Computational de novo design of a four-helix bundle protein—DND_4HB

Grant S. Murphy,[1] Bharatwaj Sathyamoorthy,[2] Bryan S. Der,[3]
Mischa C. Machius,[4] Surya V. Pulavarti,[2,5] Thomas Szyperski,[2,5]
and Brian Kuhlman[3,6]*

[1]Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-3290
[2]Department of Chemistry, State University of New York at Buffalo, Buffalo, New York 14260
[3]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7260
[4]Center for Structural Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599
[5]Northeast Structural Genomics Consortium, Buffalo, New York 14260
[6]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Received 4 August 2014; Revised 30 September 2014; Accepted 30 September 2014
DOI: 10.1002/pro.2577
Published online 6 October 2014 proteinscience.org

Abstract: The de novo design of proteins is a rigorous test of our understanding of the key determinants of protein structure. The helix bundle is an interesting de novo design model system due to the diverse topologies that can be generated from a few simple α-helices. Previously, noncomputational studies demonstrated that connecting amphipathic helices together with short loops can sometimes generate helix bundle proteins, regardless of the bundle's exact sequence. However, using such methods, the precise positions of helices and side chains cannot be predetermined. Since protein function depends on exact positioning of residues, we examined if sequence design tools in the program Rosetta could be used to design a four-helix bundle with a predetermined structure. Helix position was specified using a folding procedure that constrained the design model to a defined topology, and iterative rounds of rotamer-based sequence design and backbone refinement were used to identify a low energy sequence for characterization. The designed protein, DND_4HB, unfolds cooperatively (T_m >90°C) and a NMR solution structure shows that it adopts the target helical bundle topology. Helices 2, 3, and 4 agree very closely with the design model (backbone RMSD = 1.11 Å) and >90% of the core side chain $\chi 1$ and $\chi 2$ angles are correctly predicted. Helix 1 lies in the target groove against the other helices, but is displaced 3 Å along the bundle axis. This result highlights the potential of computational design to create bundles with atomic-level precision, but also points at remaining challenges for achieving specific positioning between amphipathic helices.

Keywords: computational protein design; four-helix bundle; rosetta; de novo protein design; NMR structure

Introduction

De novo protein design is a rigorous test of our understanding of protein structure and can be used to test which features of proteins are critical for encoding well-folded structures that adopt a specific three-dimensional structure. For instance, what are the minimal design elements required to create a helix bundle protein? Early studies demonstrated that simple amphipathic peptides enriched in amino acids with high helical propensity will often

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Science Foundation; Grant number: MCB 0817857 (to T.S.); Grant sponsor: National Institutes of Health; Grant numbers: GM084042 (to B.K.) and GM094597 (to T.S).

Bharatwaj Sathyamoorthy's current address is Department of Biochemistry, Duke University, Durham, NC 27710, USA.

*Correspondence to: Brian Kuhlman, Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260. E-mail: bkuhlman@email.unc.edu

associate into multimers with high helical content, but that these complexes are unlikely to adopt unique three-dimensional structures with well-ordered packing and helix positioning.[1–5] However, if the amphipathic helices are linked together by short flexible linkers, the probability that they will adopt a more native-like structure increases significantly, but in general these designs are still highly molten.[3,5] Taking a similar approach but on a larger scale, Kamtekar *et al*. have engineered large protein libraries ($>10^6$) in which four amphipathic helices are specified using degenerate codons that code for either polar or non-polar amino acids at appropriate sites in each helix (referred to as a binary code), with short linkers rich in loop favoring residues.[4] Sequences from these libraries were shown to have some native-like features but were still molten globules. In an attempt to improve the folding quality of their library, Wei *et al*. made a second-generation library templated on the best member (N86) of their first library. In this library, they held most of the sequence fixed and combinatorially searched only a small region at the top of the bundle. At least two members from this templated library adopted a four-helix bundle with specific interactions formed between the helices[6,7] and one member from a larger library formed a domain swapped dimer.[8] This result shows that a simple binary code is sufficient to generate small native-like helical bundles, and demonstrates the importance of the hydrophobic effect in driving protein folding. It also suggests the need to limit the sequence spaced explored and the importance of a suitable starting point.

The binary code strategy provides a recipe for generating folded bundles, but the precise positions of the helices and side chains cannot be predetermined with this approach. A long-term goal for protein engineers is to be able to create novel proteins from scratch that perform important functions useful in medicine, industry, and research. As protein binding sites and functions depend on the exact positioning of side chain and backbone atoms, it will be important to develop computational methods that can design proteins with very high accuracy, perhaps with tolerances less than 1 Å. Over the last 20 years there has been significant progress in using computational methods to design proteins that adopt a predetermined structure or interaction. These approaches use an atomic-level representation of the protein to model the dominant forces in protein structure including steric repulsion, hydrogen bonding, desolvation, and torsional preferences. In a landmark article, Harbury *et al*. used computational design to create a unique four-stranded coiled coil that closely matched the design model.[9] This was the first demonstration that explicit consideration of side chain packing and rotamer preferences could be used to design helical proteins with atomic level accuracy. Since this study, computer-based methods have been used to design new α/β proteins,[10,11] protein-protein interactions,[12–14] nanocages,[15] and protein switches.[16] However, the accurate computer-based de novo design of a single chain four-helix bundle protein (not dependant on co-factors[17,18]) has not been previously reported.

Here, we examine if sequence and structure optimization methods in the modeling program Rosetta can be used to design an up-down four-helix bundle. An important step in de novo design is creating starting models for the protein backbone that adopt the target topology. In the design of a novel coiled-coil, Harbury *et al*. used analytic equations described by Crick to create a family of symmetric coiled-coil backbones.[9] Alternatively, when designing new α/β proteins with Rosetta, Kuhlman *et al*. created starting models by folding from extended peptide chains and using distance and secondary structure constraints to specify the target fold.[10,11] In order to create favorable local interactions, backbone fragments (3-mers and 9-mers) from naturally occurring proteins were used as the building blocks for folding. In these previous studies, backbone fragments were chosen simply based on desired patterns of secondary structure. Here, we further filter fragments by checking if they have starting and ending positions in three-dimensional space that are consistent with our target topology. This is particularly helpful for building the connecting loops between helices. Following fragment-based folding, we used iterative rounds of sequence design and structure refinement to identify low energy sequences. A single sequence, DND_4HB, was then chosen for experimental biophysical characterization. The protein DND_4HB is well folded and an NMR structure shows that it adopts the target left-handed four helix-bundle topology. Moreover, interactions between three of the helices were captured with very high accuracy, while one of the helices shifted by 3 Å relative to our design model.

