protein-protein interface partners where preservation of a functional surface or pocket is important. In these cases, our approach can be extended by constraining the relative spatial locations of functionally important residues, while surrounding residues are remodeled in sequence and structural space. Design with backbone flexibility will also be important for repurposing proteins to bind novel substrates and ligands. In this case, constraints can also be used to direct functional residues into desired conformations, while the surrounding sequence and backbone are optimized for the targeted new ligands.

EXPERIMENTAL PROCEDURES

Computational Methods

Fixed Backbone Protein Design Protocol

The fixed backbone protein design protocol used here is the standard fixed backbone design protocol released with Rosetta3.3. The design protocol consists of applying a side-chain packing algorithm, which uses simulated annealing to search rotamer space, using rotamers from the Dunbrack rotamer library and using the Rosetta energy function to evaluate the fitness of sequences (Leaver-Fay et al., 2011).

Flexible Backbone Protein Design Protocol

The redesign sequences were generated using a new protocol within the Rosetta framework. The protocol has two stages, fixed backbone sequence design and fixed sequence backbone and side-chain dihedral optimization. The protocol iterates between these two stages until the energy difference between cycle i and cycle i-1 is less than 1.0 Rosetta Energy Units (REU), in practice this is ~5 redesign simulations for proteins between 100 and 200 residues.

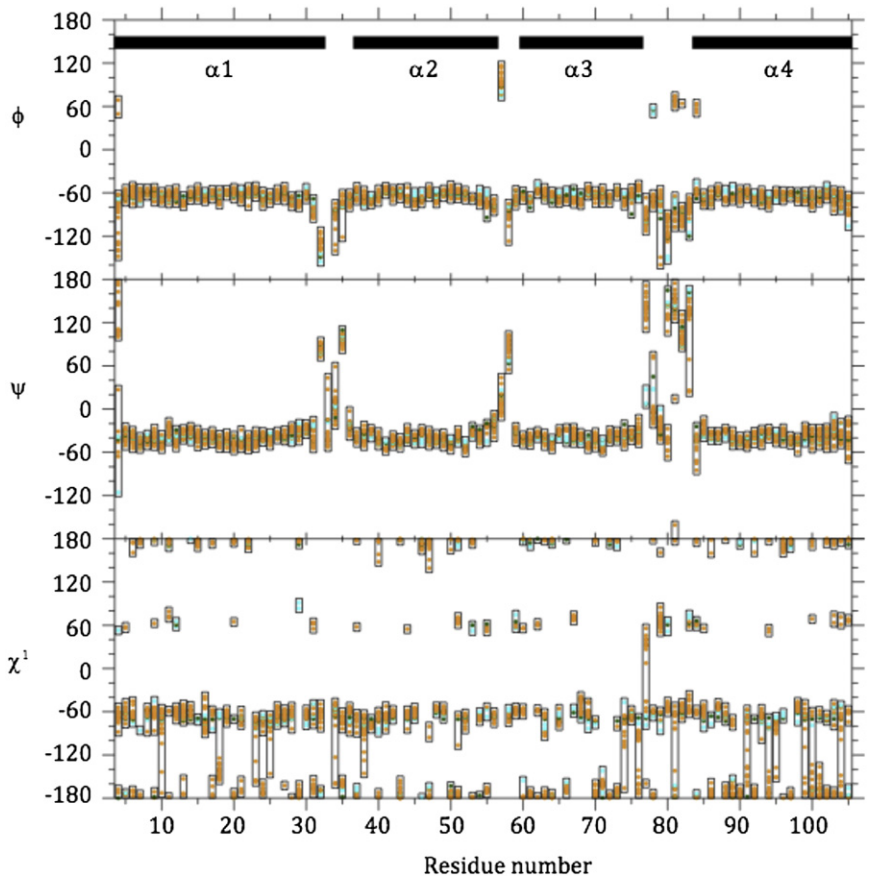


Figure 8. Comparison of DRNN NMR Structural Ensemble, DRNN X-Ray Crystal Structure, and DRNN Design Model in ϕ, ψ, and χ^1 Space

The values of the ensemble of conformers representing the NMR solution structure are shown in orange with boxes drawn around the observed range. The values observed for the two chains of the X-ray structure are shown in blue, and the values for the design model are shown in green. The black bars at the top indicate the location of the α helices. See also Figure S6.

The fixed backbone sequence design step uses the standard Rosetta side-chain packing algorithm described above and elsewhere. The fixed sequence backbone and side-chain dihedral optimization employs the Rosetta structure-optimization protocol used in structure prediction and refinement.

Computational Protein Design Experiments

Four different types of computational experiments were performed: (1) fixed backbone design where all amino acids were allowed at design positions (FBAA), (2) fixed backbone design where the native amino acid was not allowed at design positions (FBNN), (3) flexible backbone where all amino acids were allowed at design positions (DRAA), and (4) flexible backbone design where the native amino acid was not allowed at design positions (DRNN).

Core Redesign of the CheA Four-Helix Bundle

To redesign the core residues of the CheA four-helix bundle, 38 positions were identified as buried or partially buried. These positions have at least 15 neighbors each within 10 Å, where a neighbor is defined by the distance between Cβ atoms on residues i and j. Positions identified as core residues were visually inspected to remove any non-buried surface positions with a high number of neighbors. During this visual inspection, all attempts were made to include all partially buried side-chain positions, excluding positions identified as being in a loop by the DSSP algorithm (Kabsch and Sander, 1983). During the design stage, the 38 designable core positions were allowed to change amino-acid identity as described for each type of protein design experiment. An additional seven surface positions were allowed to design and mutate to any amino acid identity. The remaining 60 positions were not allowed to change amino acid identity but were free to change rotamer state. The possible rotamer states for each amino acid type are taken from the Dunbrack backbone-dependent rotamer library (Dunbrack, 2002). The 38 core designable positions were given more rotamer freedom, allowing additional sampling of rotamer states, the side-chain chi angles where given 12 extra rotamer states at ± 0.25, 0.50, 0.75, 1.00, 1.25, and 1.50 standard deviations from the most favorable dihedral angles for each rotamer. The seven designable and 60 surface positions were given extra rotamer states at ± 0.5 and 1.0 standard deviation from the most favorable rotamer states. All positions were free to sample ϕ, ψ, ω, and all dihedral χ angles during backbone and side-chain perturbation and minimization. A total of 25,000 design simulations were performed for each computational protein design experiment.

Selection of Designed Sequences for Experimental Characterization

The 25,000 designed sequences were ranked by their quality of core packing, as measured by RosettaHoles, sequences with scores less than 0.5 (0.4 for FBAA and FBNN) were pruned (Shefler and Baker, 2009). Sequences where the core design positions were predominately of a single amino-acid type, greater than 50% of the design positions, were pruned. This filter eliminates sequences where the protein core is composed primarily of only a few amino-acid types, mostly alanine and leucine. The 50 lowest-scoring models, based on total Rosetta energy, were evaluated for their secondary structure propensities using the secondary structure prediction server Jpred 3 (Cole et al., 2008). All 50 design models were predicted to have similar secondary structures compared to the design model and the native CheA. The ten lowest-energy models were subjected to structure prediction using Rosetta's structure prediction method. This filter evaluates if the designed sequence is predicted to adopt the desired fold, all designed sequences recovered the desired fold. The ten lowest-energy sequences for each experiment were evaluated by eye and one sequence from each experiment was chosen for experimental characterization. The sequence chosen from the DRNN experiment was also the lowest-scoring sequence out of the 25,000 designed sequences generated in that experiment.

Protein Expression and Purification

A codon-optimized gene for each designed sequence, and a modified version of the wild-type CheA was purchased from Genscript, lowercase letters are due to cloning and capital letters are the designed sequences.

> 1TQG_MOD_WT
mGSHQEYLQQFVDETKEYLQNLNDTLDELEKNPEDMELINEAFRALHTLK
EMAETMGFSSMAKLCHTLENILDKARNSEIKITSDLLDKIKDGVDMITRMV
DKIVS
gsylvprgslehhhhhh*
> FBAA
mGSHQEYLQKFADEAKELLQNINDFLKELEKNPEDMEMINKVLRAFHTLKE
LAETMGFSSMAKMAHTAANLADKAANSEIKITSDLLDKLKDMADMLTRFV
DKLVS

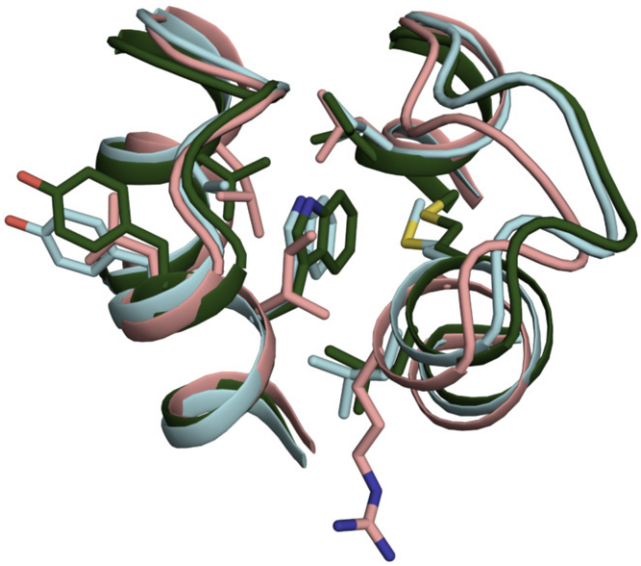


Figure 9. Comparison of Wild-Type Template, DRNN Design Model, and Crystal Structure

The wild-type template (salmon), DRNN design model (green), and the DRNN X-ray crystal structure (cyan) compared in the region of W39 (helix layer B shown in Figures 3B, 5B, and 6B).

