Understanding Protein Folding through Computational Design and Experimental Validation of First Ever De novo All β-Sheet Proteins for Plant Biotechnology Applications

Tamuka Chidyausiku

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2019

Committee:
David Baker
Suzanne Hoppins
Justin Kollman
Thelma Madzima
Sharona Gordon

Program Authorized to Offer Degree:
Department of Biochemistry

University of Washington

Abstract

Understanding Protein Folding through Computational Design and Experimental Validation of *de novo* All β-Sheet Proteins for Plant Biotechnology Applications

Tamuka Chidyausiku

Chair of the Supervisory Committee:
David Baker
Department of Biochemistry

β-sheet proteins carry out critical functions in biology, and hence are attractive scaffolds for computational protein design. Through study of loops connecting unpaired β-strands (β-arches), we have identified a series of structural relationships between loop geometry, side chain directionality and β-strand length that arise from hydrogen bonding and packing constraints on regular β-sheet structures. We use these rules to de novo design jellyroll structures with double-stranded β-helices formed by eight antiparallel β-strands. The nuclear magnetic resonance structure of a hyperthermostable design closely matched the computational model, demonstrating accurate control over the β-sheet structure and loop geometry. Our results open the door to the design of a broad range of non-local β-sheet protein structures.

Chapter 1. INTRODUCTION

β-sheet protein domains are ubiquitous in nature, carrying out a wide range of functions: transporting hydrophobic molecules, recognition and enzymatic processing of carbohydrates, and scaffolding of virus capsids and antibodies, among others. Although β-sheet protein scaffolds are well suited for incorporating new functions, their design from first principles remains an outstanding challenge. Recent progress in de novo protein design has enabled the accurate design of many hyperstable and structurally diverse proteins, but so far, other than short β-sheet peptides1,2,3, all exhibit either all-α or mixed-αβ folds4. The design of these last has been considerably facilitated by the derivation of a set of rules describing constraints on the backbone geometry of the loops connecting secondary structure elements5, but all-β proteins contain additional features that are less well understood. All-β-sheet structures are particularly challenging to design from scratch6 because a larger fraction of the interactions are non-local (that is, between residues distant along the linear sequence), leading to slower folding rates7, and because β-strands, particularly at the edges of β-sheets, can aggregate into amyloid-like structures. Hence, few β-sheet protein design studies have sought to generate new backbone structures8,9 and, except for a recent β-barrel structure with primarily local strand pairings10, those designs confirmed by high-resolution structure determination have relied heavily on sequence information11,12and backbone structures13,14 from naturally occurring β-sheet proteins.

So far, the de novo design of β-sheet loop connections has been limited to β-hairpins (two antiparallel β-strands interacting via backbone hydrogen bonding and connected through a loop), which is the most local strand

pairing possible and, in principle, the fastest to fold. However, these structures lack a critical feature of non-local globular all-β structures: loops connecting β-strands not paired to each other, also known as β-arches15. These loops connect distinct β-sheets and pair β-strands with larger sequence separation, and they are essential for enabling the protein fold complexity observed in antibodies, β-solenoids, jellyrolls and Greek key-containing structures generally. Here we set out to identify the general principles for designing non-local β-sheet structures.

Chapter 2. CONSTRAINTS ON B-ARCH GEOMETRY

We undertook investigation of the constraints on the backbone geometry of β-strands and connecting loops that arise from hydrogen bonding and the requirement for a compact hydrophobic core. We studied side chain directionality patterns of the two β-strand residues adjacent to β-arch loops (Fig. 1a, left) in naturally occurring protein structures, defining the side chain orientation of the β-strand residue preceding the loop as 'concave' (represented by ↓) if its CαCβ vector is parallel to the vector d from the first to the second β-strand, and 'convex' (represented by ↑) if the CαCβ vector is antiparallel to d. For the residue following the loop, the side chain pattern is described in the same way, but instead using the vector from the second to the first β-strand ($-d$) as a reference (Fig. 1a). This results in four possible β-arch loop side chain orientation patterns: ↑↑, ↑↓, ↓↑ and ↓↓. We analyzed the side chain patterns and the local backbone geometry (as described with ABEGO torsion bins16) of 5,061 β-arch loops from a non-redundant database of natural protein structures (torsion bins A and B are the α-helix and extended regions; G and E regions are the positive ϕ angle equivalents of A and B; and O is the *cis*-peptide bond conformation; Supplementary Fig. 1). We found that

all four side chain orientation patterns frequently occur, and, in contrast to other types of loop connections (that is, $\alpha\beta$, $\beta\alpha$ and β-hairpins)[5], there was no correlation between β-arch loop length and side chain pattern. Instead, each loop ABEGO type, because of the way in which it twists and bends the polypeptide chain[16], is associated with a specific flanking residue side chain pattern (Fig. 1b). The most frequently observed turn types (between 1 and 5 amino acids) for each side chain pattern are listed in Fig. 1c; for example, ABB, BBGB, BABB and BGB are the most frequent loop types for the patterns ↓↓, ↓↑, ↑↓ and ↑↑, respectively. The next level of non-local interaction complexity in all-β folds involves strand pairing (parallel or antiparallel) between two β-arches forming a β-arcade (Fig. 1d), a common structural motif in naturally occurring β-solenoids[15,17]. Because the β-arch loops are stacked in register, the side chains adjacent to one β-arch loop are likely to have the same orientation as the side chains adjacent to the second β-arch loop; analysis of naturally occurring β-arcades confirms that the side chain patterns of the two β-arch loops are indeed correlated (Fig. 1d, middle)