Results

Generating starting structures using biased ab initio folding

To create backbone coordinates that would be the starting point for sequence optimization simulations, we used Rosetta's ab initio structure prediction protocol for folding a protein from an extended chain.[19] This protocol pieces together short fragments (3-mers and 9-mers) from high-resolution structures in search of conformations that place polar and non-polar amino acids in an appropriate environment and have favorable packing between secondary structural elements. In structure prediction, the protein data bank (PDB) is searched for fragments that have similar sequences to the query sequence. This

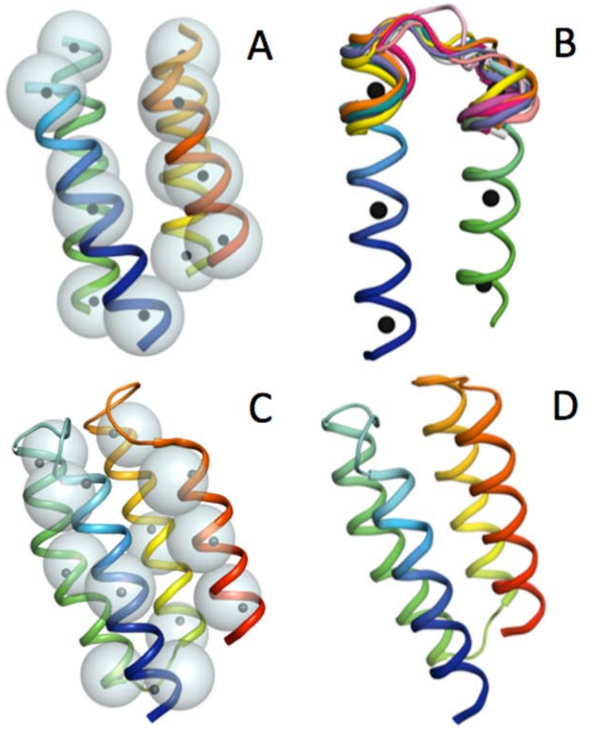

Figure 1. Starting structures and design models. To generate starting structures for design, an initial helix bundle is assembled without loops by aligning idealized helices (rainbow with helix axis points in black) to average-normalized helical positions (large grey spheres) (A). Bridge fragments that connect adjacent helices in the bundle are identified by RMSD alignment (B) and are used with axis constraints (large grey spheres) to bias fragment assembly. Fragment assembly and flexible backbone design were used to produce the DND_4HB design model (C), where DND_4HB's axis points (small black spheres) are within 3.5 Å of the axis constraints (large grey spheres) (C). The lowest energy DND_4HB forward folding model showing that the DND_4HB sequence is optimized for a left-handed four-helix bundle but loop 3 (orange loop) may adopt an alternate conformation (D).

helps ensure that the local structural elements in the predicted models are compatible with the sequence of the query protein. However, in de novo design, the target sequence is unknown. Previous efforts in de novo protein design with Rosetta picked fragments based solely on the desired secondary structure of each residue in the protein. We used this approach to pick fragments for helical regions of the bundle, but we used a more structurally explicit approach for picking loop fragments that span the connections between helices.

To identify loop fragments that would favor our target topology we started by building a model of four helices not connected by loops, called a "template bundle." The individual helices were built using idealized helical torsion angles ($\phi = -57$, $\psi = -47$), and the helices were placed near each other in relative orientations similar to that observed in naturally occurring four-helix bundles [Fig. 1(A)]. The PDB was then searched for fragments that have take-off and landing residues that align well with the start and end of the relevant helices [Fig. 1(B)]. These low scoring fragments became part of our move set in the ab initio folding experiments. Additionally, this process helped us determine what length loops to use for each connection in each template bundle. For the template bundle that produced DND_4HB, the most common length loop fragment that closed the gap between helix 1 and helix 2 was four residues, while for the 2:3 and 3:4 connections the most prevalent length loops were two and six residues, respectively.

To build starting backbones that adopt a desired topology, previous efforts in de novo design with Rosetta have made use of distance constraints between atom pairs to bias the folding simulations. In these cases, the target folds were α/β proteins for which the topology could largely be defined by specifying which residue pairs form backbone-backbone hydrogen bonds in the β-sheet. Similar constraints cannot be used for a helical bundle. Instead, before starting the folding simulation we defined the desired position of each helix with a set of three axis points that represent in three-dimensional space the desired location for the beginning, middle, and end of each helix [Fig. 1(A)]. These points were derived from the same template bundle model, with disconnected helices, that we used to pick loop fragments. During folding, distance constraints between these axis "target points" and the four helices in the model were used to bias the simulation toward that target fold.

During ab initio folding simulations, the protocol strives to bury hydrophobic amino acids and expose polar amino acids. Since at the start of de novo design there is not a defined sequence, we constructed naive sequences that were compatible with the target fold. Naive sequences are randomly generated sequences that are compatible with a target fold at the level of hydrophobicity. Naive sequences were based on the same template bundle used to pick loop fragments and define the target topology. Residues that were buried in this model were set to a random hydrophobic amino acid, while exposed positions were set to a random hydrophilic amino acid. A new naive sequence was generated for every starting structure produced.

With naive sequences, fragments, and constraints in hand, ab initio folding was used to build starting structures. Before pursuing a complete set of models, we first examined the impact of using the biased fragments and constraints during the folding simulations. Figure 2 shows the fraction of starting structures that adopt the desired topology using Rosetta's standard fragment assembly method using traditional fragments and no constraints (20%), using traditional fragments with axis constraints

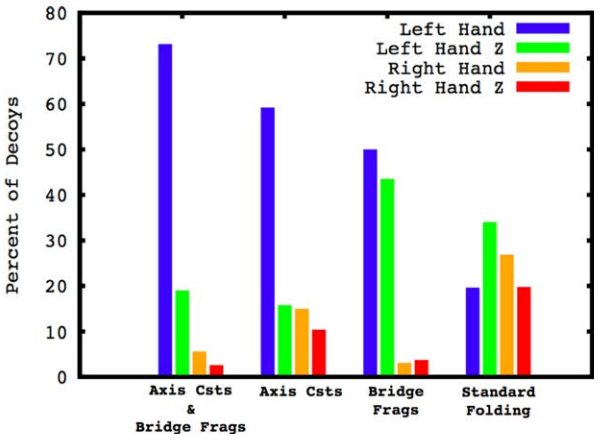

Figure 2. Percent of decoys with correct left-handed topology. Decoys were generated using Rosetta's fragment assembly protocol using axis constraints and bridge fragments, using only axis constraints, using only bridge fragments, or using the standard folding protocol. Each decoy was assigned as having a left-handed (blue), left-handed Z (green), right-handed (orange), or right-handed Z (red) topology. The left-hand four-helix bundle (blue) is the desired topology.

(59%), using bridge fragments without constraints (50%), and using bridge fragments with constraints (73%). We used this last method to generate ~100,000 starting structure models compatible with a left-handed four-helix bundle.

Iterative sequence and backbone refinement

Each of the starting models derived from ab initio folding served as input into a flexible backbone design protocol that iterated between sequence optimization and structure refinement in search of low energy sequence-structure pairs.[20] Sequence optimization was performed using a simulated annealing protocol with backbone dependent rotamers as the move set.[21] Structure refinement was performed using the FastRelax protocol in Rosetta, which iterates between repacking side chains and performing quasi-Newton minimization of torsional degrees of freedom while ramping in five steps the strength of the repulsive component of the Lennard-Jones term from 1/10th up to full strength, cycling from low- to high-strength repulsion three times.[22] Up to five rounds of sequence optimization and backbone refinement were used for each starting structure. In general, the refined models did not deviate significantly from the starting structures (average RMSD = 1.5 Å).