See also Figure S9.

gsylvprgslehhhhhh*
> FBNN
mGSHQEYIQKVADELKEHFQNINDFIKEMEKNPEDMEKVNKIQREFHTAK
EIFETMGFSSAAKIAHTAHNLADKSSNSEIKITSDLIDKLKDYADMLTRFMD
KLVS
gsylvprgslehhhhhh*
> DRAA
mGSHDEYRKKAADELKELLQNINDVLDELEKNPEDMEKINKAQRLFHTIK
DKAQTMGFSSAAKYAHTGENIADKAANSEIKITSDLLDKLKDYADMITREL
DKYVS
gsylvprgslehhhhhh*
> DRNN
mGSHQEYIKKVTDELKELIQNVNDDIKEVEKNPEDMEYWNKIYRLVHTMKE
ITETMGFSSVAKVLHTIMNLVDKMLNSEIKITSDLIDKVKKKLDMVTRELDK
KVS
gsylvprgslehhhhhh*

Each gene was supplied as 4 μg of lyophilized DNA in pUC57 vector. The gene of interest was amplified from the parent vector using polymerase chain reaction (PCR), purified using a PCR-clean-up kit from Fermentas, double digested with NdeI and XhoI from NEB, purified again using a PCR-clean-up kit, and finally ligated into a pET-21 b(+) vector from Novagen that had been prepared by double-digesting with NdeI and XhoI and using a Fermentas gel-extraction clean-up kit. The ligation reaction product was transformed into XL-10 Gold cells from Stratagene.

Each protein was expressed in BL21 (DE3) pLysS cells from Stratagene. Cells were grown in LB media with 100 μg/ml ampicillin at 37°C to an OD_{600} of 0.6 and induced with 0.5 mM IPTG for 12 hours at 16°C. Cells were centrifuged at 4500 x g for 30 minutes and cell pellets were resuspended in 0.5 M NaCl, 0.2 M Na_2HPO_4/NaH_2PO_4 pH 7.0, 10% (v/v) glycerol, 1% (v/v) Triton X-100, dithiothreitol, and treated with DNAase, RNase, benzamidine, and phenylmethanesulfonylfluoride after three rounds of sonication. The cell lysate was cleared twice by centrifugation at 18,000 x g for 30 minutes. The supernatants were then filtered using a 0.22 μM filter from Millipore. The supernatant was purified by immobilized-metal affinity chromatography using a HisTRAP column from GE Healthcare. The elution was concentrated to 2 ml and further purified by size exclusion chromatography using a Superdex S75 column from GE Healthcare. For the FBNN sequence, induction conditions with IPTG concentrations ranging from 0.1 mM to 0.5 mM and induction ranging from 4 to 12 hours were tested. Ultimately, the FBNN sequence did not generate soluble protein.

Circular Dichroism

CD data were collected on a Jasco J-815 CD spectrometer. Far-UV CD scans were collected using a cuvette with a pathlength of 1 mm at concentrations between 10 and 20 μM protein in 50 μM sodium phosphate at pH 7.4 and 20°C. Thermal denaturation of samples was conducted between 4°C and 97°C while measuring the CD signals at 208 and 222 nm.

Chemical denaturation by guanidine hydrochloride (GdnHCl) was induced by mixing 15 μM designed protein in 0 M GdnHCl with 15 μM designed protein in 7.8 M GdnHCl. Great care was taken to ensure the concentration of designed protein in each sample was the same. The protein calculation was calculated using predicted extinction coefficients. The GdnHCl concentration was monitored by the change in refractive index. Thermodynamic parameters were calculated assuming and observing that the unfolding of the designed protein was a reversible two-state process by fitting both the thermal and chemical denaturations to the Gibbs-Helmholtz equation (Kuhlman and Raleigh, 1998).

Nuclear Magnetic Resonance Spectroscopy

The NMR samples of U-^{13}C, ^{15}N-DRNN and 5% ^{13}C, U-^{15}N-DRNN were prepared at concentrations of ~1.0 mM in 90% H_2O/10% D_2O solution containing 20 mM sodium phosphate (pH 6.5). An isotropic overall rotational correlation time of about 5 ns was inferred from averaged ^{15}N spin relaxation times, indicating that DRNN is monomeric in solution.

The following spectra were recorded for U-^{13}C, ^{15}N-DRNN at 25°C on a Varian INOVA 750 spectrometer (total measurement time: 6.5 days) equipped with a conventional ^{1}H{^{13}C, ^{15}N} probe: 2D [^{15}N, ^{1}H]-HSQC, aliphatic and aromatic 2D constant-time [^{13}C, ^{1}H]-HSQC, 3D HNCO, HNCACB, CBCA(CO)NH, HBHA(CO)NH, HN(CA)CO, aliphatic (H)CCH, (H)CCH-TOCSY Cavanagh J (2007) Protein NMR spectroscopy: principles and practice. Academic Press., and simultaneous 3D ^{15}N/$^{13}C^{aliphatic}$/$^{13}C^{aromatic}$-resolved [^{1}H, ^{1}H]-NOESY (mixing time 70 ms) (Shen et al., 2005). For 5% ^{13}C, U-^{15}N-DRNN, aliphatic 2D constant-time [^{13}C, ^{1}H]-HSQC spectra were acquired as described (Penhoat et al., 2005) at 25°C on a Varian INOVA 600 spectrometer (total measurement time: 12 hours) equipped with a conventional ^{1}H{^{13}C, ^{15}N} probe to obtain stereo-specific assignments for Val and Leu isopropyl groups (Neri et al., 1989).

All NMR spectra were processed using PROSA (Güntert et al., 1992) and analyzed using CARA (Keller, 2004). Sequence-specific backbone (HN, N, C^α, H^α, and CO) and H^β/C^β resonance assignments were obtained by using the program AutoAssign (Moseley et al., 2001; Zimmerman et al., 1997). Resonance assignment of side chains was accomplished using 3D (H)CCH, 3D (H)CCH-TOCSY, and 3D ^{15}N/$^{13}C^{aliphatic}$/$^{13}C^{aromatic}$-resolved [^{1}H, ^{1}H]-NOESY. Overall, for residues 1–113, sequence-specific resonance assignments were obtained for 95.2% of backbone and 95.7% of side-chain resonances assignable with the NMR experiments listed above (Table S3). Chemical shifts were deposited in the BioMagResBank (BMRB ID: 17612). ^{1}H-^{1}H upper distance limit constraints for structure calculation were obtained from 3D ^{15}N/$^{13}C^{aliphatic}$/$^{13}C^{aromatic}$-resolved [^{1}H, ^{1}H]-NOESY, and backbone dihedral angle constraints for residues located in well-defined regular secondary structure elements were derived from chemical shifts using the program TALOS+ (Cornilescu et al., 1999).

Automated NOE assignment was performed iteratively with CYANA (Güntert et al., 1997; Herrmann et al., 2002), and the results were verified by interactive spectral analysis. Stereospecific assignments of methylene protons were performed with the GLOMSA module of CYANA, and the final structure calculation was performed with CYANA followed by refinement of selected conformers in an "explicit water bath" (Linge et al., 2003) using the program CNS (Brünger et al., 1998). Validation of the 20 refined conformers was performed with the Protein Structure Validation Software (PSVS) server (Bhattacharya et al., 2007). The NMR structure was deposited in the PDB (PDB ID: 2LCH).

Protein Crystallization and X-Ray Crystallography

Crystallization of the designed protein was performed using the hanging-drop vapor-diffusion method at 20°C. Crystals formed in a drop consisting of 0.5 μl

of protein (20 mg/ml in 100 mM ammonium acetate) and 0.5 μl of well solution (0.2 M magnesium acetate and 20% (w/v) PEG 3350. Prior to data collection, crystals were cryo-protected by transferring them into well solution supplemented with 15% (v/v) ethylene glycol before plunging into liquid nitrogen. Crystals diffracted X-rays to a resolution of better than 1.8 Å, exhibited the symmetry of space group P1 with cell parameters of a = 25.6 Å, b = 43.9 Å, c = 47.7 Å, $\alpha = 63.89^\circ$, $\beta = 80.02^\circ$, $\gamma = 87.00^\circ$, and contained two molecules in the asymmetric unit (solvent content = 36%). Diffraction data were collected at 100 K at the Advanced Proton Source GM/CA CAT 23IDB beamline. The diffraction data were processed using HKL2000 (Otwinowski and Minor, 1997). The crystal suffered from directional diffraction anisotropy. This was corrected using an automated webserver (Strong et al., 2006).

The structure was determined by molecular replacement using the program Phaser (McCoy et al., 2007); the computationally designed model was used as a search model. To test for model bias, side-chain atoms were not included in the search model. After molecular replacement and an initial round of refinement the designed side-chain positions were clearly visible in F_o-F_c and $2F_o$-F_c electron density maps. Iterative rounds of refinement were conducted with Refmac5 (Vagin et al., 2004) from the CCP4 suite (Winn et al., 2011) interspersed with manual adjustments to the model using the program COOT (Emsley et al., 2010). The final model contains two molecules in the asymmetric unit with all residues defined in the electron density, except for residue 1 in chain A and residues 1–3 in chain B. Ramachandran statistics for the final DRNN structure model show that the backbone dihedral angles of all residues are in the favored region (Table S2). The structure was deposited in the protein data bank as PDB code 3U3B.