Chapter 3. JELLYROLL DESIGN PRINCIPLES

The double-stranded β-helix can be regarded as a long β-hairpin wrapped around an axis perpendicular to the direction of β-strands, with β-helical turns formed by the pairing between β-arcades (Fig. 2a). In the compact folded structure, two antiparallel β-sheets pack against each other in a sandwich-like arrangement, with the first strand paired to the last, and all β-strands are connected through β-arch loops except for the central β-hairpin. We aimed to design β-helices with three β-arcades forming two antiparallel four-stranded β-sheets, with the eight β-strands connected through six β-arches and one β-hairpin. The non-local

character of the structure grows from the first β-arcade, which starts from the central β-hairpin, to the last one, where the N and C termini are paired. The analysis from Fig. 1 leads to strong constraints on the construction of β-sheet backbone structures, as the side chain directionality patterns of the β-strands and loops are coupled in several ways. First, the directionality patterns of the loops preceding and following each β-strand are coupled to the length of the strand (Fig. 2b): for example, a β-strand with an even number of residues that is preceded by a ↑↑ loop must be followed by a ↓↑ or a ↓↓ loop, but not a ↑↑ or ↑↓ loop, owing to the alternating pleating of β-strands. Second, because the β-arcades of the β-helix have paired β-strands and β-arch loops, the side chains adjacent to one β-arch loop must have the same orientation as the paired side chains adjacent to the second β-arch loop (Fig. 1d). Owing to the antiparallel orientation of the β-arcades, ↓↓ and ↑↑ loops are compatible with loops of the same type, but ↑↓ loops are only compatible with ↓↑ loops (Fig. 1d). Third, the twist and curvature of the two β-sheets of the β-helix are constrained by the hydrogen-bonding register between β-arcades 1 and 3 (herein called β-arcade register), and within β-strand pairs S_3/S_8 and S_4/S_7, as shown in Fig. 2c.

De novo design of protein structures

We constructed double-stranded β-helix protein backbones by Monte Carlo fragment assembly using blueprints (representations of the target protein topologies specifying the ordering, lengths and backbone torsion bins of secondary structure elements and loop connections[5]) in conjunction with backbone hydrogen-bonding constraints specifying all pairings between β-strands. We explored strand lengths between 5 and 7 residues and the most commonly observed β-arch loops between 3 and 5 residues (Fig. 1c). The central β-hairpin was designed with two-residue loops following the ββ rule[5]. The register shifts between pairs of β-

strands from different β-arcades (1 and 3) were allowed to range from 0 to 2 and the β-arcade register shifts between 0 and 4; strand pairs within the same β-arcade were kept in register. A total of 3,673 combinations were enumerated, of which 1,853 had mutually compatible strand lengths and loop types consistent with the constraints summarized in the previous paragraph. For each of these internally consistent blueprints, we used Rosetta to build thousands of protein backbones. The resulting ensemble of backbone structures has considerable structural diversity; those with all strands in register had narrow, sandwich-like structures (Fig. 2d), while those with large register shifts had wider, barrel-like structures (Fig. 2e).

For each generated backbone, we carried out flexible-sequence design calculations[18,19] to identify low-energy amino acid identities and side chain conformations providing close complementary packing, side chain–backbone hydrogen bonding in β-arch loops (to pre-organize their conformation and facilitate folding), and high sequence–structure compatibility. We favored inward-pointing charged or polar amino acids at the four edge strands to minimize aggregation propensity[20]. Loop sequences were designed with consensus profiles obtained from fragments with the same backbone ABEGO torsion bins[21]. Because the very large size of the space sampled by our design procedure limits convergence on optimal sequence–structure pairs, we carried out a second round of calculations starting from the blueprints yielding the lowest-energy designs, intensifying sampling at both the backbone and sequence level. For a subset of designs, we introduced disulfide bonds between paired β-strand positions with high sequence separation (for example, between the first and last β-strands) and optimal orientation (see Methods): disulfide bonds distant in primary sequence decrease the entropy of the unfolded state and therefore enhance the thermodynamic

stability of the native state. To assess compatibility of the top-ranked designed sequences with their structures, we characterized their folding energy landscape with biased forward folding simulations[21], and those with substantial near-native sampling were subsequently assessed by Rosetta ab initio structure prediction calculations[22,23]. Designs with funnel-shaped energy landscapes (where the designed structure is at the global energy minimum and has a substantial energy gap with respect to alternative conformations) were selected for experimental characterization. Ab initio structure prediction of natural β-sheet proteins tends to oversample local contacts[24,25] (that is, favors β-hairpins over β-arches), but we succeeded in designing sequences with the β-arches sufficiently strongly encoded that they folded in silico to near the designed target structure.