Selecting sequences for computational refolding and experimental characterization

We evaluated designed sequences based on total Rosetta energy, number of unsatisfied buried polar atoms, quality of packing using the RosettaHoles(v1) method,[23] and predicted secondary structure using JPRED.[24] The DND_4HB sequence was the lowest energy sequence produced with a total Rosetta energy of −162 and did not contain any unsatisfied buried polar atoms.

The DND_4HB designed model had high quality packing with a RosettaHoles score of 0.66. The JPRED secondary structure prediction server predicted the sequence to have four helices. Figure 1(C) shows a ribbon diagram of the design model of DND_4HB and the target axis constraints and the design model axis points.

Computational refolding

To assess the preference of the DND_4HB sequence for the target fold, we used Rosetta's structure prediction and full atom refinement methods to identify low energy conformations. Refolding of the DND_4HB sequence without biased fragments and constraints shows that the sequence adopts a left handed four-helix bundle but loop three may prefer an alternate conformation that is still consistent with the desired topology [Fig. 1(D) and Supporting Information Fig. 1]. The forward folding experiment also indicated that phenylalanine 54 may pack in an alternate conformation.

Biophysical characterization of DND_4HB

DND_4HB was overexpressed as soluble protein in *Escherichia coli* at a variety of induction temperatures and IPTG concentrations with yields greater than >15 mg/L. DND_4HB eluted as a single peak from a size exclusion column with an apparent molecular weight of ~12 kD, which is consistent with the predicted size as a monomer. Purified protein remained soluble at concentrations greater than 1 m*M*. Circular dichroism (CD) experiments showed that DND_4HB is α-helical, with strong minima present at the characteristic α-helix minima at 208 nm and 222 nm [Fig. 3(A)]. The stability of DND_4HB was determined by monitoring the CD signal at 208 nm and 222 nm as a function of temperature and guanidine hydrochloride (Gdn-HCl) [Fig. 3(B,C)]. In the absence of Gdn-HCl the unfolding transition begins at 80°C but is not complete by 100°C. To determine values for m, T_{m}, ΔH°, $\Delta C_{\mathrm{p}}^{\circ}$, and ΔG°, a Gibbs-Helmholtz surface was constructed by fitting several thermally induced denaturations in the presence of varying amounts of Gdn-HCl to the Gibbs-Helmholtz equation modified to account for the effect of denaturant concentration [Fig. 3(D), Methods, and Eq. (1)].

$$\Delta G{=}\Delta \mathrm{H}{-}\mathrm{T}\Delta \mathrm{S}{-}\mathrm{m}[GdnHCl] \qquad (1)$$

From this analysis, DND_4HB was determined to have a T_{m} value of 96°C and a ΔG° of folding of −4.9 kcal/mol. Additionally, parameters for

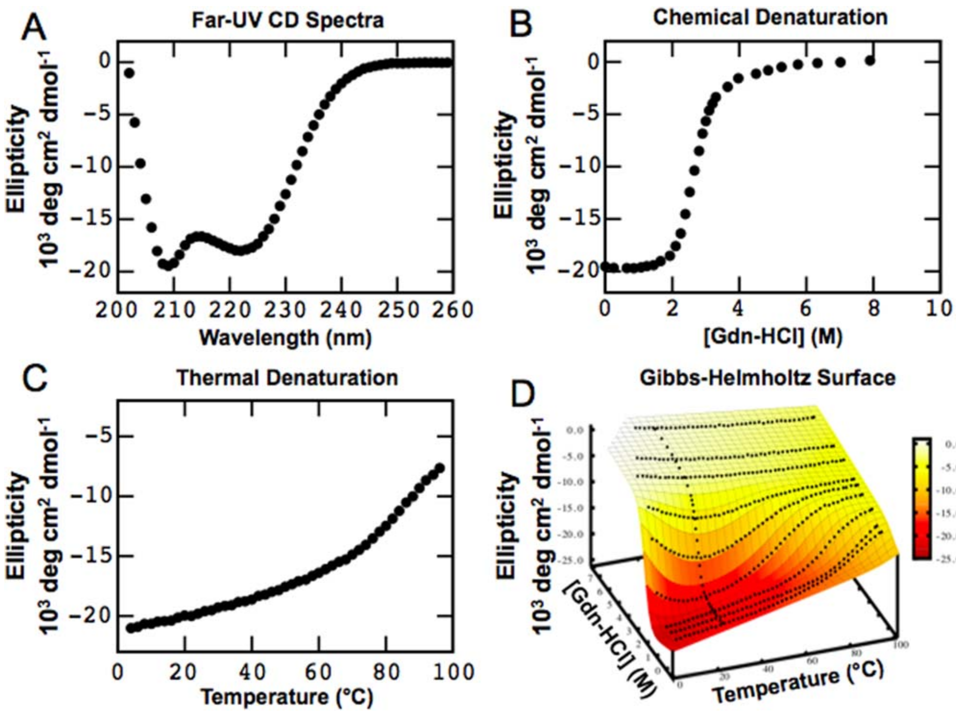

Figure 3. DND_4HB biophysical characterization. Far-UV CD of DND_4HB showing characteristic helix minima at 208 and 222 nm (A). CD signal at 222 nm versus concentration of Gdn-HCl (B) and versus temperature (C). CD signal at 222 nm versus temperature and Gdn-HCl with a global fit (mesh) to Eq. (1) (D).

$\Delta H^\circ = -52$ kcal/mol (25°C), $\Delta C_p{}^\circ = 0.7$ kcal/mol deg, and $m = 1.9$ kcal/(mol M) were calculated from the fit of the Gibbs-Helmholtz surface.

NMR spectroscopy of DND_4HB

Good signal dispersion was observed in one-dimensional ^{1}H NMR spectra recorded for unlabeled DNB_4HB and subsequently also in heteronuclear resolved two-dimensional NMR experiments recorded for ^{15}N-labeled and ^{15}N,^{13}C-labeled DNB_4B, which confirmed the finding inferred from CD that the designed protein is well folded. Moreover, DNB_4HB turned out to be highly soluble indicating that NMR-based structure determination appeared to be feasible. Hence, we acquired a comprehensive set of higher-dimensional NMR experiments for resonance assignment and structure determination (see Methods section).

NMR solution structure of DND_4HB

Protein DND_4HB was nominated as a PSI:Biology community outreach target assigned to the Northeast Structural Genomics Consortium (http://www.nesg.org; NESG target ID OR188). The two-dimensional [^{15}N, ^{1}H]-HSQC spectrum of DND_4HB (Fig. 4) shows that a homogeneous NMR sample containing well-folded DND_4HB was obtained. Furthermore, the correlation time for isotropic reorientation estimated from average ^{15}N spin relaxation times ($\tau_c = \sim 8.5$ ns; in agreement with 8.2 ns obtained from hydrodynamic calculations using the program HYDRONMR[25] confirmed that DND_4HB is monomeric in solution, as seen previously by size exclusion chromatography. A high-quality NMR solution structure was obtained (Supporting Information Table S1) and deposited into the protein data bank (PDB ID: 2lse).

