Theoretical and experimental studies for de novo design of “difficult-to-describe” all-α protein structures

December 2021

Koya Sakuma

The Graduate University for Advanced Studies, SOKENDAI
School of Physical Sciences
Department of Structural Molecular Science

Table of Contents

Chapter 01:
General Introduction

Protein structure and function

Protein performs various functions in the cells and underlies most of the biological phenomena. The function of proteins ranges from catalyzing the chemical reactions to forming cytoskeletons, and their functional diversity is believed to be based on the diversity of protein structures. For example, enzymes that catalyze specific chemical reactions require complementary ligand pockets that selectively accommodate the substrate molecules. As the size and shape of pockets are defined by the structure of the enzyme, the structure of proteins has very crucial roles in determining the specificity of the reaction [1]. Therefore, protein structures have been believed to define their functions.

Protein structure and folding

When seen as a group of covalently bonded atoms, the structure of a protein is just a polymer. Chemically, a protein is just a hetero polypeptide chain composed of amide-bonded amino-acid residues. However, this is just one aspect of the protein structure. One of the most astonishing aspects of naturally occurring proteins is that the polypeptide chains spontaneously fold into specific three-dimensional structures to perform their work. This phenomenon in which polypeptide chains fold into specific conformations is called “protein folding”, which makes proteins distinct from random heteropolymers [2,3].

Anfinsen's dogma

The specific conformation the protein folds into is called the native state. The native states vary between protein species, which are specified by their amino-acid sequences. The native state of protein is believed to be the state with lowest conformational free-energy as long as the folding process obeys equilibrium thermodynamics. This is called “Thermodynamic principle of protein folding” or Anfinsen's dogma [2,4].

Structure prediction

Owing to these physico-chemical backgrounds, there has been much effort to predict the three-dimensional structure of proteins from the amino-acid sequence. The formalism of the problem is very simple; for a given amino acid sequence, predict the native state that the sequence is most likely to fold into. In short, an amino-acid sequence is given to the solver, and the solver returns the plausible three-dimensional structure that the amino-acid sequence is likely to fold into. This problem, structure prediction, is a very fundamental problem in protein science and also has many outcomes if practically solved [5–7]. For example, accurate computational structure prediction of drug-target proteins that are difficult to crystallize may lead to accelerated design of drug molecules.

Protein design as inverse problem of structure prediction

Protein design, the main subject of this thesis, is formalized as the inverse problem of the structure prediction problem [8]. Namely, a target backbone structure is given to the solver, and the solver returns an amino-acid sequence that is likely to fold into the target structure. This seems to be a very clear definition of protein design problem.

However, this formalism on protein design is oversimplified and too formal. The most important question that need to be asked here is “where does the target backbone structure come from?” In other words, who determines what to design? Structure prediction has no analogous question to this. In structure prediction, the target amino-acid sequence comes from sequence databases that store vast numbers of experimentally determined genomic DNA sequences. It is generally easier to obtain the nucleotide sequences of the genes than to solve the three-dimensional structure of the protein experimentally, and this makes protein structure prediction a well-defined problem. In contrast, it remains ambiguous where the target structures to design come from. This is the key viewpoint to understand protein design as something more than the inverse problem of protein structure prediction.

Redesign of naturally occurring protein structures

In the early history of protein design, the design was performed exactly following the formalism explained above; they took naturally occurring protein structure as inputs, removed the sidechain atoms, and then computationally designed the amino acid sequences that fold into the target structure [9,10]. However, in the modern context of protein design, such protocols depending on the naturally occurring backbone structure tend to be called “redesign” rather than just design. This is because another paradigm of design, de novo design, has been widely accepted in the field of protein design.

De novo design

De novo protein design is a more drastic concept for protein design. In de novo protein design, backbone structures are built from scratch, and not stolen from naturally occuring proteins [11]. Once the backbone structures are built by some means, designers can perform the amino-acid sequence designs as if they were taken from natural proteins. In this way, they successfully design completely new proteins. This offers higher flexibility in selection of design targets and therefore would greatly enhance the variety of designed structures. Such ability to design completely new proteins may enable the design of customized proteins such as artificial enzymes and artificial antibodies for example. However, there are limited numbers of methods to generate de novo backbone structures [12–15], which may be bottlenecking the diversity of designed proteins. Further development of design protocols is required to unleash the potential of de novo protein design

Outline of this thesis

In this thesis, the author focused on the strategies to build the backbone structures and explore how new approaches can extend the repertoire of design protein structures.

In section 2, the author aimed to comprehensively construct three-helix bundle structures. The purpose of this section is to learn what it is like to comprehensively design small tertiary structures. The author first performed a statistical analysis of helix-loop-helix fragments, and identified αα-hairpin motifs specifically related to left- or right-handedness of helix-helix packing. Using these motifs as building blocks, the author thoroughly performed backbone-building simulation of all of the possible small three-helix bundle structures. As is expected without simulations, the length of the second α-helix plays a significant role in the compaction of three-helix bundle structure. Performing these backbone-building simulations, the author identified the combinations of the loop types and the helix-lengths that result in tight compaction of three helix-bundles required to form a hydrophobic core. The author also performed amino-acid sequence designs for those thoroughly enumerated the compact three-helix bundle structures, and found that they were highly designable. The author also made this comprehensive set of three-helix bundle structures publically available, hoping that this structural library allows other designers to skip the rebuilding of similar topologies and enable them to focus on their specialized design tasks.

In section 3, the author took two simple four-helix bundle structures as examples and critically investigated the reason why diverse all-alpha protein structures have not been designed so far in the current framework of de novo protein design. The author identified that “double-meaning” of GBB loop critically lowers the purity of fragments and leads to low efficiency of backbone building in BlurpintBDR. The author concluded that the blueprint method is not suitable for the design of all-alpha proteins.

In section 4, the author developed new strategies for design of all-α proteins to overcome the limitation of current methods pointed out in Section 2. The author guessed that “difficult-to-describe” structures were also difficult to draft out. Therefore the author intended to skip the step of making blueprints, and directly model backbone structures which are ready for amino-acid sequence designs. First, the author started from the classification of typical helix-loop-helix fragments identified as fundamental building blocks for building backbone structures. The author generated literally all the possible combinations of these building blocks and the connecting α-helix lengths, and evaluated their compaction and clashes, and composed a myriad of globular all-α backbone decoys. Statistical analysis of these decoys clarified that the conformation generated by this strategy largely covers the conformational space that has not been sampled by previous de novo designs. Then the author searched for the backbone topologies that attracted my attention as if the author looked over a catalog of protein backbone structures, and then picked some up for sequence design. For the five backbone topologies the author got interested in, amino-acid sequence design was performed. The author performed experimental validation of foldability of these designed proteins. For the most promising designs, the three-dimensional structures were solved in collaboration with Dr. Naohiro Kobayashi and Dr. Toshihiko Sugiki at Riken and Osaka university. The structure of a design protein Elsa was solved in collaboration with Murata group at Chiba university and they clarified the protein forms domain swapped dimer form in the crystal.

In section 5 and 6, the author seeked for the applications of the backbone building technique developed in section 4. In section 5, the author created a massive library of de-novo designed mini all-α proteins to examine the value of the structural diversity made accessible by the backbone building method. The library encodes 294 distinct topologies by 7,350 amino acid sequences, whose structural diversity would be useful to design functional proteins. In section 6, the author designed idealized versions of globin topology to show that the typical building blocks can provide sufficient structural complexity covering the most famous example of “difficult-to-describe” topology. The author was able to design the amino-acid sequences that are predicted to fold into the target globin-like topologies. Though experimental validations for these library and globin-like designs are yet to be done, these results highlight the applicability of the structural diversity and complexity provided by the backbone building method developed in section 4.

[1] A.D. McLachlan, Protein Structure and Function, Annu. Rev. Phys. Chem. **23** 165–192 (1972). DOI: 10.1146/annurev.pc.23.100172.001121.

[2] K.A. Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl, The protein folding problem, Annu. Rev. Biophys. **37** 289–316 (2008). DOI: 10.1146/annurev.biophys.37.092707.153558.

[3] K.A. Dill, J.L. Maccallum, P. Folding, The Protein-Folding Problem , 50 Years On, 1042–1047 (2012).

[4] C.B. Anfinsen, Principles that Govern the Folding of Protein Chains, J. Phys. A Math. Theor. **44** 1689–1699 (2011). DOI: 10.1088/1751-8113/44/8/085201.

[5] J. Moult, A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction, Curr. Opin. Struct. Biol. **15** 285–289 (2005). DOI: 10.1016/j.sbi.2005.05.011.

[6] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W.R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning, Nature. **577** 706–710 (2020). DOI: 10.1038/s41586-019-1923-7.

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature. **596** 583–589 (2021). DOI: 10.1038/s41586-021-03819-2

[8] C. Pabo, Designing Proteins And Peptides, Nature. **301** 200 (1983).

[9] B.I. Dahiyat, C.A. Sarisky, S.L. Mayo, De novo protein design: Towards fully automated sequence selection, J. Mol. Biol. **273** 789–796 (1997). DOI: 10.1006/jmbi.1997.1341.

[10] A.G. Street, S.L. Mayo, Computational protein design, Structure. **7** 105–109 (1999). DOI: 10.1016/S0969-2126(99)80062-8.

[11] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, Design of a Novel Globular Protein Fold with Atomic-Level Accuracy, **302** 1364–1369 (2003).

[12] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T.B. Acton, G.T. Montelione, D. Baker, Principles for designing ideal protein structures, Nature. **491** 222–227 (2012). DOI: 10.1038/nature11600.

[13] T.M. Jacobs, B. Williams, T. Williams, X. Xu, A. Eletsky, J.F. Federizon, T. Szyperski, B. Kuhlman, Design of structurally distinct proteins using strategies inspired by evolution, Science (80-.). **352** 687–690 (2016). DOI: 10.1126/science.aad8036.

[14] E. Marcos, B. Basanta, T.M. Chidyausiku, Y. Tang, G. Oberdorfer, G. Liu, G.V.T. Swapna, R. Guan, D.A. Silva, J. Dou, J.H. Pereira, R. Xiao, B. Sankaran, P.H. Zwart, G.T. Montelione, D. Baker, Principles for designing proteins with cavities formed by curved b sheets, Science (80-.). **355** 201–206 (2017). DOI: 10.1126/science.aah7389.

[15] B. Basanta, M.J. Bick, A.K. Bera, C. Norn, C.M. Chow, L.P. Carter, I. Goreshnik, F. Dimaio, D. Baker, An enumerative algorithm for de novo design of proteins with diverse pocket structures, Proc. Natl. Acad. Sci. U. S. A. **117** 22135–22145 (2020). DOI: 10.1073/pnas.2005412117.

Chapter 02 :
Enumeration and comprehensive design of three-helix bundle structures composed of typical αα-hairpins

Abstract

The design of protein structures from scratch requires special attention to the combination of the types and lengths of the secondary structures and the loops required to build highly designable backbone structure models. However, it is difficult to predict the combinations that result in globular and protein-like conformations without simulations. In this study, the author used single-chain three-helix bundles as simple models of protein tertiary structures and sought to thoroughly investigate the conditions required to construct them, starting from the identification of the typical αα-hairpin motifs. First, by statistical analysis of naturally occurring protein structures, the author identified three αα-hairpins motifs that were specifically related to the left- and right-handedness of helix-helix packing. Second, specifying these αα-hairpins motifs as junctions, the author performed sequence-independent backbone-building simulations to comparatively build single-chain three-helix bundle structures and identified the promising combinations of the length of the α-helix and αα-hairpins types that results in tight packing between the first and third α-helices. Third, using those single-chain three-helix bundle backbone structures as template structures, the author designed amino acid sequences that were predicted to fold into the target topologies, which supports that the compact single-chain three-helix bundles structures that the author sampled show sufficient quality to allow amino-acid sequence design. The enumeration of the dominant subsets of possible backbone structures for small single-chain three-helical bundle topologies revealed that the compact foldable structures are discontinuously and sparsely distributed in the conformational space. Additionally, although the designs have not been experimentally validated in the present research, the comprehensive set of computational structural models generated also offers protein designers the opportunity to skip building similar structures by themselves and enables them to quickly focus on building specialized designs using the prebuilt structure models. The backbone and best design models in this study are publicly accessible from following URL:
https://doi.org/10.5281/zenodo.4321632

Background

Designing protein structures from scratch requires the careful selection of the length and types of the secondary structures and the loop types; further, the global structure and the local structural motifs need to be consistent. However, it is difficult to predict the combinations of building blocks that result in globular and protein-like conformations without simulations. Therefore, it would be beneficial for protein designers to limit the building blocks to the typical ones and enumerate the dominant subsets of their possible combinations in order to find promising combinations that result in highly designable backbone structures. In addition, once such conformational enumeration is performed, their results can be shared with other designers, and would enhance further design studies by allowing them to skip resampling the similar structures.

The αα-hairpin is a well-known structural motif, which consists of two adjacent α-helices and a loop region in between [1]. The loop region allows two flanking α-helices to pack into antiparallel arrangements, and the steep turn leads to tight non-local contacts between the two adjacent α-helices. Although the loop regions in general show non-repetitive structures and their conformations are more complicated than secondary structures, a few of them show clear patterns and can be classified into several subtypes, and are thus regarded as local motifs [2, 3]. Several pioneering studies have identified certain typical conformations of loops that are specifically related to αα-hairpins [1, 4, 5] and utilized them for design [6].

In this study, the author considered single-chain three-helix bundles as the simplest tertiary structures and investigated the conditions required to consistently construct them. The single-chain three-helix bundle is composed of three α-helices and two connecting loop regions that fold into hairpin conformations, causing neighboring α-helices to pack tightly into a compact antiparallel bundle configuration. Consequently, the third α-helix can be packed parallel to the first helix. The single-chain three-helix bundle structures are frequently observed in naturally occurring proteins and also have been designed artificially as well [7, 8]. Of note, the design of the single-chain three-helix bundle was one of the earliest efforts in the de novo protein design [8]. The design of helical bundles or multiple-chain coiled-coils is nowadays one of the largest fields in the protein design study, allowing diverse α-helix arrangements [9]. It is now clear what residue-residue non-local interactions can cause tight packing between α-helices [10] and result in various helix-bundle arrangements [11], which originates from analysis and design of coiled-coil structures [12–14]. Such knowledge for helical bundle designs have recently led to design of antibody-like and interleukin-mimicking artificial proteins [15–17] and programmable heterodimers [18]. However, many of previous works focus on the interface design between α-helices and do not pay much attention to the detailed conformations of loops that optimally connect individual α-helices. Therefore, when compared to coiled-coils and peptide assemblies that are composed of several independent chains, it still remains unclear and undocumented which combinations of the α-helix lengths and loop types result in compact single-chain helical bundle structures. Understanding the dominant subsets of possible conformational spaces allowed for single-chain helical bundles will be fundamentally important and even informative to efficiently design pharmacologically valuable artificial proteins. To this end, we aimed to understand which combinations of αα-hairpins and α-helix lengths can result in compact single-chain three-helical bundle structures, considering the αα-hairpins as the fundamental building blocks.

Results and discussion

Specific αα-hairpin loops determine the handedness of helix-helix packing

To identify typical hairpin motifs, the author performed a statistical analysis of helix-loop-helix fragments and found that shorter loops are present in greater frequency (Figure S2-1). To focus on hairpins rather than general helix-loop-helix fragments, the author defined helix-orientation vectors [19] and calculated their crossing angles (Figure S2-1). On applying the condition that the helix-helix crossing angles θ_{HH} are less than 60° in the fragment dataset, the author found a decrease in the population of single-residue loops, as the short helix-loop-helix prefers corners or kinks rather than hairpins. The author focused on the more frequent short hairpin fragments and extracted 2, 3, and 4 residue length loops for subsequent analysis.

To investigate the preferable loop conformations related to the specific handedness of helix-helix packing in naturally occurring protein structures, the author assigned a 5-state coarse-grained representation for the backbone torsion angle, i.e., ABEGO representations (Figure S2-2) for each fragment and evaluated their statistical information [2]. ABEGO is a five-state coarse-grained representation of polypeptide backbone dihedral angles; Ramachandran map is divided into four sections and labelled by single letters A, B, E and G, to enable the representation of dihedral angle series by character strings. The A region roughly corresponds to the conformation of α-helix, and the B region corresponds roughly to the β-strand conformation. The G region corresponds to left-handed α-helix, and the E region represents the rest of the Ramachandran map. The O state correspondds to the cis-conformation of peptide bond, which are almost negligible in this paper. Then the author sorted the backbone torsion types specified by the ABEGO representations by their population and found that the hairpins showed limited conformations (Figure 2-1). For example, the GB and BB loops occupied more than 90% of the top five frequent populations among two-residue loops.

To identify hairpins that show specific handedness in helix-helix packing, the author defined helix-helix dihedral angles φ_{HH} and calculated the ratio of left- (L-) and right- (R-) types among the helix-helix dihedral angle distribution (Figure 2). The author selected the most populated loop types that showed R/L or L/R ratios higher than 5.0 in each class of loop lengths as representative hairpin species. This resulted in the selection of the GB, GBB, and BAAB loops; the author did not select BAB loop because their inter-helix dihedral angles were broadly distributed, resulting in both left- and right-handed helix-helix packing (Figure S2-3).

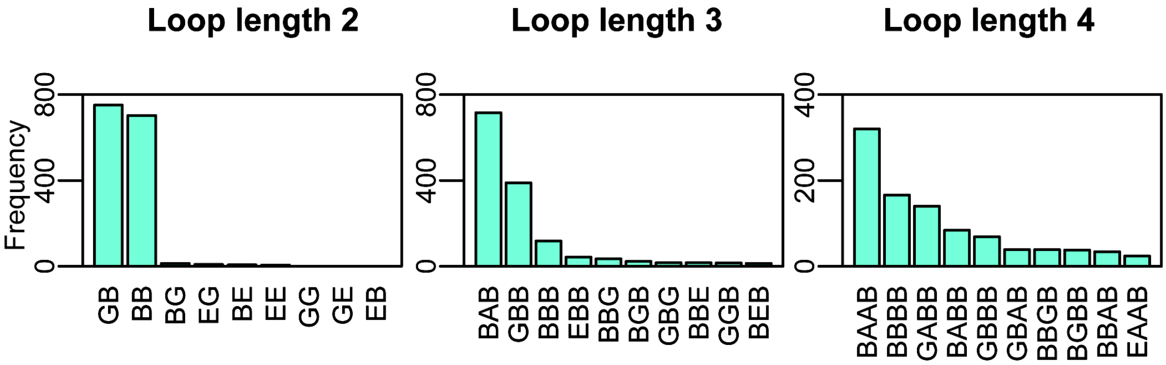


Figure 2-1: Identification of typical αα-hairpin motifs by the ABEGO representation for the loop conformations. The distributions of the top-10 typical hairpin conformations for two, three, and four-residue length loops that were identified by ABEGO. Several specific series of backbone torsion angles are strongly preferred in the hairpin loop region.

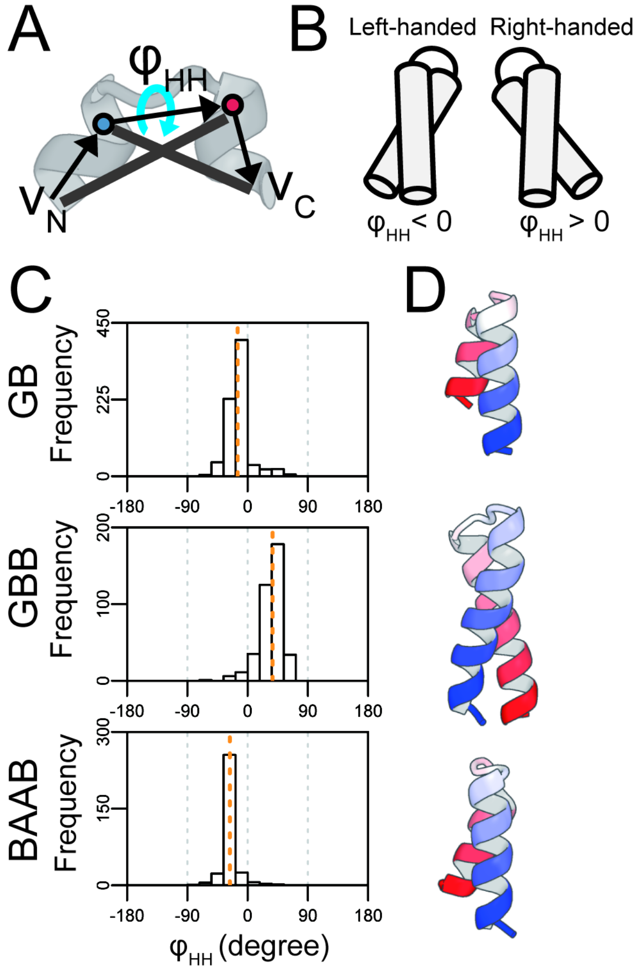


Figure 2-2: The handedness of helix-helix packing forced by typical αα-hairpin motifs. (A) Definition of φ_{HH} (B) Definition of packing handedness (C) The

distribution of φ_{HH} for GB, GBB, and BAAB hairpins. The horizontal axis represents the values of the helix-helix dihedral angle φ_{HH}, and the vertical axis corresponds to the number of loop fragments in the dihedral bins. The barplots are binned every 18°, and the orange dotted line indicates the median value of the distribution. (D) Representative structures for the GB, GBB, and BAAB hairpins. The structures show the GB and BAAB hairpins are related to left-handed packing, and GBB is related to right-handed packing. The α-helices are shown as cartoons and are colored in the blue-white-red gradient from the N to the C-terminus.

The author extracted the structures whose φ_{HH} was nearest to the median of the angle distribution as the class representatives. The representative structure and the distribution of φ_{HH} clarified that GB loops are closely related to the L-type handedness of helix-helix packing motifs (Figure 2-2). Similarly, the GBB loop was related to the R-type and the BAAB loop to L-type packing. The handedness of helix packing for the GB, GBB, and BAAB loops was broadly consistent with a previous report [20] and the overall tendency did not change when the author performed the same analysis for different dataset (Figure S2-4). The author also checked the clustering quality by sequence alignments for each hairpin structure, and observed typical periodic patterns of hydrophobic residues in the flanking helix regions (Figure S2-5). The author concluded that classification using the ABEGO patterns worked well to extract hairpin motifs related to the specific handedness of helix-helix packing.

Previous studies on αα-hairpins reported both L and R types of helix-helix packing can result from GB or GBB hairpins [1, 4]. However, the author observed that these loops indeed strongly bias the handedness of helix-helix packing. This does not imply that a single ABEGO-level representation can always specify the single handedness of helix-helix packing; for example, BAABB loop can result in both the L and R type packing (Figure S2-3). However, certain hairpin conformations such as GB, GBB, and BAAB can strongly determine the handedness of the packing of two flanking α-helices, and are an example of a pair of local and nonlocal structural motifs that are consistently incorporated into a single tertiary structure. As the spatial arrangements and orientation of two α-helices connected by a loop region are stereochemically determined by the backbone dihedral angles in the loop region, preference to specific handedness of helix-helix packing can be attributed to the rigid conformation of the loop region. Hydrogen bond analysis using DSSP[21] revealed that intra-loop backbone-backbone hydrogen-bond network energetically stabilizes such typical loop conformations, making the loop conformation rigid enough to relate the local conformation of loops to the specific geometry helix-loop-helix fragments (Figure S6–S8).

Sequence-independent backbone-building simulations clarify the condition for building compact single-chain three-helix bundles

The length of the second α-helix is expected to play a crucial role in the construction of compactly packed single-chain three-helical bundle structures since extension of α-helix leads to large repositioning of the following segments. As one turn of the α-helix requires 3.6 residues, compact bundle structures may appear for every increase of 3 or 4 residues. However, it remains unclear which exact combination of loops and helix-length results in a compact single-chain three-helix bundle structure. Therefore, the author performed comparative sequence-independent fragment-assembly simulations to identify which combination of loops and helix lengths result in tight packing between the first and third α-helices.

The set of backbone dihedral angles of the fragments are roughly specified in the ABEGO representation (referred to as "blueprint" [22]) and are used in fragment-picking before the fragment-assembly simulations. Hereafter this type of fragment assembly simulation guided by the blueprints is referred to as backbone-building simulations. Using the GB, GBB, and BAAB loops identified in the previous section, the author constructed blueprint files for various types of single-chain three-helix structures and systematically scanned the length of the second α-helix. Next, 2500 trajectories of backbone-building simulations were performed for each of these blueprints [23, 24]. The author prepared ideal single-chain three-helical bundle decoys using CC-builder for the reference structures [25], and calculated the template modeling score (TM-score) of the final structure from each trajectory that was referenced by the decoys to quantify the success ratio of the backbone-building simulations [26]. Importantly, the author used two reference decoys for each blueprint-based folding simulation because single-chain three-helical bundles can take two types of helix configurations; i.e., a clockwise (CW) or counter-clockwise (CCW) arrangement of three α-helices (Figure S2-9).

The results of the backbone-building simulations are summarized in Figure 2-3. The simulations showed three important features that are summarized here by taking the results for the helix-GB-helix-GB-helix simulations as examples. First, the length of the second alpha-helix plays a crucial role in the construction of compactly packed single-chain three-helical bundle structures; the success ratio of the backbone-building simulations was obviously related to the periodicity of the α-helix structure. For example, CW bundles can be efficiently generated with the second α-helix with lengths of 10, 14, and 17 residues for helix-GB-helix-GB-helix blueprints. Similarly, the second α-helix with lengths of 9, 12, 16 residues resulted in CCW bundles. Here, the peaks were separated in every three or four residues, which was consistent with the canonical α-helix structure that requires 3.6 residues per turn. Second, certain combinations of loops and helix-lengths do not yield well-packed helix bundles. For example, the blueprint with a 15 residue helix in the middle cannot fold into a compact helical bundle. This is because the number of turns in the second α-helix is unable to pack the first and third helices closely, causing them to be apart from each other (Figure S2-10). Such a blueprint has a local conformation that is inconsistent with the global structure of compact single-chain three-helical bundles. Third, the position of the peaks oscillates between CW and CCW bundles as the length of the second helix increases. The switch between a CW bundle to the neighboring CCW bundle is very sharp and sometimes requires an increase/decrease of a single residue. For example, the blueprint with a 16-residue helix in the middle preferentially results in the CCW bundle structure, and an increase of one residue results in a preference for the CW bundle. Overall, the results of the backbone-building simulations agree with the qualitative expectations that were guided by the periodicity of the α-helix structure, and provide further detailed information on the exact combination of loop types and helix-lengths that result in compact bundle conformations. These results were not affected when different threshold and reference structures were used for analysis (Figures S2-11 and S2-12).

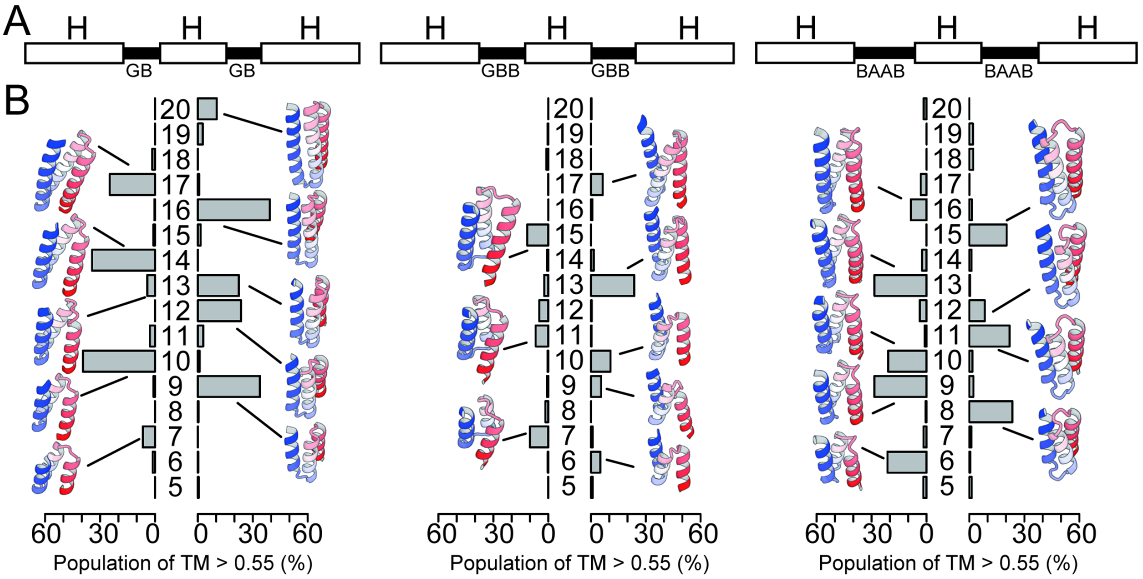


Figure 2-3: The length of the second helix is highly responsible for the compaction of single-chain three-helix bundles. (Top) The blueprints of single-chain three-helix bundles. The white bars indicate the α-helices, and the black bars indicate the loop regions, where integer H denotes the variable length of the second α-helix. The alphabets beneath the loop represent the ABEGO of the loop region specified in the blueprint. (Bottom) Bar graphs to summarize the foldability of each blueprint with the variable length of the second α-helix, H, scanned from 5 to 20 residues.The bars represent the population of folded structures that showed a TM-score higher than 0.55 as referenced by ideal CW (left) or CCW (right) three-helical topologies. The vertical axis represents the length of the second α-helix H. **T**he structures shown beside the bars are the representative snapshots from the

backbone-building simulations with highest TM-scores.