SUPPLEMENTAL INFORMATION

Supplemental Information includes nine figures and three tables and can be found with this article online at doi:10.1016/j.str.2012.03.026.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health Grants RO1GM073960 (to B.A.K.) and U54 GM094597 (to T.S.) and an award from the W.M. Keck Foundation.

Received: October 26, 2011
Revised: February 15, 2012
Accepted: March 30, 2012
Published online: May 24, 2012

REFERENCES

Apgar, J.R., Hahn, S., Grigoryan, G., and Keating, A.E. (2009). Cluster expansion models for flexible backbone protein energetics. J. Comput. Chem. *30*, 2402–2413.

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

Borgo, B., and Havranek, J.J. (2012). Automated selection of stabilizing mutations in designed and natural proteins. Proc. Natl. Acad. Sci. USA *109*, 1494–1499.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Cavanagh, J. (2007). Protein NMR spectroscopy: principles and practice. (Academic Press).

Cole, C., Barber, J.D., and Barton, G.J. (2008). The Jpred 3 secondary structure prediction server. Nucleic Acids Res. *36* (Web Server issue), 197–201.

Cornilescu, G., Delaglio, F., and Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J. Biomol. NMR *13*, 289–302.

Correia, B.E., Ban, Y.E., Friend, D.J., Ellingson, K., Xu, H., Boni, E., Bradley-Hewitt, T., Bruhn-Johannsen, J.F., Stamatatos, L., Strong, R.K., and Schief, W.R. (2011). Computational protein design using flexible backbone remodeling and resurfacing: case studies in structure-based antigen design. J. Mol. Biol. *405*, 284–297.

Dahiyat, B.I., and Mayo, S.L. (1997). Probing the role of packing specificity in protein design. Proc. Natl. Acad. Sci. USA *94*, 10172–10177.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J. Mol. Biol. *332*, 449–460.

Dantas, G., Corrent, C., Reichow, S.L., Havranek, J.J., Eletr, Z.M., Isern, N.G., Kuhlman, B., Varani, G., Merritt, E.A., and Baker, D. (2007). High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. J. Mol. Biol. *366*, 1209–1221.

Davis, I.W., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2004). MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res. *32* (Web Server issue), W615-9.

Davis, I.W., Raha, K., Head, M.S., and Baker, D. (2009). Blind docking of pharmaceutically relevant compounds using RosettaLigand. Protein Sci. *18*, 1998–2002.

DeGrado, W.F., and Nilsson, B.O. (1997). Engineering and design Screening, selection and design: standing at the crossroads in three dimensions. Curr. Opin. Struct. Biol. *7*, 455–456.

Desjarlais, J.R., and Handel, T.M. (1999). Side-chain and backbone flexibility in protein core design. J. Mol. Biol. *290*, 305–318.

Dill, K.A. (1990). Dominant forces in protein folding. Biochemistry *29*, 7133–7155.

Dunbrack, R.L., Jr. (2002). Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. *12*, 431–440.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501.

Friedland, G.D., Linares, A.J., Smith, C.A., and Kortemme, T. (2008). A simple model of backbone flexibility improves modeling of side-chain conformational variability. J. Mol. Biol. *380*, 757–774.

Fung, H.K., Floudas, C.A., Taylor, M.S., Zhang, L., and Morikis, D. (2008). Toward full-sequence de novo protein design with flexible templates for human beta-defensin-2. Biophys. J. *94*, 584–599.

Georgiev, I., and Donald, B.R. (2007). Dead-end elimination with backbone flexibility. Bioinformatics *23*, i185–i194.

Gordon, D.B., Marshall, S.A., and Mayo, S.L. (1999). Energy functions for protein design. Curr. Opin. Struct. Biol. *9*, 509–513.

Grigoryan, G., and Degrado, W.F. (2011). Probing designability via a generalized model of helical bundle geometry. J. Mol. Biol. *405*, 1079–1100.

Grigoryan, G., Ochoa, A., and Keating, A.E. (2007). Computing van der Waals energies in the context of the rotamer approximation. Proteins *68*, 863–878.

Güntert, P., Dötsch, V., Wider, G., and Wüthrich, K. (1992). Processing of multidimensional NMR data with the new software PROSA. J. Biomol. NMR *6*, 619–629.

Güntert, P., Mumenthaler, C., and Wüthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. *273*, 283–298.

Harbury, P.B., Tidor, B., and Kim, P.S. (1995). Repacking protein cores with backbone freedom: structure prediction for coiled coils. Proc. Natl. Acad. Sci. USA *92*, 8408–8412.

Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T., and Kim, P.S. (1998). High-resolution protein design with backbone freedom. Science *282*, 1462–1467.

Havranek, J.J., and Baker, D. (2009). Motif-directed flexible backbone design of functional interactions. Protein Sci. *18*, 1293–1305.

Hecht, M.H., Richardson, J.S., Richardson, D.C., and Ogden, R.C. (1990). De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. Science *249*, 884–891.

Herrmann, T., Güntert, P., and Wüthrich, K. (2002). Protein NMR structure determination with automated NOE assignment using the new software

CANDID and the torsion angle dynamics algorithm DYANA. J. Mol. Biol. *319*, 209–227.

Hill, R.B., and DeGrado, W.F. (2000). A polar, solvent-exposed residue can be essential for native protein structure. Structure *8*, 471–479.

Hu, X., Wang, H., Ke, H., and Kuhlman, B. (2007). High-resolution design of a protein loop. Proc. Natl. Acad. Sci. USA *104*, 17668–17673.

Kabsch, W., and Sander, C. (1983). How good are predictions of protein secondary structure? FEBS Lett. *155*, 179–182.

Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., and Hecht, M.H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. Science *262*, 1680–1685.

Keller, R. (2004). The computer aided resonance assignment tutorial (Goldau, Switzerland: Cantina Verlag).

Kuhlman, B., and Raleigh, D.P. (1998). Global analysis of the thermal and chemical denaturation of the N-terminal domain of the ribosomal protein L9 in H2O and D2O. Determination of the thermodynamic parameters, deltaH(o), deltaS(o), and deltaC(o)p and evaluation of solvent isotope effects. Protein Sci. *7*, 2405–2412.

Kuhlman, B., and Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA *97*, 10383–10388.

Kuhlman, B., O'Neill, J.W., Kim, D.E., Zhang, K.Y., and Baker, D. (2002). Accurate computer-based design of a new backbone conformation in the second turn of protein L. J. Mol. Biol. *315*, 471–477.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science *302*, 1364–1368.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. *487*, 545–574.

Lim, W.A., Hodel, A., Sauer, R.T., and Richards, F.M. (1994). The crystal structure of a mutant protein with altered but improved hydrophobic core packing. Proc. Natl. Acad. Sci. USA *91*, 423–427.

Linge, J.P., Williams, M.A., Spronk, C.A., Bonvin, A.M., and Nilges, M. (2003). Refinement of protein structures in explicit solvent. Proteins *50*, 496–506.

Lovejoy, B., Choe, S., Cascio, D., McRorie, D.K., DeGrado, W.F., and Eisenberg, D. (1993). Crystal structure of a synthetic triple-stranded α-helical bundle. Science *259*, 1288–1293.

Malakauskas, S.M., and Mayo, S.L. (1998). Design, structure and stability of a hyperthermophilic protein variant. Nat. Struct. Biol. *5*, 470–475.

Mandell, D.J., and Kortemme, T. (2009). Backbone flexibility in computational protein design. Curr. Opin. Biotechnol. *20*, 420–428.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J. Appl. Cryst. *40*, 658–674.

Minor, D.L., Jr., and Kim, P.S. (1994). Measurement of the beta-sheet-forming propensities of amino acids. Nature *367*, 660–663.

Moseley, H.N., Monleon, D., and Montelione, G.T. (2001). Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Methods Enzymol. *339*, 91–108.

Munson, M., Balasubramanian, S., Fleming, K.G., Nagi, A.D., O'Brien, R., Sturtevant, J.M., and Regan, L. (1996). What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. Protein Sci. *5*, 1584–1593.

Murphy, P.M., Bolduc, J.M., Gallaher, J.L., Stoddard, B.L., and Baker, D. (2009). Alteration of enzyme specificity by computational loop remodeling and design. Proc. Natl. Acad. Sci. USA *106*, 9215–9220.

Myers, J.K., Pace, C.N., and Scholtz, J.M. (1995). Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci. *4*, 2138–2148.

Neri, D., Szyperski, T., Otting, G., Senn, H., and Wüthrich, K. (1989). 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 ^{13}C labeling. Biochemistry *28*, 7510–7516.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Macromolecular Crystallography *276*, 307–326.

Penhoat, C.H., Li, Z., Atreya, H.S., Kim, S., Yee, A., Xiao, R., Murray, D., Arrowsmith, C.H., and Szyperski, T. (2005). NMR solution structure of Thermotoga maritima protein TM1509 reveals a Zn-metalloprotease-like tertiary structure. J. Struct. Funct. Genomics *6*, 51–62.