Comparison of the DND_4HB NMR structure and the computationally predicted structure is the most rigorous test of the success of our design. We compared the predicted structure and the experimental structure by calculating several metrics: root mean square deviation (RMSD) values for backbone heavy atoms N, Cα, and C', by comparing ϕ, ψ, and χ_1 dihedral angles, and by identifying NOE-derived ^{1}H$-^1$H upper-distance limit constraints which are violated in the design model.

The RMSD value calculated for all backbone heavy atoms between the DND_4HB design model and the mean coordinates of the 20 conformers is 2.53 Å, and the RMSD to the most similar of the 20 conformers representing the solution structure is 2.32 Å. The corresponding superposition of the design model with the lowest energy NMR conformer revealed that helices 2 to 4 align more closely with the design model than helix 1. Helix 1 is shifted 3 Å along the long axis of the bundle (Fig. 6). As a result, the RMSD value obtained after superposition of only helices 2 to 4 (residues 26–39, 46–60, 70–81) is 1.11 Å.

The comparison of ϕ, ψ, and χ dihedral angles of the design model with the corresponding range

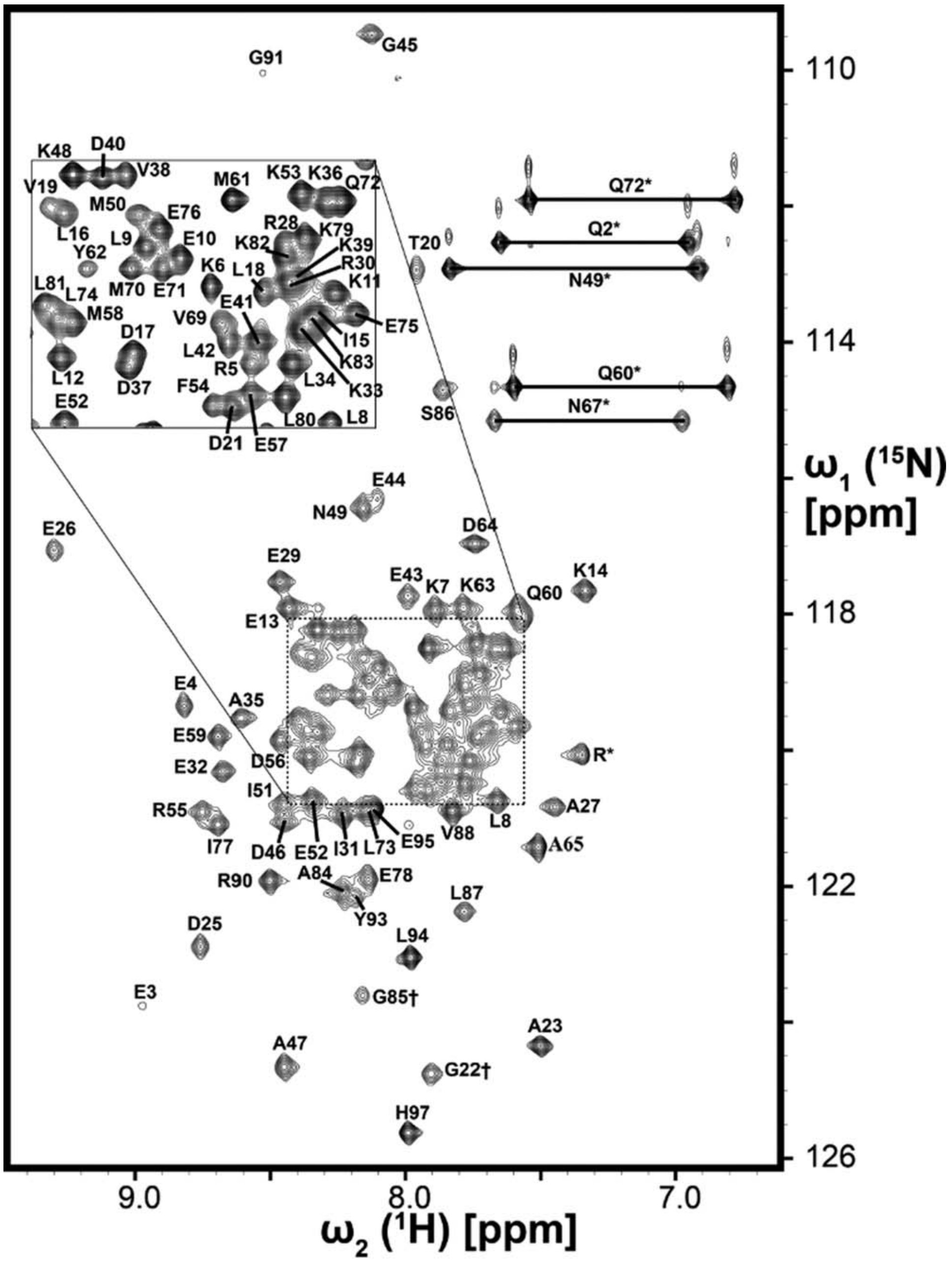

Figure 4. 2D [^{15}N,^{1}H] HSQC spectrum of DND_4HB. The [^{15}N,^{1}H] HSQC spectrum of DND_4HB (~2 m*M* in 50 m*M* sodium phosphate and 50 m*M* NaCl at pH 6.5) recorded at 600 MHz ^{1}H resonance frequency and at 25°C which shows very good signal dispersion and completeness of signal detection (>95%). Resonance assignments are indicated using one-letter amino acid code. Signals arising from side chains (Asn H$^{\delta 2}$/Nδ2,Gln H$^{\epsilon 2}$/N$^{\epsilon 2}$, and Arg H$^{\epsilon}$/N$^{\epsilon}$) are labeled with (*) and folded signals are designated with (†) next to the residue number. Signals arising from the last four residues of the C-terminal His purification tag were not sequence specifically assigned.

observed in the 20 conformers representing the NMR solution structure (Fig. 7) likewise documents the high accuracy of the design model. First, 97% of ϕ angles and 94% of ψ angles in the design model are within $\pm 15°$ of the corresponding angle in the NMR ensemble. Second, 88% of $\chi 1$ and 77% of $\chi 2$ angles are within $\pm 15°$ of the corresponding angle in the NMR ensemble. In the core of the protein (residues 8, 9, 12, 15, 16, 19, 27, 31, 34, 35, 38, 42, 50, 51, 54, 58, 61, 70, 73, 74, 77, 80, 81, and 84) the agreement is even higher: 91% (22 of 24 residues) in the design model have $\chi 1$ and $\chi 2$ angles within $\pm 15°$ of the corresponding angle in the NMR ensemble. Methionine 50 is the only core residue found to be in an obviously different rotamer state in the NMR structure.