Chapter 4. EXPERIMENTAL CHARACTERIZATION

For experimental characterization, we chose 19 designs with funnel-shaped energy landscapes ranging between 70 and 94 amino acids (Supplementary Table 1). BLAST searches[26,27] indicated that the designed sequences had little or no similarity with native proteins (lowest Expect (*E*) values ranging from 0.003 to >10; Supplementary Table 2). Synthetic genes encoding the designs (design names are BH_n, where BH stands for β-helix, n stands for the design number and the _ss suffix is used if disulfide bonds are present) were obtained; the proteins were expressed in *Escherichia coli* and purified by affinity chromatography. Sixteen of the designs expressed well and were soluble, and two (BH_10 and BH_11) were monomeric (Supplementary Fig. 2) by size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) (most of the non-monomeric designs were either dimers or soluble aggregates). Both monomeric designs had far-ultraviolet circular

dichroism spectra (CD) at 25 °C characteristic of β proteins, a melting temperature (T_m) above 95 °C, and well-ordered structures according to 2D ^{1}H-^{15}N heteronuclear single-quantum coherence (HSQC) spectra (Fig. 3a–c and Supplementary Fig. 3). For both designs, the number of NMR peaks matched the number of expected amide resonances based on the protein sequence, but the higher stability of BH_10 under the conditions of the NMR experiments made it a better candidate for NMR structure determination. The two monomeric designs with well-ordered structures were among those with better-packed cores and a larger proportion of β-arch loops containing prolines and with backbone polar atoms making hydrogen bonds (Supplementary Table 3). β-arch loops that are structurally preorganized with the polar groups making internal hydrogen bonding likely favor folding to the correct topology and contribute to stability by compensating for the loss of interactions with water of polar groups in the side chains and backbone. These interactions likely disfavor the competing local strand-pairing arrangement in which the two strands form a β-hairpin; this is a very common pathology in ab initio structure prediction[25]. For the most stable dimeric design (BH_6), we introduced disulfide bonds to stabilize protein regions having contacts with large sequence separation (for example, between the N- and C-terminal strands), but this did not succeed in yielding stable monomers. Addition of an α-helix to the C terminus (one of the two extremes of the β-helix), as a capping domain protecting the strand edges from intermolecular pairing, also failed to yield stable monomers, even in combination with disulfide bonds. This suggests that the sequence of the core β-sheet must strongly encode its structure independently of disulfide bonds or protecting domains aimed at increasing stability.

NMR structure of a de novo–designed β-helix

We succeeded in solving the structure of BH_10 by 4D NMR spectroscopy (Fig. 3d, Table 1 and Supplementary Fig. 4), using the 4D-CHAINS/AutoNOE-Rosetta automated pipeline for resonance assignments and structure calculation[28], and found it to be in very close agreement with the computational model (Cα r.m.s. deviation (r.m.s.d.) of 0.84 Å, averaged over 10 NMR models). The overall topology is accurately recapitulated, including all strand pairings, register shifts and loop connections, as supported by 132 long-range nuclear Overhauser effects (NOEs) between backbone amide and side chain protons (Supplementary Fig. 5). The designed aliphatic and aromatic side chain packing in the protein core, as well as salt bridge interactions across the two β-sheet surfaces, was also accurately reproduced; three salt bridges between the two paired β-arcades and one within the third β-arcade are well supported by the observed NOEs (Supplementary Fig. 6). The agreement in both the backbone conformation and hydrogen-bonding interactions of the loops forming the three β-arcades is remarkable given that these elements are the most flexible parts of the structure and therefore difficult to design owing to sampling bottlenecks. The β-arcades were designed with pairs of β-arch loops that interact via backbone-backbone hydrogen bonds (owing to the complementarity between their backbone conformations) stabilizing loop pairing and avoiding burial of polar backbone atoms (see Supplementary Fig. 7 for the BH_10 loop sequences and side chain patterns). For example, β-arcade 1 is formed by BBG and ABB loops, and the buried backbone NH group of the G position in the former makes a hydrogen bond with the buried backbone C=O of the neighboring loop (Fig. 3e). The other two β-arcades were designed with one β-arch loop containing buried and fully hydrogen-bonded asparagines (four hydrogen bonds in total) that stabilize both loop pairing and the local β-arch conformation (of ABABB

loops). By design, the asparagine side chain geometry was further stabilized with hydrophobic stacking interactions from the two β-arch loops of the same arcade. The high degree of convergence of the designed rotamer in the NMR ensemble illustrates the high structural preorganization of this particular motif (Fig. 3f). The amino acid sequence of BH_10 is unrelated to any sequence in the NCBI nr database (BLAST found one hit with insignificant sequence similarity; *E* value 6.3). We searched the Protein Data Bank (PDB) for similarities in structure (using the Dali server[29] with the lowest-energy NMR model as the query structure) or sequence (with HHpred[30] for sensitive profile-based sequence search), and identified matches similar in fold but containing additional and irregular secondary structures and longer loops. These matches are all homodimers with sheet-to-sheet interface packing (Supplementary Fig. 8) or domains integrated in larger structures, in sharp contrast to the BH_10 monomer.