Pokala, N., and Handel, T.M. (2001). Review: protein design—where we were, where we are, where we're going. J. Struct. Biol. *134*, 269–281.

Sammond, D.W., Bosch, D.E., Butterfoss, G.L., Purbeck, C., Machius, M., Siderovski, D.P., and Kuhlman, B. (2011). Computational design of the sequence and structure of a protein-binding peptide. J. Am. Chem. Soc. *133*, 4190–4192.

Schweiker, K.L., and Makhatadze, G.I. (2009). Protein stabilization by the rational design of surface charge-charge interactions. Methods Mol. Biol. *490*, 261–283.

Shapovalov, M.V., and Dunbrack, R.L., Jr. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure *19*, 844–858.

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

Shen, Y., Atreya, H.S., Liu, G., and Szyperski, T. (2005). G-matrix Fourier transform NOESY-based protocol for high-quality protein structure determination. J. Am. Chem. Soc. *127*, 9085–9099.

Strong, M., Sawaya, M.R., Wang, S., Phillips, M., Cascio, D., and Eisenberg, D. (2006). Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. USA *103*, 8060–8065.

Su, A., and Mayo, S.L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Sci. *6*, 1701–1707.

Vagin, A.A., Steiner, R.A., Lebedev, A.A., Potterton, L., McNicholas, S., Long, F., and Murshudov, G.N. (2004). REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallogr. D Biol. Crystallogr. *60*, 2184–2195.

Walsh, S.T., Sukharev, V.I., Betz, S.F., Vekshin, N.L., and DeGrado, W.F. (2001). Hydrophobic core malleability of a de novo designed three-helix bundle protein. J. Mol. Biol. *305*, 361–373.

Willis, M.A., Bishop, B., Regan, L., and Brunger, A.T. (2000). Dramatic structural and thermodynamic consequences of repacking a protein's hydrophobic core. Structure *8*, 1319–1328.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. *67*, 235–242.

Yin, S., Ding, F., and Dokholyan, N.V. (2007). Modeling backbone flexibility improves protein stability estimation. Structure *15*, 1567–1576.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R., and Montelione, G.T. (1997). Automated analysis of protein NMR assignments using methods from artificial intelligence. J. Mol. Biol. *269*, 592–610.

Structure, Volume *20*

Supplemental Information

Increasing Sequence Diversity

with Flexible Backbone Protein Design:

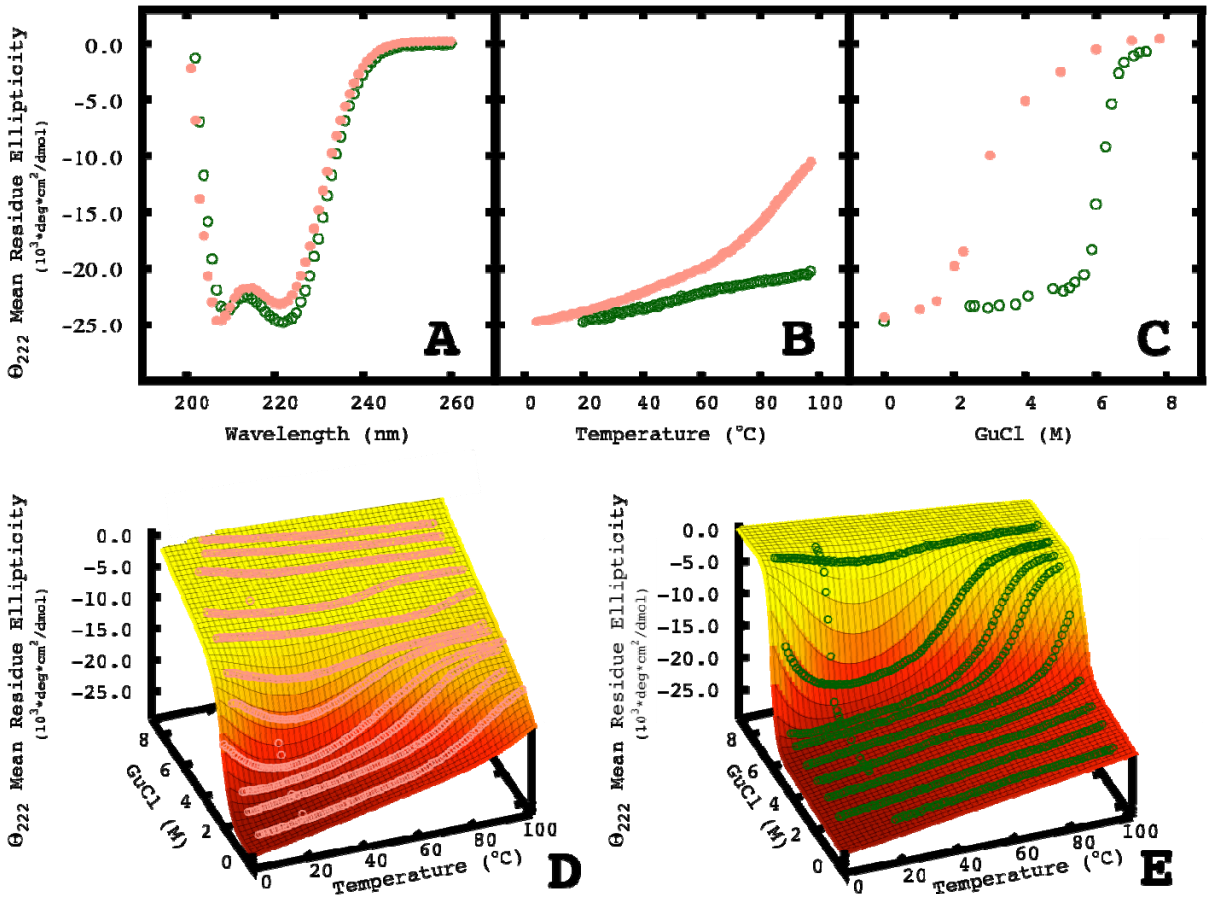
The Complete Redesign of a Protein Hydrophobic Core

Grant S. Murphy, Jeffrey L. Mills, Michael J. Miley, Mischa Machius, Thomas Szyperski, and Brian Kuhlman