The high similarity of NMR structure and design model is further evidenced by the finding that out of the 1586 NOE-derived ^{1}H$-^1$H upper distance limit constraints >96% are satisfied by the

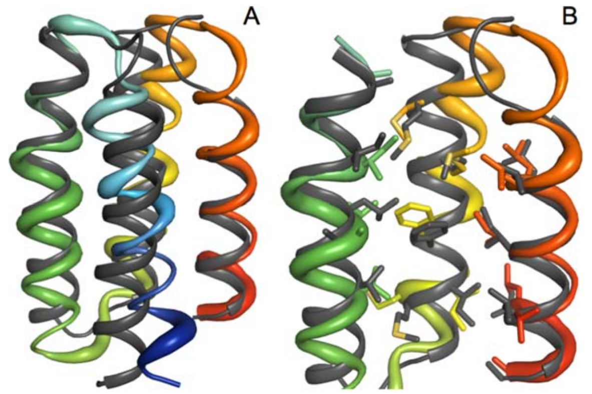

Figure 5. Comparison of DND_4HB design model and NMR structure. (A) The global similarity of the DND_4HB design model (grey) and experimental NMR structure (rainbow). Helix 1 (blue) is translated 3 Å relative to the design model. (B) The close alignment of helices 2 (green), 3 (yellow), and 4 (red) with the side chains of the core region around phenylalanine 54 shown as sticks.

DND_4HB design model, i.e., that less than 4% (63) are violated by more than 1 Å. Additionally, 124 of 126 dihedral constraints are satisfied by the DND_4HB design model. The two dihedral violations and 41 of the NOE violations are due to the incorrect modeling of loop 3. Of the remaining 22 NOE violations, 11 are violations between phenylalanine 54 from residues on helix 1 and helix 2. Despite these 11 violations, 91.2% (114) of phe54's NOE constraints are correctly satisfied. Another five violations are between residues on helix 1 and 2, and the last six violations are between helix 2 and 3.

Discussion

As predicted, DND_4HB adopts a left-handed four-helix bundle and the relative positioning and packing of helices 2, 3, and 4 very closely matches the design model (i.e., nearly within the resolution of the NMR structure). However, the modeling did not precisely determine the placement of helix 1 which is translated ~3Å along the long axis of the bundle. In the design model and the NMR structure, helix 1 residues Leu11, Ile14, and Val18 pack into large hydrophobic depressions formed by helices three and four (Fig. 6). The Rosetta energy function prefers the packing arrangement observed in the design model. However, the Rosetta energy difference between the design model and Rosetta models derived from the NMR structure is quite small (~5 REUs).

Furthermore, the design model has short loops connecting helices 1 and 2, and helices 2 and 3, and the predicted conformations for these loops are similar to the NMR structure. A longer six-residue loop was designed for the connection between helix 3 and 4, and the conformation of this loop was not correctly predicted: in the NMR structure the last three residues of the loop adopted a helical conformation and thus extended the N-terminus of helix 4. Despite this discrepancy in the loop, the packing between helix 3 and 4 is very similar in the NMR structure and the design model. Notably, in native proteins it has been shown that loops connecting regular secondary structure elements can vary dramatically without affecting the packing arrangement of the secondary structure elements.[26]

Computational refolding—referred to as "forward folding"—is a particularly attractive approach for evaluating de novo sequences before conducting wet-lab experiments. Frequently, when small (~100 residues) natural proteins with known structures are computationally folded using Rosetta's structure prediction method, a low energy and low RMSD population is identified. This has also been shown to be the case in double blind

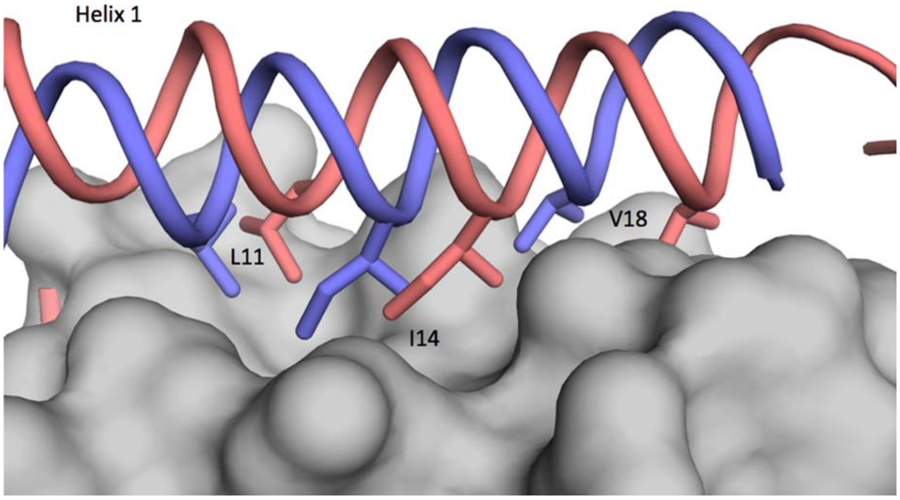

Figure 6. Packing of helix 1 in the DND_4HB design model and NMR structure. Helix 1 is displaced 3 Å along the long axis of the protein in the NMR structure (salmon) compared with the design model (design model). Helix 1 core residues are shown as sticks and labeled with helices 3 and 4 from the NMR structure shown as a gray surface. Helix 2 is not shown for clarity.