Taken together, these results indicate that the appropriate combination of local loop motifs and the length of the secondary structures are relatively rare among the possible combinations, especially under approximation that the loops and α-helices are semi-rigid under ABEGO constraints on backbone dihedral angles. In the author’s simple simulations for single-chain three-helix bundle structures, approximately half of the blueprints were able to generate compact bundle conformations. As the foldable combinations of building blocks are rare and sparsely distributed even for simple single-chain three-helical bundles, valid combinations for more complicated topologies are expected to become rarer and more difficult to find. The author expects that the possibility of obtaining foldable combinations will decrease exponentially as the number of secondary structures increases, and it will be difficult to hypothesize as to which combinations may result in a compact, globular, and protein-like structure without exhaustive sampling in the conformational space.

Other types of blueprints, such as for helix-GBB-helix-GBB-helix, and helix-BAAB-helix-BAAB-helix showed results similar to the GB-blueprint. Interestingly, the “phase” of the peak oscillation was inverted between the GBB and BAAB-blueprints, whereas the positions of the peaks were similar to each other, reflecting the local handedness of the hairpin structures. The former results in a CCW bundle when the second helix has 13 residues, and the latter yields a CW bundle in the same conditions. These observations that the local handedness of hairpins can control the global chirality of the topology may be informative for efficiently diversifying the shapes of design proteins. Additionally, the blueprints showing a mixture of hairpins with different handedness failed to pack the first and third α-helices because their crossing angles do not cancel out (Figure S2-10 and S2-13–2-17).

Amino acid sequence design suggests that the enumerated globular single-chain three-helix bundle structures may be designable.

As the backbone model generated in the previous section lacked any information on amino acid sequences, the author performed sequence designs using Rosetta [24] to check if the compact single-chain three-helix bundle structures are designable as concrete amino acid sequences. The author selected 27 backbone structures that are listed in Figure 3 and performed amino acid sequence designs for these backbone structures. The author designed ~7000–9000 sequences for each backbone structure and observed that the interfaces between the first and third α-helices recovered the sequence motifs for helix-helix packing (Figure S2-18). The results show that the relative arrangements of the first and third α-helices sampled in the sequence-independent backbone-building simulations are realistic enough to mold the typical amino acid sequences observed in helix-helix packing motifs. The optimal combinations of local properties such as the hairpin types and α-helix lengths lead to the successful recovery of non-local features. From these ensembles of design models, the author selected the most foldable sequences for each topology using sequence-dependent fragment assembly simulations [27] (Figure 2-4 and Figure S2-19–S2-21). In most of the simulation settings, the lowest-score models agreed well with the design models and recovered local hairpin structures well (Figure S2-22–S2-27 and Table S2-1). The author also performed negative-control designs in which the loop regions of the up-down helix bundles have atypical conformations, such as EE, BEB, and BEEE. The best-effort designs for these backbone models were indistinguishable in terms of par-residue Rosetta scores from the designs with typical hairpin motifs (Table S2-2). However, they were not able to efficiently fold into the target topology in the sequence dependent fragment-assembly simulations (Figure S2-28). This result suggests that the compact up-down bundle structures with typical hairpins have higher designability than the ones composed of atypical hairpins. For these best design models, the author performed blast search using blastp against a non-redundant sequence database [28, 29] and confirmed that three were no similar sequences found in the database.

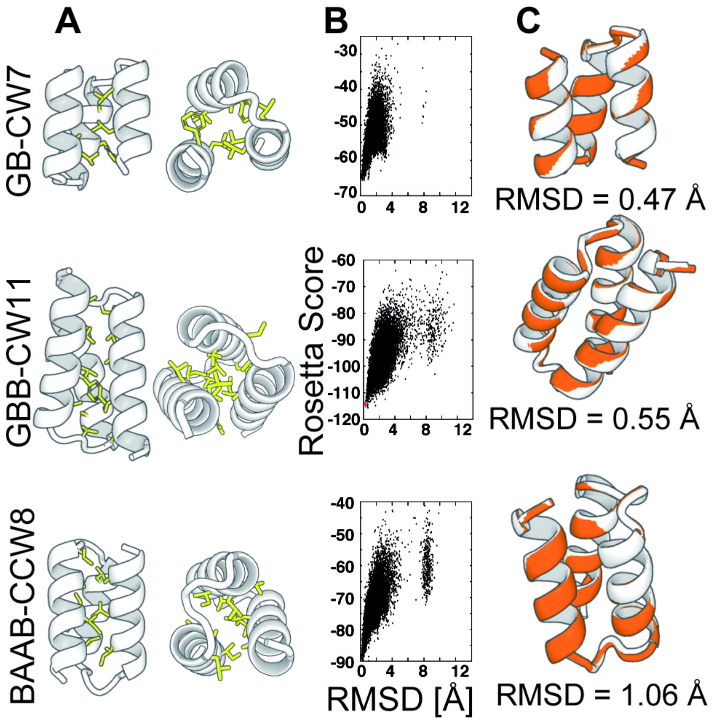


Figure 2-4: The representative structures and the results of sequence-dependent folding simulations of the designed single-chain three-helix bundles: GB-CCW9, GBB-CW11, and BAAB-CCW8. (A) The side-view and top-view of the designed structures with the α-helix shown as a cartoon and the hydrophobic side-chains represented as sticks. (B) The results of the folding and relax simulations. The vertical axis represents the Rosetta score, and the horizontal axis represents the root-mean-square deviation from the target structures. The black dots correspond to the final snapshots of the fragment-assembly folding simulations

starting from extended conformations, and the red dots correspond to the final snapshots of the relax simulations starting from the native conformations. These designs are predicted to fold into the target topologies because the trajectories of the folding simulations can reach the near-native ensembles. (C) The lowest score models in folding simulations (orange) superimposed onto the design models (white). The predicted models and design models agree well, which suggests the designed amino acid sequences fold well into the target conformations.

To confirm the stability of designed proteins independently from the Rosetta score function and fragment assembly simulations, the author utilized molecular dynamics simulations of designed proteins models and assessed their quality [30, 31]. The author performed molecular dynamics simulations for the 27 best design models using GPU-accelerated GROMACS 2020.6 [32, 33] alongside Amber 15FB force field [34]. For each design, the author performed 10 trajectories of 100 ns molecular dynamics simulations with explicit TIP3P water models. After energy minimization and equilibration, the author performed 100 ns of the production run under the pressure of 1 bar and temperature of 300 K. The simulation showed the most of the design models can stay within 5 Å in the root-mean-square deviations (RMSD) of Cα coordinate from the designed structures for 100 ns (Figure S2-29–S2-34). The only exceptions were 7th trajectory of GB-CW7 and 3rd and 8th trajectories of GBB-CCW6 (Figure S2-29 and S2-31), which resulted in partial unfolding of the structures. These structures may be unstable probably because they have too small hydrophobic cores to maintain the designed topologies. Overall, the molecular dynamics simulations showed that the designed proteins were stable enough to keep the native conformation in the solution state. These results provided independent validation for the designability of the backbone structures we sampled by fragment assembly with the Rosetta score function.

As the author designed up-down types of helical bundles, whose substructures can be regarded as antiparallel and parallel coiled coils, the author confirmed whether the best-designed sequences can be recognized as coiled-coils by DeepCoil [35–37]. Interestingly, the predicted probability to observe coiled-coil arrangements within the designed structure increased as the length of the design proteins increased (Figure S2-35–S2-40). Approximately, when the length of the second α-helix is longer than 15 amino-acid residues, the probability to recognize the sequence as coiled-coil becomes higher than the significance threshold. This suggests that these up-down helical bundles can be regarded as coiled coils when the chain lengths are large enough. Therefore, although the author designed amino-acid sequences without considering the structures are related to coiled coils, sequence design techniques for coiled coils can be repurposed to the sequence design of large helical bundles and may result in more optimized helix-helix packing, which may lower the computational cost of design and improve the yield of successful design sequences. On the other hand, smaller helical bundles failed to be predicted as coiled coils. This does not immediately imply that such small helical bundles are not designable; such small helical bundles were designed in a previous study [38]. Therefore, the sequence design of such small helical bundles should be performed without considering the structure as coiled-coils. This analysis suggested that optimal design methods may be able to be selected depending on the size of target helical bundles. It is also interesting whether parametrically designed multiple-chain coiled coils can be redesigned into single-chain helical bundles by designing the loops connecting the α-helices; the question is whether the designer can find appropriate loop conformations to connect the α-helices [39] without frustrations between local and nonlocal interactions [40, 41].

Finally, to detect knobs-into-holes structure in the author’s designs, the author performed structure analysis using SOCKET [42]. According to SOCKET, knob-into-holes structures were observed roughly in two-third of the author’s designs (Figure S2-41-S2-44). In addition, SOCKET detected coiled-coil structure in 4 of the author’s designs, although the author did not intend to design coiled-coil-like substructures in the author’s design scheme. This also suggests that the design of helical bundles shares many similar aspects with the design of coiled coils, and the rich and matured protocols for coiled-coil design can be imported into the design of single-chain helical bundles.

Conclusion

In this study, the author used single-chain three-helix bundles as simple models of protein tertiary structures and investigated the conditions required to construct them, and aimed to understand the mechanisms by which these local and nonlocal motifs are consistently incorporated into a single three-dimensional structure. First, the author showed that the GB- and BAAB-hairpins are related to left-handed helix-helix packing, whereas the GBB-hairpins are related to right-handed packing. Second, by enumerating the combinations of the hairpin types and the helix length, the author identified the combinations of helix-length and loop types that resulted in successful compaction of single-chain three-helix bundle structures. As the author has enumerated most of the backbone structures that are potentially obtainable for these simple topologies under the condition that the hairpins are limited to GB, GBB, or BAAB, and the lengths of the second α-helix are less than 20 residues, no other single-chain up-down three-helical bundle structures are plausible in this subspace of the structural space. Combined with the observation that the populations of loops are strongly biased towards a limited number of typical conformations, such enumeration can cover most of the possible conformational space.

The author also showed that the backbone structures composed of such short hairpin motifs may be highly designable by amino acid sequence design and sequence-dependent folding simulations, although experimental validation for these designed proteins should be done elsewhere. In addition, Molecular dynamics simulations supported that the designed proteins are stable in solution, which suggests that designed proteins do not have internal frustrations between local and nonlocal interactions. Using programs to detect coiled-coil sequences and structures, the author also found that the designed sequences and structures can be recognized as coiled-coil when the sequences are long enough. This implies sequence design methods based on sequence periodicity of coiled coils, which is usually utilized in design of multi-chained coiled coils or peptide assemblies, can be repurposed for the design of single-chain up-down helical bundles to realize optimized helix-helix packing.

Though the author's analyses are limited to the simplest class of tertiary structure, single-chain up-down three-helical bundles, the author has shown that the enumerative exploration into the conformational space can clarify the appropriate combinations of building blocks. The author also showed such exploration can yield transferable structural resources for protein design that can be shared with other protein designers. As such enumeration does not need to be done twice, data sharing among designers would promote advances in the protein design fields. To this end, the 27 types of backbone structure that the author enumerated and the best sequences that the author designed are now publicly available at https://doi.org/10.5281/zenodo.4321632

Method

Initial dataset preparation

A collaborator of the author composed a subset of the ECOD database (version 238) whose sequence redundancy was reduced by 40% sequence identity [43]. Next, secondary structures were assigned using DSSP [21], and a total of 39,938 helix-loop-helix substructures were extracted having loop lengths that were less than or equal to 10. The author discarded the structures whose α-helices have less than or equal to 9 residues.

The author prepared another dataset of PDB structures whose sequence redundancy was reduced by 25% sequence identity with resolution lower than 3.0 A using Pisces server[44], and obtained 29,149 helix-loop-helix structures. These structure were used to check the effect of resolution cut-off for the geometric analysis of helix-loop-helix fragments (Figure S2-4)

ABEGO-level dataset preparation

The backbone dihedral angles were translated into 5 state coarse-grained ABEGO representations (Figure S2-2). ABEGO is a coarse-grained representation of polypeptide backbone dihedral angles, where the Ramachandran map is divided into four sections and labeled by single letters A, B, E and G. The O state corresponds to the cis-conformation of the peptide bond, which is almost negligible in this paper. The A region roughly corresponds to the conformation of α-helix, and the B region corresponds roughly to the β-strand conformation. The G region corresponds to left-handed α-helix, and the E region represents the rest of the Ramachandran map.

As it is ambiguous whether the dihedral angle of A in ABEGO representation is a loop region or α-helix termini, the author removed the loop fragments that start/end with A of ABEGO. The author only included fragments that started/ended with B, E, and G. After this data pruning, the author obtained 19,844 helix-loop-helix fragments. The author checked that the removal of fragments that started/ended with

A did not largely change the distribution of the frequent loop types (Figure S2-45). All of the date processing was performed with in-house R and python programs.

Definition of the geometrical features of helix-loop-helix fragments

For the final and first single turn on the N/C-terminal α-helices of the helix-loop-helix fragments, the vectors $\boldsymbol{v}_{\mathrm{N}}$ and $\boldsymbol{v}_{\mathrm{C}}$ representing the orientation of these α-helices were defined as per Krissinel et al. [19]. Additionally, the author defined the loop orientation vector $\boldsymbol{v}_{\mathrm{L}}$ as starting from the final/first Cα coordinates of the N-/C-terminal α-helices. Next, the author defined 2-geometric features using $\boldsymbol{v}_{\mathrm{N,}}\boldsymbol{v}_{\mathrm{C}}$, and $\boldsymbol{v}_{\mathrm{L}}$: (1) the helix-helix crossing angle θ_{HH}, (2) the helix-helix dihedral angle φ_{HH}. θ_{HH} is the crossing angle between two N/C-terminal α-helices (Figure S1) defined by the arc-cosine of the inner-product of $\boldsymbol{v}_{\mathrm{N,}}$ and $\boldsymbol{v}_{\mathrm{C}}$. φ_{HH} is the inter-helix dihedral angle between two α-helices defined by $\boldsymbol{v}_{\mathrm{N}}$, $\boldsymbol{v}_{\mathrm{L}}$, and $\boldsymbol{v}_{\mathrm{C}}$ (Figure 2-2). As the author focused on αα-hairpins, the author only collected fragments that satisfied the condition that θ_{HH} was less than 60° before performing the rest of the analysis.

Sequence-independent fragment-assembly simulations: Backbone-building simulations

Sequence-independent fragment assembly simulations, which the author referred to as backbone-building simulations, were performed using Rosetta BluePrintBDR [24] similarly as in Lin et al. [23]. The blueprint files were generated manually and were used in fragment picking to specify the backbone torsion in the ABEGO representation. For each site of proteins, 200 fragments were picked from the default structure library. For each blueprint, simulations were repeated for 2500 trajectories, and the final snapshots from the trajectories were used for structural analysis. A parameter set, fldcen.wts, was used as weight parameters for BluePrintBDR simulations.

In the analysis, two ideal decoy structures of single-chain three-helical bundles were used as references to calculate the TM-scores using TM-align [26]. The author prepared two types of reference decoys, i.e., clockwise (CW) and counter-clockwise (CCW) bundles originating from the ideal decoy structures that were generated by CC-builder [25]. The parameters for CC-builder were as follows; oligomeric state 3; radius 6.75, 8.1, 9.0, 9.9, and 11.25 for x0.75, x0.90, x1.00, x1.10, and x1.25 radius variant of helix bundles; pitch 300; interface Angle 20. Based on these decoys built by CC-builder, the author manually modified their helix-orientation to up-down-up and packing chirality by re-sorting and mirroring the Cα coordinate and superimposing ideal α-helices onto the mirrored helix arrangements (Figure S2-9). In the data analysis, the snapshots showing TM-scores higher than 0.55 were counted as folded into three-helical bundle structures [45]. To check the robustness against the change in reference structures, the author systematically modified the diameter of reference helix-bundles by the magnitude of 0.75, 0.9, 1.10, and 1.25. The author also changed the threshold of TM-score (0.50, 0.55, and 0.60) to check the robustness of the results. These parameters were found not to largely change the results (Figure S11 and S12).

Construction of negative-control helix-bundle structures

Based on the anti-parallel part of decoy structures described above, the loop regions were modeled using Modeller [46] and six types of helix-helix hairpins that have atypical EE, BEB, and BEEE conformations were selected. Then the respective hairpin structures were repeated to form three-helix bundles composed of atypical hairpins. The severe steric clashes between alpha-helices were removed using Foldit-standalone [47].

Amino acid sequence design and sequence-dependent folding simulations

Amino acid sequence design was performed using the Rosetta flxbb protocol [24] starting from the backbone structure that showed the best TM-score in the previous sequence-independent folding simulations. Score Talaris2014 was used in all designs and folding simulations including negative-control designs. In the loop region, amino acid profiles were constructed using similar loop structure fragments (RMSD < 2 Å) and used as constraints for residue types, similarly to Marcos et al [48]. In addition, the specification on the residue types was refined based on the buriedness of the backbone atoms using in-house programs. The text files, i.e. the so-called “resfiles" specifying the final residues set were attached as supplementary files. The author performed 10,000 design trials for each backbone model and obtained ~7,000–9,000 design sequences that passed the secondary structure filter. The author selected the best 5–10 sequences using the fragment-quality score. The author defined the fragment quality score as the average of the logarithm of the number of fragments with RMSD lower than 1.5 Å from the design model, similarly to Marcos et al [48].

The author performed sequence-dependent fragment-assembly folding simulations [27] to identify the best design sequences. Sequence dependent fragment assembly simulations, which the author denoted “folding” simulations, were performed using AbinitoRelax binary in Rosetta suite with 200 3-mer and 9-mer fragments collected by psi-blast search in the default structure library. Near-native sampling simulations, which the author denoted “relax“ simulations, were performed by relax binary in Rosetta suite to sample near-native conformation starting from the designed structure models. 20,000 trajectories of fragment-assembly folding simulations were performed for each design protein, and their ability to fold into the target structures was evaluated by the shapes of the energy landscapes.

MD simulations

All of the simulations were performed using GROMACS 2020.6 [32, 33] with Amber force field ff15FB [34]. First, the author performed the in-vacuo energy minimization by the steepest descent for 500,000 steps, and the energy-minimized protein structures were solvated by TIP3P water models. The initial box size was set to 6 nm x 6 nm x 6 nm, which was large enough for all types of designs. The Na+ and Cl- ions were introduced to the system at the concentration of 0.1 mol/L. Depending on the total charge of the designed proteins, additional Na+ or Cl- ions were added to the system so that the system has zero net charges. The whole system was energy-minimized again for 500,000 steps.

With the step size of 2.0 fs, the whole system was equilibrated by 100 ps of NVT and NPT simulations under harmonic constraint for heavy atoms. Then 100 ns of production runs were performed without any external constraints to the system under 1bar of pressure and 300 K of temperature. The production runs of the MD simulations were performed with LINCS constraint algorithm for the bonds between hydrogen atoms and heavy atoms. The temperature of the system was controlled to 300 K by the V-rescale algorithm (modified Berendsen thermostat) with the time constant of 0.1 ps. The pressure was controlled to 1.0 bar by Parrinello-Rahman algorithm with the time constant of 2 ps. The electrostatic part of the force field was calculated using the particle mesh Ewald scheme with the order of 4.

Availability of data and materials

The backbone models and best design models in this study are publicly accessible from following URL: https://doi.org/10.5281/zenodo.4321632

Best-effort negative control designs are available from the following URL : https://zenodo.org/record/5512549

The in-house python and R script for ABEGO analysis is disclosed in GitHub:

For ABEGO-based analysis: https://github.com/yakomaxa/ssdoublet

Other scripts to reproduce the research are also disclosed in GitHub:

For backbone building: https://github.com/yakomaxa/bbdesign_template

For MD simulations: https://github.com/yakomaxa/MD_gromacs_template

References

1. Efimov A V. Structure of α-α-hairpins with short connections. Protein Eng Des Sel. 1991;4:245–50.

2. Wintjens RT, Rooman MJ, Wodak SJ. Automatic classification and analysis of αα-turn motifs in proteins. J Mol Biol. 1996;255:235–53.

3. Oliva B, Bates PA, Querol E, Avilés FX, Sternberg MJE. An automated classification of the structure of protein loops. J Mol Biol. 1997;266:814–30.

4. Brazhnikov E, Efîmov A. Structure of a-hairpins with short connections in globular proteins. Mol Biol. 2001;35:100–8.

5. Efimov A V. Standard structures in proteins. Prog Biophys Mol Biol. 1993;60:201–39.

6. Lahr SJ, Engel DE, Stayrook SE, Maglio O, North B, Geremia S, et al. Analysis and design of turns in α-helical hairpins. J Mol Biol. 2005;346:1441–54.

7. Schneider JP, Lombardi A, DeGrado WF. Analysis and design of three-stranded coiled coils and three-helix bundles. Fold Des. 1998;3:29–40.

8. Walsh STR, Cheng H, Bryson JW, Roder H, Degrado WF. Solution structure and dynamics of a de novo designed three-helix bundle protein. Proc Natl Acad Sci U S A. 1999;96:5486–91.

9. Rhys GG, Wood CW, Beesley JL, Zaccai NR, Burton AJ, Brady RL, et al.

Navigating the Structural Landscape of de Novo α-Helical Bundles. J Am Chem Soc. 2019;141:8787–97.

10. Zhang SQ, Kulp DW, Schramm CA, Mravic M, Samish I, Degrado WF. The membrane- and soluble-protein helix-helix interactome: Similar geometry via different interactions. Structure. 2015;23:527–41. doi:10.1016/j.str.2015.01.009.

11. Dawson WM, Martin FJ, Rhys GR, Shelley KL, Brady RL, Woolfson D. Coiled coils 9-to-5: Rational de novo design of alpha-helical barrels with tunable oligomeric states. Chem Sci. 2021;9:4132.

12. Crick FHC. The packing of α-helices: simple coiled-coils. Acta Crystallogr. 1953;6:689–97.

13. Woolfson DN. The design of coiled-coil structures and assemblies. Adv Protein Chem. 2005;70:79–112.

14. Lupas AN, Gruber M. The structure of α-helical coiled coils. Adv Protein Chem. 2005;70:37–8.

15. Silva DA, Yu S, Ulge UY, Spangler JB, Jude KM, Labão-Almeida C, et al. De novo design of potent and selective mimics of IL-2 and IL-15. Nature. 2019;565:186–91. doi:10.1038/s41586-018-0830-7.

16. Linsky TW, Vergara R, Codina N, Nelson JW, Walker MJ, Su W, et al. De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. Science. 2020;:eabe0075. doi:10.1126/science.abe0075.

17. Cao L, Goreshnik I, Coventry B, Case JB, Miller L, Kozodoy L, et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science. 2020;370:426–31.

18. Chen Z, Boyken SE, Jia M, Busch F, Flores-Solis D, Bick MJ, et al. Programmable design of orthogonal protein heterodimers. Nature. 2019;565:106–11. doi:10.1038/s41586-018-0802-y.

19. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr Sect D Biol Crystallogr. 2004;60 12 I:2256–68.

20. Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard BL, et al. Rational design of α-helical tandem repeat proteins with closed architectures. Nature. 2015;528:585–8.

21. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen‐bonded and geometrical features. Biopolymers. 1983;22:2577–637.

22. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, et al. Principles for designing ideal protein structures. Nature. 2012;491:222–7. doi:10.1038/nature11600.

23. Lin YR, Koga N, Tatsumi-Koga R, Liu G, Clouser AF, Montelione GT, et al. Control over overall shape and size in de novo designed proteins. Proc Natl Acad Sci U S A. 2015;112:E5478–85.

24. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM, Khare SD, Koga N, et al. Rosettascripts: A scripting language interface to the Rosetta

Macromolecular modeling suite. PLoS One. 2011;6:1–10.

25. Wood CW, Woolfson DN. CCBuilder 2.0: Powerful and accessible coiled-coil modeling. Protein Sci. 2018;27:103–11.

26. Zhang Y, Skolnick J. TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33:2302–9.

27. Bradley P, Misura KMS, Baker D. Biochemistry: Toward high-resolution de novo structure prediction for small proteins. Science. 2005;309:1868–71.

28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

29. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36 Web Server issue:5–9.

30. Childers MC, Daggett V. Insights from molecular dynamics simulations for computational protein design. Mol Syst Des Eng. 2017;2:9–33.

31. Ludwiczak J, Jarmula A, Dunin-Horkawicz S. Combining Rosetta with molecular dynamics (MD): A benchmark of the MD-based ensemble protein design. J Struct Biol. 2018;203:54–61. doi:10.1016/j.jsb.2018.02.004.

32. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation. Comput Phys Commun. 1995;91:43–56.

33. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1–2:19–25.

34. Wang LP, McKiernan KA, Gomes J, Beauchamp KA, Head-Gordon T, Rice JE, et al. Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. J Phys Chem B. 2017;121:4023–39.

35. Ludwiczak J, Winski A, Szczepaniak K, Alva V, Dunin-Horkawicz S. DeepCoil - A fast and accurate prediction of coiled-coil domains in protein sequences. Bioinformatics. 2019;35:2790–5.

36. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018;430:2237–43. doi:10.1016/j.jmb.2017.12.007.

37. Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinforma. 2020;72:1–30.

38. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017;357:168–75.

39. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods. 2009;6:551–2. doi:10.1038/nmeth0809-551.

40. Gō N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng. 1983;12:183–210.

41. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of Protein Folding: The Energy Landscape Perspective. Annu Rev Phys Chem. 1997;48:545–600.

42. Walshaw J, Woolfson DN. SOCKET: A program for identifying and analysing coiled-coil motifs within protein structures. J Mol Biol. 2001;307:1427–50.

43. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. ECOD: An Evolutionary Classification of Protein Domains. PLoS Comput Biol. 2014;10.

44. Wang G, Dunbrack RL. PISCES: A protein sequence culling server. Bioinformatics. 2003;19:1589–91.

45. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010;26:889–95.

46. Šali A, Blundell TL. Comparative Protein Modelling by Satisfaction of Spatial Restraints. J Mol Biol. 1993;234:779–815. doi:https://doi.org/10.1006/jmbi.1993.1626.

47. Kleffner R, Flatten J, Leaver-Fay A, Baker D, Siegel JB, Khatib F, et al. Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta. Bioinformatics. 2017;33:2765–7.

48. Marcos E, Basanta B, Chidyausiku TM, Tang Y, Oberdorfer G, Liu G, et al. Principles for designing proteins with cavities formed by curved b sheets. Science. 2017;355:201–6.

Supplementary Material for "Enumeration and comprehensive in-silico modeling of three-helix bundle structures composed of typical αα-hairpins"

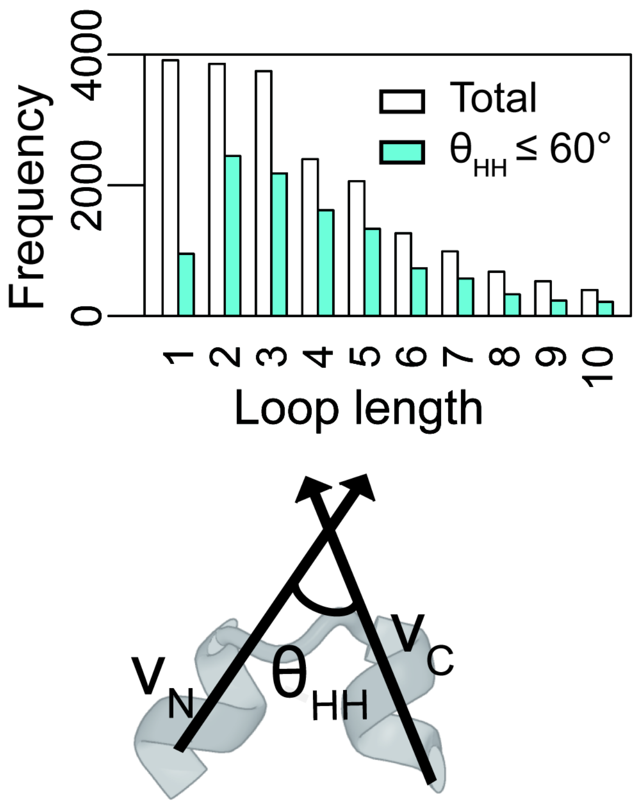


Figure S2-1. The Distributions of loop lengths. (Top) The distributions of loop lengths in the general helix-loop-helix fragment (white) and αα-hairpins conditioned by θ_{HH} < 60° (cyan). Loop length is the number of residues between two flanking α-helices. (Bottom) The definition of the helix-helix crossing angle θ_{HH}. Shorter loops were generally preferred but the population of single-residue loops diminished when θ_{HH} was less than or equal to 60°. Single-residue loops were too short to form hairpin structures.