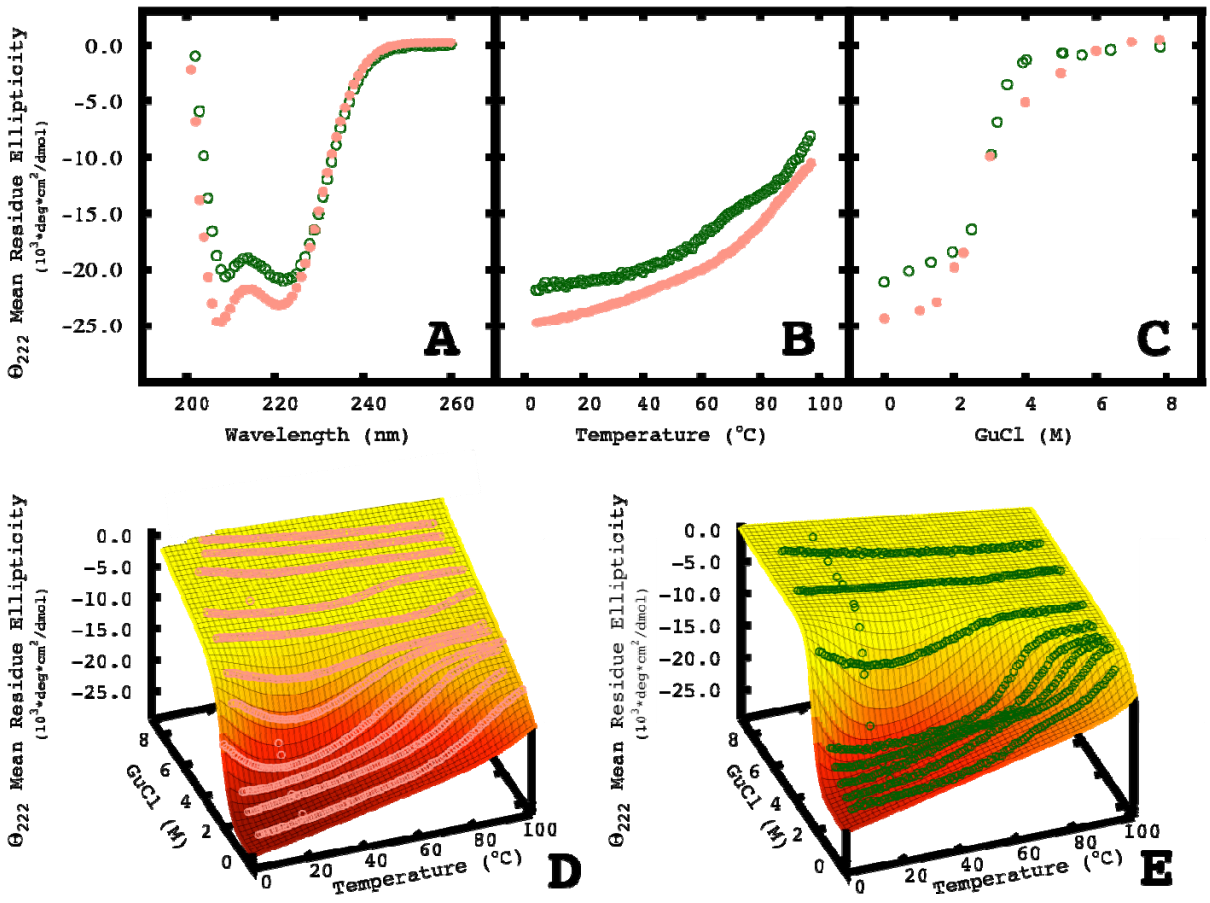
Inventory of Supplemental Information

- **Supplemental Figure 8: Comparison of low energy sequences from FBAA, FBNN, DRAA, and DRNN design experiments. This figure is a series of sequence logos to illustrate the types of sequences being designed in each experiment presented. See also Figure 2.**
- **Supplemental Figure 3: Comparison of wild-type and FBAA design model.**
- **This figure shows the differences between the wild-type protein and the FBAA model. See also Figure 3.**
- **Supplemental Figure 4: Comparison of wild-type and FBNN design model.**
- **This figure shows the differences between the wild-type protein and the FBNN model. See also Figure 3.**
- **Supplemental Figure 5: Comparison of wild-type and DRAA design model.**
- **This figure shows the differences between the wild-type protein and the DRAA model. See also Figure 3.**
- **Supplemental Table 1: Change in number of buried hydrophobic atoms and the Ramachandran and Rotamer Favorability for design models and DRNN X-ray structure. This table shows the difference in number of hydrophobic atoms buried in each design model or solved structure and the favorability of the Ramachandran space and rotamer space sampled. See also Figure 3.**
- **Supplemental Table 2: Crystallographic data collection and refinement statistics for DRNN. See also Figure 5.**
- **Supplemental Figure 7: DRNN and Wild-Type Difference Plot of Ramachandran and Rotamer Probability. This figure shows the differences in favorability between the wild-type protein and the DRNN structure. See also Figure 7.**
- **Supplemental Table 3: DRNN NMR Structure Statistics. See also Figure 7.**
- **Supplemental Figure 6: Comparison of DRNN design model and NMR solution structure. This figures gives a more in depth comparison of the DRNN design model and the DRNN NMR solution structure. See also Figure 8.**
- **Supplemental Figure 9: Raw counts of Cα atoms vs. distance to DRNN X-tal structure. This figure compares the distance between equivalent Cα atoms between the DRNN model, the DRNN X-ray crystal structure, and the wild-type protein X-ray crystal structure. See also Figure 9.**
- **Supplemental Figure 1: Biophysical characterization of FBAA and wild-type template. This figure shows the biophysical data for the FBAA sequence. See also Table 1**
- **Supplemental Figure 2: Biophysical characterization of DRAA and wild-type template. This figure shows the biophysical data for the DRAA sequence. See also Table 1**
- **Rosetta Command Lines**
  These command lines could be used to run the same experiments or modified to run similar experiments.

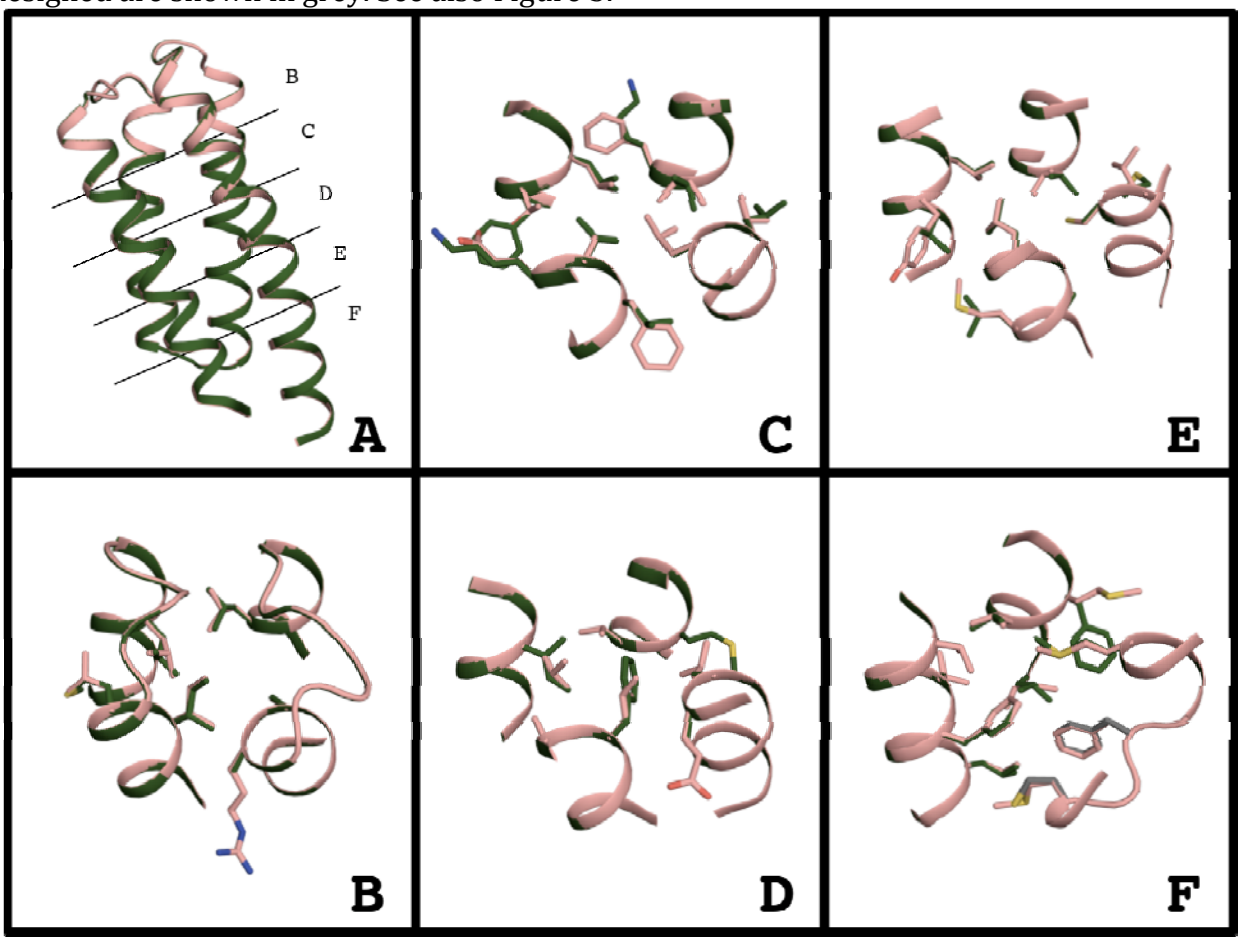
Supplemental Figure 1: Biophysical characterization of FBAA and wild-type template.
Far-UV Circular Dichroism (A), Thermal Denaturation (B), and Chemical Denaturation (C) of DRNN (green) and wild-type (salmon). Global fits (mesh) of thermal and chemical denaturation data for wild-type (D) and FBAA (E) obtained by fitting the data to the Gibbs-Helmholtz equation. All experiments were carried out at 10-20 μM protein concentration in 50 μM sodium phosphate at pH 7.4 and 20°C. See also Table 1.



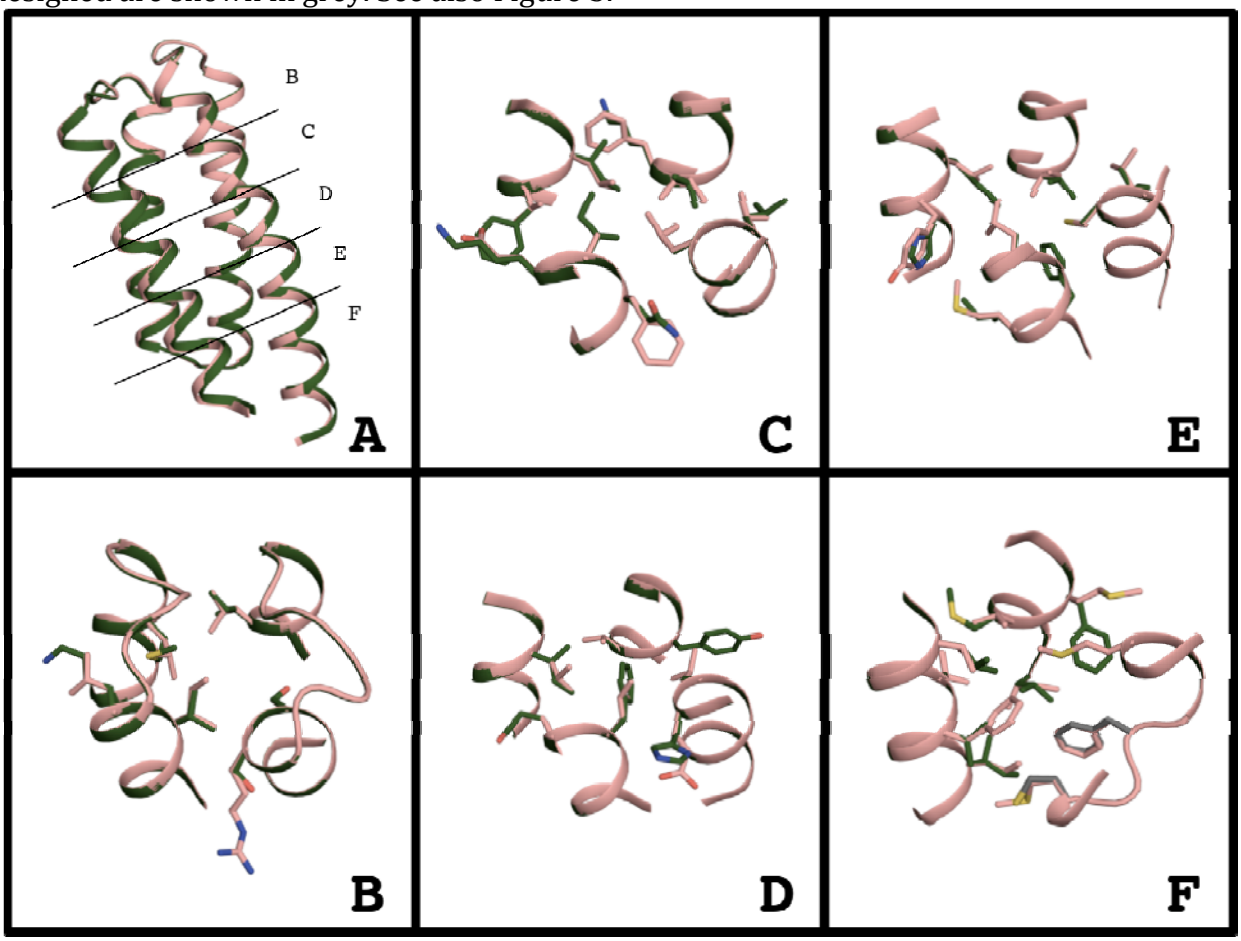
Supplemental Figure 2: Biophysical characterization of DRAA and wild-type template.
Far UV Circular Dichroism (A), Thermal Denaturation (B), and Chemical Denaturation (C) of DRNN (green) and wild-type (salmon). Global fits (mesh) of thermal and chemical denaturation data for wild-type (D) and DRAA (E) obtained by fitting the data to the Gibbs-Helmholtz equation. All experiments were carried out at 10-20 μM protein concentration in 50 μM sodium phosphate at pH 7.4 and 20°C. See also Table 1.