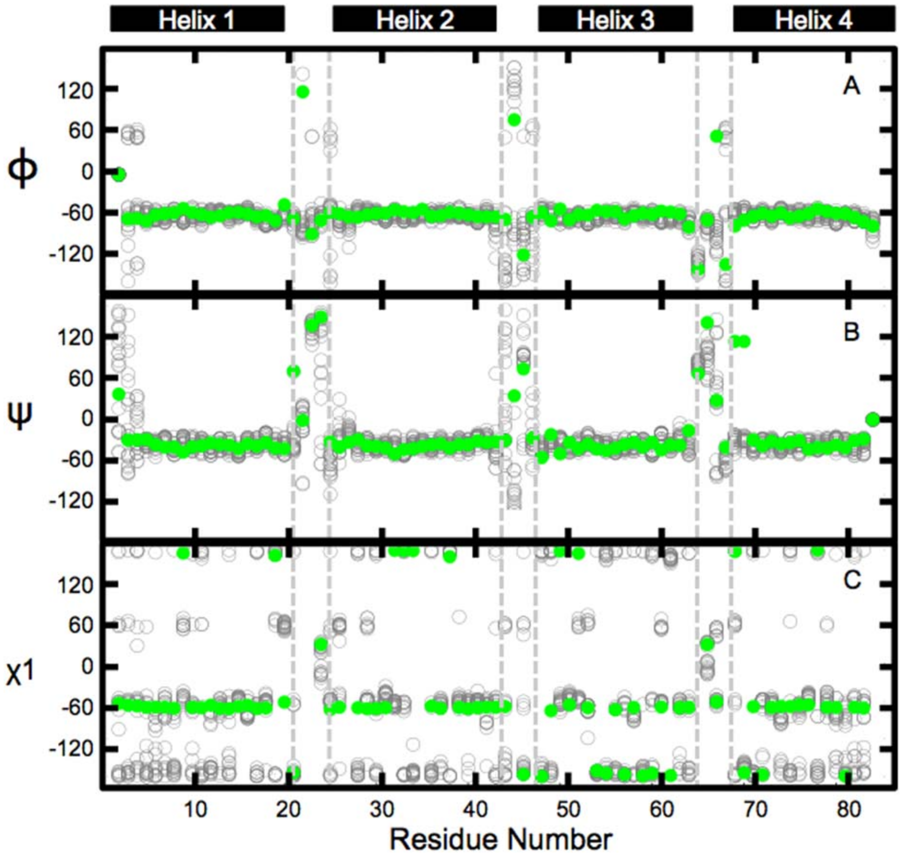

Figure 7. DND_4HB design model versus NMR ensemble in ϕ, ψ, and χ_1 space. DND_4HB NMR solution structure (grey) and design model (green) ϕ, ψ, and χ_1 angles versus residue position (A, B, and C, respectively) are shown. Helix positions are indicated by black bars and loops by dashed lines.

predictions, such as in the CASP competitions.[19] Structure prediction of a de novo sequence then may give additional information about the de novo sequences preference for the desired topology or alternative topologies. If a designed sequence is shown to adopt multiple protein topologies in a computational folding experiment, this behavior may be a sign that the sequence is frustrated and not ideal for either state. However, if a designed sequence is shown to behave like a natural protein, with a low energy, low RMSD population, then this is positive evidence that it may be well behaved in the laboratory. The value of this approach was powerfully demonstrated by Koga *et al*. in the de novo design of a set of α/β proteins.[11]

In the forward folding experiments with DND_4HB, the low scoring models consistently adopted a left-handed helical bundle. This was the primary reason that we decided to test DND_4HB in the laboratory. However, there are two notable differences between our design model and the lowest energy forward folding models; a conformational change in loop 3 and altered packing of phenylalanine 54. In the design model, residues 68 to 70 are part of loop 3, however, in the forward folding model, residues 68 to 70 are helical and are the beginning of helix 4. The NMR structure agrees with the forward folding model with residues 68 to 70 being helical. In contrast, the core packing around residue phenylalanine 54 in the experimental NMR structure more closely matches the design model, RMSD 1.11 Å, compared with the forward folding model, RMSD 1.97 Å, where phenylalanine 54 samples a different rotamer (RMSDs calculated for helices 2, 3, and 4).

Interestingly, the Rosetta energy function assigns essentially the same energy, ∼−160 REUs, to all of these structures—the original design model, the lowest energy forward folding model, and a Rosetta derived NMR structure, indicating that the current Rosetta energy function cannot discriminate the energetic differences due to these subtle structural changes.

The difficulty in capturing these subtle but important structural and energetic differences is strong evidence for the need to generate improved conformational sampling methods and more precise energy functions. Many advancements in conformational sampling methods and energy function precision are coming from careful observation of protein characteristics on large datasets, such as the RosettaHoles method[23] and the Backrub motion.[27]

Additionally, these datasets are being used to inform and train new energy functions, such as Rosetta's newest energy function.[28]

In summary, our results confirm previous efforts in de novo computational design that indicate that combining protein folding and refinement protocols with rotamer-based sequence optimization is an effective protocol for designing well-folded globular proteins.[9–11,20] Computer-based design has now been used to create new helical bundles, coiled-coils, and a variety of α/β topologies. In the future, it will be exciting to see if these or similar approaches can be used to create all novel β-proteins, such as β-barrels, β-sandwiches, and β-propellers.

Methods

Computational procedures

The computational de novo design of proteins can be separated into three steps: (i) generation of protein backbone starting models, (ii) sequence design and refinement, and (iii) selection of de novo sequences for experimental testing.

Generating protein backbone starting models

The generation of protein backbone starting models using Rosetta's fragment assembly requires a naive sequence and fragments of the desired secondary structure. We built idealized helices ($\phi = -57$, $\psi = -47$, $\omega = 180$) of various lengths and assembled them into bundles without loops. Individual helices were placed in the template bundle by first calculating average helix positions from a set of naturally occurring four-helix bundle motifs (PDB ID: 1rj1, 1x90, 1yo7, 2qsb, 2zrr). These bundles are all left-handed four-helix bundles but have helices of various lengths. To determine average positions of each helix, we calculated the helical axis of each helix[29] and took the n-terminal point, mid-point, and c-terminal point as a reduced presentation of each helix. The helical axis points for each corresponding helix were then normalized to a constant helix length (10 residues) and RMSD superimposed. With the RMSD superimposed coordinates, we calculate the average position of each n-terminal, mid, and c-terminal axis point. These average points become the axis restraints used to assemble template bundles and used in fragment assembly. To generate backbone models we selected a random length, between 12 and 20 amino acids, for each helix and RSMD aligned a model helix to appropriately scaled axis points. An additional degree of freedom is present in the rotation of each helix about its own axis. We sampled this degree of freedom by randomly assigning a residue on each helix to be in the core, and then rotating the helix to align this residue's Cα-Cβ bond vector with the center of the template bundle. Using the template bundle, we generated naive sequences by assigning positions as buried or solvent exposed based on solvent accessible surface and their Cα-Cβ bond vector. Buried positions are assigned as hydrophobic residues and positions that are surface exposed are assigned as polar residues. To identify favorable loop lengths, we collected fragments of high-resolution structures present in the protein databank with secondary structures assignments of five residues of helix, two to eight residues of loop, followed by five residues of helix. We investigated the ability of these fragments to close the gap between adjacent helices by RMSD aligning the helical residues of the bridge fragment with the template bundle. This allowed us to identify favorable loop lengths for particular template bundles and to identify favorable fragments to use during fragment assembly.

To generate full length helix bundle models (with loops), we used Rosetta's fragment assembly protocol with naive sequences, traditional 3-mer and 9-mer fragments based on desired secondary structure, bridge fragments, and axis constraints. The axis constraints were implemented as a spatial distance constraint that applies a penalty to the Rosetta score function when a helix axis point is >3.5 Å away from the template bundle's axis point, if the point is within 3.5 Å then a penalty is not applied. Applying the penalty in this manner ensures that models are biased toward the target topology but allows the Rosetta energy function to optimize local interactions in an unbiased manner in the vicinity of the target state. The models produced by this method do not have optimized sequences or structures; the next stage of the procedure is sequence design and structure refinement.