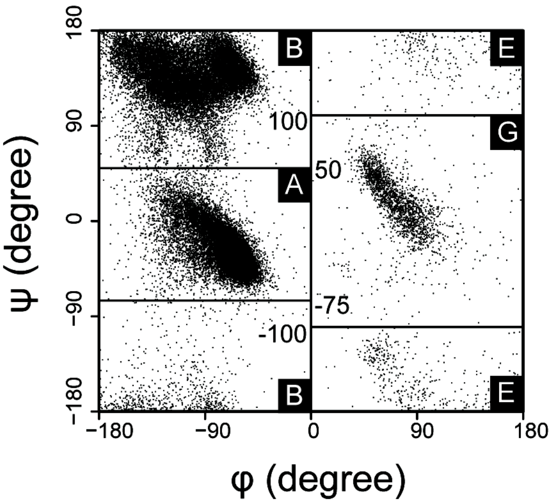


Figure S2-2. The definition of ABEGO, a coarse-grained backbone torsion representation. ABEGO is a five-state coarse-grained representation of polypeptide backbone dihedral angles; Ramachandran map is divided into four sections and labelled by single letters A, B, E and G, to enable the representation of dihedral angle series by character strings. The A region roughly corresponds to the conformation of α-helix, and the B region corresponds roughly to the β-strand conformation. The G region corresponds to left-handed α-helix, and the E region represents the rest of the Ramachandran map. The O state correspondds to the cis-conformation of peptide bond, which are almost negligible in this paper.

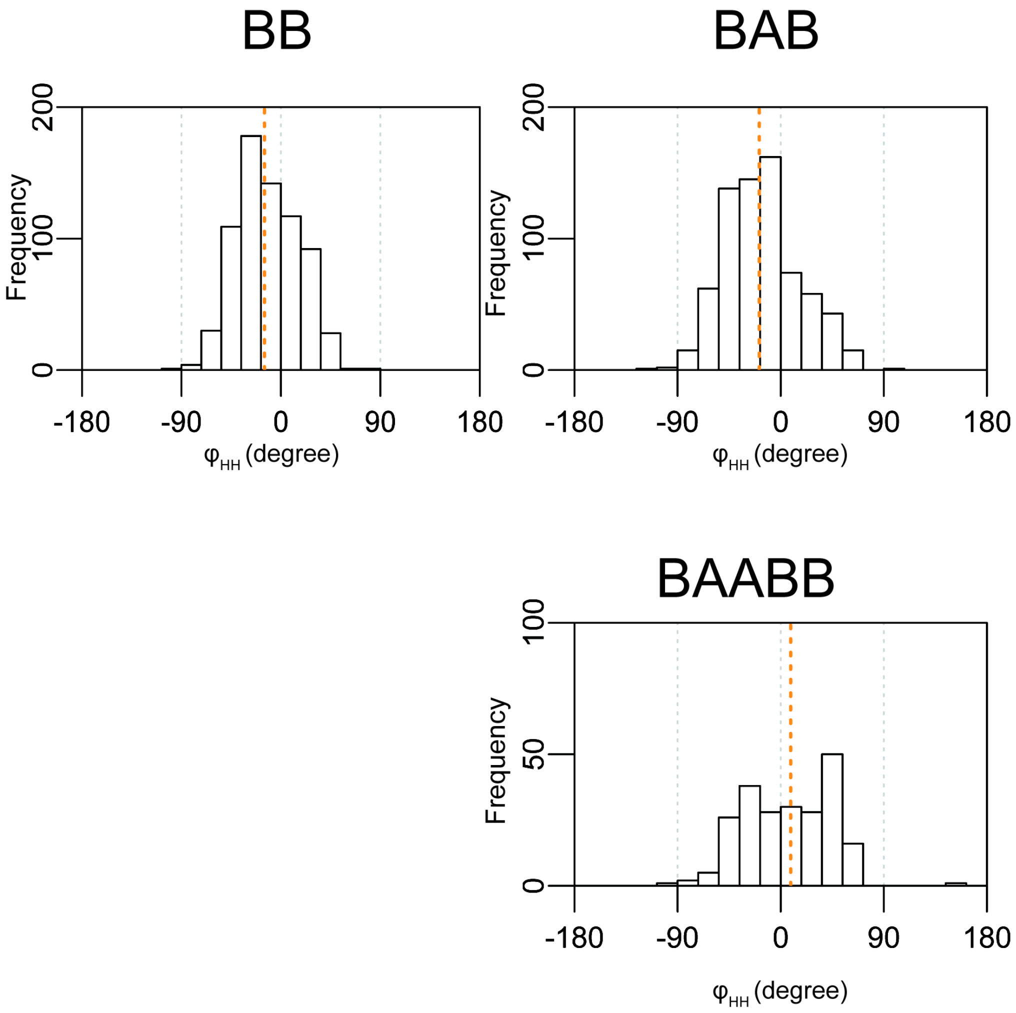


Figure S2-3. The distribution of φ_{HH} for BB, BAB, and BAABB loops. BB, BAB, and BAABB loops exhibited undetermined distributions of φ_{HH}, showing broad spectra around φ_{HH} = 0. The orange dotted line indicates the median of the distributions. These show that these BB, BAB, and BAABB hairpins cannot specify the handedness of the helix-helix packing, which the author omitted in the later analysis.

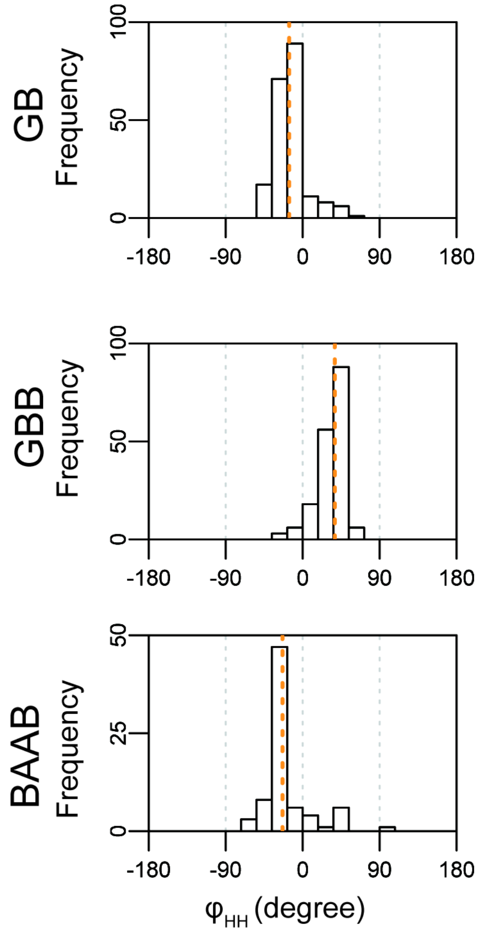


Figure S2-4: The distribution of φ_{HH} for GB, GBB, and BAAB loops using high-resolution structure dataset.

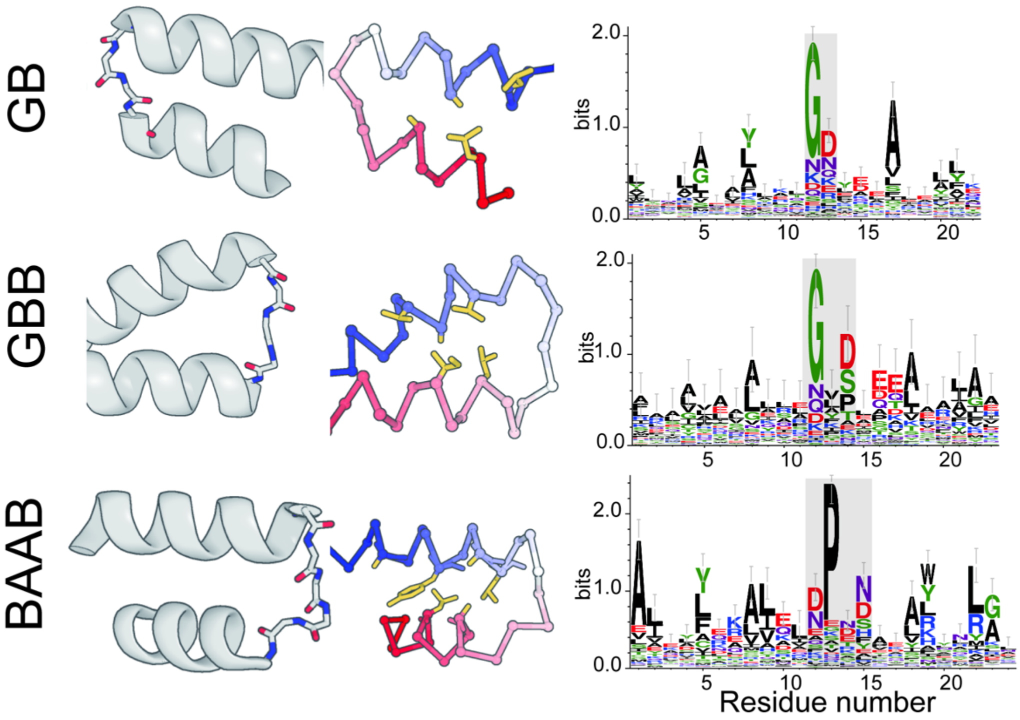


Figure S2-5. Packing between two α-helices connected by typical hairpins. (Left) The representative structure of typical hairpins. The α-helices are represented as cartoons, and loop regions are shown as sticks. (Middle) Hydrophobic residues enable tight packing between two α-helices in the representative structures. The hydrophobic residues are shown as orange sticks, and the backbones are shown as Cα-traces. (Left) The structure-based sequence alignments are shown as the sequence logo for the hairpin and the flanking ten residues of α-helices. The gray-shaded boxes indicate hairpin loop regions. There are hallmarks of hydrophobic helix-helix packing motifs in the α-helix regions, where small and hydrophobic residues periodically appear in the sequences for tight Van-der-Waals contacts.

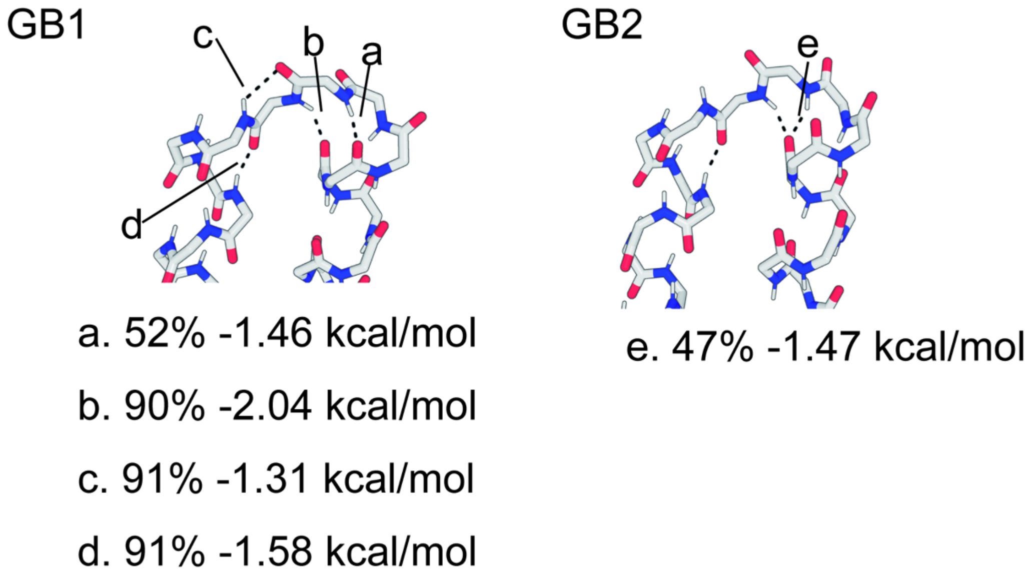
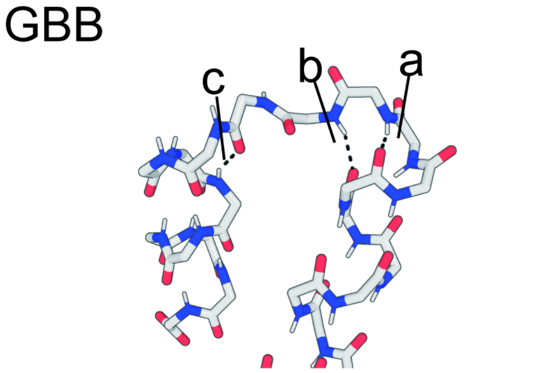


Figure S2-6: Hydrogen bond patterns in GB-loop. The percentage indicates the ratio of hydrogen bond formation in the dataset. The energy values represent the average bonding energy for each hydrogen bond estimated by DSSP. The GB-loop typically has 3 or 4 intra-loop hydrogen bonds that stabilize the loop conformation.

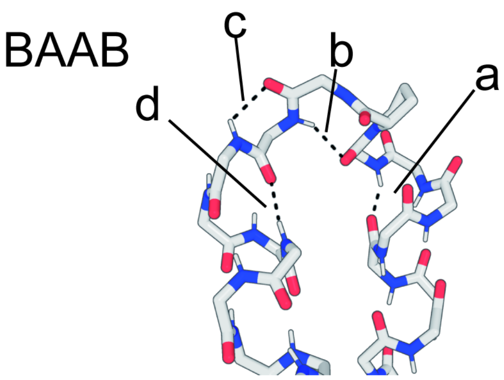


a. 87% -1.00 kcal/mol

b. 85% -1.99 kcal/mol

c. 96% -2.31 kcal/mol

Figure S2-7: Hydrogen bond patterns in GBB-loop. The percentage indicates the ratio of hydrogen bond formation in the dataset. The energy values represent the average bonding energy for each hydrogen bond estimated by DSSP. The GBB-loop typically has 3 intra-loop hydrogen bonds that stabilize the loop conformation.



a. 94% -2.21 kcal/mol

b. 98% -1.61 kcal/mol

c. 87% -1.00 kcal/mol

d. 82% -1.66 kcal/mol

Figure S2-8: Hydrogen bond patterns in BAAB-loop. The percentage indicates the ratio of hydrogen bond formation in the dataset. The mean energy values represent the average bonding energy for each hydrogen bond estimated by DSSP. The GBB-loop typically has 4 intra-loop hydrogen bonds that stabilize the loop conformation.

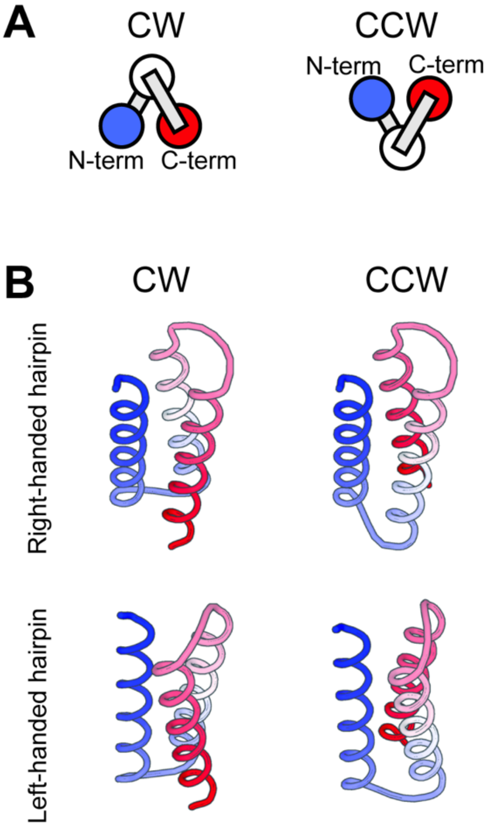


Figure S2-9. Two possible chiral forms of three-helix bundles and decoy structure used in the evaluation of sequence-independent folding simulations. (A) Three-helical bundles can have two types of chirality in their overall structures: Clockwise (CW) and Counterclockwise (CCW). Please note that this is independent of the local hairpin motifs. Circles indicate the α-helix viewed from the top, and bars indicate the connecting loops. (B) There are four possible three-helical bundle structures when local handedness of αα-hairpins are also considered. The author used the right-handed decoys for GBB-bundle folding simulations, and the left-handed decoys for the GB and BAAB-bundle simulations.

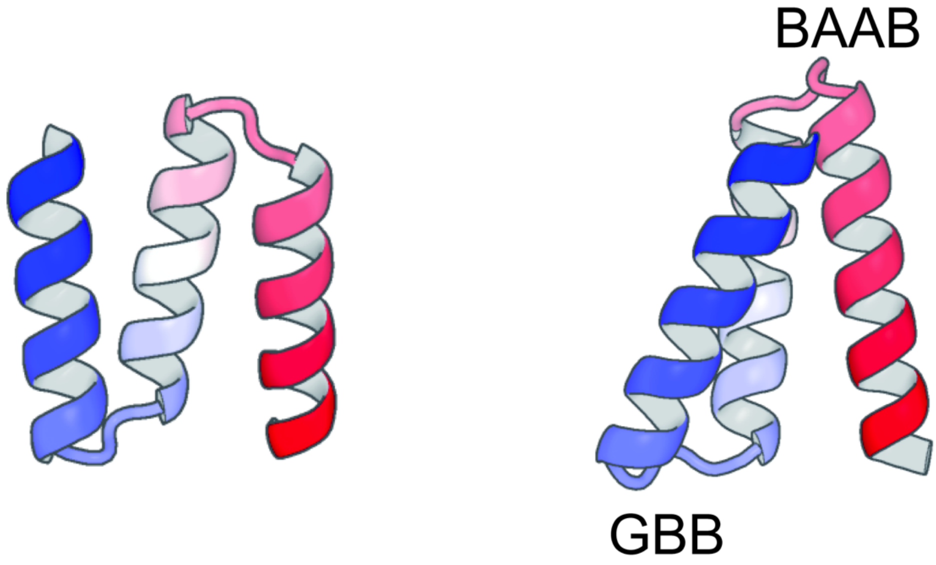


Figure S2-10. Two examples of poorly packed three-helical bundle structure. (Left) When two GB-hairpins are connected by a15 residue helix in the middle, the first and third helices are placed apart and result in an extended confirmation lacking contacts between the first and third helix. This type of extended structure lacks the would-be-hydrophobic-core region and therefore is not considered designable. (Right) An example of the structure with left- and right-handed mixed hairpins. The first and third α-helix can not pack when the hairpins with different handedness are mixed, and do not yield compact and globular three-helical bundle structures. See also Figure S15-S18, where such mixed-loop simulations are shown to be unable to generate either CW or CCW compact bundle structures.

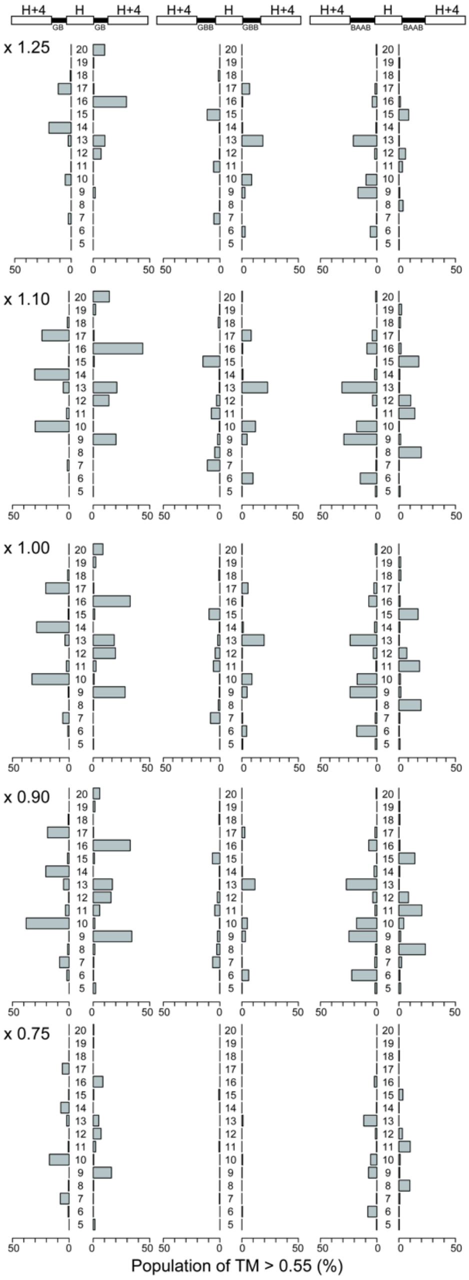


Figure S2-11: Effect of reference structures for TM-score calculations in the analysis on backbone-building simulations. The value at the upper left of each panel represents the magnification factor for the diameter of reference helix-bundles.This figure corresponds to figure 2 in main text, and shows that the results are not severely affected by the change in reference structures for TM-score calculation.

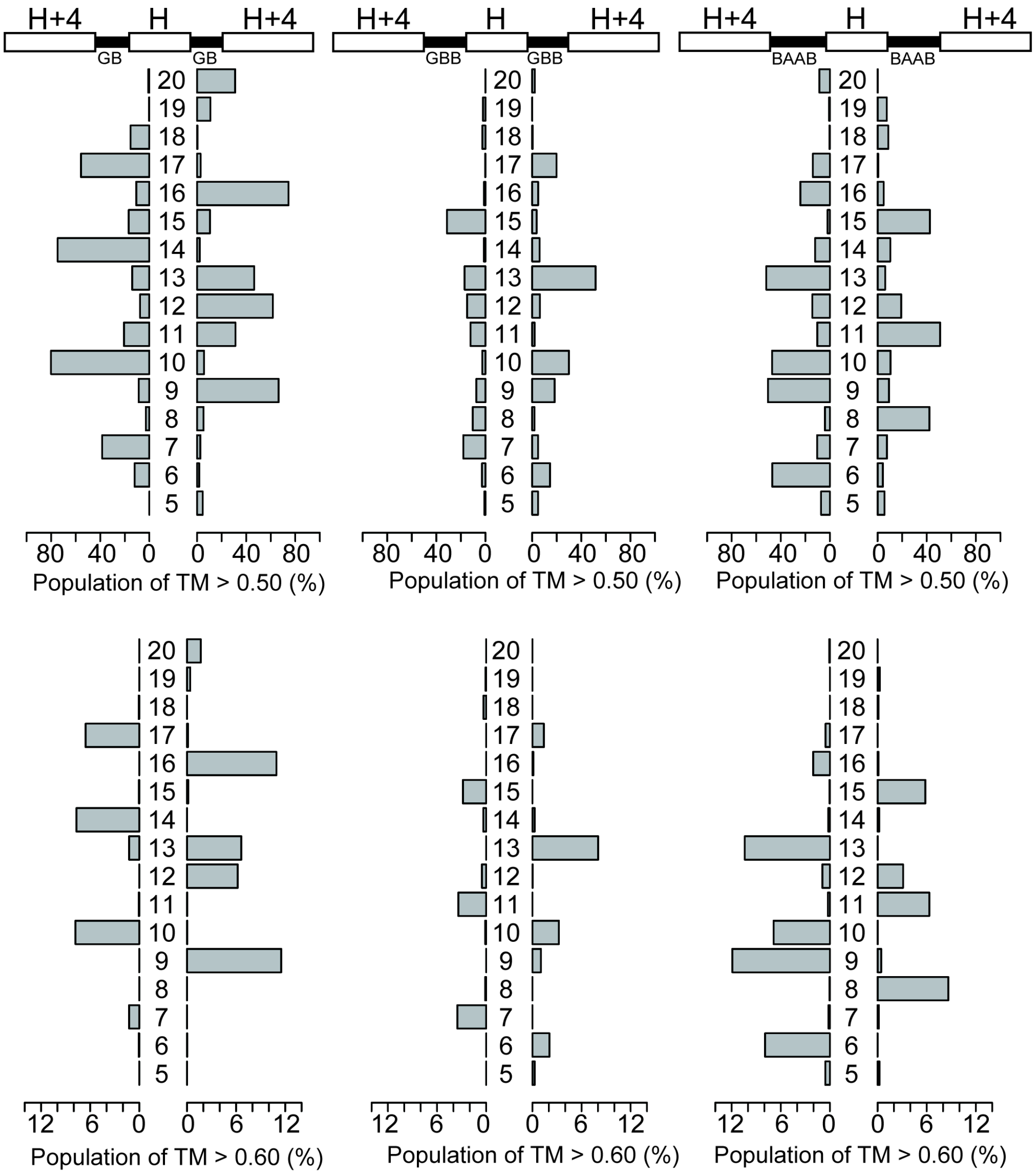


Figure S2-12: Effect of cut-off for TM-score in the analysis on backbone-building simulations. Two different thresholds 0.50 and 0.60 are used instead of the original threshold 0.55. This figure corresponds to figure 2 in the main text, and shows that the results are not severely affected by the change in TM-score threshold..

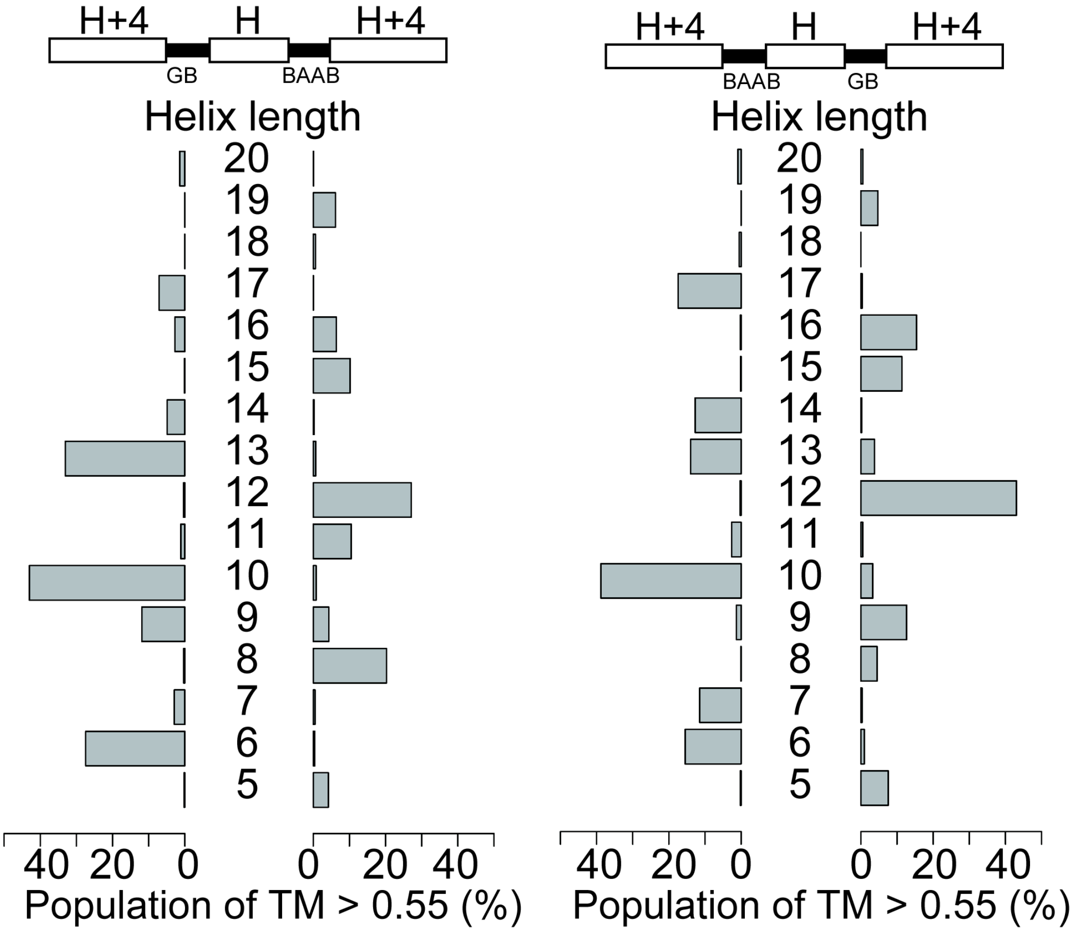


Figure S2-13. The blueprints and result of backbone-building simulation with blueprints containing two different loop types. (left) the GB-BAAB blueprint (Right) the BAAB-GB blueprints. As the GB and BAAB hairpins are both left-handed type hairpins, the helix-helix crossing angle cancels out so that the mixture of the GB and BAAB-loops can yield compact three-helix bundles in similar manners as GB-GB or BAAB-BAAB blueprints.

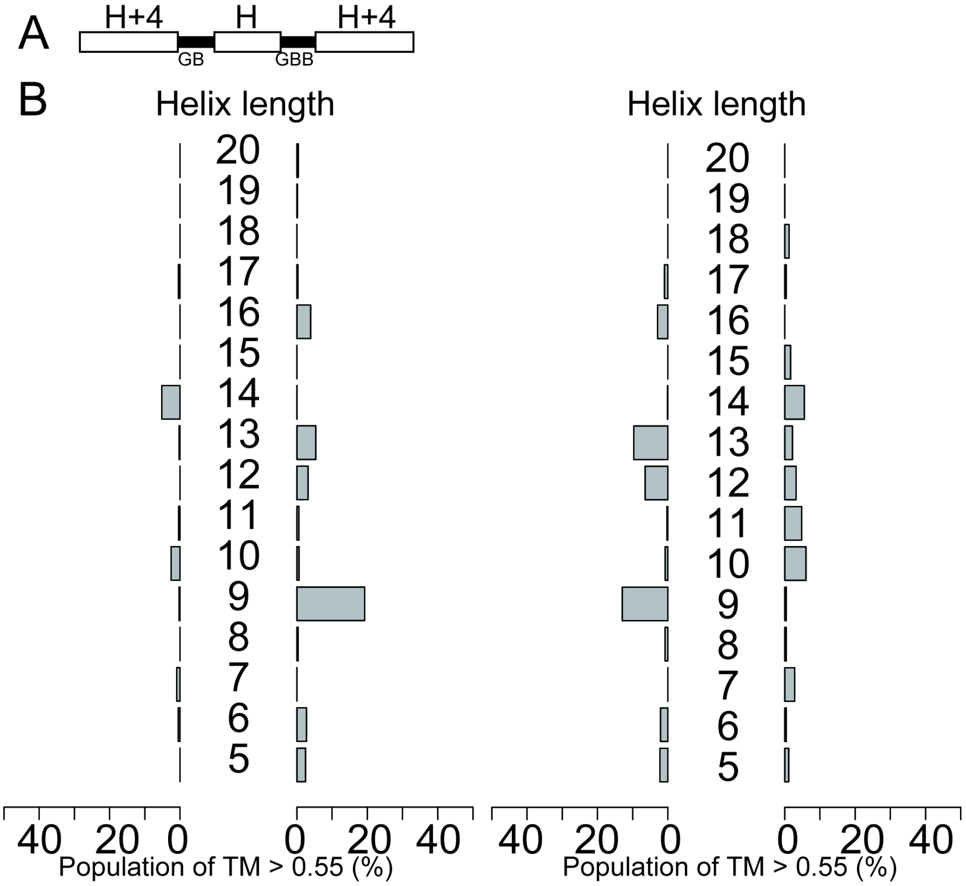


Figure S2-14. Backbone-building simulations with blueprints containing two different loop types. (A) The GB-GBB blueprint (B) Results of GB-GBB blueprint simulation referenced by CW decoy (left) GB-GBB blueprint simulation referenced by CCW decoy(right). Almost no patterns can be observed compared to the consistent blueprint such as GB-GB blueprints because combination of GB and GBB cannot cancel out the helix-helix packing angles.