Contact order and sequence determinants of the BH_10 fold

The non-local character of BH_10 is of particular note: a large fraction of the contacting residues are distant along the linear sequence, with extensive strand pairing between the N- and C-terminal β-strands. The contact order of the structure (that is, the average separation along the linear sequence of residues in contact in the 3D structure) is higher than that for any previous single-domain protein designed de novo (Fig. 3g,h). Proteins with high contact order fold more slowly than those with low contact order as there is a greater loss in chain entropy for forming the first native interactions, and they tend to form long-lived non-native structures that can oligomerize or aggregate31. We have overcome the challenges in designing non-local structures by focusing on backbones

lacking internal strain and having maximal internal coherence, and programming β-strand orientation with highly structured loops.
One of the challenges in achieving high contact order through β-arches is to disfavor formation of more sequence-local β-hairpins. To evaluate in silico how each of our design features contributes to favoring β-arches over β-hairpins, we generated folding energy landscapes for a series of mutants of BH_10 in which loop hydrogen bonding, side chain packing of loop neighbors and loop local geometry were disrupted one at a time. For all conformations generated, we classified all the β-strand connections as β-arch or β-hairpin depending on strand-pairing formation, and calculated the overall frequency of β-hairpin formation for each pair of consecutive β-strands. As shown in Supplementary Fig. 9, disruption of packing within or between β-arch loops, removal of side chain–backbone hydrogen-bonding interactions and reducing loop geometry encoding by eliminating prolines all increase sampling of competing β-hairpin conformations, and thus substantially decrease sampling of β-arches and the target designed structure.

Chapter 5. CONCLUSION

The design of all-β globular proteins from first principles has remained elusive for two decades of protein design research. We have successfully designed a double-stranded β-helix de novo, as confirmed by the NMR structure of the design BH_10, based on a series of rules describing the geometry of β-arch loops and their interactions in more complex β-arcades. Our work also achieves two related milestones: the first accurate design of an all-β globular protein with exposed β-sheet edges, and the most non-local structure yet designed from scratch. Comparison of successful and failed designs suggests that folding and stabilization of

the monomeric structure (and implicitly disfavoring competing topologies with more local strand pairings) is bolstered by loops containing side chain–backbone and backbone-backbone hydrogen bonds together with well-packed mixed aliphatic/aromatic side chains in the protein core, inward-pointing polar amino acids at strand edges and salt bridges between paired strands. Previous design studies on β-propellers[11] or parallel β-helices[12] have used naturally occurring backbone structures and consensus sequence information on the target fold families; this approach, while powerful, sheds less light on the key principles underlying β-sheet structure construction and does not allow the programming of new backbone geometries. The β-helix fold designed here is well suited for incorporating metal, ligand-binding and active sites, as illustrated by the broad functional diversity of cupin protein domains, which are the closest naturally occurring structural analogs. With the basic design principles now understood, our de novo design strategy should enable the construction of a wide range of β-helix structures tailored to a broad range of target ligands.

Initial advances in protein design were from algorithms that allowed rapid identification of a very-low-energy sequence for a given backbone structure. In recent years, progress has come from the realization that the requirements of burying hydrophobic residues in a core away from solvent while avoiding the burial of backbone polar groups without compensating hydrogen bonds, together with torsional restrictions on the peptide backbone, considerably constrain overall globular protein backbone geometry, particularly for β-sheet-containing proteins: it is much harder than originally expected to construct new backbones that have these properties. The de novo design of β-sheet-containing proteins advanced considerably following the elucidation of β-sheet design principles for construction of backbones meeting the above constraints

while having desired geometries: for example, principles for controlling the chirality of β-hairpins[5], reducing strain in β-strands with glycine kinks[10], and combining β-bulges and register shifts to curve β-sheets[21]. The design rules described here are a considerable further advance as they provide control over β-arch connections between distinct β-sheets, and should enable the design of a broad range of β-protein families beyond the β-barrel and β-helix, with considerable medical and biotechnological potential; for example, the immunoglobulin fold widely utilized for binding and loop scaffolding in nature is topologically very similar to the double-stranded β-helices designed here, with a larger proportion of β-hairpins over β-arches.