Sequence design and refinement

To de novo design sequences for the models produced by fragment assembly, we used a two-stage flexible backbone protein design protocol that iterates between cycles of (1) fixed backbone sequence optimization and (2) constant sequence backbone and side chain optimization. This iterative process continues until the energy between cycles i and $i+1$ is less than 1.0 Rosetta Energy Units (REU). This method was previously used to completely redesign the core of a naturally occurring four-helix bundle.[20] During the design stage, we limited buried positions to hydrophobic amino acids and surface exposed positions to polar amino acids. The output of this process is an atomic model of a helix bundle, the Rosetta energies, a RosettaHoles score, and a count of the number of buried unsatisfied hydrogen bond partners.

Selection of de novo sequences

To select de novo sequences for experimental characterization from the models produced during the flexible backbone design stage, we considered total Rosetta energy, core packing, number of buried unsatisfied polar atoms, secondary structure prediction using JPRED, and the ability of a designed sequence to be computationally refolded into the target state. We investigated the 10% lowest energy sequences with packing greater than 0.5, as measured by RosettaHoles(V1). RosettaHoles gives a score of 0 to 1, with larger scores indicating better packing. X-ray crystal structures with resolutions of 2.0 Å or better have RosettaHoles scores of >0.5. We also used the JPRED secondary structure prediction server to determine if a designed sequence was predicted to adopt four helices.[24] We also evaluated a sequence's ability to computationally refold, that is, for the predicted state to be correctly identified, low energy and low RMSD compared with the design model, using the Rosetta fragment assembly method with non-biased fragments and without axis constraints. This step can identify sequences that show preferences for more than one topology, for instance a four-helix bundle that has favorable energy for both the left and right-handed topologies. Sequences that passed the low energy metric, packing metric, JPRED server, and refolding metric were evaluated visually to determine which sequence will be expressed and characterized.

Experimental procedures

Cloning, expression, and purification. A codon-optimized gene for the de novo sequence DND_4HB was purchased from Genscript, lowercase letters are due to cloning and capital letters are the designed sequence.

>DND_4HB

mQEERKKLLEKLEKILDEVTDGAPDEARERIE
KLAKDVKDELEEGDAKNMIEKFRDEMEQMYKDA
PNAVMEQLLEEIEKLLKKAgsylvprgslehhhhhh*

The gene was supplied as 4 μg of lyophilized DNA in puc57 vector and was amplified out of 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, and purified again using a PCR-clean up kit, and finally ligated into pET-21 b(+) vector from Novagen, which had been previously been double digested with NdeI and XhoI and purified from an agarose gel using a Fermentas gel-extraction clean-up kit. The ligation reaction product was transformed into XL-10 Gold cells from Stratagene. Success of the cloning and transformation was verified by sequencing.

DND_4HB 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 m*M* IPTG for 12 to 16 h at 16°C. Cells were recovered from liquid culture by spinning at 4500*g* for 30 min in a centrifuge. The resulting cell pellets were resuspended in 0.5*M* NaCl, 0.2*M* Na_2HPO4/NaH_2PO4 at pH 7.0, 10% (v/v) glycerol, 0.1% (v/v) triton, 1 m*M* dithioreitol, followed by three rounds of sonication on ice. After sonication, the sample was treated with DNAse, RNAse, benzamidine, and phenylmethanesulfonylfluoride. The cell lysate was cleared twice by centrifugation at 18,000*g* for 30 min. The supernatant was then filtered using 0.22 μ*M* filters from Millipore. DND_4HB was purified from the supernatant using a HisTRAP from GE Healthcare. The elution peak was concentrated to 2 mL and further purified on a Superdex S75 gel filtration column.

Circular dichroism

CD data were collected on a Jasco J-815 CD spectrometer. Far-UV CD scans were collected using a 1 mm cuvette at concentrations between 30 and 40 μ*M* protein in 50 m*M* sodium phosphate at pH 7.4 and 20°C. Thermal denaturation of samples was conducted between 4°C and 97°C while measuring CD signal at 208 nm and 222 nm. Chemical denaturation by guanidine hydrochloride (GdnCl) was done by titrating a sample of 30 μ*M* DND_4HB protein in 0*M* GdnCl into a sample of 30 μ*M* DND_4HB with 7.8*M* GdnCl. The GdnCl concentration was monitored by refractive index. Thermodynamic parameters were calculated assuming that the folding of the designed protein was a two-state process and by fitting both the thermal and chemical denaturations to the Gibbs-Helmholtz equation using gnuplot's nonlinear least squares fitting routine.

Nuclear magnetic resonance spectroscopy

In order to acquire heteronuclear $^{13}C/^{15}N$-resolved NMR spectra, designed protein DND_4HB was grown and purified as described above, except that cells were harvested by centrifugation at OD_{600} of 0.6 and then washed and transferred to minimal media with uniformly labeled ^{13}C glucose and ^{15}N ammonium chloride. Subsequently, protein overexpression was induced by adding 0.5 m*M* IPTG.

NMR samples of [U-^{13}C,^{15}N]-labeled DND_4HB and biosynthetically-directed fractionally [10% ^{13}C,U-^{15}N]-labeled[30] DND_4HB were prepared at concentrations of ~2.0 m*M* in 90% H_2O/10% D_2O containing 50 m*M* sodium phosphate and 50 m*M* NaCl (pH 6.5). An isotropic overall rotational correlation time of ~8.5 ns was inferred from averaged ^{15}N spin relaxation times, indicating that DND_4HB is monomeric in solution.

The comparably high protein concentration of 2 m*M* allowed recording all NMR data for resonance assignment and structure determination with a total measurement time of only 2 days. The following spectra were recorded for [U-^{13}C, ^{15}N]-DND_4HB at

25°C on Varian INOVA 600 and 750 spectrometers equipped with cryogenic ^{1}H[^{13}C,^{15}N] probes: 2D [^{15}N,^{1}H] HSQC, aliphatic and aromatic 2D constant-time [^{13}C,^{1}H] HSQC, 3D HNCO, (4,3)D HNN$\underline{\mathrm{C}}^{\alpha}\underline{\mathrm{C}}^{\alpha\beta}$, (4,3)D $\underline{\mathrm{C}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha}$(CO)NHN (4,3)D $\underline{\mathrm{H}}^{\alpha}\underline{\mathrm{C}}^{\alpha}$(CO)NHN, aliphatic and aromatic (4,3)D $\underline{\mathrm{HC}}$CH,[31,32] 3D H(CC-TOCSY-CO)NHN[33] and simultaneous 3D ^{15}N/^{13}C$^{\mathrm{aliphatic}}$/^{13}Caromatic-resolved [^{1}H, ^{1}H]-NOESY (mixing time 70 ms, measurement time 2 days).[34] For [10% ^{13}C, U-^{15}N]-DND_4HB, aliphatic 2D constant-time [^{13}C,^{1}H]-HSQC spectra were acquired in ~12 h as described[35] at 25°C on a Varian INOVA 600 spectrometer (total measurement time: 12 h) equipped with a cryogenic probe ^{1}H[^{13}C,^{15}N] probe in order to obtain stereo-specific assignments for Val and Leu isopropyl groups.[30]