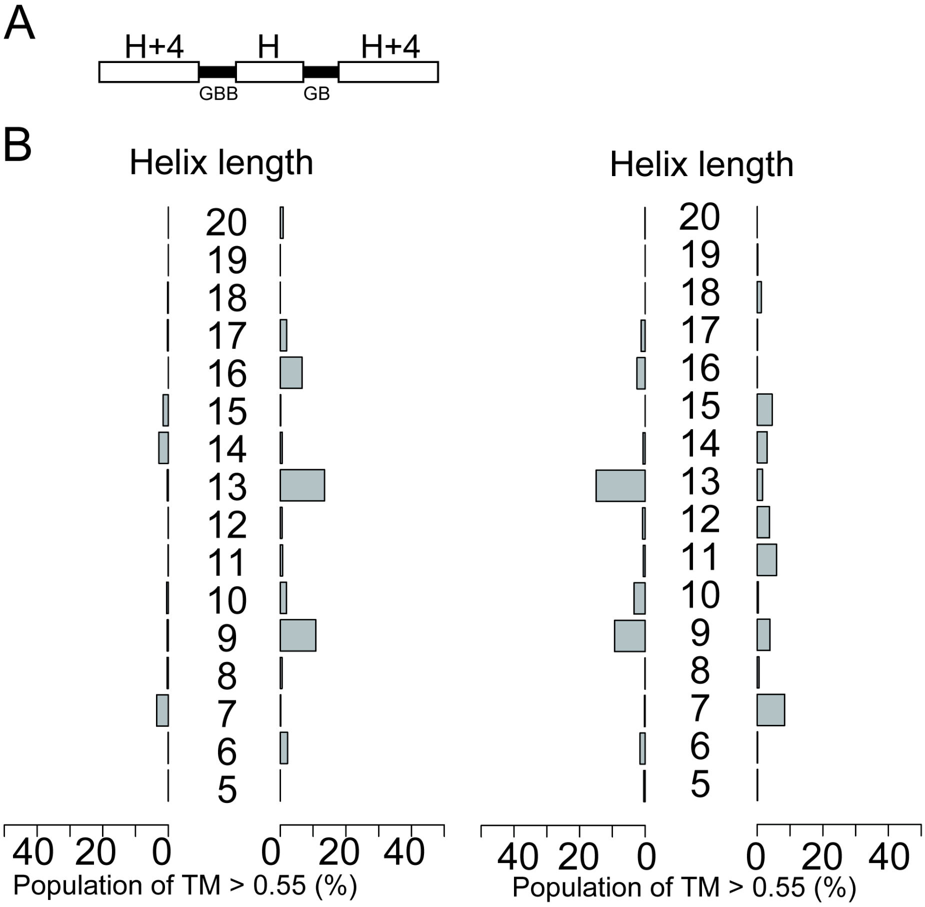


Figure S2-15. Backbone-building simulations with blueprints containing two different loop types. (A) The GBB-GB blueprint (B) Results of GBB-GB blueprint simulations referenced by CW decoy (left) GBB-GB blueprint simulations referenced by CCW decoy (right). Almost no patterns can be observed compared to the consistent blueprint such as GB-BAAB blueprint because combination of GB and GBB cannot cancel out the helix-helix packing angles

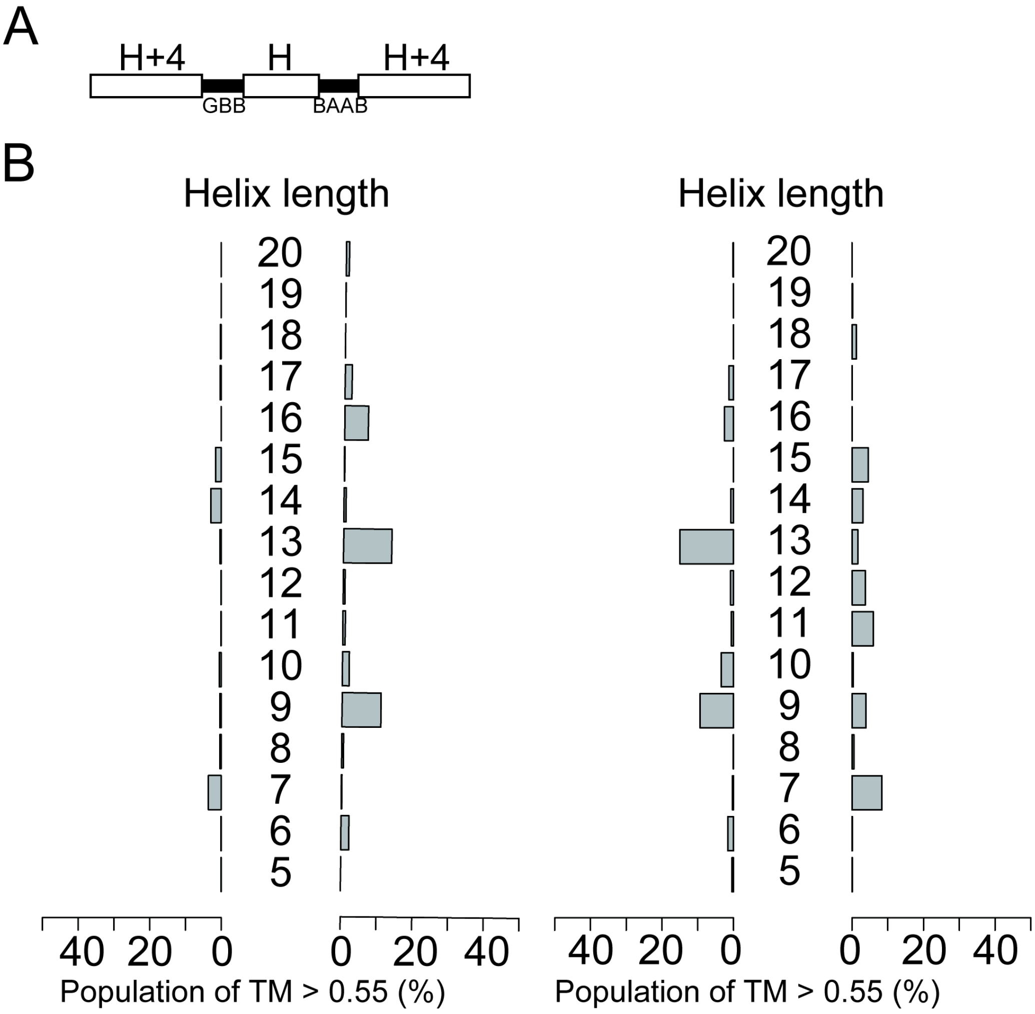


Figure S2-16. Backbone-building simulations with blueprints containing two different loop types. (A) The GBB-BAAB blueprint (B) Results of GBB-BAAB blueprint simulations referenced by CW decoy (left) GB-GBB blueprint simulation referenced by CCW decoy (right). Almost no patterns can be observed compared to the consistent blueprint such as GB-GB blueprints because combination of BAAB and GBB cannot cancel out the helix-helix packing angles.

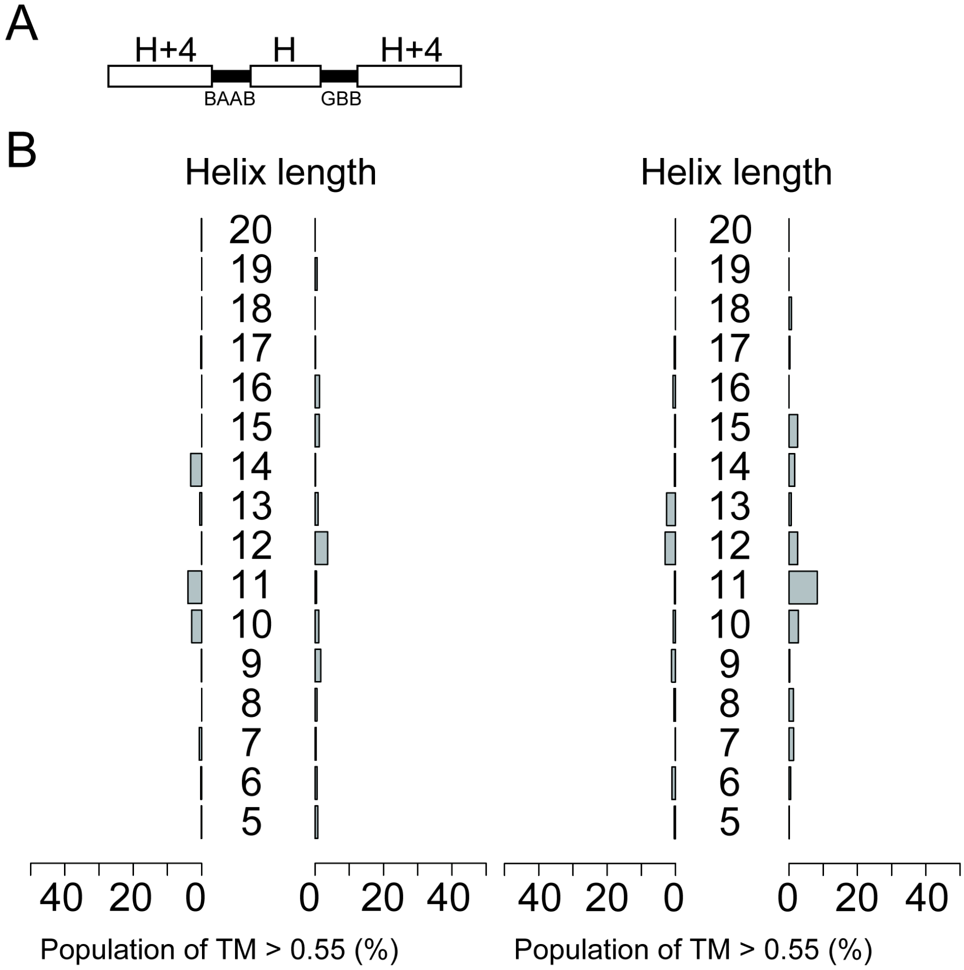


Figure S2-17. Backbone-building simulations with blueprints containing two different loop types. (A) The BAAB-GBB blueprint (B) Results of BAAB-GBB blueprint simulation referenced by CW decoy (left) BAAB-GBB blueprint simulation referenced by CCW decoy (right). Almost no patterns can be observed compared to the consistent blueprint such as GB-GB blueprints because combination of BAAB and GBB cannot cancel out the helix-helix packing angles.

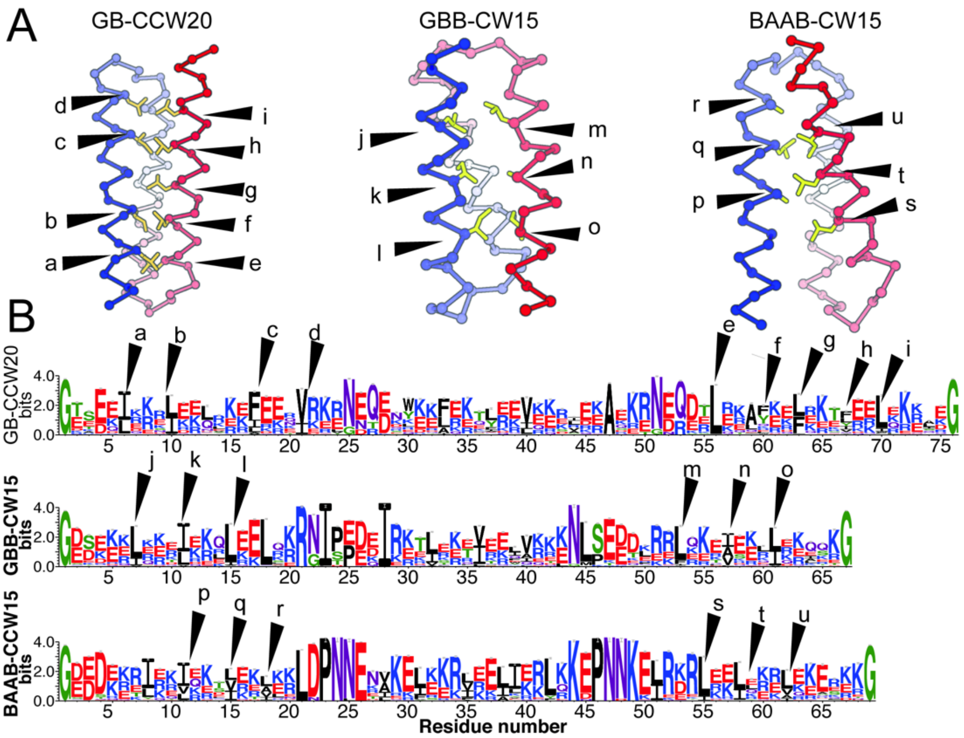


Figure S2-18. The side-chain packing in the interface between first and third α-helices in the structures of designed three-helix bundles, GB-CCW20, GBB-CW15, and BAAB-CCW15. (A) The structures of representative three-helix bundles designed. The chains are colored in blue-white-red gradient from N-term to C-term, where the first/third α-helix is colored approximately in blue/red. The side-chain atoms are represented as yellow sticks, and Cα atoms are represented as spheres. (B) The sequence profiles of all of the designed sequences aligned. The alphabets indicate which residue in the structures corresponds to the site in the profiles. Hydrophobic residues appear in every three or four residues and form tight packing between first and third α-helices.

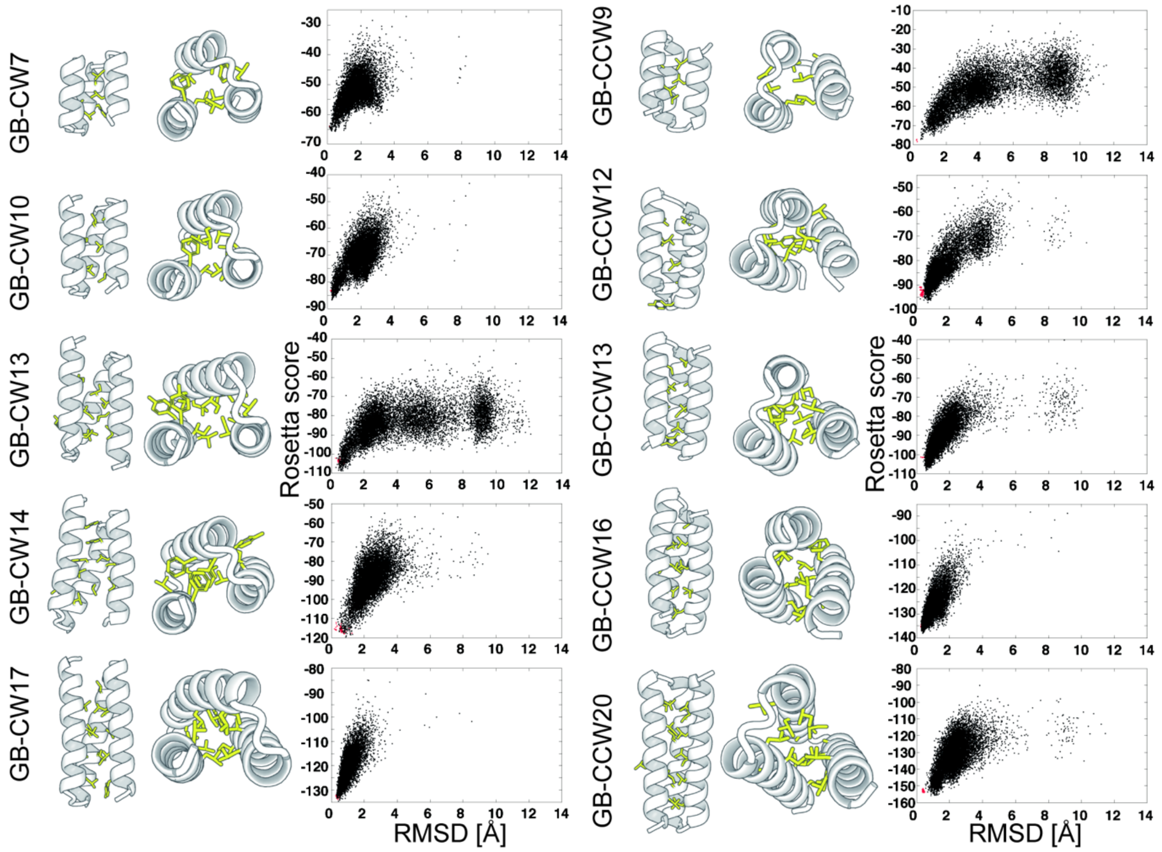


Figure S2-19. Structures and folding-funnels of GB-bundles. (Left) The side-view of the designed structures with α-helix shown as cartoon and hydrophobic side-chains represented as sticks. (Center) The top-view of the designed structures. (Right) The result of folding simulations. The vertical axis represents the Rosetta score, and the horizontal axis represents the RMSD from the target structures. The black dots correspond to the final snapshots of the fragment-assembly folding simulations starting from extended conformations, and red dots correspond to the final snapshots of relax-simulation starting from native conformations.

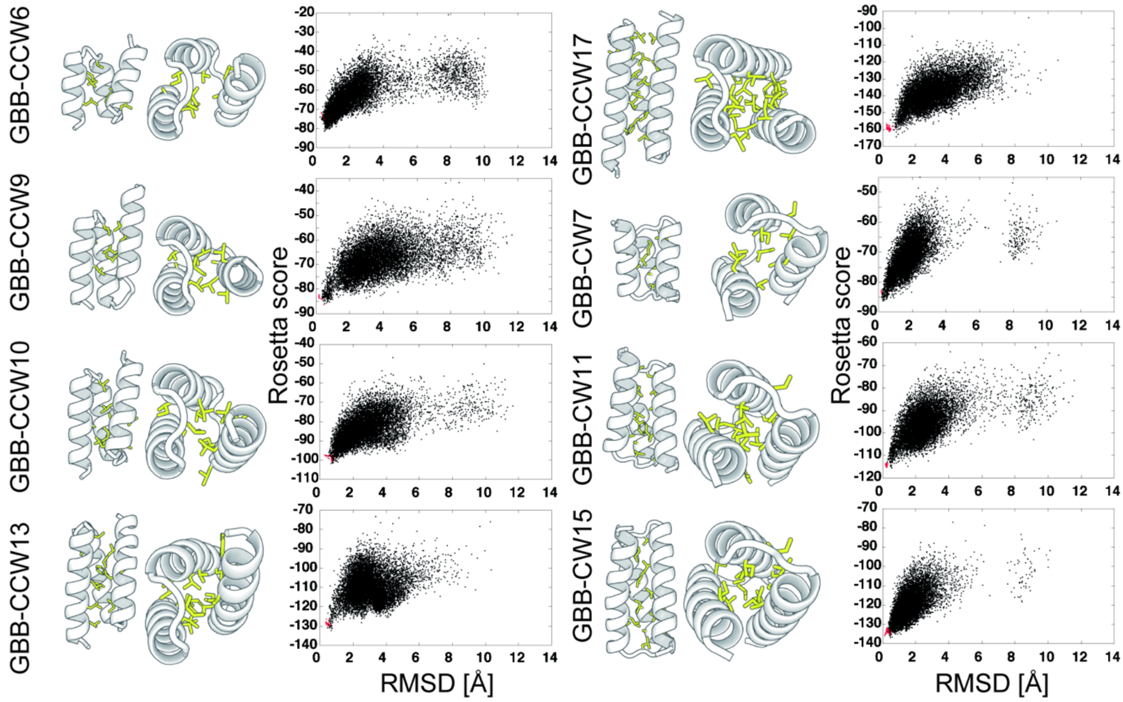


Figure S2-20. Structures and folding-funnels of GBB-bundles. (Left) The side-view of the designed structures with α-helix shown as cartoon and hydrophobic side-chains represented as sticks. (Center) The top-view of the designed structures. (Right) The result of folding simulations. The vertical axis represents the Rosetta score, and the horizontal axis represents the RMSD from the target structures. The black dots correspond to the final snapshots of the fragment-assembly folding simulations starting from extended conformations, and red dots correspond to the final snapshots of relax-simulation starting from native conformations.

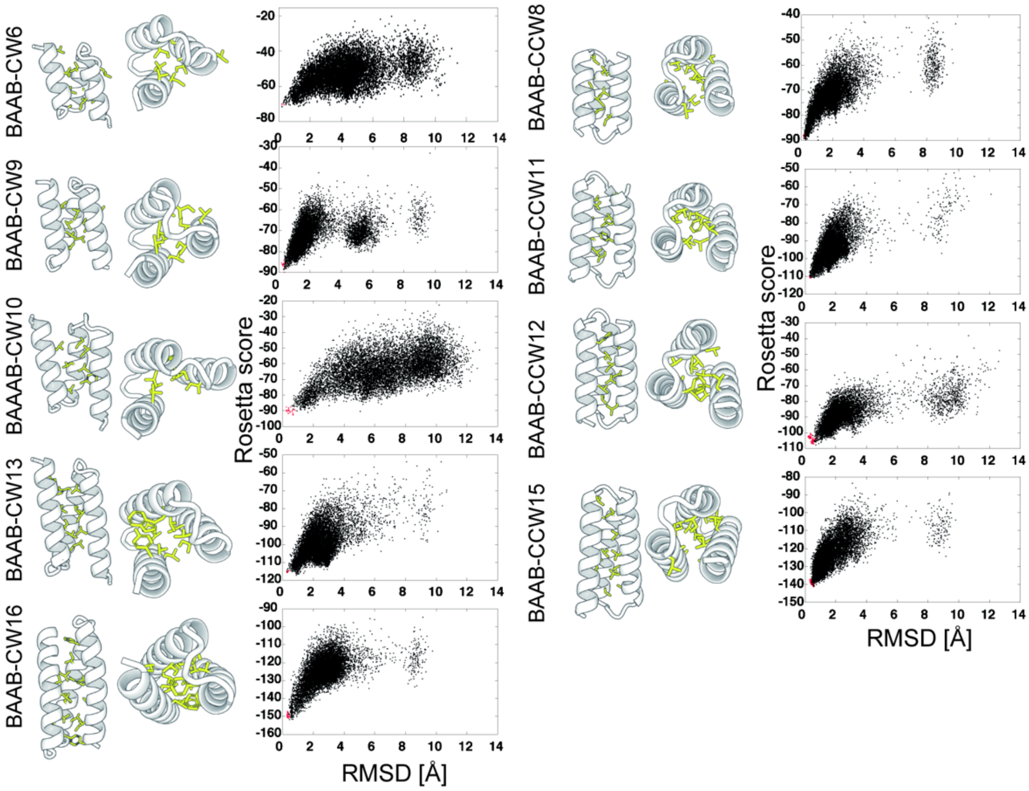


Figure S2-21. Structures and folding-funnels of BAAB-bundles. (Left) The side-view of the designed structures with α-helix shown as cartoon and hydrophobic side-chains represented as sticks. (Center) The top-view of the designed structures. (Right) The result of folding simulations. The vertical axis represents the Rosetta score, and the horizontal axis represents the RMSD from the target structures. The black dots correspond to the final snapshots of the fragment-assembly folding simulations starting from extended conformations, and red dots correspond to the final snapshots of relax-simulation starting from native conformations.

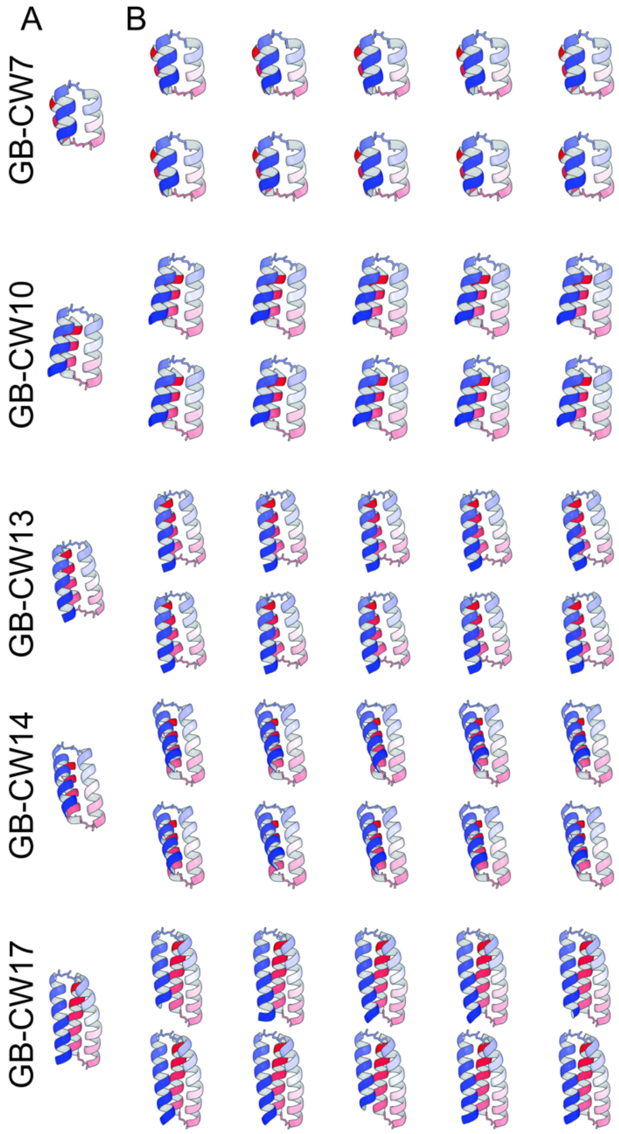


Figure S2-22. Comparison of the design structure composed of the GB-hairpins (A) and the 10 lowest score predictions by sequence-dependent folding simulations (B). Loops are shown as sticks in order to show the detailed conformations. The predictions precisely recovered the local conformations in most of the lowest score models. Overall topologies of predicted models agree with design models.

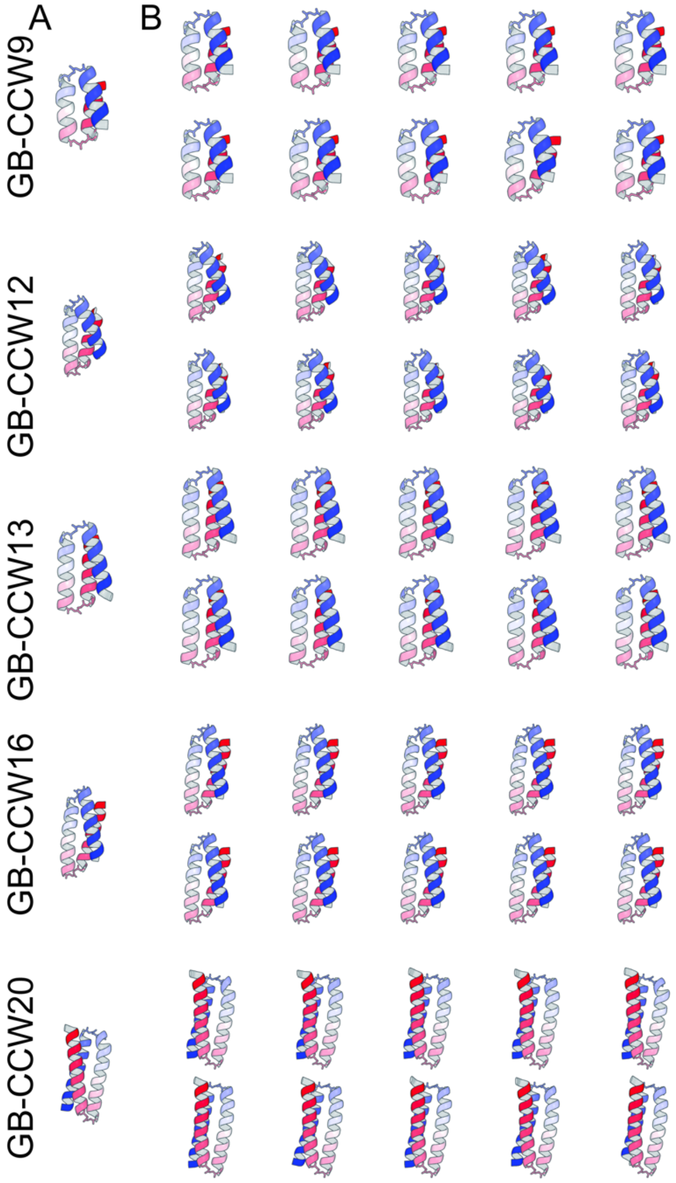


Figure S2-23. Comparison of the design structure composed of the GB-hairpins (A) and the 10 lowest score predictions by sequence-dependent folding simulations (B). Loops are shown as sticks in order to show the detailed conformations. The predictions precisely recovered the local conformations in most of the lowest score models. Overall topologies of predicted models agree with design models.