FIGURES & TABLES

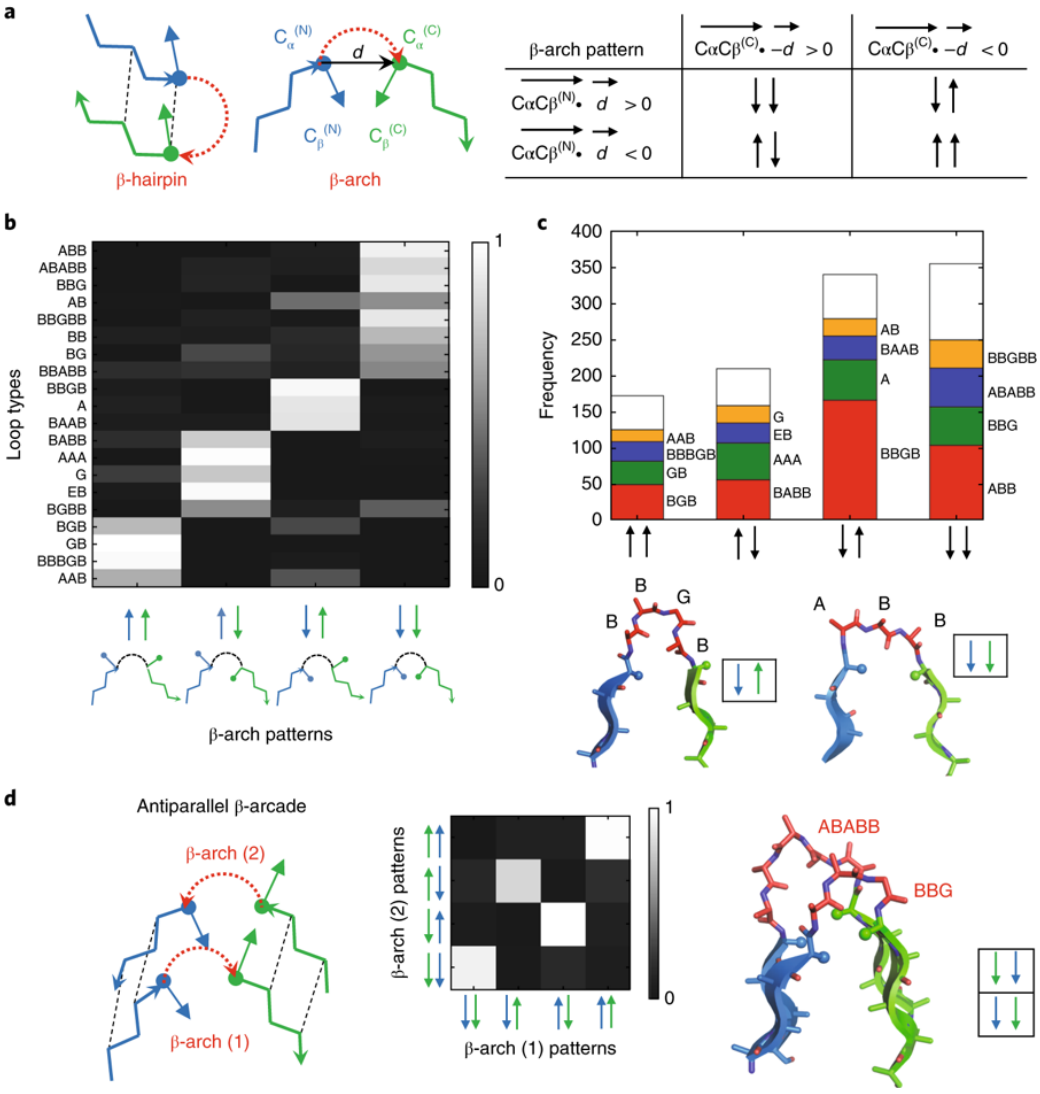

Fig. 1: Constraints on β-arch geometry.

a, Side chain directionality in the β-arch. Comparison between β-hairpin and β-arch (left); the CαCβ and *d* vectors used to define the orientation of the two adjacent side chains are indicated. The four possible side chain directionality patterns are on the right. **b**, Turn type dependence of β-arch side chain patterns. Loops on the *y* axis are described by their

ABEGO torsion bins (Supplementary Fig. 1). Most of the loops adopt only one of the four possible side chain patterns. **c**, Frequency of the most common loops for each of the four β-arch side chain patterns. There are strong preferences: for example, BBGB is strongly associated with the ↓↑ pattern, whereas ABB is strongly associated with the ↓↓ pattern (bottom). Only loops with bending <120° (see Methods) and containing between 1 and 5 amino acids were considered in this analysis. **d**, β-arcades consist of two stacked β-arches with in-register strand pairing (left). Because strand pairs of the β-arcade are in register, the side chains adjacent to one β-arch loop must have the same orientation as the paired side chains that are adjacent to the second β-arch loop, and therefore not all loop pairs are allowed (middle). Example of a β-arcade formed by two common β-arches with compatible side chain patterns (right).