All NMR spectra were processed using PROSA[36] and analyzed using CARA.[37] Sequence-specific backbone (HN, N, C$^{\alpha}$, H$^{\alpha}$, and CO) and H$^{\beta}$/C$^{\beta}$ resonance assignments were obtained by using the program AutoAssign.[38,39] Resonance assignment of side chains was accomplished using (4,3)D $\underline{\mathrm{HC}}$CH, 3D H(CC-TOCSY-CO)NH, and simultaneous 3D ^{15}N/^{13}C$^{\mathrm{aliphatic}}$/^{13}Caromatic-resolved [^{1}H, ^{1}H]-NOESY. Overall, for residues 1 to 93, sequence-specific resonance assignments were obtained for 98% of backbone and 100% of side chain resonances assignable with the NMR experiments listed above (Supporting Information Table S1). Chemical shifts were deposited in the BioMagResBank (BMRB ID: 18429). ^{1}H−^{1}H upper distance limit constraints for structure calculation were obtained from simultaneous 3D ^{15}N/^{13}C$^{\mathrm{aliphatic}}$/^{13}Caromatic-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+.[40,41]

Automated NOE assignment was performed iteratively with CYANA,[42–44] and the results were verified by interactive spectral analysis. Stereo-specific 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"[45] using the program CNS.[46] Validation of the 20 refined conformers was performed with the Protein Structure Validation Software (PSVS) server.[47] The NMR structure was deposited in the PDB (PDB ID: 2LSE).

References

1. Betz SF, Raleigh DP, DeGrado WF, Lovejoy B, Anderson D, Ogihara N, Eisenberg D (1995) Crystallization of a designed peptide from a molten globule ensemble. Fold Des 1:57–64.
2. Eisenberg D, Wilcox W, Eshita SM, Pryciak PM, Ho SP, DeGrado WF (1986) The design, synthesis, and crystallization of an alpha-helical peptide. Proteins 1: 16–22.
3. Hecht MH, Richardson JS, Richardson DC, Ogden RC (1990) De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. Science 249:884–891.
4. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262:1680–1685.
5. Regan L, DeGrado WF (1988) Characterization of a helical protein designed from first principles. Science 241:976–978.
6. Wei Y, Kim S, Fela D, Baum J, Hecht MH (2003) Solution structure of a de novo protein from a designed combinatorial library. Proc Natl Acad Sci USA 100: 13270–13273.
7. Go A, Kim S, Baum J, Hecht MH (2008) Structure and dynamics of de novo proteins from a designed superfamily of 4-helix bundles. Protein Sci 17:821–832.
8. Arai R, Kobayashi N, Kimura A, Sato T, Matsuo K, Wang AF, Platt JM, Bradley LH, Hecht MH (2012) Domain-swapped dimeric structure of a stable and functional de novo four-helix bundle protein, WA20. J Phys Chem B 116:6789–6797.
9. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS (1998) High-resolution protein design with backbone freedom. Science 282:1462–1467.
10. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368.
11. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D (2012) Principles for designing ideal protein structures. Nature 491:222–227.
12. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332: 816–821.
13. Der BS, Machius M, Miley MJ, Mills JL, Szyperski T, Kuhlman B (2012) Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. J Am Chem Soc 134:375–385.
14. Stranges PB, Machius M, Miley MJ, Tripathy A, Kuhlman B (2011) Computational design of a symmetric homodimer using beta-strand assembly. Proc Natl Acad Sci USA 108:20562–20567.
15. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, Andre I, Gonen T, Yeates TO, Baker D (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336:1171–1174.
16. Ambroggio XI, Kuhlman B (2006) Computational design of a single amino acid sequence that can switch between two distinct protein folds. J Am Chem Soc 128:1154–1161.
17. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ, DeGrado WF (2005) Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. J Am Chem Soc 127:1346–1347.
18. Bender GM, Lehmann A, Zou H, Cheng H, Fry HC, Engel D, Therien MJ, Blasie JK, Roder H, Saven JG, DeGrado WF (2007) De novo design of a single-chain diphenylporphyrin metalloprotein. J Am Chem Soc 129:10732–10740.
19. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science 309:1868–1871.
20. Murphy GS, Mills JL, Miley MJ, Machius M, Szyperski T, Kuhlman B (2012) Increasing sequence

diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core. Structure 20:1086–1096.

21. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97:10383–10388.
22. Tyka MD, Keedy DA, Andre I, Dimaio F, Song Y, Richardson DC, Richardson JS, Baker D (2011) Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol 405:607–618.
23. Sheffler W, Baker D (2009) RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Sci 18:229–239.
24. Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36:W197–W201.
25. Garcia de la Torre J, Huertas ML, Carrasco B (2000) HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. J Magn Reson 147:138–146.
26. Gilbreth RN, Esaki K, Koide A, Sidhu SS, Koide S (2008) A dominant conformational role for amino acid diversity in minimalist protein-protein interfaces. J Mol Biol 381:407–418.
27. Davis IW, Arendall WB 3rd, Richardson DC, Richardson JS (2006) The backrub motion: how protein backbone shrugs when a sidechain dances. Structure 14:265–274.
28. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, et al. (2013) Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol 523:109–143.
29. Sugeta H, Miyazawa T (1967) General method for calculating helical parameters of polymer chains from bond lengths, bond angles and internal-rotation angles. Biopolymers 5:673.
30. Neri D, Szyperski T, Otting G, Senn H, Wuthrich 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 13C labeling. Biochemistry 28:7510–7516.
31. Kim S, Szyperski T (2003) GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125:1385–1393.
32. Atreya HS, Szyperski T (2004) G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignment. Proc Natl Acad Sci U S A 101:9642–9647.
33. Cavanagh J, Fairbrother WJ, Palmer III AG , Rance M, Skelton NJ (2006) Protein NMR spectroscopy: principles and practice. San Diego: Academic Press.
34. Shen Y, Atreya HS, Liu G, Szyperski T (2005) G-matrix Fourier transform NOESY-based protocol for high-quality protein structure determination. J Am Chem Soc 127:9085–9099.
35. Penhoat CH, Li Z, Atreya HS, Kim S, Yee A, Xiao R, Murray D, Arrowsmith CH, 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.
36. Guntert P, Dotsch V, Wider G, Wuthrich K (1992) Processing of multidimensional NMR data with the new software PROSA. J Biomol NMR 2:619–629.
37. #.Keller R (2004) The computer aided resonance assignment tutorial. Cantina: Verlag.
38. Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269:592–610.
39. Moseley HN, Monleon D, Montelione GT (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Methods Enzymol 339:91–108.
40. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302.
41. 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 44:213–223.
42. Guntert P, Mumenthaler C, Wuthrich K (1997) Automated NOE assignment was performed iteratively with CYANA. Methods Mol Biol 278:353–378.
43. Guntert P, Mumenthaler C, Wuthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273:283–298.
44. Herrmann T, Guntert P, Wuthrich 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.
45. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M (2003) Refinement of protein structures in explicit solvent. Proteins 50:496–506.
46. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, et al. (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystrallogr D 54: 905–921.
47. Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. Proteins 66:778–795.