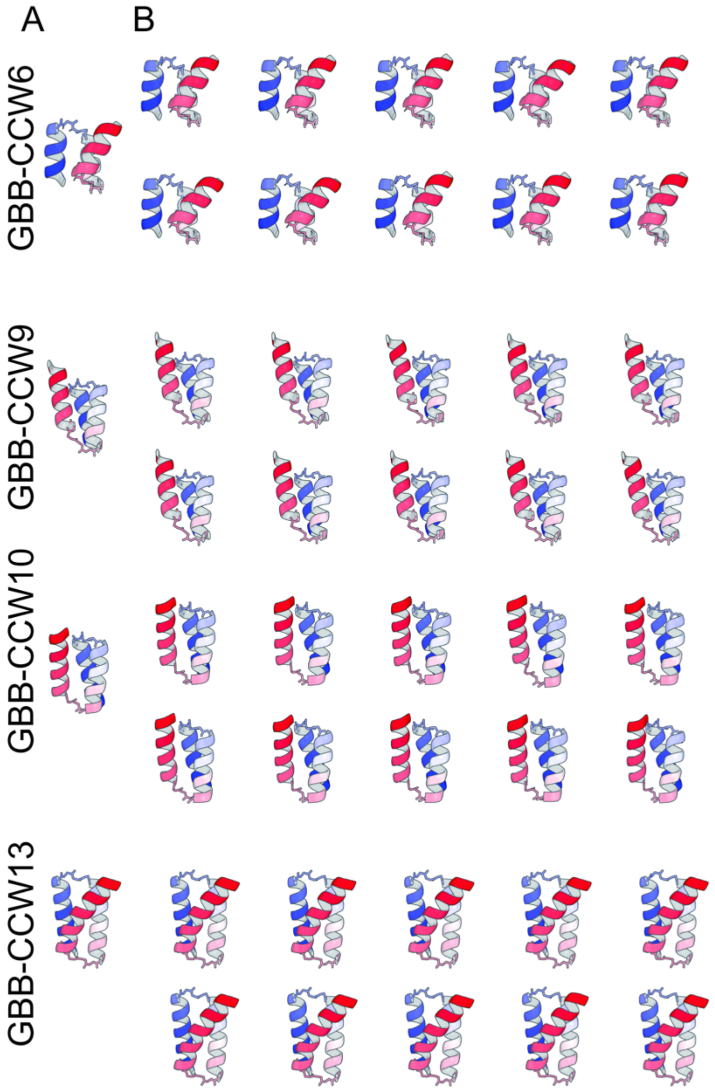


Figure S2-24. Comparison of the design structure composed of the GBB-hairpins (A) and the 10 lowest score predictions by sequence-dependent folding simulations (B). Loops are shown as sticks in order to show the detailed conformations. The predictions precisely recovered the local conformations in most of the lowest score models. Overall topologies of predicted models agree with design models.

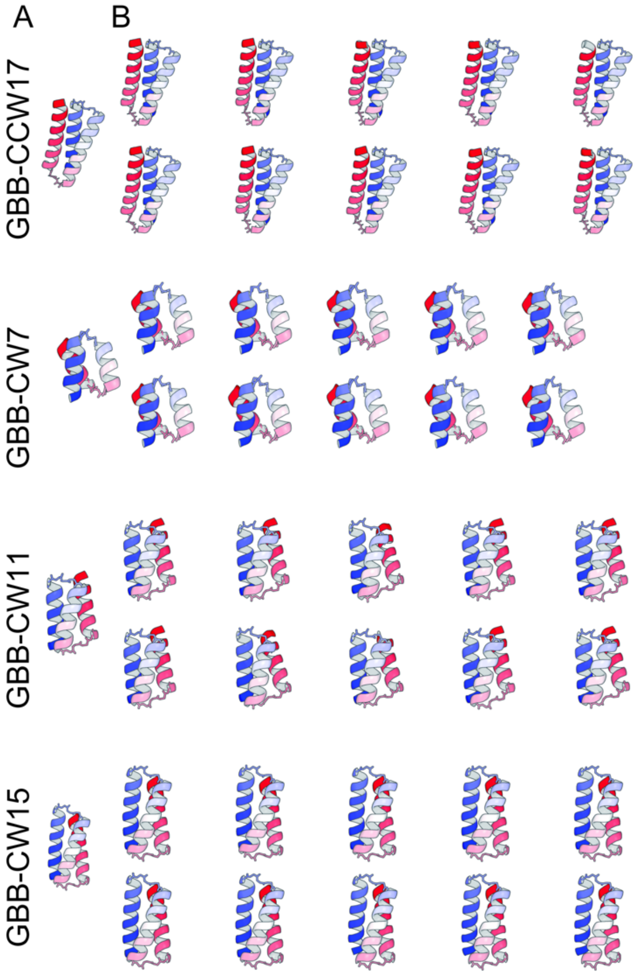


Figure S2-25. Comparison of the design structure composed of the GBB-hairpins (A) and the 10 lowest score predictions by sequence-dependent folding simulations (B). Loops are shown as sticks in order to show the detailed conformations. The predictions precisely recovered the local conformations in most of the lowest score models. Overall topologies of predicted models agree with design models.

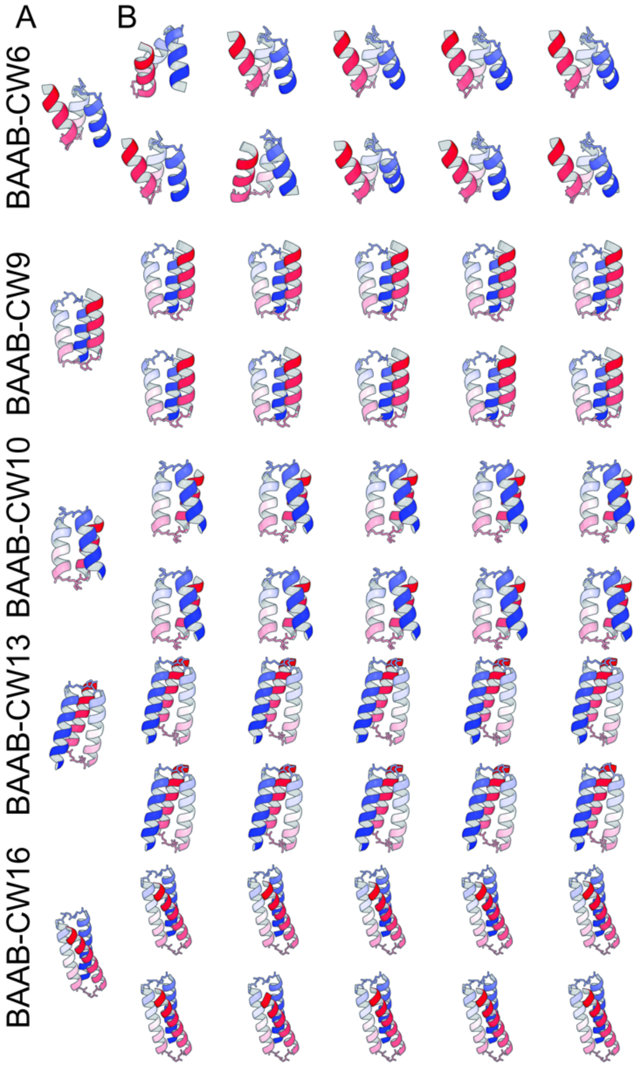


Figure S2-26. Comparison of the design structure composed of the BAAB-hairpins (A) and the 10 lowest score predictions by sequence-dependent folding simulations (B). Loops are shown as sticks in order to show the detailed conformations. The predictions precisely recovered the local conformations in most of the lowest score models. Overall topologies of predicted models agree with design models.

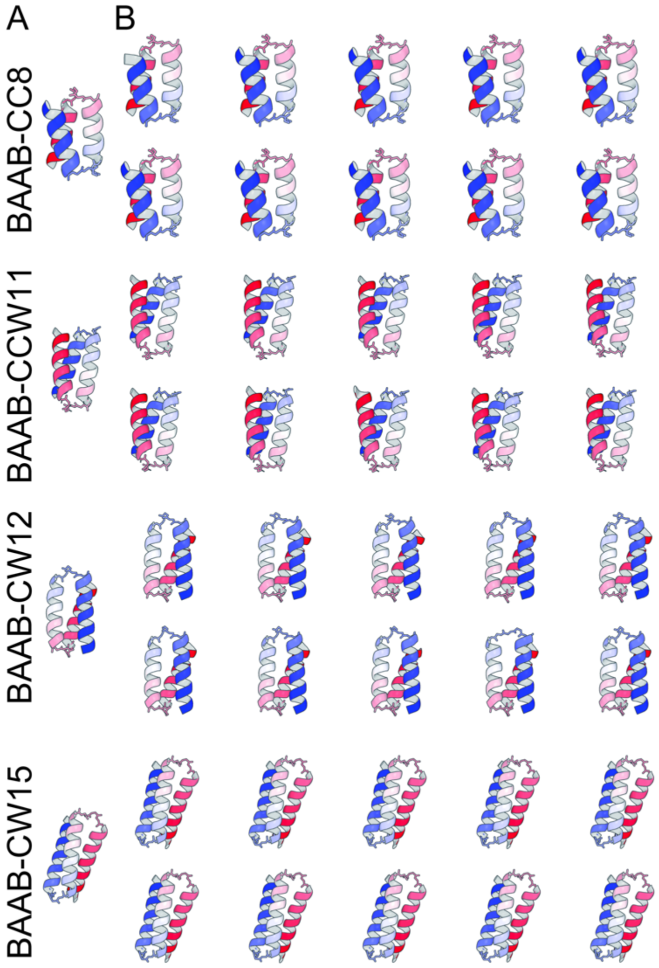


Figure S2-27. Comparison of the design structure composed of the BAAB-hairpins (A) and the 10 lowest score predictions by sequence-dependent folding simulations (B). Loops are shown as sticks in order to show the detailed conformations. The predictions precisely recovered the local conformations in most of the lowest score models. Overall topologies of predicted models agree with design models.

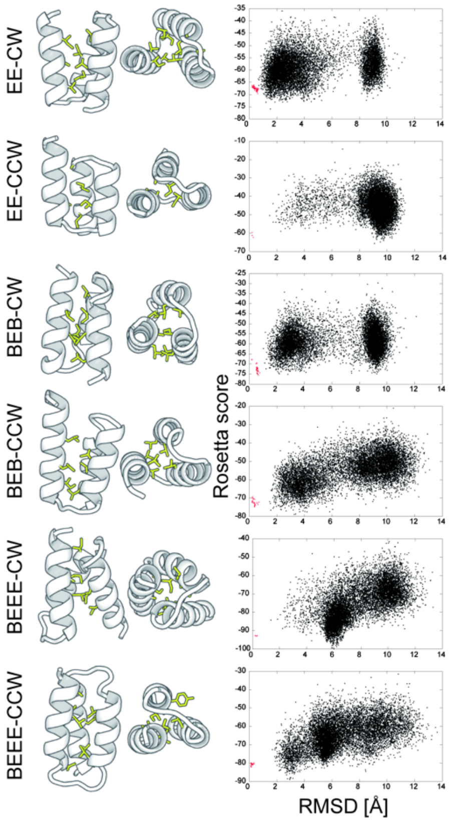


Figure S2-28: Structures and folding-funnels of best-effort-design three-helix bundles composed of atypical hairpin structures. (Left) The side-view of the designed structures with α-helix shown as cartoon and hydrophobic side-chains represented as sticks. (Center) The top-view of the designed structures. (Right) The result of folding simulations. The vertical axis represents the Rosetta score, and the horizontal axis represents the RMSD from the target structures. The black dots correspond to the final snapshots of the fragment-assembly folding simulations starting from extended conformations, and red dots correspond to the final snapshots of relax-simulation starting from native conformations. These folding-funnels are flat-bottomed showing that the folding simulations were unable to reach the near-native structures.

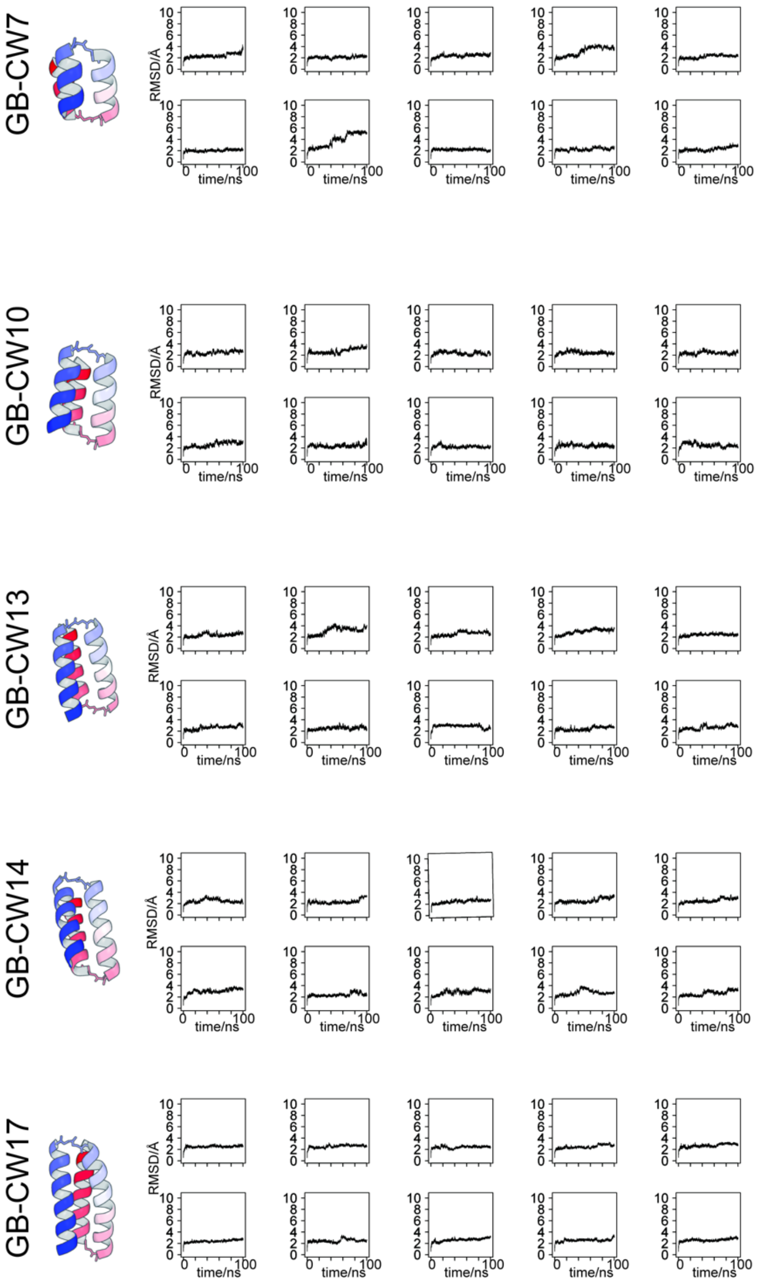


Figure S2-29: Time series of RMSD in 10 trajectories of 100 ns molecular dynamics simulations. (Left) Structure of designed proteins (Right) Time series of Cα RMSD referenced by the designed protein structure. Horizontal axis represents times in ns, and veritcal axis represents RMSD in Å.

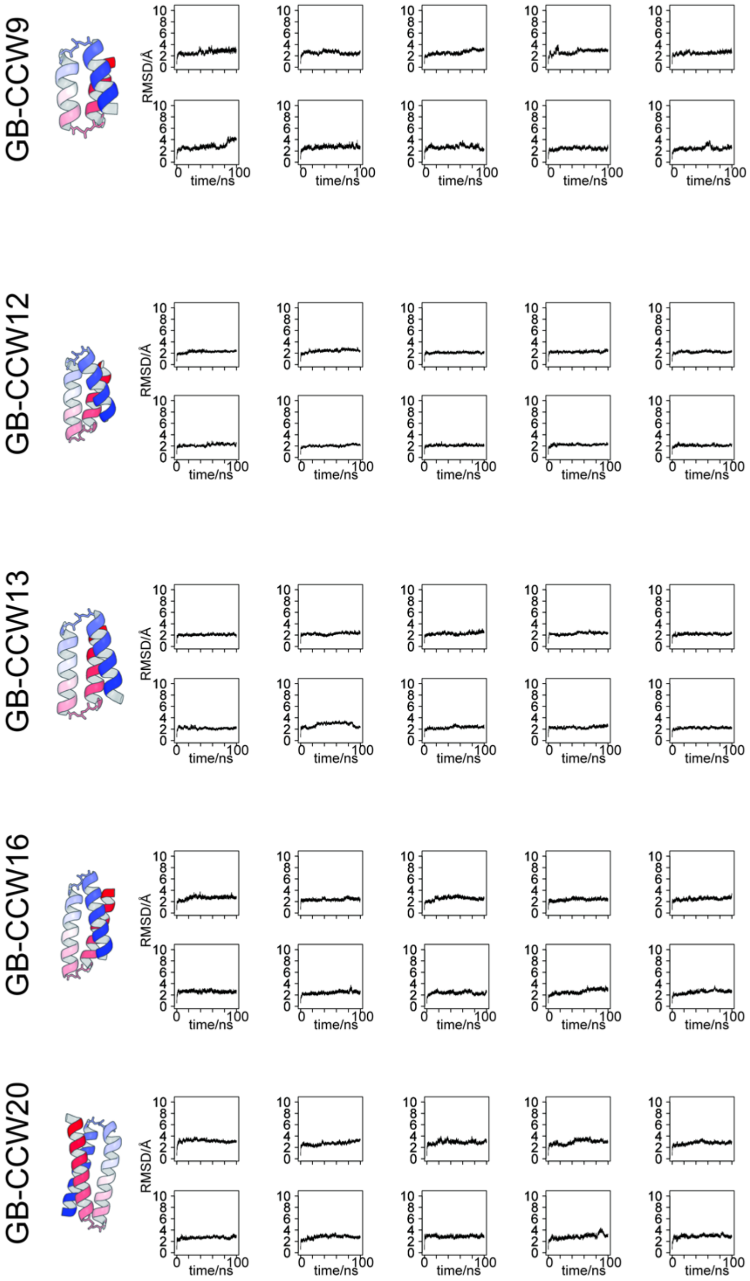


Figure S2-30: Time series of RMSD in 10 trajectories of 100 ns molecular dynamics simulations. (Left) Structure of designed proteins (Right) Time series of Cα RMSD referenced by the designed protein structure. Horizontal axis represents times in ns, and veritcal axis represents RMSD in Å.

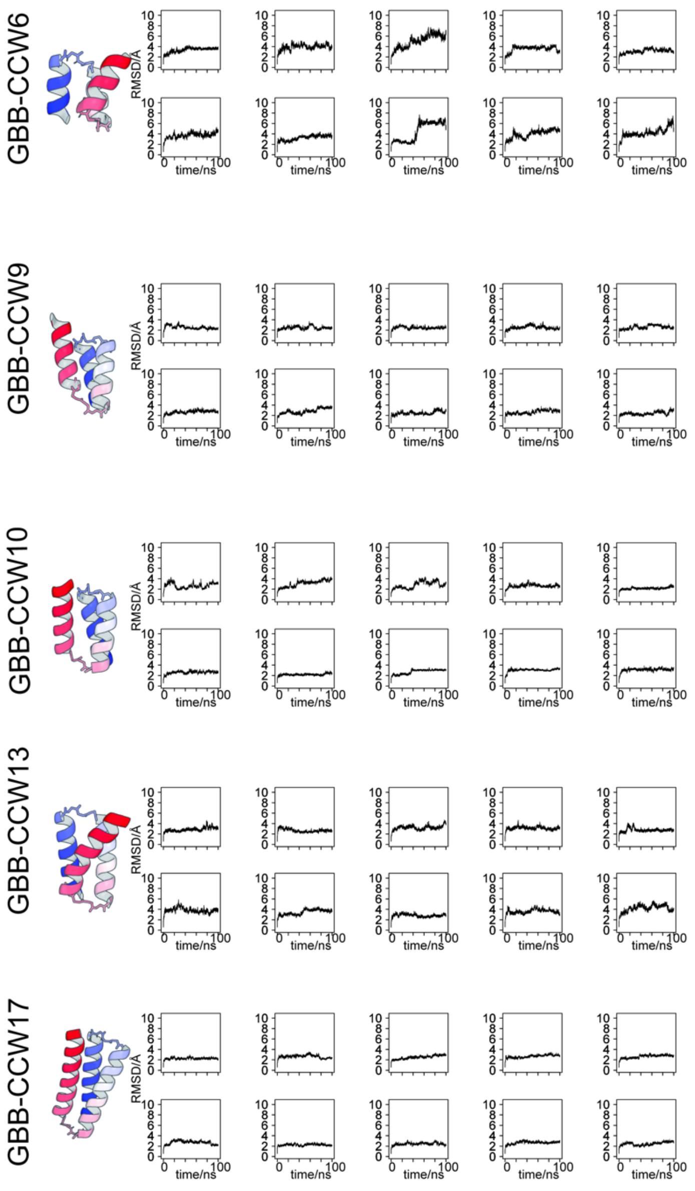


Figure S2-31 Time series of RMSD in 10 trajectories of 100 ns molecular dynamics simulations. (Left) Structure of designed proteins (Right) Time series of Cα RMSD referenced by the designed protein structure. Horizontal axis represents times in ns, and veritcal axis represents RMSD in Å.

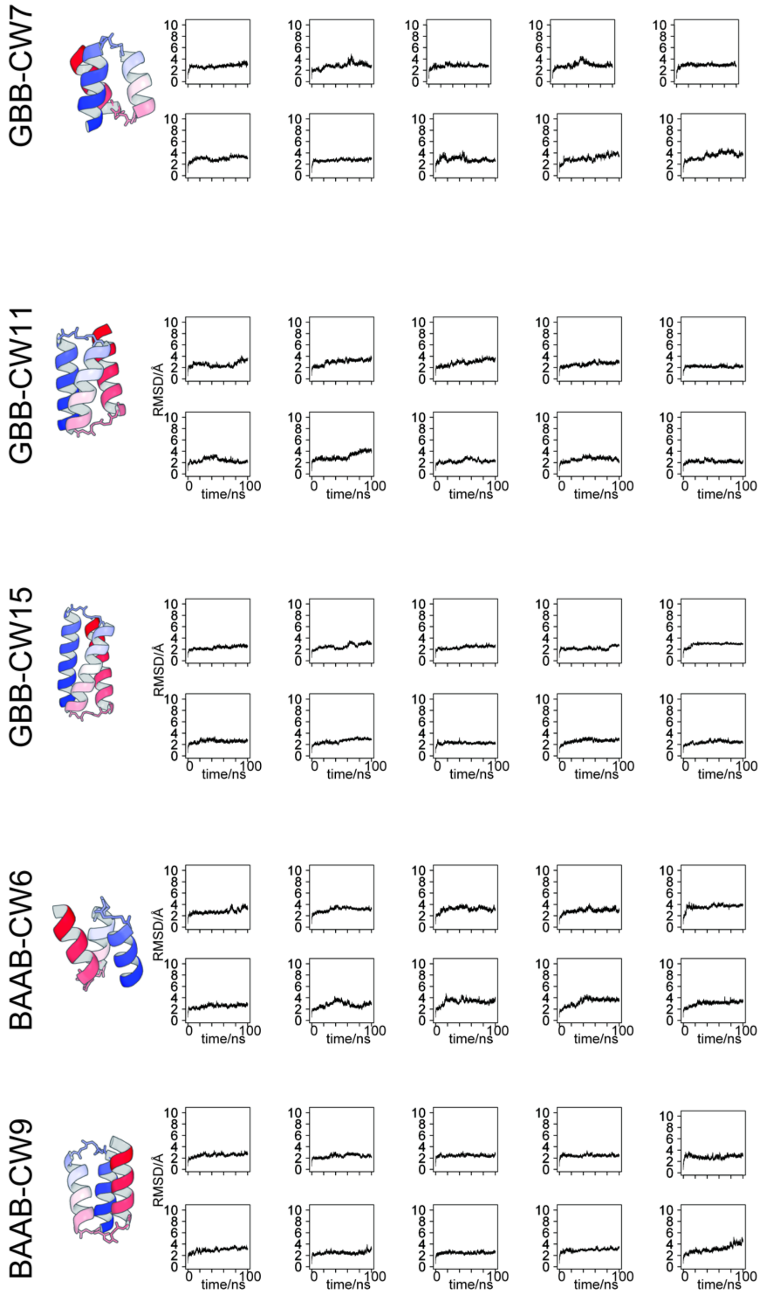


Figure S2-32: Time series of RMSD in 10 trajectories of 100 ns molecular dynamics simulations. (Left) Structure of designed proteins (Right) Time series of Cα RMSD referenced by the designed protein structure. Horizontal axis represents times in ns, and veritcal axis represents RMSD in Å.

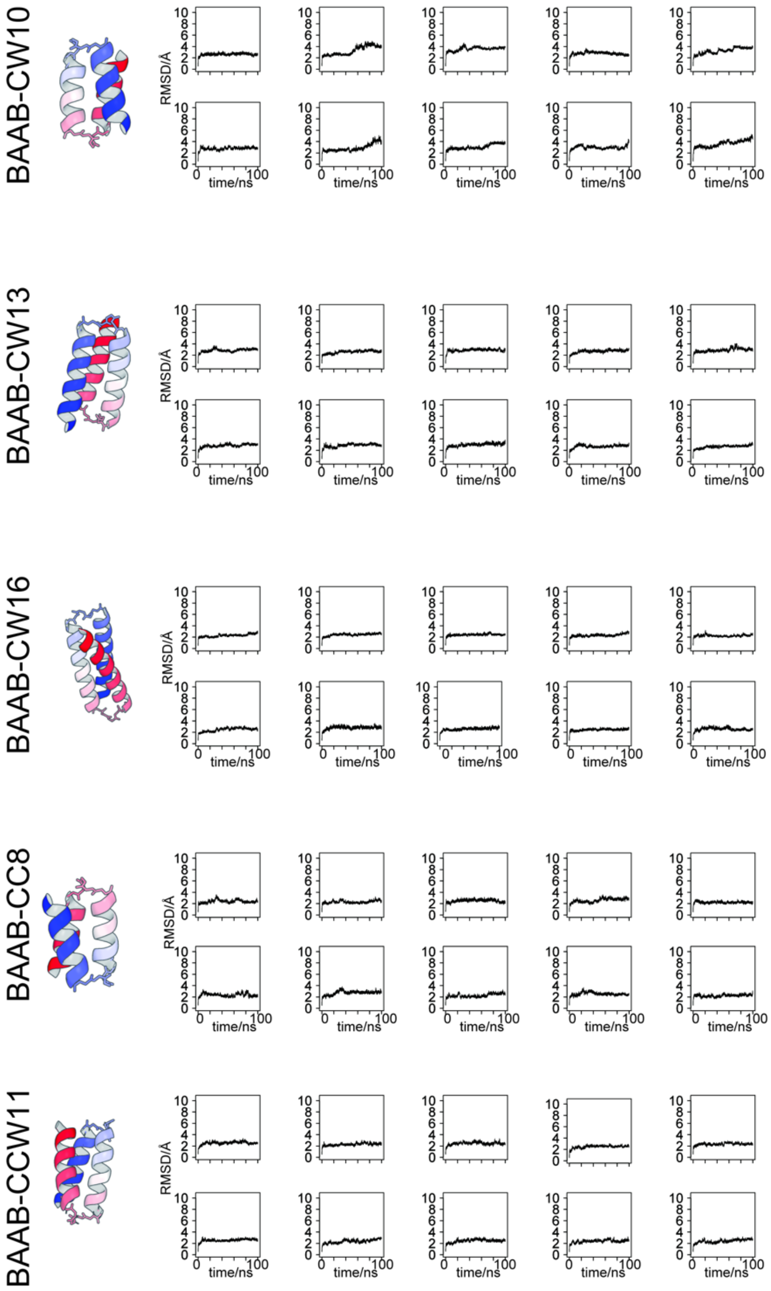


Figure S2-33: Time series of RMSD in 10 trajectories of 100 ns molecular dynamics simulations. (Left) Structure of designed proteins (Right) Time series of Cα RMSD referenced by the designed protein structure. Horizontal axis represents times in ns, and veritcal axis represents RMSD in Å.