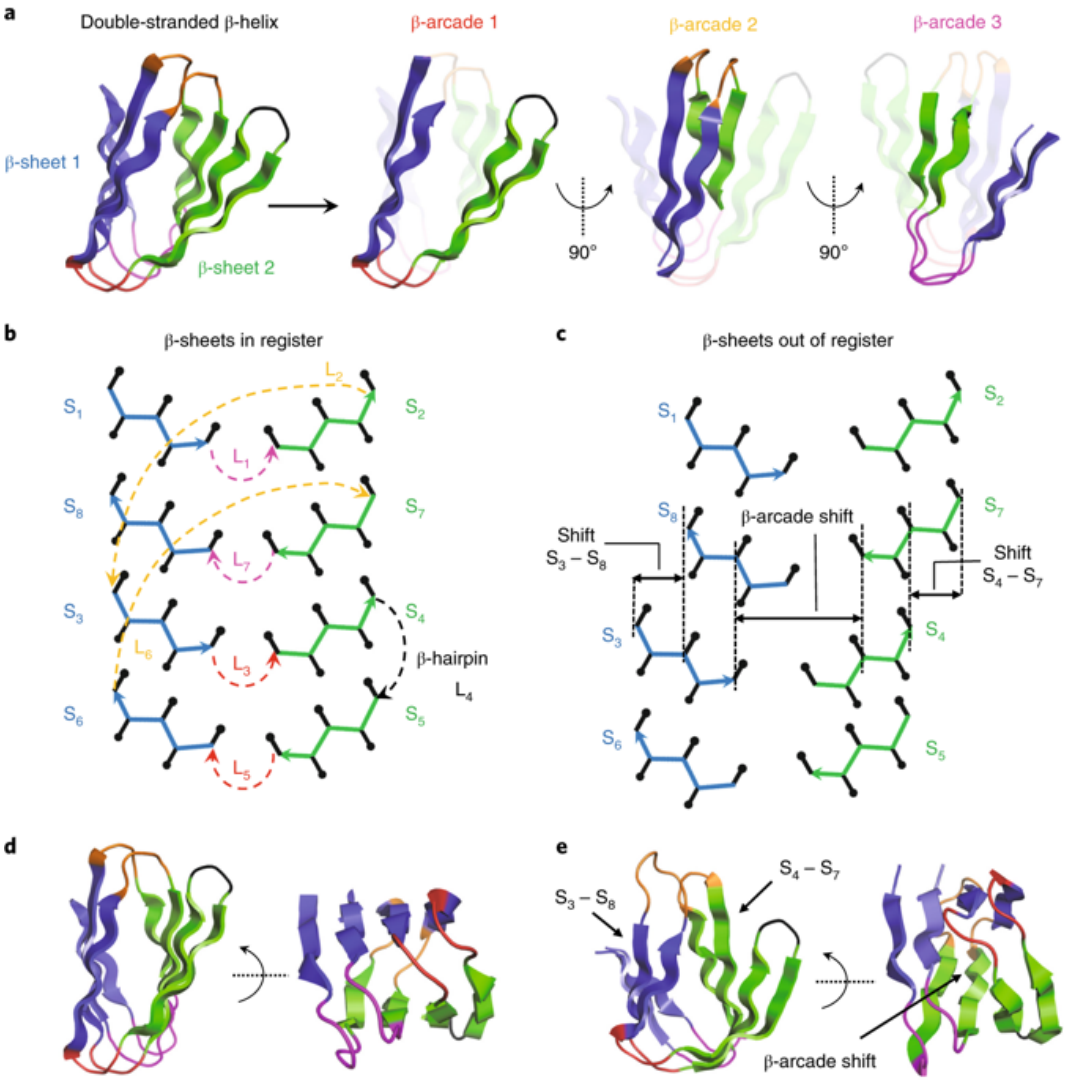

Fig. 2: Double-stranded β-helix topology specification.

a, The double-stranded β-helix fold consists of two four-stranded antiparallel β-sheets (in blue and green) with six β-arch and one β-hairpin connection. Pairs of β-arches forming the three β-arcades are highlighted (right). β-arch loops belonging to the same β-arcade are displayed with the same color throughout the figure (β-arcades 1, 2 and 3 in red, orange and magenta, respectively). **b**, Topology diagram of a designed double-stranded β-helix with all β-strand pairs in register. The Cα traces of the first and second β-sheets are colored in blue and green, respectively. Side chain Cβ positions oriented toward the inner and outer faces of the β-helix are represented with up and down black arrows with rounded tips, respectively. β-arch loops are colored as in **a**. **c**, Definition of β-arcade register shift varied during conformational sampling. The β-arcade register shift (between β-arcades 1 and 3) is determined by the register of β-strand pairs S_3/S_8 and S_4/S_7, and the

lengths of β-strands S_3, S_4, S_8 and S_7 (see Methods). In this example, β-strand pairs S_3/S_8 and S_4/S_7 each have a two-residue register shift, resulting in an overall β-arcade register shift of four residues. Loops are omitted to facilitate visualization. **d**, Example of a design model with all β-strand pairs in register forming a sandwich-like structure. **e**, Example of a design model with register shifts between β-arcades 1 and 3 (magenta and red) forming a barrel-like structure.