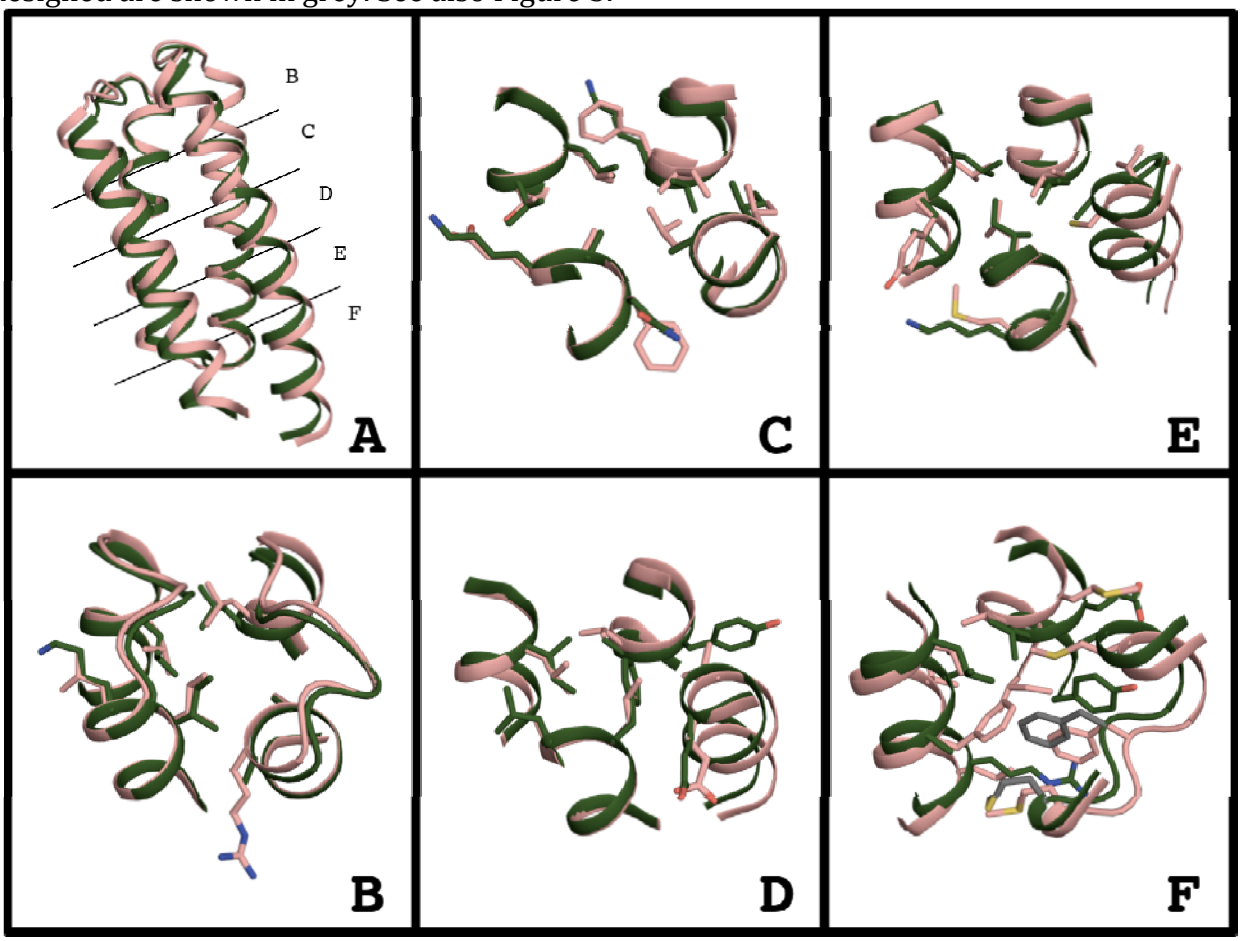
Supplemental Figure 3: Comparison of wild-type and FBAA design model.
The design and the wild-type bundle can be divided into 5 layers of interacting side-chains. Panel A shows the global view of the side-chain layers. Panels B-F show the layers with wild-type in salmon and FBAA in green, positions that were not designed are shown in grey. See also Figure 3.



Supplemental Figure 4: Comparison of wild-type and FBNN design model.
The design and the wild-type bundle can be divided into 5 layers of interacting side-chains. Panel A shows the global view of the side-chain layers. Panels B-F show the layers with wild-type in salmon and FBNN in green, positions that were not designed are shown in grey. See also Figure 3.

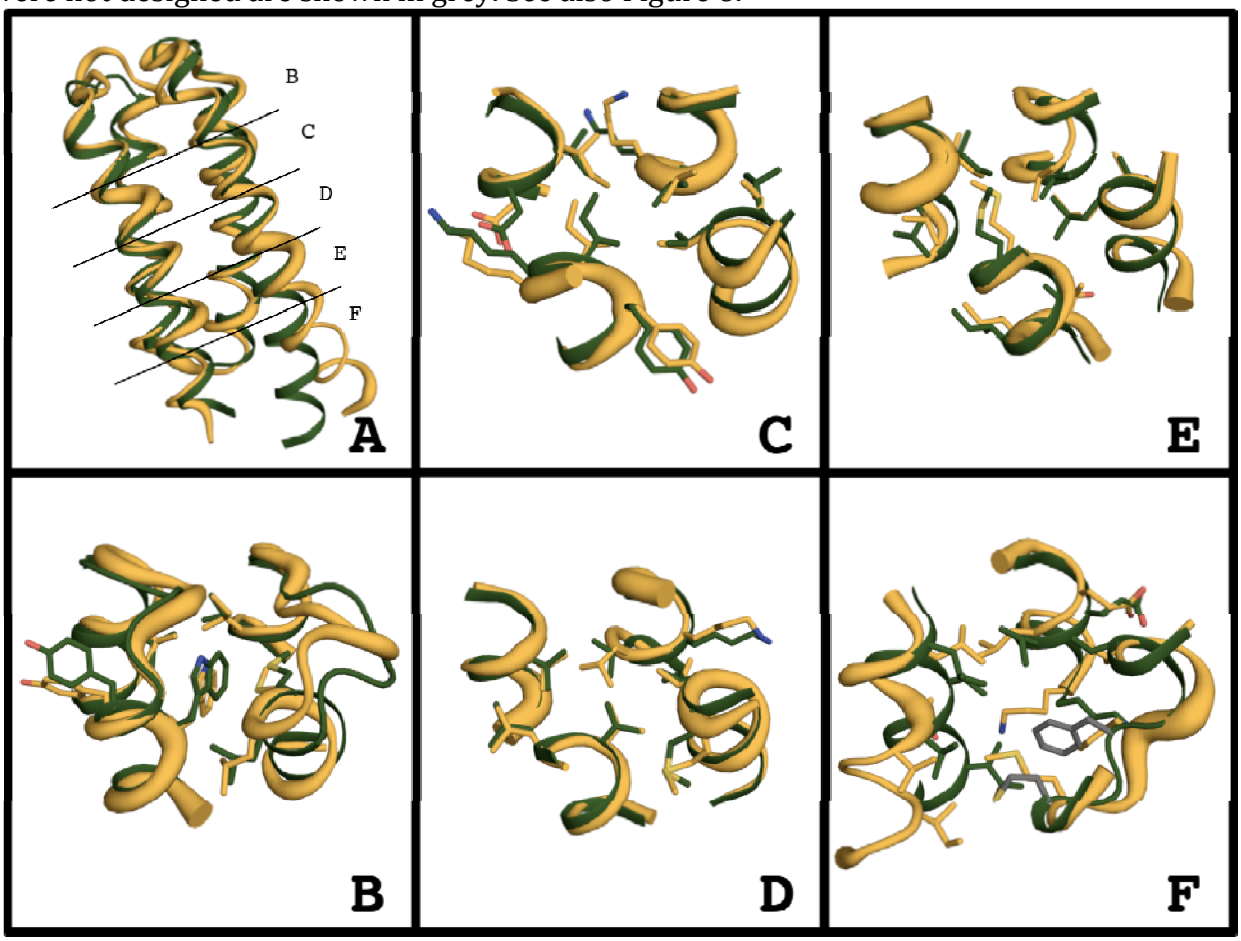


Supplemental Figure 5: Comparison of wild-type and DRAA design model.
The design and the wild-type bundle can be divided into 5 layers of interacting side-chains. Panel A shows the global view of the side-chain layers. Panels B-F show the layers with wild-type in salmon and DRAA in green, positions that were not designed are shown in grey. See also Figure 3.