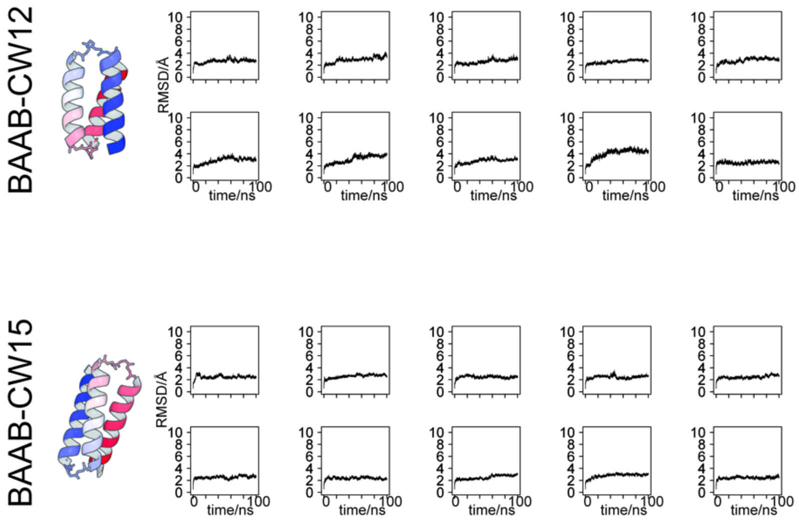


Figure S2-34: Time series of RMSD in 10 trajectories of 100 ns molecular dynamics simulations. (Left) Structure of designed proteins (Right) Time series of Cα RMSD referenced by the designed protein structure. Horizontal axis represents times in ns, and veritcal axis represents RMSD in Å.

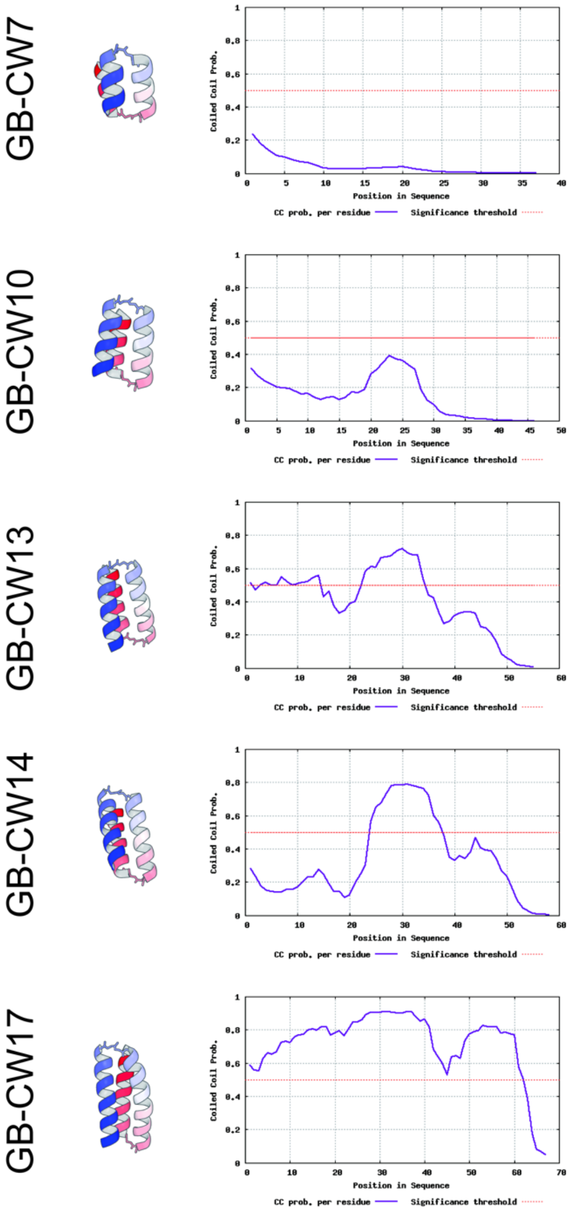


Figure S2-35: Probabilities that the designed sequence have coiled-coil arrangement of α-helix predicted by DeepCoil. (Left) Designed protein structures (Right) Predicted probability that the design sequences have coiled-coil arrangement of α-helix predicted by DeepCoil. Horizontal axis represents residue number, and the vertical axis represents the probability that the sequence is recognized as coiled-coil by DeepCoil.

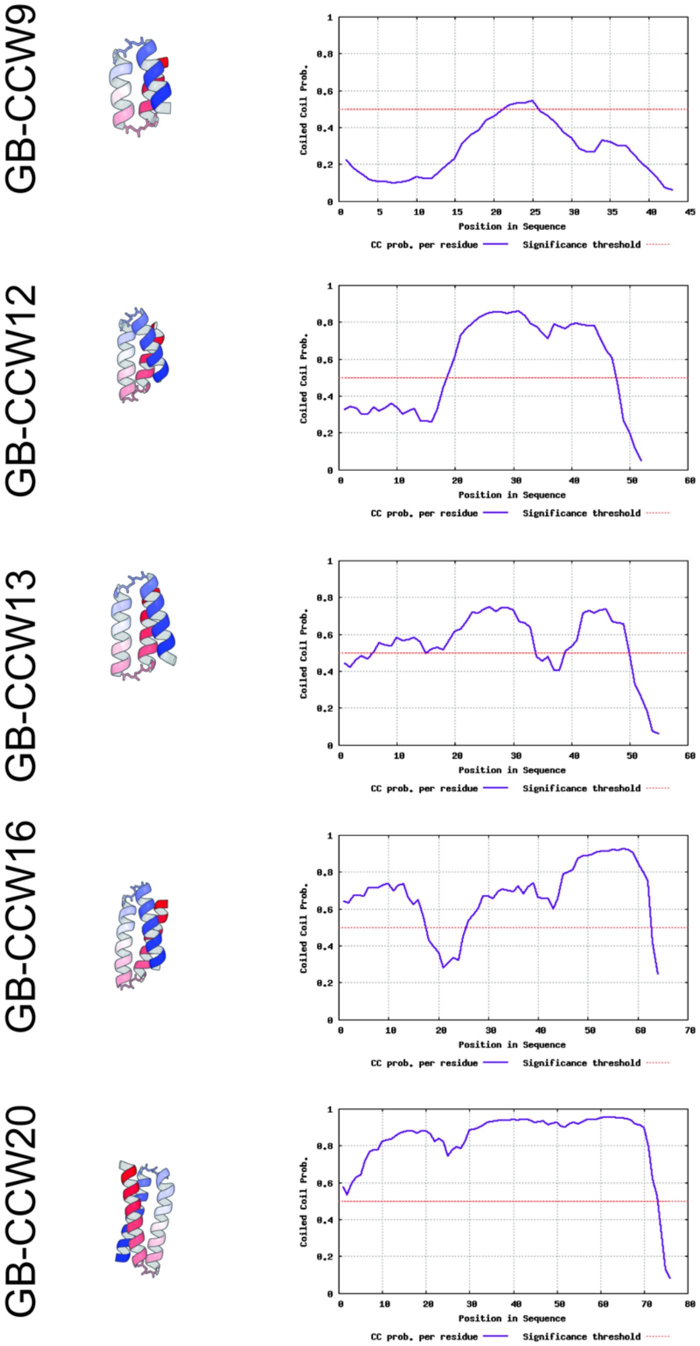


Figure S2-36: Probabilities that the designed sequence have coiled-coil arrangement of α-helix predicted by DeepCoil. (Left) Designed protein structures (Right) Predicted probability that the design sequences have coiled-coil arrangement of α-helix predicted by DeepCoil. Horizontal axis represents residue number, and the vertical axis represents the probability that the sequence is recognized as coiled-coil by DeepCoil.

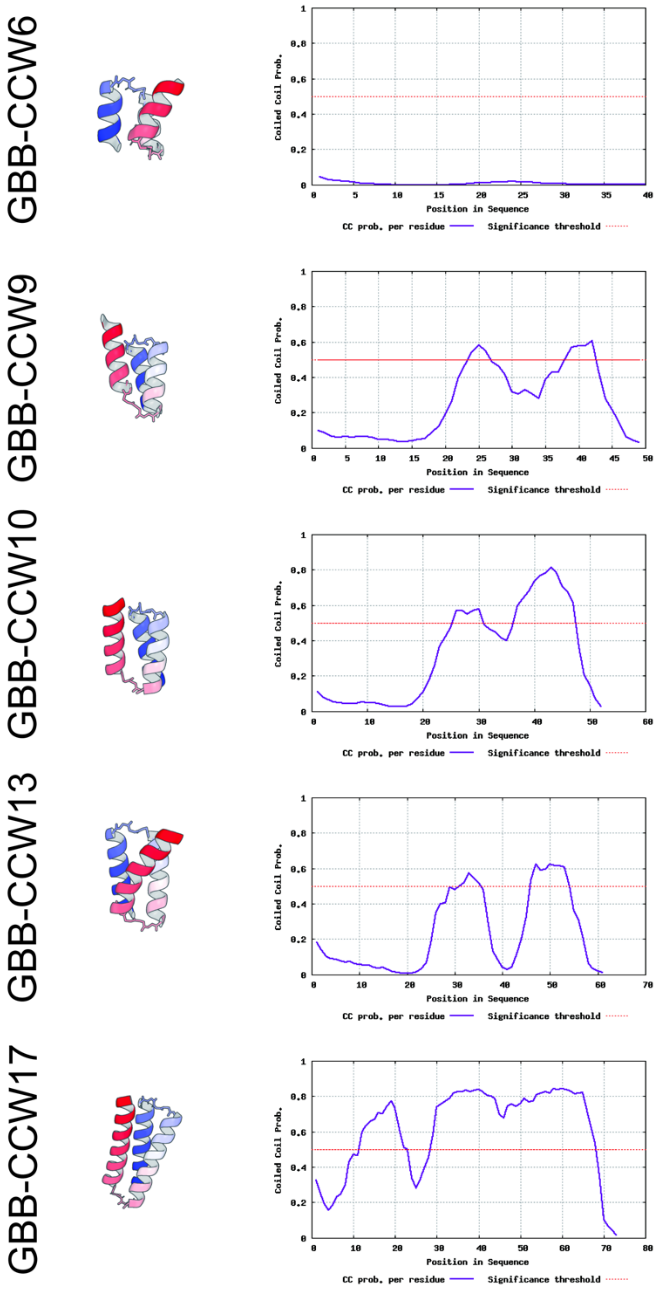


Figure S2-37: Probabilities that the designed sequence have coiled-coil arrangement of α-helix predicted by DeepCoil. (Left) Designed protein structures (Right) Predicted probability that the design sequences have coiled-coil arrangement of α-helix predicted by DeepCoil. Horizontal axis represents residue number, and the vertical axis represents the probability that the sequence is recognized as coiled-coil by DeepCoil.

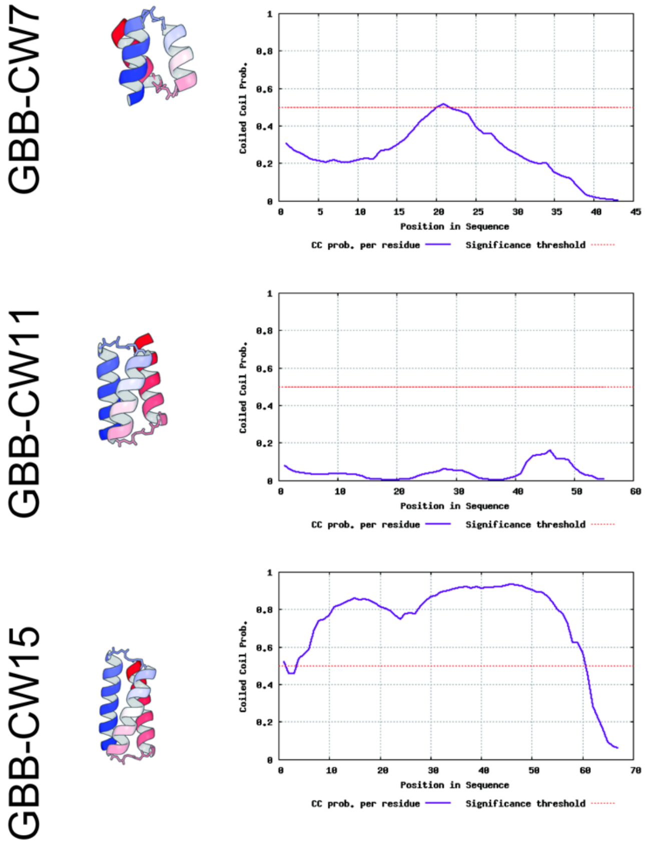


Figure S2-38: Probabilities that the designed sequence have coiled-coil arrangement of α-helix predicted by DeepCoil. (Left) Designed protein structures (Right) Predicted probability that the design sequences have coiled-coil arrangement of α-helix predicted by DeepCoil. Horizontal axis represents residue number, and the vertical axis represents the probability that the sequence is recognized as coiled-coil by DeepCoil.

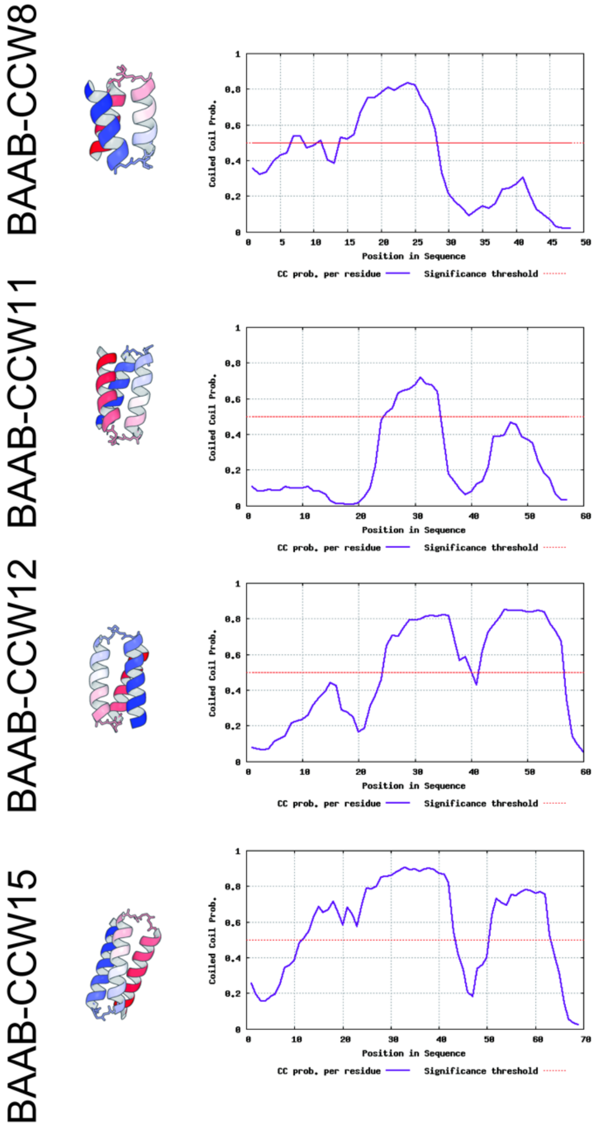


Figure S2-39: Probabilities that the designed sequence have coiled-coil arrangement of α-helix predicted by DeepCoil. (Left) Designed protein structures (Right) Predicted probability that the design sequences have coiled-coil arrangement of α-helix predicted by DeepCoil. Horizontal axis represents residue number, and the vertical axis represents the probability that the sequence is recognized as coiled-coil by DeepCoil.

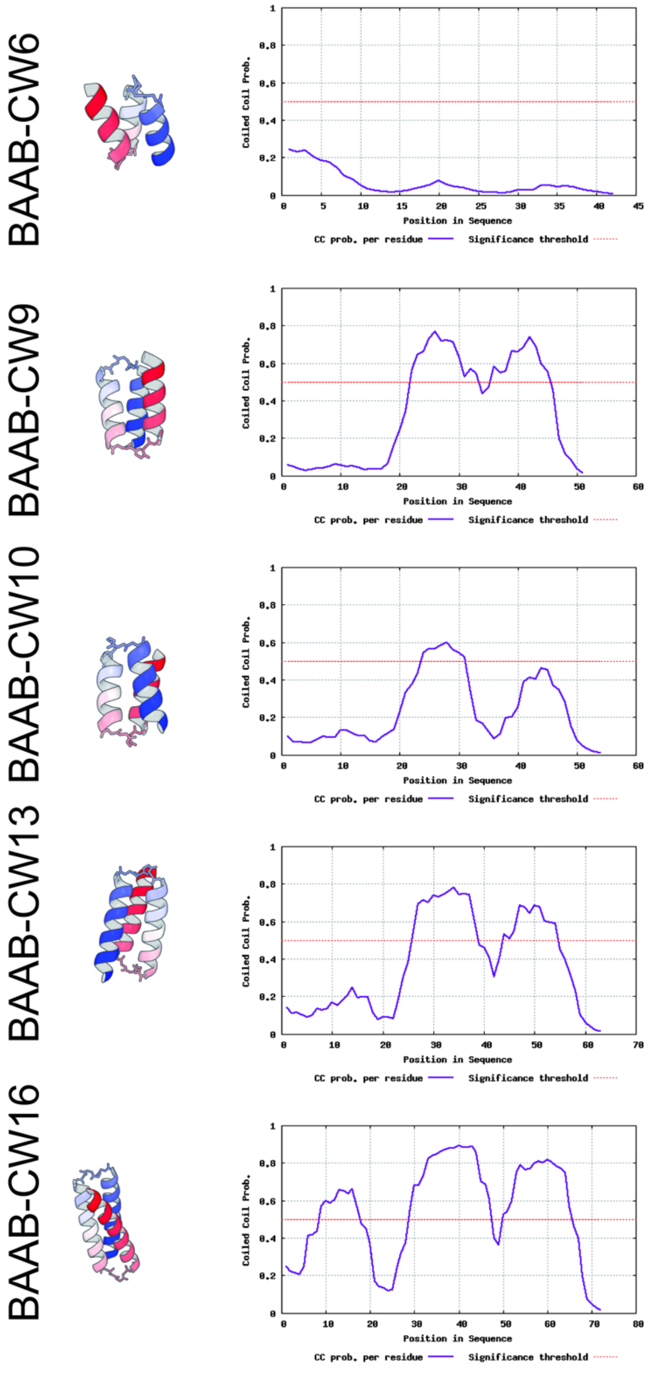


Figure S2-40: Probabilities that the designed sequence have coiled-coil arrangement of α-helix predicted by DeepCoil. (Left) Designed protein structures (Right) Predicted probability that the design sequences have coiled-coil arrangement of α-helix predicted by DeepCoil. Horizontal axis represents residue number, and the vertical axis represents the probability that the sequence is recognized as coiled-coil by DeepCoil.

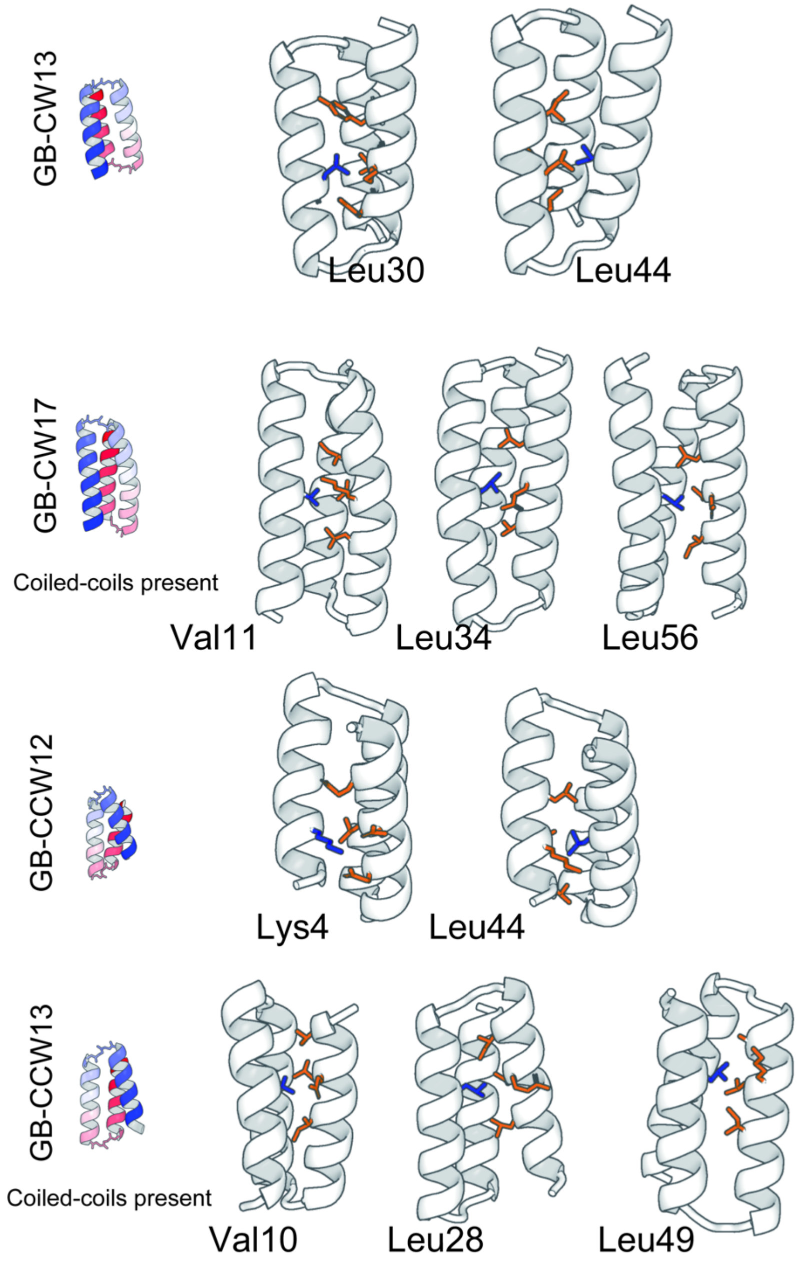


Figure S2-41: Knobs-into-holes in the author's design structures detected by SOCKET. (Left) Design structures (Right) Structure explaining knobs-into-holes sub-structures. The knob residues are colored in blue, and the hole residues are colored in orange. The residue name of knob residue is indicated below the structures.

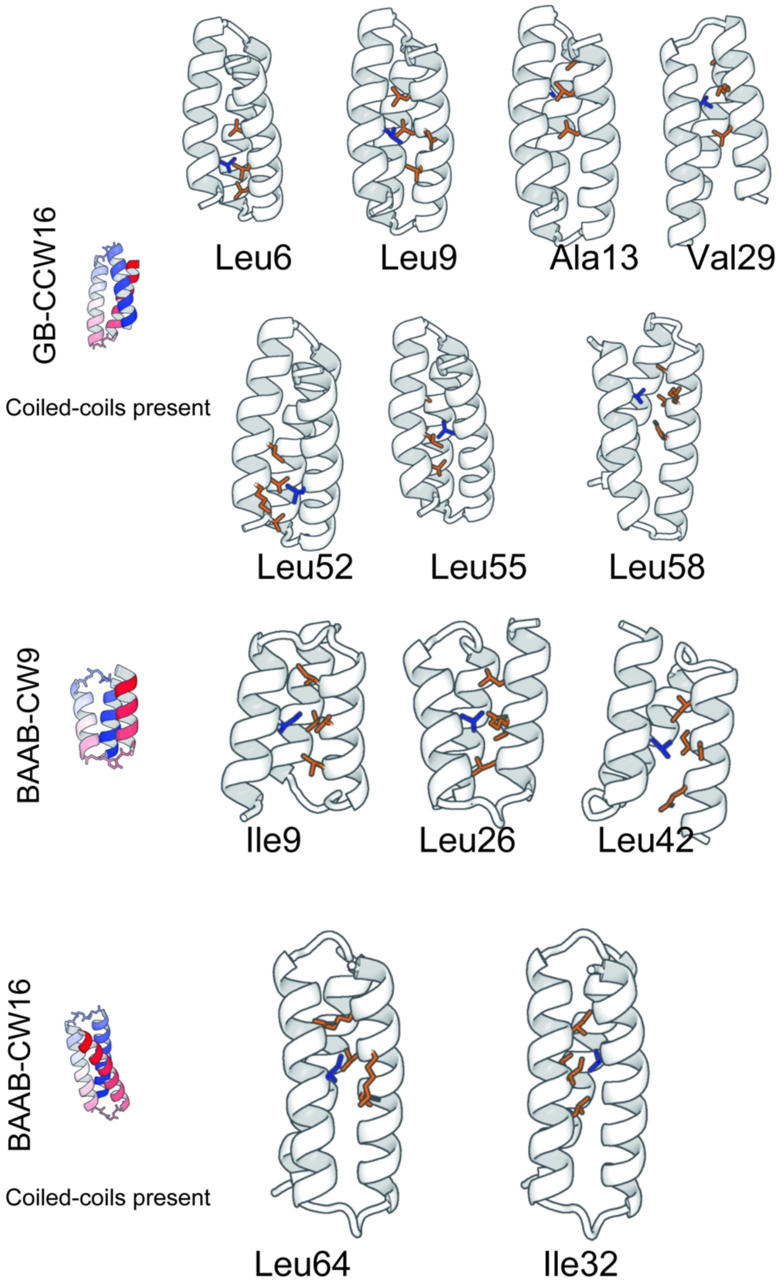


Figure S2-42: Knobs-into-holes in the author's design structures detected by SOCKET. (Left) Design structures (Right) Structure explaining knobs-into-holes sub-structures. The knob residues are colored in blue, and the hole residues are colored in orange. The residue name of knob residue is indicated below the structures.

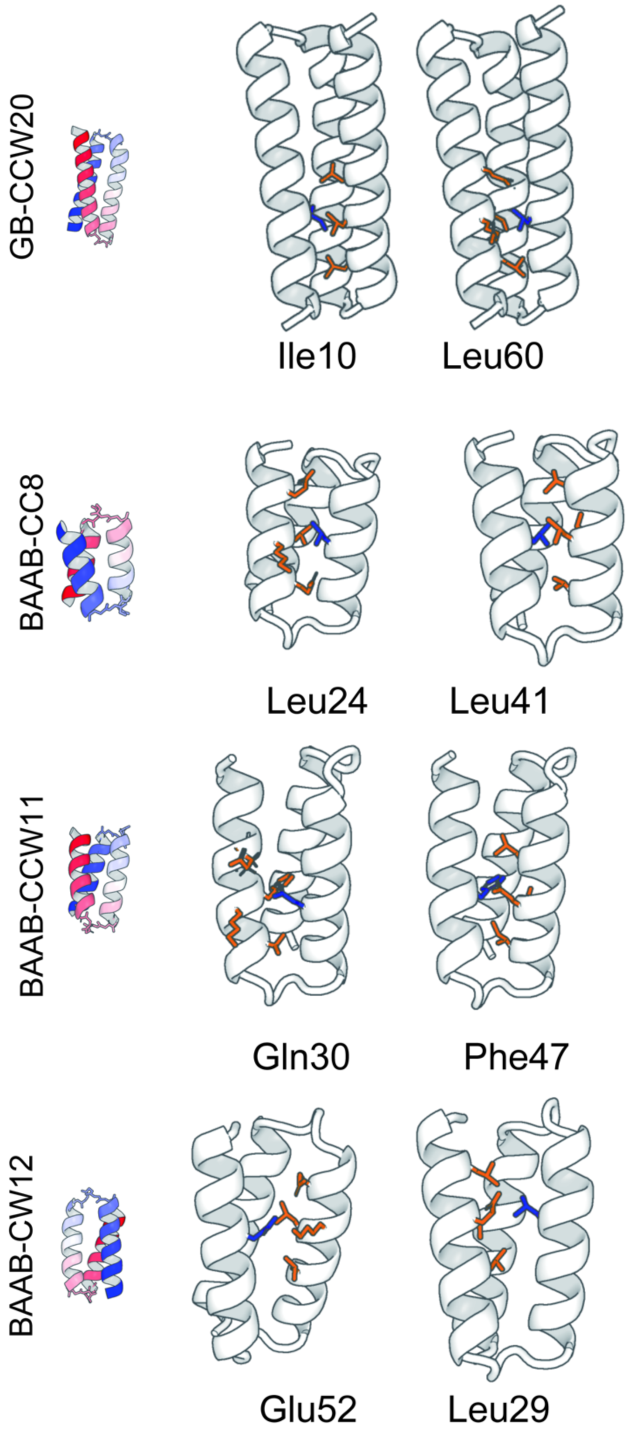


Figure S2-43: Knobs-into-holes in the author's design structures detected by SOCKET. (Left) Design structures (Right) Structure explaining knobs-into-holes sub-structures. The knob residues are colored in blue, and the hole residues are colored in orange. The residue name of knob residue is indicated below the structures.

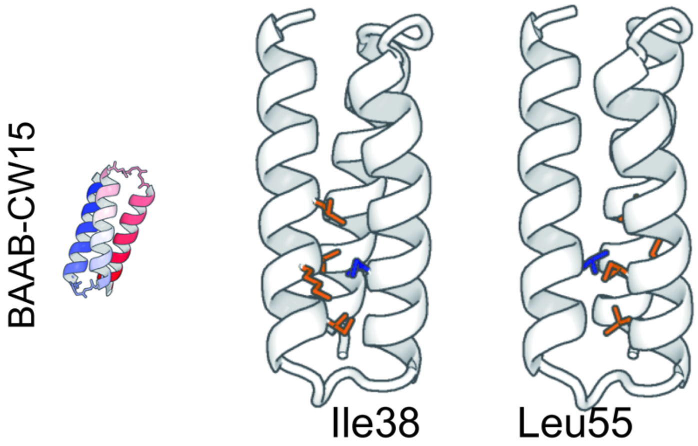


Figure S2-44: Knobs-into-holes in the author's design structures detected by SOCKET. (Left) Design structures (Right) Structure explaining knobs-into-holes sub-structures. The knob residues are colored in blue, and the hole residues are colored in orange. The residue name of knob residue is indicated below the structures.