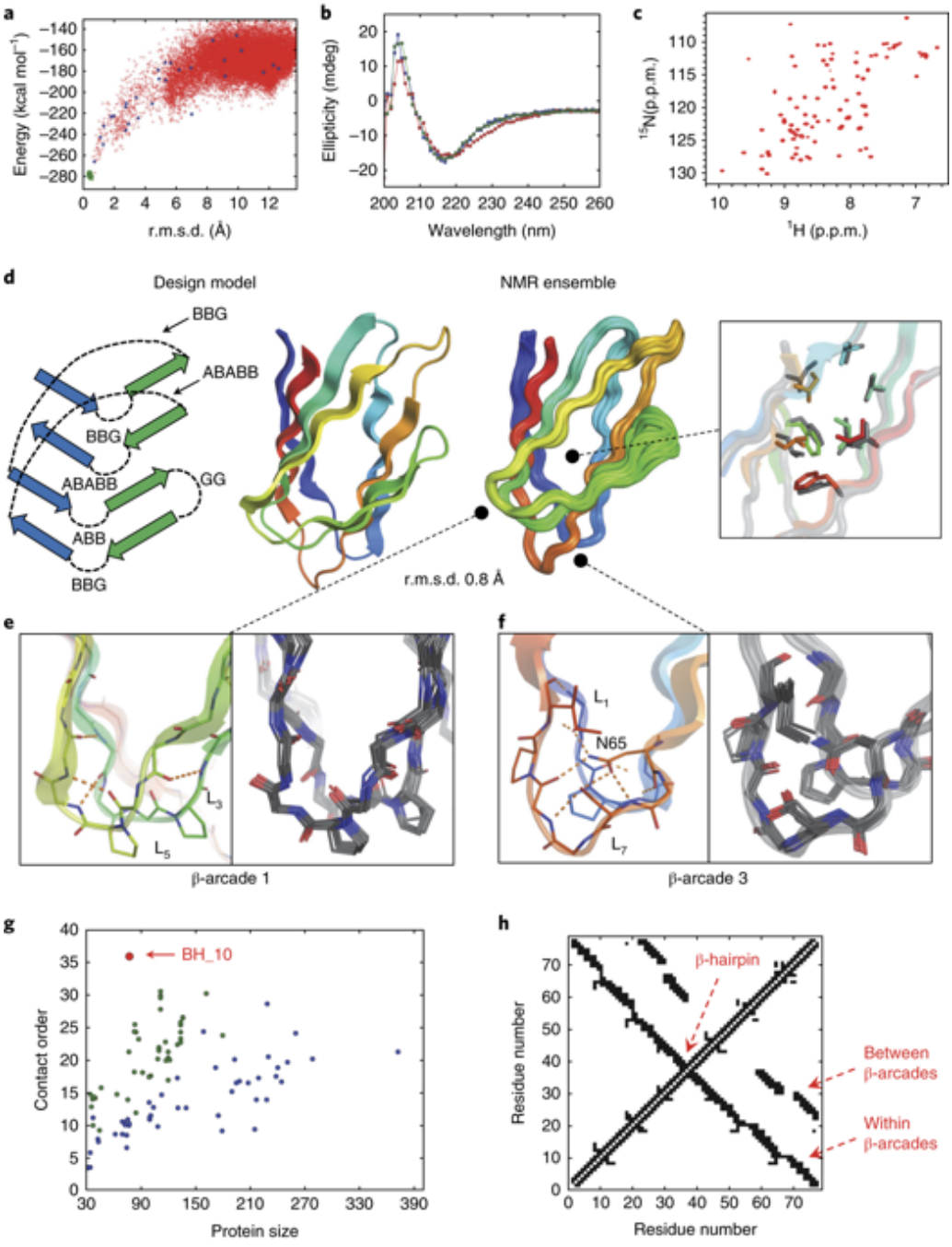

Fig. 3: The NMR structure of BH_10 is nearly identical to the design model.

a, Calculated BH_10 folding energy landscape. Each dot represents the lowest-energy structure obtained from ab initio folding trajectories starting from an extended chain (red dots), biased forward folding trajectories (blue dots) or local relaxation of the designed structure (green dots); the *x* axis is the Cα r.m.s.d. from the designed model and the *y* axis is the Rosetta all-atom energy. **b**, Far-ultraviolet CD spectra (blue, 25 °C; red, 95 °C; green, 25 °C after cooling). **c**, ^{1}H-^{15}N HSQC spectra obtained at 37 °C at a ^{1}H field of 800 MHz. **d**, NMR structure in comparison with the design model. Comparison of core side chain rotamers (NMR structure in gray and design in rainbow) (inset). Topology diagram showing the ABEGO torsion bins of all loop connections (left). Atomic coordinates for the design model are in Supplementary Dataset 1. **e**, Backbone hydrogen bonding of β-arcade 1 is well preserved across the NMR ensemble. **f**, Side chain interactions of N65 with backbone and side chains form a hydrogen-bonded network in β-arcade 3 that is well recapitulated in the NMR ensemble. **g**, Contact order of de novo–designed protein domains confirmed by high-resolution structure determination; all-α (blue), αβ (green) and all-β (red). The BH_10 design stands out with a contact order of 35.8 for a chain length of 78 residues. The domains are listed in Supplementary Tables 4 and 5. **h**, Contact map illustrating the large sequence separation of the contacts present in the BH_10 topology.

Table 1 NMR and refinement statistics for BH_10	
From: De novo design of a non-local β-sheet protein with high stability and accuracy	
	BH_10 (PDB 6E5C)
NMR distance and dihedral constraints	
Distance constraints	
Total NOE	659
Intraresidue	272
Inter-residue	387
Sequential ($\|i-j\|=1$)	222
Medium range ($2\leq\|i-j\|\leq4$)	33
Long range ($\|i-j\|\geq5$)	132
Intermolecular	0
Hydrogen bonds	0
Total dihedral angle restraints	156
ϕ	78
ψ	78
Structure statistics	
Violations (mean ± s.d.)	
Distance constraints (Å)[a]	0.30 ± 0.46
Dihedral angle constraints (°)[b]	9.30 ± 2.49
Max. dihedral angle violation (°)[b]	47.59
Max. distance constraint violation (Å)[a]	1.32
Deviations from idealized geometry	
Bond lengths (Å)	0.00 ± 0.00
Bond angles (°)	0.00 ± 0.00
Impropers (°)	0.00 ± 0.00
Average pairwise r.m.s.d. (Å)[c]	
Heavy	0.61 ± 0.13
Backbone	0.51 ± 0.11

a. Distance constraint violations in the structural ensemble were calculated using a 7-Å universal upper distance bound for the NOE restraints assigned by AutoNOE-Rosetta.

b. Dihedral angle restraints were derived from TALOS-N. The violations were calculated for the core secondary structural regions of the ten lowest-energy models using a 15° cutoff beyond TALOS-N-predicted dihedral angles.

c. Pairwise r.m.s.d. was calculated among ten refined models for a core secondary structural region defined by residues 2–8, 11–18, 21–28, 32–36, 39–43, 46–53, 59–65 and 71–75.