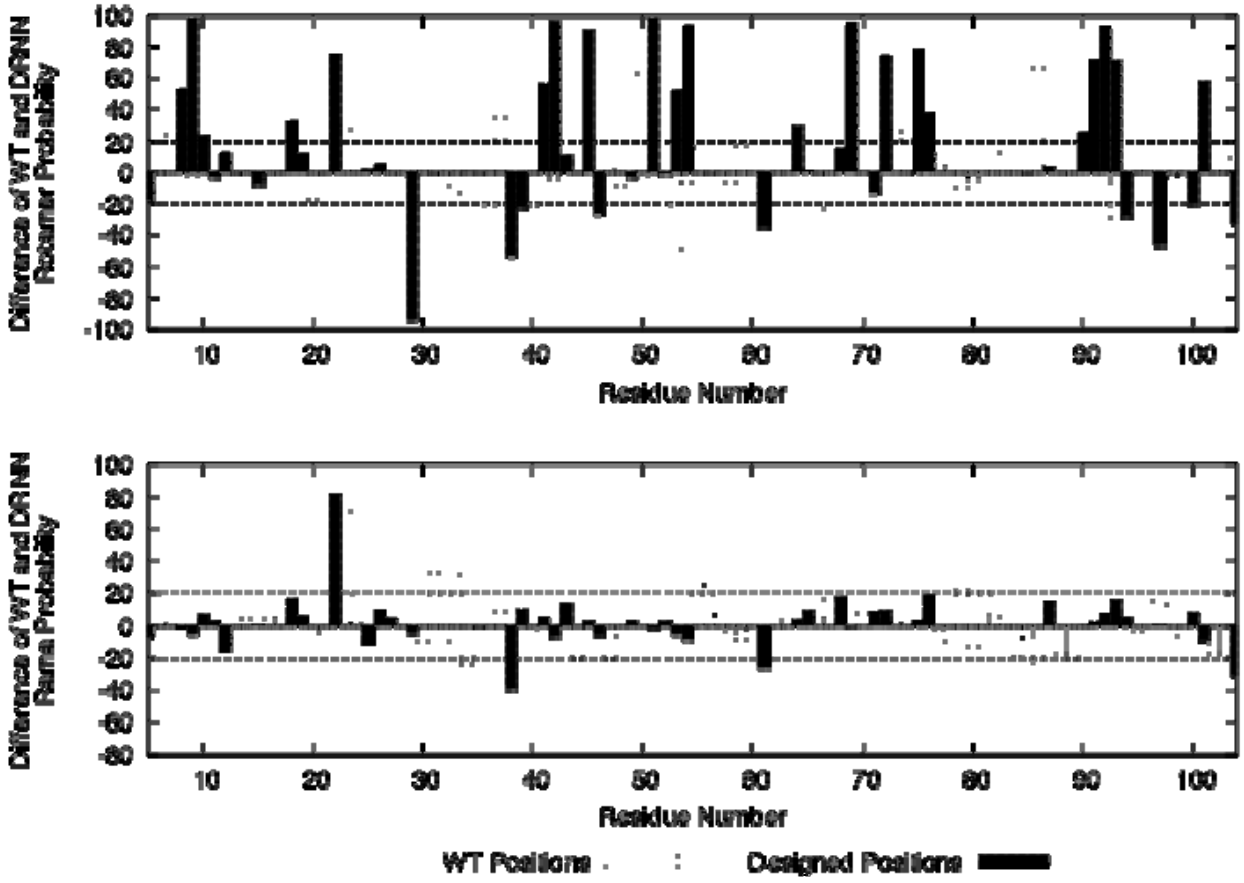


Supplemental Figure 6: Comparison of DRNN design model and NMR solution structure.

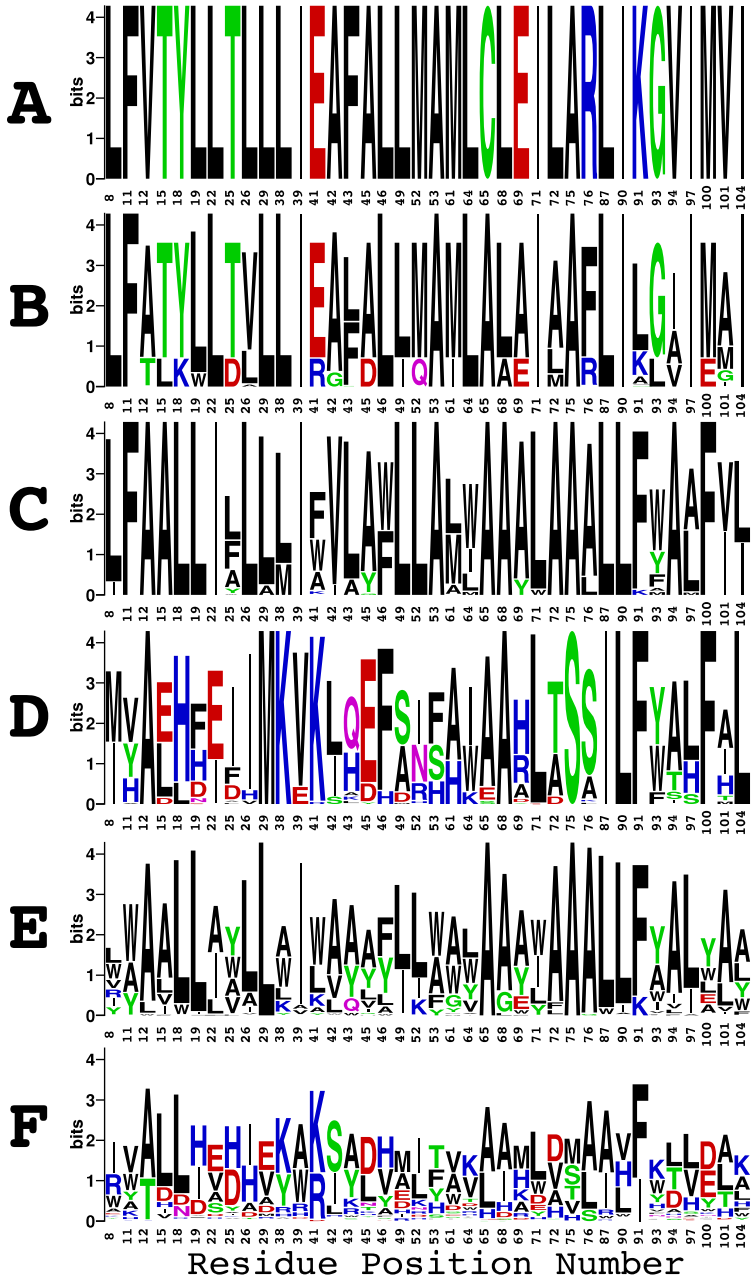
The DRNN design model (green) and NMR solution structure (orange) shown in a global view (A) and as the five layers that make the bundle core (B-F), positions that were not designed are shown in grey. See also Figure 8.