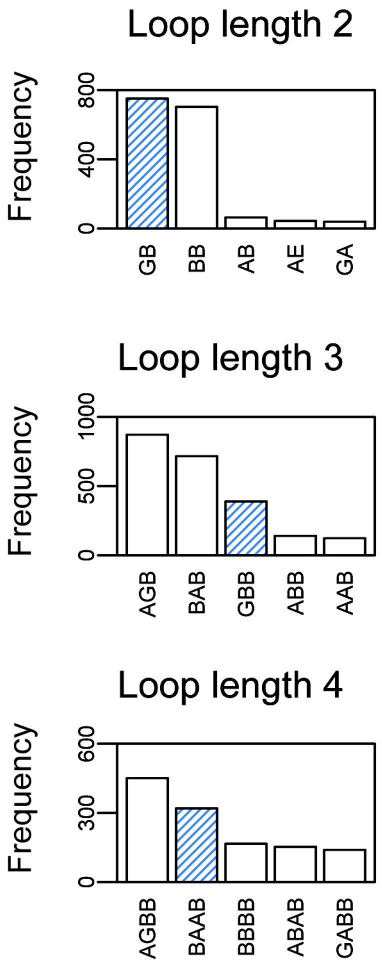


Figure S2-45. The population statistics of the hairpins in the ABEGO representation including loops starting/ending with A.

Structure	Loop 1	Loop 2
GB-CW7	GB (GB)	GB (GB)
GB-CW10	GB (GB)	GB (GB)
GB-CW13	GB (GB)	GBA (GB)
GB-CW14	GB (GB)	GB (GB)
GB-CW17	GB (GB)	GB (GB)
GB-CCW9	GB (GB)	GB (GB)
GB-CCW12	GB (GB)	GB (GB)
GB-CCW13	GB (GB)	AGB (GB)
GB-CCW16	GB (GB)	GB (GB)
GB-CCW20	GB (GB)	GB (GB)
GBB-CW7	GBB (GBB)	GBB (GBB)
GBB-CW11	GBB (GBB)	GBB (GBB)
GBB-CW15	GBB (GBB)	GBB (GBB)
GBB-CCW6	GBB (GBB)	GBB (GBB)
GBB-CCW9	GBB (GBB)	GBB (GBB)
GBB-CCW10	GBB (GBB)	GBB (GBB)
GBB-CCW13	GBB (GBB)	GBB (GBB)
GBB-CCW17	AGBB (GBB)	GBB (GBB)
BAAB-CW6	B (BAAB)	BAAB (BAAB)
BAAB-CW9	BAAB (BAAB)	BAAB (BAAB)
BAAB-CW10	BAAB (BAAB)	BAAB (BAAB)
BAAB-CW13	BAAB (BAAB)	BAAB (BAAB)
BAAB-CW16	BAAB (BAAB)	BAAB (BAAB)
BAAB-CCW8	BAAB (BAAB)	BAAB (BAAB)
BAAB-CCW11	BAAB (BAAB)	BAAB (BAAB)
BAAB-CCW12	BAAB (BAAB)	BAAB (BAAB)
BAAB-CCW15	BAAB (BAAB)	BAAB (BAAB)

Table S2-1: Comparison of backbone torsion angles between the lowest energy prediction structure from the sequence-dependent fragment assembly simulations for each design protein. The torsion angles observed in the lower energy structures are represented in the ABEGO representations, and their target torsion angles are represented in ABEGO in the parentheses. DSSP was used for assignment of secondary structure boundaries. Most of the lowest energy structures recovered the same local conformations as design models. See also Figure S22—S26 for comparison of their overall topologies.

Structure	Mean score per residue (a.u.)
GB-CCW09	-2.38
GB-CCW12	-2.37
GB-CCW13	-2.41
GB-CCW16	-2.66
GB-CCW20	-2.61
GB-CW07	-2.28
GB-CW10	-2.04
GB-CW13	-2.36
GB-CW14	-2.66
GB-CW17	-2.65
GBB-CCW06	-2.25
GBB-CCW09	-2.27
GBB-CCW10	-2.33
GBB-CCW13	-2.49
GBB-CCW17	-2.57
GBB-CW07	-2.39
GBB-CW11	-2.42
GBB-CW15	-2.52
BAAB-CCW08	-2.38
BAAB-CCW11	-2.53
BAAB-CCW12	-2.38
BAAB-CCW15	-2.51
BAAB-CW06	-2.17
BAAB-CW09	-2.29
BAAB-CW10	-2.31
BAAB-CW13	-2.46
BAAB-CW16	-2.57
EE-CW	-2.00
EE-CCW	-2.11
BEB-CW	-2.08
BEB-CCW	-2.09
BEEE-CW	-2.15
BEEE-CCW	-2.02

Table S2-2: Mean Rosetta score par residue for best-effort design models. The structures were relaxed using Relax protocol of Rosetta and 1000 near-native structures were generated and their scores were calculated. Score Talaris2014 was used in the simulations and scoring. Scores were averaged over the 1000 structures and the number of residues of respective structures.

Chapter 03 :
Limitations of the ABEGO representation: ambiguity between αα-corner and αα-hairpin

Abstract

ABEGO is a coarse-grained representation for polypeptide backbone dihedral angles. The Ramachandran map is divided into four segments denoted as A, B, E, and G to represent the local conformation of polypeptide chains in the character strings. Although the ABEGO representation is widely used in backbone building simulation for de novo protein design, it cannot capture minor differences in backbone dihedral angles, which potentially leads to ambiguity between two structurally distinct fragments. Here, the author shows a nontrivial example of two local motifs that could not be distinguished by their ABEGO representations. The author found that two well-known local motifs αα-hairpins and αα-corners are both represented as α-GBB-α and thus indistinguishable in the ABEGO representation, although they show distinct arrangements of the flanking α-helices. The author also found that α-GBB-α motifs caused a loss of efficiency in the ABEGO-based fragment-assembly simulations for de novo protein backbone design. Nevertheless, the author was able to design amino-acid sequences that were predicted to fold into the target topologies that contained these α-GBB-α motifs, which suggests such topologies that are difficult to build by ABEGO-based simulations are designable once the backbone structures are modeled by some means. The finding that certain local motifs bottleneck the ABEGO-based fragment-assembly simulations for construction of backbone structures suggests that finer representations of backbone torsion angles are required for efficiently generating diverse topologies containing such indistinguishable local motifs.

Introduction

Proteins are polymers, and using idealized bond lengths and bond angles, the conformation of a polypeptide chain can be represented as a series of backbone dihedral angle triplets (φ, ψ, and ω) [1]. Provided that all peptide bonds have *trans* conformations with ω of approximately 180°, the two-dimensional plot of φ and ψ called the Ramachandran map can have sufficient information to specify the residue-wise conformations of a polypeptide chain. To construct coarse-grained representations of backbone conformations, the Ramachandran map can be divided into subsections to cluster similar backbone conformations into the same class. A widespread approach is to define a four-state representation dividing the map into four segments and assigning the single letters A, B, E, and G to the regions (Figure 3-1) [2]. This enables the rough backbone structures to be expressed by character strings and is beneficial in structure-informatics analyses. Broadly, the A region corresponds to α-helices and the B region to β-strands. For regions with positive φ, the G region corresponds to the left-handed α-helix and the E region represents the remaining map. With an additional state O corresponding to the *cis*-conformation of the peptide bond, this five-state discrete representation can cover the conformational space of polypeptide chains in a coarse-grained manner. This five-state coarse-grained representation of the polypeptide chain conformation is termed the ABEGO representation, which is the main focus of the current study.

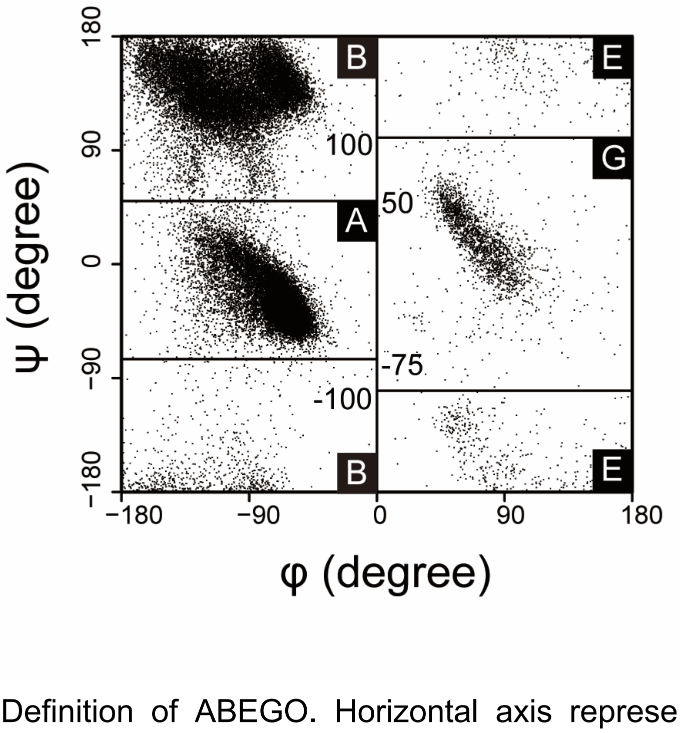


Figure 3-1: Definition of ABEGO. Horizontal axis represents φ and vertical axis represents ψ angle of polypeptide backbone structure. Ramachandran plot is divided into four sections named A, B, E, and G. The values of phi and psi angles for the borderline are indicated on the left or right of the border lines. The state O is not defined in this diagram because it represents cis-peptide.

An important application of the ABEGO representation is the *de novo* design of protein backbone structures [3–15]. In this protocol, designers specify the target topology using ABEGO sequences, select structure fragments that satisfy the desired ABEGO sequences, and perform fragment-assembly simulations to build the atomistic backbone structures with the desired topology. Hereafter, these fragment-assembly simulations guided by ABEGO specification are referred to as ABEGO-based backbone-building simulations. This approach is widely accepted in *de novo* protein design and has been used to construct a variety of topologies ranging from small α-helical bundles to TIM barrels [3–15]. Therefore, this ABEGO-based approach can be taken as a de facto standard approach to generate backbone structures for de novo protein design.

However, ABEGO representation is a coarse-grained representation of backbone dihedral angles that sometimes fail to distinguish two different conformations, which may cause troubles in ABEGO-based backbone building simulations. In this study, the author shows a non-trivial example of two famous local motifs that are indistinguishable by their ABEGO representation and points out that the ambiguity between these two motifs can lead to loss of efficiency in the ABEGO-based backbone building simulations. Clarifying the limitations of the ABEGO representation will motivate further development of more sophisticated representation for backbone conformation and backbone-building methods.

Materials and methods

Analysis of helix–loop–helix fragments

A collaborator of the author composed a set of 29,397 non-redundant domain structures, which were a subset of the Evolutionary Classification Of protein Domains database (version develop238) culled by 40% sequence identity [16]. Next, secondary structures were assigned using the DSSP [17], and helix-loop-helix fragments were extracted. The fragments whose helix have residues less than and equal to nine residues were discarded. The ABEGO representations of backbone torsion were assigned using in-house Python scripts according to the definition shown in Figure 3-1. Next the fragments possessing the GBB loop were extracted. In total, 318 αα-corner and 317 αα-hairpin fragments were obtained, which were illustrated in Figures 3-2, 3-3 and Supplementary Figures S3-1, S3-4, and S3-5. I calculated the all-to-all Cα root mean square deviation (RMSD) within these GBB fragments and performed k-medoid clustering with k = 2. The cluster representatives were extracted and used for reference structure in Supplementary Figure S3-8. Next, the author identified the helix–helix crossing angle (Supplementary Figure S3-1) using the helix orientation vector defined by Krissinel *et al*. [18] and confirmed that the clustering can clearly separate αα-corners and αα-hairpins (Figure 3-2).

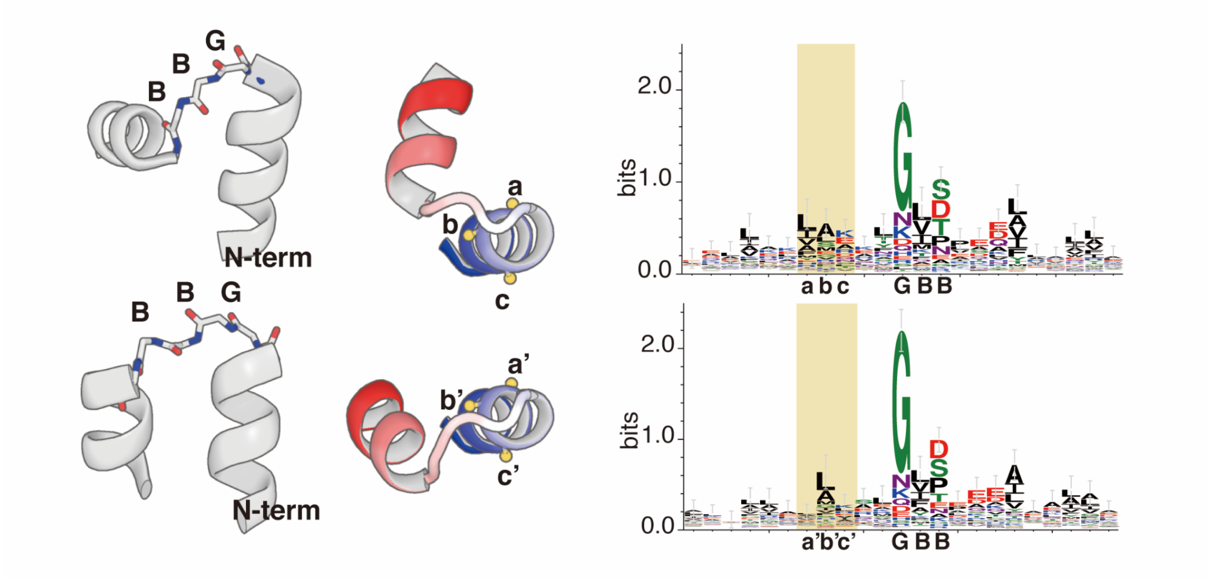


Figure 3-2: Comparison of αα-corner and αα-hairpins. They have similar backbone torsions but provide distinct contact patterns between two flanking α-helices. (Left) The overall structures of αα-corner and αα-hairpins. The loop regions are shown as sticks and colored in CPK-scheme. The α-helices are shown in the cartoon representation. (Center) αα-corner and αα-hairpins offer different environments for nearby residues. Each fragment is colored in blue-white-red gradient from N- to C-terminal. The orange sphere represents Cβ atoms on N-terminal α-helical segments. The Cβ a corresponds to a’, b to b’, and c to c’. See that position a is more buried than position a’, and similarly b is more exposed than b’. (Right) Sequence logo for αα-corner and αα-hairpins. The region shaded in orange corresponds to the residues whose Cβ atoms are colored in orange in the center panel. The alphabets beneath the logos indicate residue positions for the region on the N-terminal of loops, and the ABEGO backbone torsion angle representation for the loop regions. As conformations largely differ between αα-corners and αα-hairpins, the variance in the amino-acids compositions are most recognizable in the orange-shaded regions, which correspond to the flanking sequence rather than loop region.

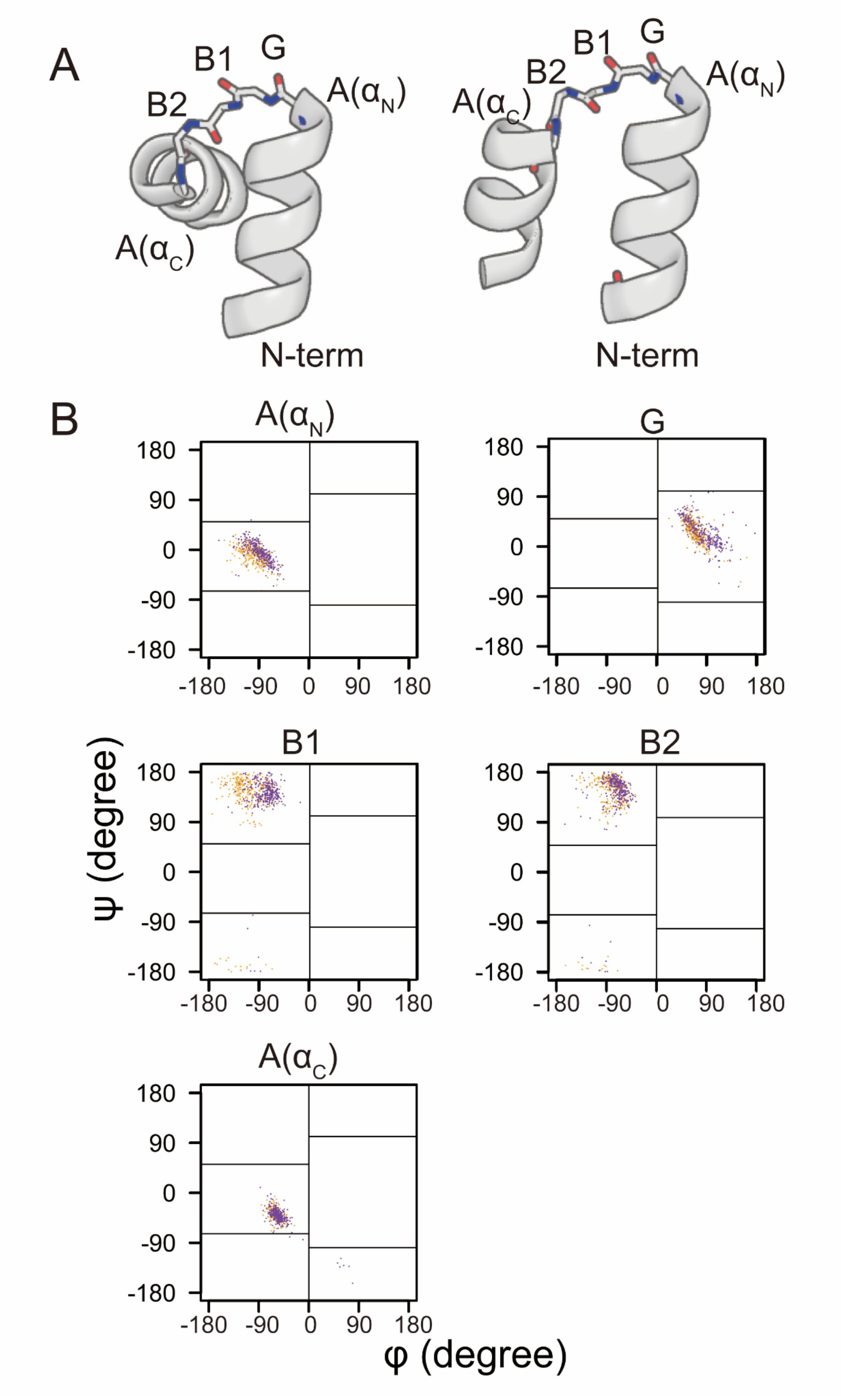


Figure 3-3: Identification of the residues responsible for the diversification between αα-corners and αα-hairpins. (A) Structure of αα-corner and αα-hairpin and assignment of site names. The loop regions are shown in sticks. (B) The Ramachandran plots for site A(α$_N$), G, B1, B2, and A(α$_C$). The orange/purple dots correspond to data from αα-corners/αα-hairpins. B1 site shows most divergent dihedral angles between αα-corners and αα-hairpins.

ABEGO analysis of helix-loop-helix fragments

From the non-redundant domain structure set which was a subset of ECOD database [16] whose sequence similarity was reduced by 40% sequence identity, 39,938 helix-loop-helix fragments were extracted. For the fragments with α-helices longer than 10 residues, ABEGO sequences were assigned for the loop regions. The ABEGO types of these fragments were counted and used to make Figure 4A.

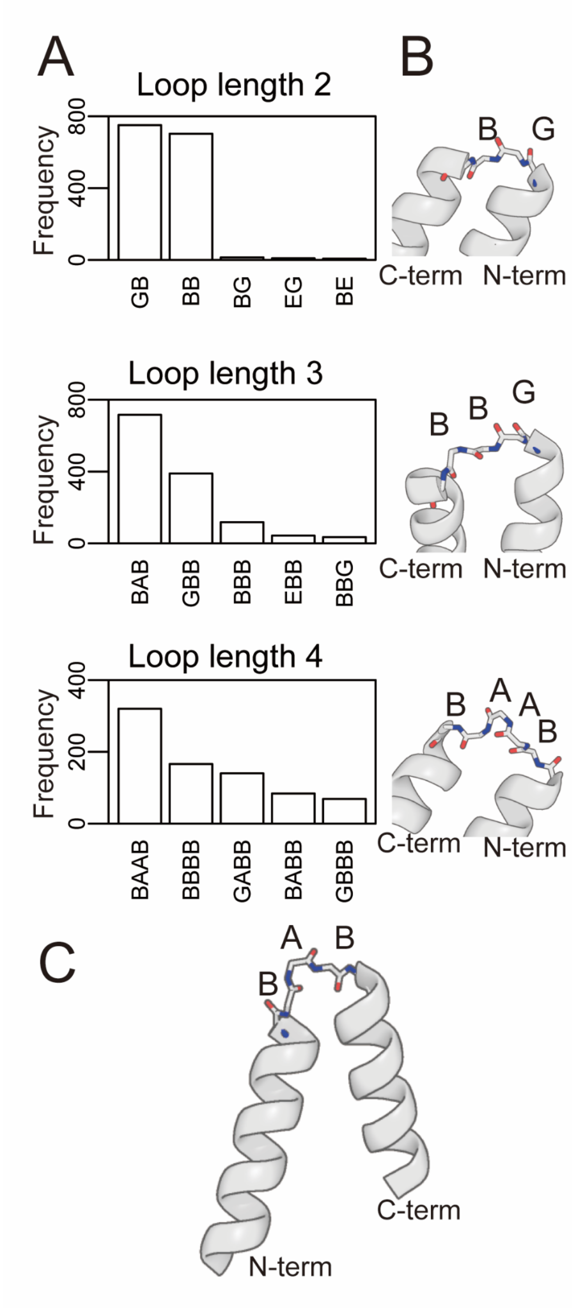


Figure 3-4: Statistical analysis of helix-loop-helix fragments revealed GB, GBB, and BAAB loops are most frequent αα-hairpins. (A) Histogram of ABEGO types for length 2, 3, and 4 loops. (B) Structures of GB, GBB, and BAAB αα-hairpins. (C) Although BAB-loop is the most frequent loop types in the statistics of length 3 loops, BAB loop is a v-shaped loop rather than αα-hairpins. For this reason BAB-loop was not used in this study.

Construction of target structures

The author composed the GBB, GB, and BAAB up-down bundles as well as the GBB orthogonal bundle by manually grafting the helix–loop–helix fragments using PyMOL (The PyMOL Molecular Graphics System, version 2.0 Schrödinger, LLC.) and removed severe steric clashes using Foldit [19]. The constructed backbone structures were used as templates for the ABEGO specifications, and the reference structures for the ABEGO-based backbone-building simulations. These structures were also used as template backbones for amino acid sequence design by Rosetta.

Backbone-building simulations

Sequence-independent fragment assembly simulations, termed ABEGO-based backbone-building simulations, were performed using Rosetta BluePrintBDR [20], as described by Lin *et al*. [6]. Blueprint files were generated based on the target backbone structure that was manually built in advance, and the files were used for fragment selection to specify the backbone torsion in the ABEGO representation. For each ABEGO specification, simulations were repeated for 10,000 trajectories, and the final snapshots from the trajectories were used for structural analysis. During the analysis, the Cα RMSDs of each structure referenced by the target backbone structures were calculated.

Amino acid sequence design and sequence-dependent folding simulations

The author performed amino acid sequence designs using the Rosetta flxbb protocol [20] starting from the backbone structure that was built manually. To enhance the efficiency of sequence design, amino acid profiles were constructed for the loop region using similar loop structure fragments (Cα RMSD < 2 Å) and were used as constraints for the residues used, as described by Marcos *et al*. [4]. The specifications of the residues were refined based on the buriedness of the backbone atoms using in-house programs. The author performed 10,000 design trials for each backbone model, selected the best sequences based on the fragment-quality score, and performed sequence-dependent fragment-assembly folding simulations [21] to identify the best design sequences. The author defined the fragment-quality score as the average of the logarithm of the number of fragments that had a Cα RMSD value lower than 1.5 Å in the design model. A total of 20,000 trajectories for folding simulations were obtained for each design protein to check the foldability.

Results and Discussion

αα-corners and αα-hairpins are indistinguishable in ABEGO representation

First, the author investigated a nontrivial example in which ABEGO representation could not distinguish two structurally different local motifs. Using structural informatics analysis, the author identified two distinct types of helix–loop–helix fragments that were indistinguishable based on their ABEGO sequences. Conformations of both motifs were represented as α-GBB-α in their ABEGO representation, but they result in distinct overall structures and sequence preferences (Figures 3-2 and Supplementary Figure S3-1). The first α-GBB-α motif is traditionally classified as an αα-corner that results in an almost orthogonal crossing angle between two flanking α-helices [22], and the second is called an αα-hairpin, which results in a steep hairpin turn for tightly packing adjacent α-helices into an antiparallel configuration [23]. By making Ramachandran-plots for each site in the loop region, the author found that the first B site (B1) showed most divergent torsion angles between αα-corners and αα-hairpins (Figure 3-3). The author also confirmed that αα-hairpin can be transformed into αα-corner by systematically changing the value of dihedral angle φ at the site B1 from -70° to -150° (Supplementary Figure S3-2). From these observations, the author divided the region B into two sub-regions S and P by the line of φ= -90° so that the αα-hairpins and αα-corners were separated from each other (Supplementary Figures S3-3, S3-4, and S3-5). This extension of ABEGO representation can deal with B region in finer resolution, and would be helpful to specify the conformation more precisely. However, as the original ABEGO representation does not take the heterogeneity of the B region into account, αα-hairpins and αα-corners are taken as identical in their ABEGO representation and are therefore indistinguishable in the coarse-grained representation.

α-GBB-α units cause loss of efficiency in ABEGO-based backbone building simulations

Next, the author sought to identify whether the ambiguity between the αα-hairpin and αα-corner in the original ABEGO representation causes loss of efficiency in ABEGO-based backbone-building simulations. The author first performed statistical analysis of loop regions and found that GB and BAAB loops are most frequent short αα-hairpin fragments in addition to GBB loop (Figure 3-4). The author manually generated six types of four-helix up-down bundle structures using these hairpin motifs: GBB, GB, and BAAB bundles with right-handed or left-handed topologies (Figure 3-5). Based on these decoy structures, the backbone dihedral angles were roughly specified by the ABEGO representations (Supplementary Figure S3-6) to select the fragments satisfying the specification, and ABEGO-based backbone-building simulations were performed [6,20]. Although the simulations for the GB bundles successfully recovered the original four-helix up-down bundle topologies, the ABEGO-based backbone-building simulations for the GBB bundles failed to efficiently generate the target topology (Figure 3-5). The results of BAAB bundles were marginal; the behavior was better than GBB but worse than GB bundles. More specifically, GB bundles showed best result where almost all of the populations resides within 5 Å from the native structure in the Cα-RMSD; BAAB bundle showed almost one forth of the population stayed within 5 Å from the native in the Cα-RMSD; GBB performed worst, in which most of the population showed Cα-RMSD larger than 10 Å. These results were independent of the handedness of target bundle topologies; both right-handed and left-handed four-helix bundles showed similar results depending on the loop types. In the simulations for the GBB bundle, most trajectories were trapped in misfolded structures that contained GBB corner fragments (Supplementary Figure S3-7), which is undesirable for building the up-down bundles.