REFERENCES:

1. Huang PS, et al. (2016) De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nat Chem Biol 12(1):29-34.

2. Keefe AD & Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410(6829):715-718.

3. Sillitoe I, et al. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res 41(Database issue):D490-498.

4. Harbury PB, Plecs JJ, Tidor B, Alber T, & Kim PS (1998) High-resolution protein design with backbone freedom. Science 282(5393):1462-1467.

5. Kuhlman B, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302(5649):1364-1368.

6. Watters AL, et al. (2007) The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. Cell 128(3):613-624.

7. Koga N, et al. (2012) Principles for designing ideal protein structures. Nature 491(7423):222-227.

8. Rocklin GJ, et al. (2017) Global analysis of protein folding using massively parallel design, synthesis, and testing. Science 357(6347):168-175.

9. Bale JB, et al. (2016) Accurate design of megadalton-scale two-component icosahedral protein complexes. Science 353(6297):389-394.

10. Boyken SE, et al. (2016) De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. Science 352(6286):680-687.

11. Lu P, et al. (2018) Accurate computational design of multipass transmembrane proteins. Science 359(6379):1042-1046.

12. Butterfield GL, et al. (2017) Evolution of a designed protein assembly encapsulating its own RNA genome. Nature 552(7685):415-420.

13. Plaxco KW, Simons KT, Ruczinski I, & Baker D (2000) Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. Biochemistry 39(37):11177-11183.

14. Matthews CR (1987) Effect of point mutations on the folding of globular proteins. Methods Enzymol 154:498-511.

15. Krishna MM, Hoang L, Lin Y, & Englander SW (2004) Hydrogen exchange methods to study protein folding. Methods 34(1):51-64.

16. Maity H (2003) Protein hydrogen exchange mechanism: Local fluctuations. Protein Science 12(1):153-160.

17. Zhang YZ (1995) Protein and Peptide Structure and Interactions Studied by Hydrogen Exchange and NMR.

18. Pace C (1990) Measuring and increasing protein stability. Trends in Biotechnology 8:93-98.

19. Myers JK, Nick Pace C, & Martin Scholtz J (1995) Denaturantmvalues and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. Protein Science 4(10):2138-2148.

20. Rother K, Hildebrand PW, Goede A, Gruening B, & Preissner R (2009) Voronoia: analyzing packing in protein structures. Nucleic Acids Res 37(Database issue):D393-395.

21. Gangadhara BN, Laine JM, Kathuria SV, Massi F, & Matthews CR (2013) Clusters of branched aliphatic side chains serve as cores of stability in the native state of the HisF TIM barrel protein. J Mol Biol 425(6):1065-1081.

22. Gu Z, Zitzewitz JA, & Matthews CR (2007) Mapping the structure of folding cores in TIM barrel proteins by hydrogen exchange mass spectrometry: the roles of motif and sequence for the indole-3-glycerol phosphate synthase from Sulfolobus solfataricus. J Mol Biol 368(2):582-594.

23. Jha SK & Marqusee S (2014) Kinetic evidence for a two-stage mechanism of protein denaturation by guanidinium chloride. Proc Natl Acad Sci U S A 111(13):4856-4861.

24. Chevalier A, et al. (2017) Massively parallel de novo protein design for targeted therapeutics. Nature 550(7674):74-79.

25. Hosseinzadeh P, et al. (2017) Comprehensive computational design of ordered peptide macrocycles. Science 358(6369):1461-1466.

26. Hsia Y, et al. (2016) Design of a hyperstable 60-subunit protein dodecahedron. [corrected]. Nature 535(7610):136-139.

27. Kuhlman B & Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97(19):10383-10388.

28. Tanford C (1970) Protein Denaturation. 24:1-95.

29. Nozaki Y & Tanford C (1965) The Solubility of Amino Acids and Related Compounds in Aqueous Ethylene Glycol Solutions. Journal of Biological Chemistry 240(9):3568-3573.

30. Tzul FO, Schweiker KL, & Makhatadze GI (2015) Modulation of folding energy landscape by charge-charge interactions: linking experiments with computational modeling. Proc Natl Acad Sci U S A 112(3):E259-266.

31. Das RK & Pappu RV (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A 110(33):13392-13397.

32. Arrington CB & Robertson AD (2000) Correlated motions in native proteins from MS analysis of NH exchange: evidence for a manifold of unfolding reactions in ovomucoid third domain. J Mol Biol 300(1):221-232.

33. Aghera N & Udgaonkar JB (2013) The utilization of competing unfolding pathways of monellin is dictated by enthalpic barriers. Biochemistry 52(34):5770-5779.

34. Jha SK, Dhar D, Krishnamoorthy G, & Udgaonkar JB (2009) Continuous dissolution of structure during the unfolding of a small protein. Proc Natl Acad Sci U S A 106(27):11113-11118.

35. Malhotra P & Udgaonkar JB (2016) Secondary Structural Change Can Occur Diffusely and Not Modularly during Protein Folding and Unfolding Reactions. J Am Chem Soc 138(18):5866-5878.

36. Honeycutt JD & Thirumalai D (1990) Metastability of the folded states of globular proteins. Proc Natl Acad Sci U S A 87(9):3526-3529.

37. Vieille C & Zeikus GJ (2001) Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. Microbiology and Molecular Biology Reviews 65(1):1-43.