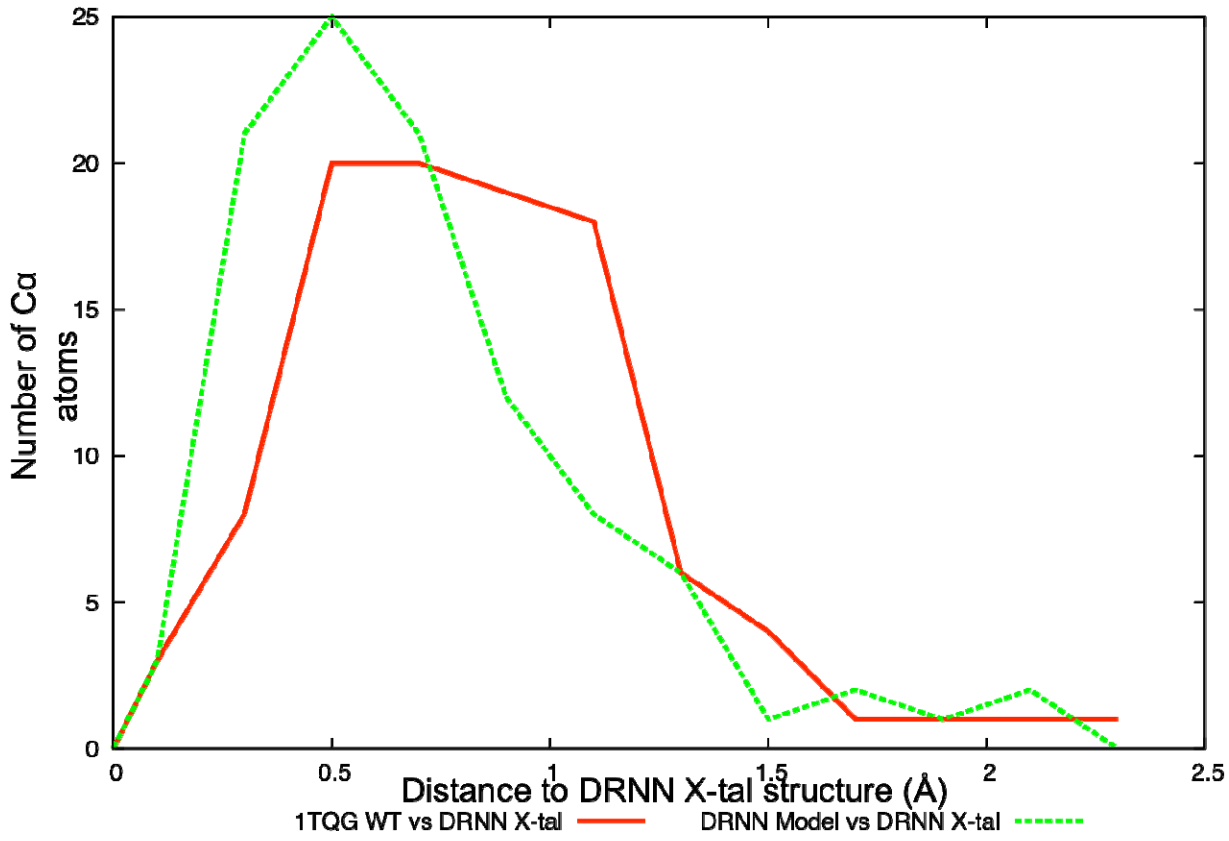
Supplemental Figure 7: DRNN and Wild-Type Difference Plot of Ramachandran and Rotamer Probability. The favorability of DRNN design rotamers is shown as the difference between the probabilities of observing the designed rotamer in structures available through the Protein Data Bank compared to the wild-type rotamer. Design positions are shown in black boxes and fixed positions are shown as open boxes. Rotamers that are more favorable in DRNN than the wild type have positive values. DRNN has slightly more favorable rotamers on average, 2-3%. The Ramachandran probabilities for DRNN are shown in the same manner and are on average slightly more favorable, 1-2%, than the wild-type protein. See also Figure 7.



Supplemental Figure 8: Comparison of low energy sequences from FBAA, FBNN, DRAA, and DRNN design experiments. The probability that an amino acid was designed at a core position is shown as bits of information for TRAD (B), FBAA (C), FBNN (D), DRAA (E), DRNN (F), and the wild type sequence is shown for comparison (A). The TRAD sequences are lowest scoring 200 sequences from the fixed backbone runs while the FBAA sequenecs are the lowest scoring fixed backbone sequences with less than 50% identity to the wild type sequence. The design relax procedure (E&F) leads to greater sequence variation than the fixed backbone procedure (C&D). See also Figure 2.



Supplemental Figure 9: Raw counts of Cα atoms vs. distance to DRNN X-tal structure.
The DRNN design model has a greater fraction of equivalent Cα atoms that are closer to the DRNN X-ray crystal structure than the 1TQG wild type X-ray crystal structure.
In total, 65% of Cα atoms are closer to the design model than to the WT structure.
See also Figure 9.



Supplemental Table 1: Change in number of buried hydrophobic atoms and the Ramachandran and Rotamer Favorability for design models and DRNN X-ray structure. See also Figure 3 and Supplemental Figures 4 & 5

	Total Number of extra hydrophobic Atoms	Change in number of 100% buried hydrophobic atoms	Change in number of hydrophobic atoms > 50% buried	Ramachandran Favorability	Rotamer Favorability
FBAA Model	3	-1	4	99%	99
FBNN Model	15	13	0	99%	99
DRAA Model	1	0	-1	99%	99
DRNN Model	25	17	10	99%	99
DRNN_XTAL	25	14	16	99%	99

Supplemental Table 2. Crystallographic data collection and refinement statistics for DRNN. See also Figure 5

Data collection		
Protein		DRNN
Wavelength (Å)		0.9794
Space group		P1
Cell dimensions	*a*, *b*, *c* (Å)	27.56, 43.96, 47.68
	α, β, γ (degrees)	63.89, 80.03, 87.01
Resolution (Å)		39.46 – 1.85 (1.87-1.85)
R_{merge} (%)		4.7 (39.4)
$I/\sigma I$		25.0 (1.8)
Unique reflections		16,156
Completeness (%)		96.9 (94.7)
Redundancy		2.30 (2.20)
Wilson B-factor (Å^2)		26.40
Refinement		
Resolution (Å)		39.46 – 1.85
No. of reflections work/free		16,156/819
R_{work} / R_{free}		0.176/0.214
No. of atoms	Protein	1735
	Ions + ligands	0
	Water	140
B-factors (Å^2)	Overall	41.29
	Protein	41.19
	Water	42.57
R.m.s. deviations	Bond lengths (Å)	0.011
	Bond angles (°)	1.068
Ramachandran	Favored (%)	100
	Generally Allowed (%)	0
	Disallowed (%)	0
Missing residues		molecule A: 1,106-113
		molecule B: 1-3,106-113

Numbers in parentheses refer to the highest-resolution shell

Supplemental Table 3: DRNN NMR Structure Statistics. See also Figure 7

Completeness of resonance assignments (%)[a]	
Backbone/side chain	95.2/95.7
Completeness of stereospecific assignments (%)	
Val and Leu isopropyl/$^{\beta}CH_2$ (%)	52.3/19.4
Conformationally-restricting distances constraints	
Intraresidue [i = j]	356
Sequential [|i - j| = 1]	395
Medium range [1 < |I - j| < 5]	380
Long range [|I - j| ≥ 5]	275
Dihedral angle constraints (ϕ/ψ)	86/86
Average number of constraints per residue	14.0
Average number of long range distance constraints per residue	2.4
CYANA target function ($Å^2$)	0.25 ± 0.07
Average number of distance constraint violations per conformer	
0.2-0.5 Å	0.2
>0.5Å	0
Average number of dihedral angle constraint violoations per conformer	
>10°	1.4
Average RMSD from mean coordinates (Å)	
Regular secondary structure elements[b], backbone heavy atoms	0.69
Regular secondary structure elements[b], all heavy atoms	1.13
Ordered residues[c], backbone heavy atoms	0.64
Ordered residues[c], all heavy atoms	1.08
Global quality scores[c] (raw/Z-score[d])	
PROCHECK G-factor (ϕ and ψ)	0.60/2.68
PROCHECK G-factor (all dihedral angles)	0.26/1.54
MOLPROBITY clash score	15.58/-1.15
Verify 3D	0.36/-1.61
ProsaII	0.92/1.12
RPF scores[e]	
Recall/precision/F-measure	0.959/0.844/0.898
DP-score	0.755
Ramachandran plot summary (%)[c]	
Most favored	96.8
Additionally allowed	3.2
Generously allowed	0.0
Disallowed	0.0

[a] Residues 1-113, excluding Pro ^{15}N, Lys and Arg side chain amino groups, hydroxyl protons, carboxyl groups of Asp and Glu, ε1 and imino groups of His, and aromatic non-protonated ^{13}C.
[b] Residues in regular secondary structure elements: 3-31, 36-56, 59-76, and 84-103.
[c] Ordered residues 5-79, 85-105.
[d] Calibrated relative to a set of high-resolution X-ray crystal structures for which the corresponding mean structure-quality score corresponds to a Z-score =0.0 (Bhattacharya et al., 2007)
[e] As described in (Huang et al., 2005)

Rosetta Command Lines

For DRAA and DRNN experiments the following command lines were used, and extra rotamers were assigned automatically as described in the methods

~/DesignRelaxApp.macosgccrelease -database ~/database/ -s *.pdb -core_design –DRNN

~/DesignRelaxApp.macosgccrelease -database ~/database/ -s *.pdb -core_design –DRAA

For FBAA experiments the standard Rosetta fixed backbone design protocol was used

~/fixbb.macosgccrelease -database ~/database/ -s *.pdb -resfile fbaa_resfile

the fbaa_resfile contained the following information

for fixed positions RES# A NATAA USE_INPUT_SC EX 1 LEVEL 4 EX 2 LEVEL 4

For designable positions RES# A ALLAA EX 1 LEVEL 6 EX 2 LEVEL 6

For FBNN experiments the standard Rosetta fixed backbone design protocol was used

~/fixbb.macosgccrelease -database ~/database/ -s *.pdb -resfile fbnn_resfile

the fbnn_resfile contained the following information

for fixed positions RES# A NATAA USE_INPUT_SC EX 1 LEVEL 4 EX 2 LEVEL 4

For designable positions RES# A NOTAA "NATIVE_RES" EX 1 LEVEL 6 EX 2 LEVEL 6

Depending on the computational resources available the flag –lin_mem_ig 10 may be need to use large number of rotamers for any of these experiments

Reference

Huang, Y.J., Powers, R., and Montelione, G.T. (2005). 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.