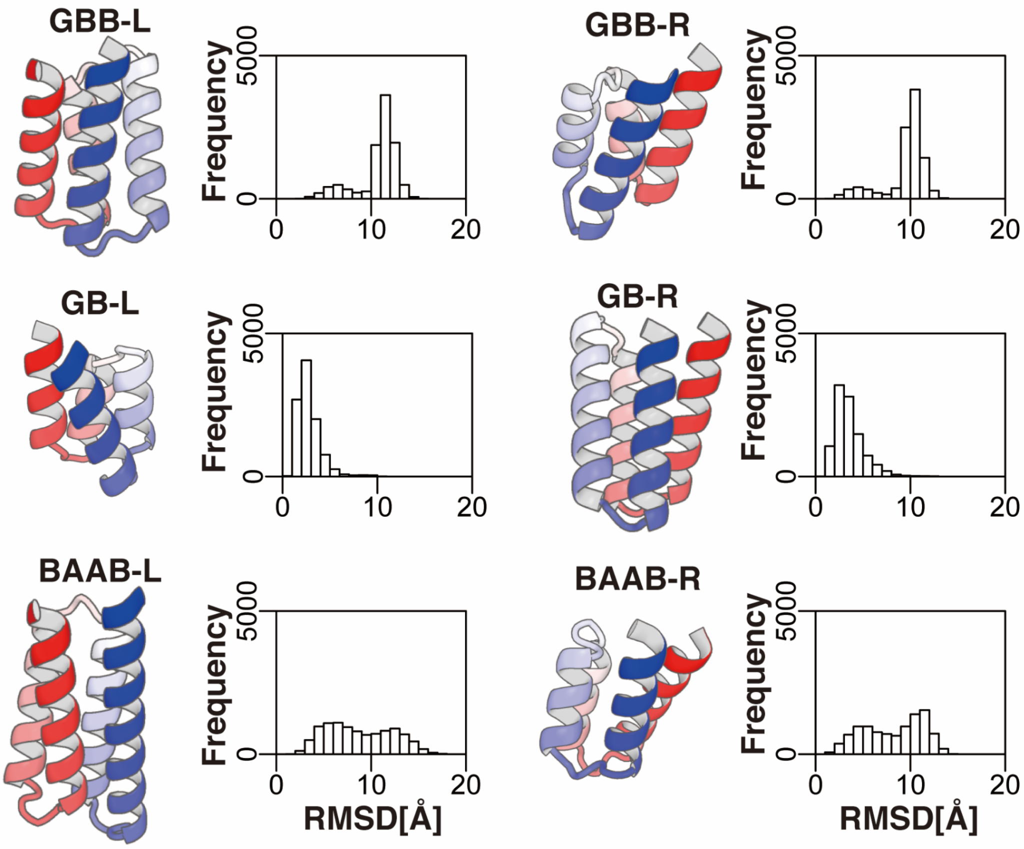


Figure 3-5: The foldability of four-helix up-down bundles. The structure four-helix up-down bundles are shown on the left of each column. The ABEGO of hairpins and handedness of bundles are indicated above each structure. The distributions of Cα RMSDs from 10^5 trajectories of backbone building simulations are shown on the left of each column. The GBB bundle has a large peak around 10 Å, which indicates the ABEGO-specification cannot force the polypeptide chain to fold into the target structure. GB and BAAB bundle show reasonably large populations on the left (Cα RMSD < 5 Å), which indicates that their ABEGO-specification is capable of letting the chain fold into the target topology.

So, why were GBB-containing structures more difficult to build in ABEGO-based backbone building simulations than GB-containing or BAAB-containing structures? To investigate this, the author looked into the contents of fragments that were picked up for the ABEGO-based backbone simulations from the structure database named filtered.vall.dat.2006-05-05. The number of fragments was 200 for each loop type. This clarified that the fragment libraries contained non-hairpin fragments in addition to the hairpins in all of three types of loop fragments (Figure 3-6). The GB fragments possessed most purified hairpin conformations, and the BAAB fragments showed long-tailed distribution of the conformation but it also had a sharp peak representing the hairpin structures. The GBB fragment library possessed the largest population of non-hairpin fragments. To estimate the population ratio of corner against hairpins in the GBB fragment library, the author gathered the fragments showing RMSDs lower than 1.5 Å from the representative αα-corner or αα-hairpin fragments. The ratio of corners against hairpins was about 4:1 in the fragment set (Supplementary Figure S3-8). This tendency is well consistent with the result of fragment assembly simulations; GB performs best, BAAB performs so-so, and GBB performs worst. As the populations ratio of corners against hairpin was almost 1:1 in the fragment library from manually curated domain database (Supplementary Figure S3-1), this bias of fragment populations toward the corner would be Rosetta-specific artifact and should be improved to allow more unbiased sampling of conformational space. However, even if the fragment set show unbiased populations of corners and hairpins, ABEGO-based fragment picking for α-GBB-α motifs results in the mixture of αα-corners and αα-hairpins and will still suffer from the unwanted fragment insertion at the loop region and lead to low sampling efficiency for GBB-containing structures. To summarize, the GBB-containing structures are difficult to build for two reasons: (1) low purity of fragments caused by double-meaning α-GBB-α motifs (2) the unbalance between αα-corner and αα-hairpins populations. More precise assembly of GBB-containing structures requires updates for the fragment picking algorithm and structure database from which fragments are picked up. This may require paying more attention on how to divide B region of ABEGO classification into subsections.

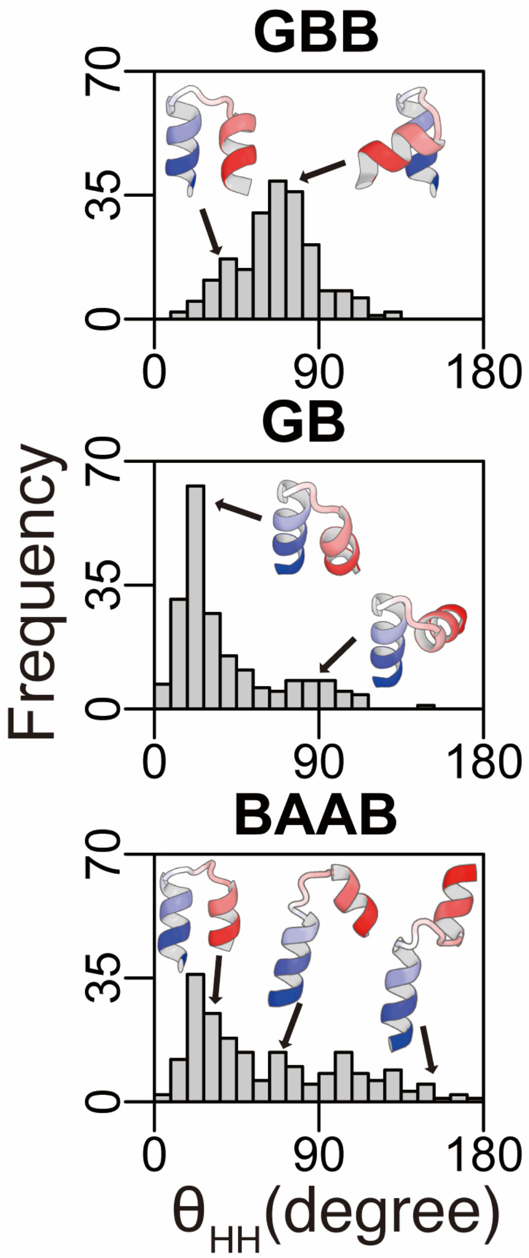


Figure 3-6: Distributions of helix-helix crossing angles in Rosetta-derived fragment library. GBB library shows a large peak at 90°, which corresponds to αα-corners. GB and BAAB libraries have the largest peak around 30°, which corresponds to the αα-hairpins. GBB fragment library is largely biased to the αα-corners so that αα-hairpins are difficult to appear in the fragment assembly simulations.

Amino-acid sequences for backbone structures composed of α-GBB-α units can be designed and predicted in-silico to fold into the target topologies

Considering the structures containing α-GBB-α fragments are difficult to compose in ABEGO-based backbone-building simulations, the author sought to identify whether they can be designed when their amino acid sequences are completely specified. Are they difficult to build again? The author performed amino acid sequence design of two distinct structures composed of α-GBB-α motifs alone using Rosetta [20]. The first structure was the four-helix up-down bundle that was described in the previous section, and the second structure was a small four-helix orthogonal bundle composed of two αα-hairpins and an αα-corner (Figures 3-7A and 3-7B). Similar to the ABEGO-based backbone-building simulations for the GBB up-down bundle, those for the GBB orthogonal bundle were also trapped in a misfolded state and showed low efficiency for achieving the target conformation (Supplementary Figure S3-9), which is consistent with the observation in GBB up-down bundles. However, by carefully designing amino acid sequences onto these structures using Rosetta, amino acid sequences that are predicted to fold into the respective target topologies can be obtained (Figures 3-7C and 3-7D). In contrast to the misfolding observed in the ABEGO-based backbone-building simulations, sequence-dependent fragment-assembly simulations successfully predicted both target topologies as having the lowest energy structures [21]. The results showed that plausible amino acid sequences can be designed once the backbone structures are built by some means even if they contain two types of α-GBB-α motifs indistinguishable in the ABEGO representation. This result indicated that the conformational space that can be covered by the amino acid sequence design is broader than the conformational space in which ABEGO-based backbone-building simulations can firmly sample. Further, a novel backbone-building methodology may be required to improve the ability to generate more diverse and complicated backbone structures.

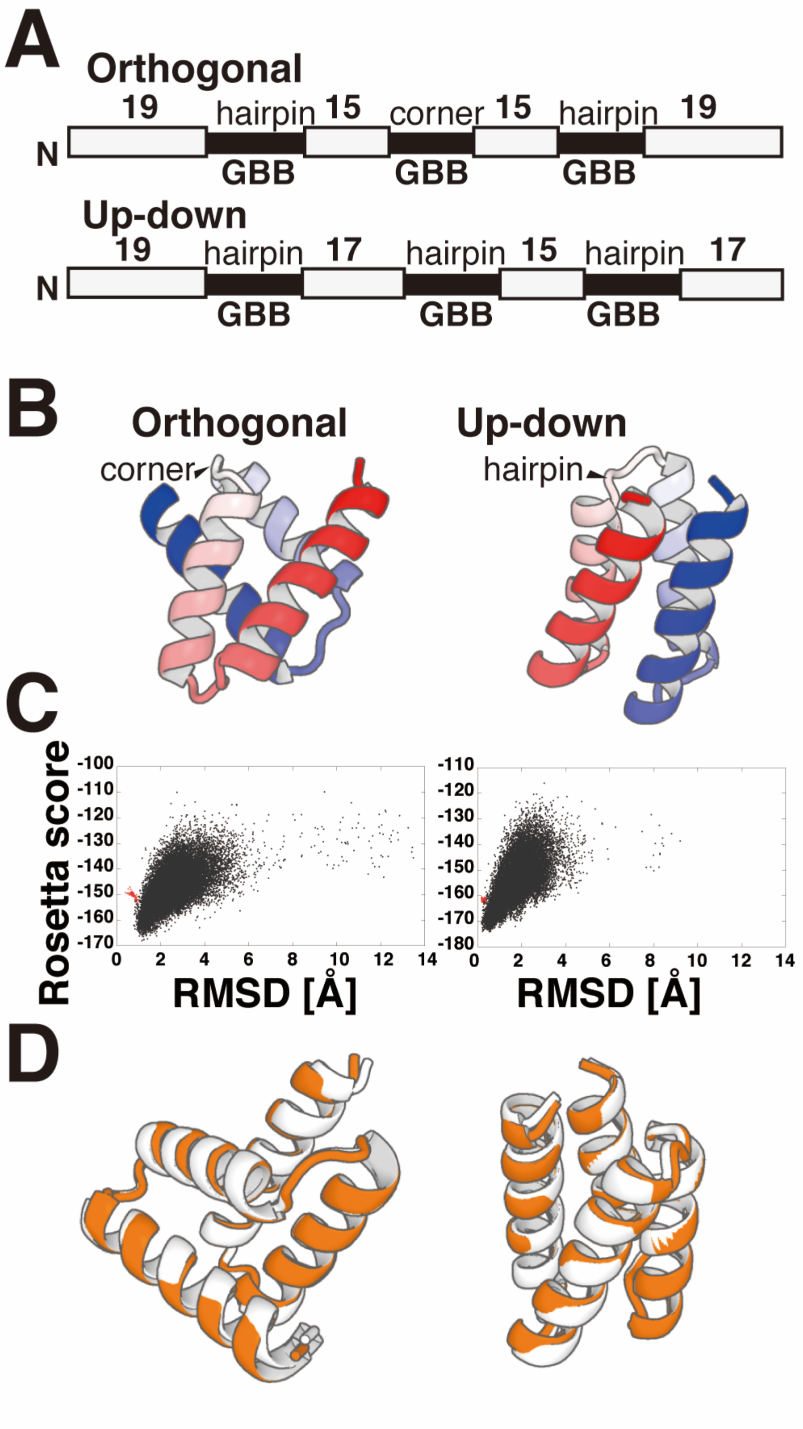


Figure 3-7: Design and sequence-dependent folding simulations of the four-helix orthogonal bundle and up-down bundles. (A) (B) Blueprints and structures of the GBB orthogonal bundles (left) and up-down bundles (right). The gray bars represent the α-helix and black bars represent loop regions. As all the loops are represented as GBB in the ABEGO representation, their intended structure types are indicated above the loop regions. (C) Energy-RMSD scatterplot from sequence-dependent folding simulations for orthogonal (left) and up-down bundle (right). Both of the designs have funneled energy landscapes, and are predicted to fold into the target

topology. (D) The superposition of the lowest energy structure (orange) onto the target structures. The lowest-energy structure from folding simulation of the orthogonal bundles showed Cα RMSD = 1.1 Å from the native. The lowest-energy structure for the up-down bundle showed Cα RMSD = 0.5 Å from the native.

Conclusion

In this study, the author showed that ABEGO is a coarse representation that can fail to distinguish different conformations, causing inefficiency in ABEGO-based backbone building for *de novo* protein design. The αα-corner and αα-hairpins are indistinguishable in the ABEGO representation because both are represented as α-GBB-α fragments. This ambiguity between these two distinct structures leads to difficulty in constructing simple four-helix bundle topologies composed of these α-GBB-α motifs.

Although the author used the two indistinguishable α-GBB-α fragments as a nontrivial example in this study, such confusion may occur for other motifs if the backbone torsion angles are represented in coarse-grained manners. Especially, the B region of ABEGO representation contains very heterogeneous conformations so that the region should be carefully divided into subsections in order to represent the subtle conformational changes. To this end, I dividied B region into the S and P subsections and proposed an extended version of ABEGO that can separate αα-hairpin and αα-corners. However, this extension is not always enough and there may be other pairs of fragments that still fail to be separated.

Interestingly, sequence design for GBB-containing backbone structures does not appear to be difficult compared to the backbone building; the author showed that two types of four-helix bundles composed of GBB fragments can be designed to be predicted to fold into the target topologies. This suggests that there are many topologies designable as amino-acid sequences which have not been tried because their backbone modeling remains difficult. In other words, difficulty in backbone modeling may be bottlenecking the design of novel artificial proteins. Therefore, novel methodologies for backbone building that can sample diverse structures unreachable by conventional structural modeling techniques may enable the design of a wide variety of protein structures. This will allow protein designers to further explore the protein structure universe and expand their design repertoires.

References

[1] Ramachandran, G.N., Ramakrishnan, C. & Sasisekharan, V., Stereochemistry of polypeptide chain configurations, *J. Mol. Biol.* **7** 95–99 (1963). DOI: 10.1016/S0022-2836(63)80023-6

[2] Wintjens, R.T., Rooman, M.J. & Wodak, S.J., Automatic classification and analysis of αα-turn motifs in proteins, *J. Mol. Biol.* **255** 235–253 (1996). DOI: 10.1006/jmbi.1996.0020

[3] Huang, P.S., Feldmeier, K., Parmeggiani, F., Velasco, D.F., Hocker, B. & Baker, D., De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy, *Nat. Chem. Biol.* **12** 29–34 (2016). DOI:

10.1038/nchembio.1966

[4] Marcos, E., Basanta, B., Chidyausiku, T.M., Tang, Y., Oberdorfer, G., Liu, G., *et al.*, Principles for designing proteins with cavities formed by curved β sheets, *Science.* **355** 201–206 (2017). DOI: 10.1126/science.aah7389

[5] Dou, J., Vorobieva, A.A., Sheffler, W., Doyle, L.A., Park, H., Bick, M.J., *et al.*, De novo design of a fluorescence-activating β-barrel, *Nature*. **561** 485–491 (2018). DOI: 10.1038/s41586-018-0509-0

[6] Lin, Y.R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A.F., Montelione, G.T., *et al.*, Control over overall shape and size in de novo designed proteins, *Proc. Natl. Acad. Sci. U. S. A.* **112** E5478–E5485 (2015). DOI: 10.1073/pnas.1509508112

[7] Basanta, B., Bick, M.J., Bera, A.K., Norn, C., Chow, C.M., Carter, L.P., *et al.*, An enumerative algorithm for de novo design of proteins with diverse pocket structures, *Proc. Natl. Acad. Sci. U. S. A.* **117** 22135–22145 (2020). DOI: 10.1073/pnas.2005412117

[8] Koepnick, B., Flatten, J., Husain, T., Ford, A., Silva, D.A., Bick, M.J., *et al.*, De novo protein design by citizen scientists, *Nature*. **570** 390–394 (2019). DOI: 10.1038/s41586-019-1274-4

[9] Wei, K.Y., Moschidi, D., Bick, M.J., Nerli, S., McShan, A.C., Carter, L.P., *et al.*, Computational design of closely related proteins that adopt two well-defined but structurally divergent folds, *Proc. Natl. Acad. Sci. U. S. A.* **117** 7208–7215 (2020). DOI: 10.1073/pnas.1914808117

[10] Rocklin, G.J., Chidyausiku, T.M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., *et al.*, Global analysis of protein folding using massively parallel design, synthesis, and testing, *Science.* **357** 168–175 (2017). DOI: 10.1126/science.aan0693

[11] Chevalier, A., Silva, D.A., Rocklin, G.J., Hicks, D.R., Vergara, R., Murapa, P., *et al.*, Massively parallel de novo protein design for targeted therapeutics, *Nature*. **550** 74–79 (2017). DOI: 10.1038/nature23912

[12] Vorobieva, A.A., White, P., Liang, B., Horne, J.E., Bera, A.K., Chow, C.M., *et al.*, De novo design of transmembrane b barrels, *Science.* **371** (2021). DOI: 10.1126/science.abc8182

[13] Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T.B., Montelione, G.T., *et al.*, Principles for designing ideal protein structures, *Nature*. **491** 222–227 (2012). DOI: 10.1038/nature11600

[14] Marcos, E., Chidyausiku, T.M., McShan, A.C., Evangelidis, T., Nerli, S., Carter, L., *et al.*, De novo design of a non-local β-sheet protein with high stability and accuracy, *Nat. Struct. Mol. Biol.* **25** 1028–1034 (2018). DOI: 10.1038/s41594-018-0141-6

[15] Romero Romero, M.L., Yang, F., Lin, Y.R., Toth-Petroczy, A., Berezovsky, I.N., Goncearenco, A., *et al.*, Simple yet functional phosphate-loop proteins, *Proc. Natl. Acad. Sci. U. S. A.* **115** E11943–E11950 (2018). DOI: 10.1073/pnas.1812400115

[16] Cheng, H., Schaeffer, R.D., Liao, Y., Kinch, L.N., Pei, J., Shi, S., *et al.*, ECOD:

An Evolutionary Classification of Protein Domains, *PLoS Comput. Biol.* **10** (2014). DOI: 10.1371/journal.pcbi.1003926

[17] Kabsch, W. & Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*. **22** 2577–2637 (1983). DOI: 10.1002/bip.360221211

[18] Krissinel, E. & Henrick, K., Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60** 2256–2268 (2004). DOI: 10.1107/S0907444904026460

[19] Kleffner, R., Flatten, J., Leaver-Fay, A., Baker, D., Siegel, J.B., Khatib, F., *et al.*, Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta, *Bioinformatics*. **33** 2765–2767 (2017). DOI: 10.1093/bioinformatics/btx283

[20] Fleishman, S.J., Leaver-Fay, A., Corn, J.E., Strauch, E.M., Khare, S.D., Koga, N., *et al.*, Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite, *PLoS One*. **6** 1–10 (2011). DOI: 10.1371/journal.pone.0020161

[21] Bradley, P., Misura, K.M.S. & Baker, D., Biochemistry: Toward high-resolution de novo structure prediction for small proteins, *Science.* **309** 1868–1871 (2005). DOI: 10.1126/science.1113801

[22] Efimov, A. V., A novel super-secondary structure of proteins and the relation between the structure and the amino acid sequence, *FEBS Lett.* **166** 33–38 (1984). DOI: 10.1016/0014-5793(84)80039-3

[23] Efimov, A. V., Structure of α-α-hairpins with short connections, *Protein Eng. Des. Sel.* **4** 245–250 (1991). DOI: 10.1093/protein/4.3.245

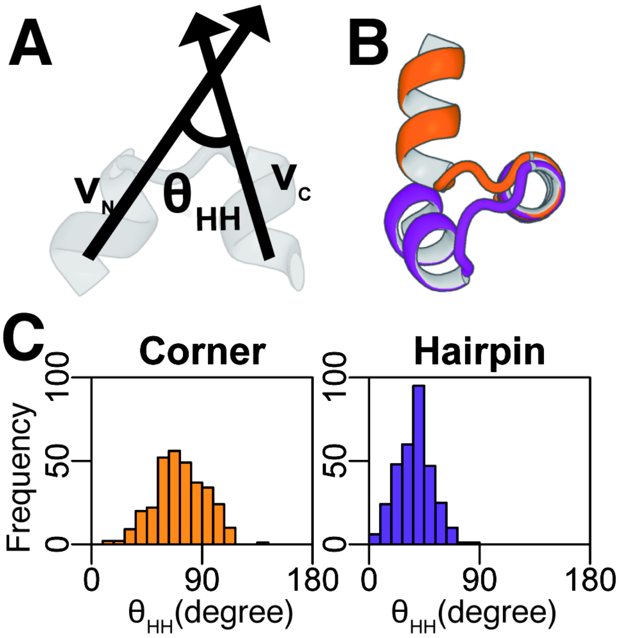


Figure S3-1: Comparison of αα-corners and αα-hairpins based on the helix-helix crossing angles. αα-corner and αα-hairpins are easily distinguished by helix-helix crossing angles. (A) Definition of helix-helix crossing angle. (B) αα-corner (Orange) and αα-hairpins (Purple) are superimposed to show their distinct conformations. Both of the structures are cluster representatives (medoids) by k-medoids clustering with k=2 taking the all-to-all Cα RMSDs as the metric. (C) The distribution of helix-helix crossing angles after k-medoids clustering with k=2. The αα-corner and αα-hairpins have distinct distribution of helix-helix crossing angles.

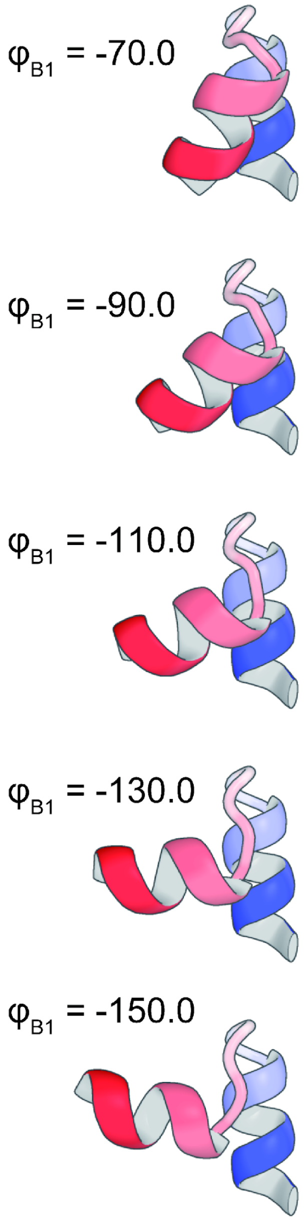


Figure S3-2: Morphing between αα-haripin and αα-corner by systematically scanning the φ angle at B1 site from -70.0 degree to -150.0 degree. This clarifies the B1 site is the key residue to diversification between αα-haripin and αα-corner.

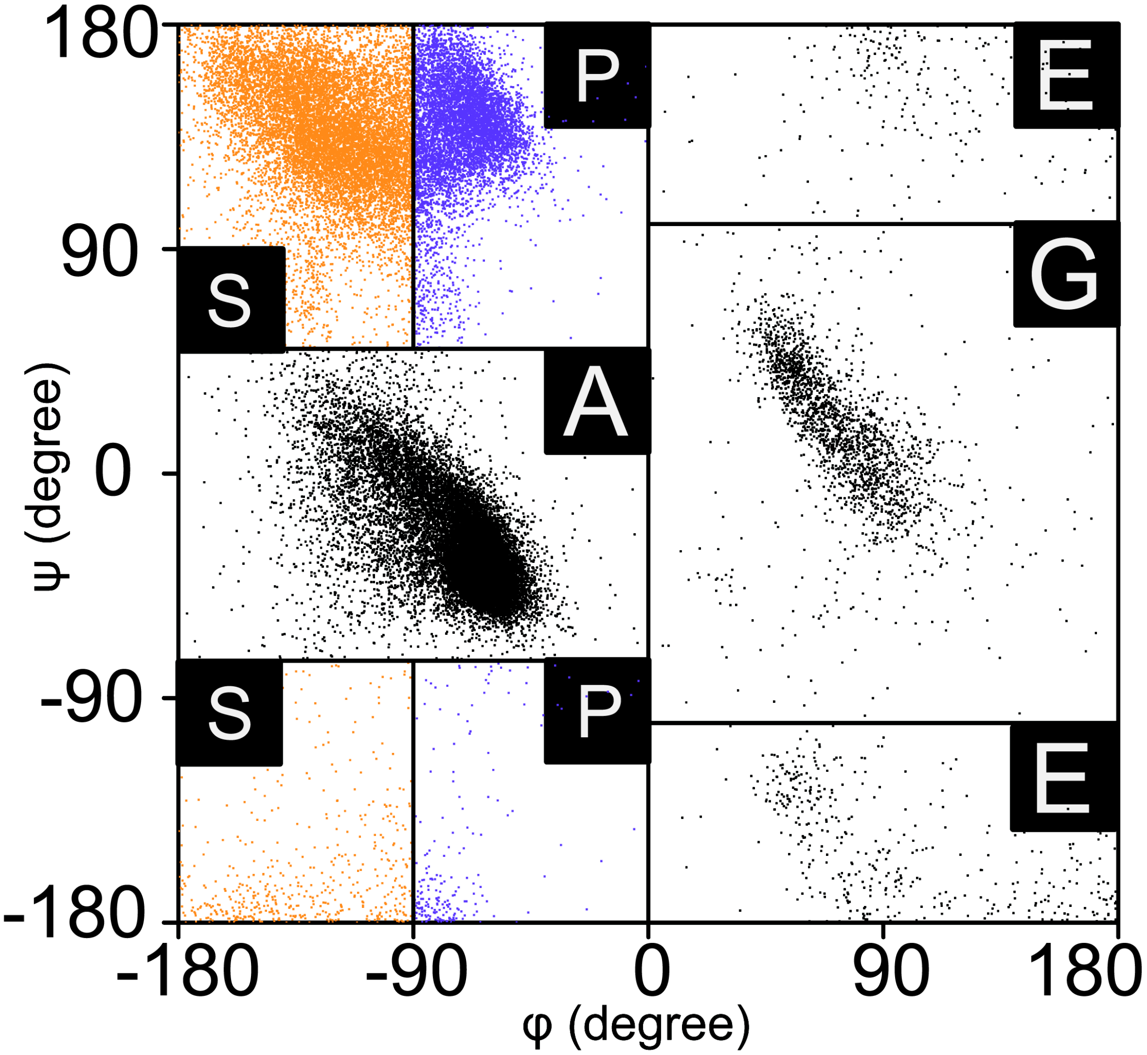


Figure S3-3: New definition of 6-state ABEGO with B region divided into S and P subsection. With these new states, GBB αα-corners and GBB αα-hairpins can be distinguished by their first B site of GBB. Broadly, αα-corners correspond to GSP, and αα-hairpins correspond to GPP.

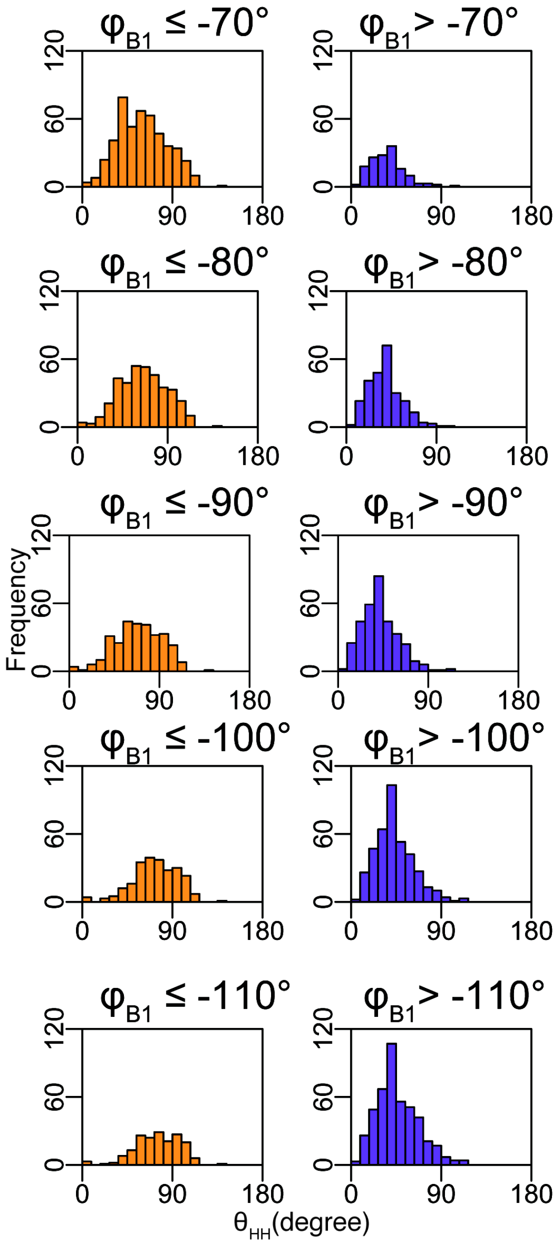


Figure S3-4: Separation of αα-corners (left) and αα-hairpins (right) by the condition that the dihedral angle φ of the first B site (φ_{B1}) in GBB is smaller/larger than -70° – -110°. Based on these plots, the border between S and P was determined to -90° so that the definition can separate hairpin and corners most sharply.

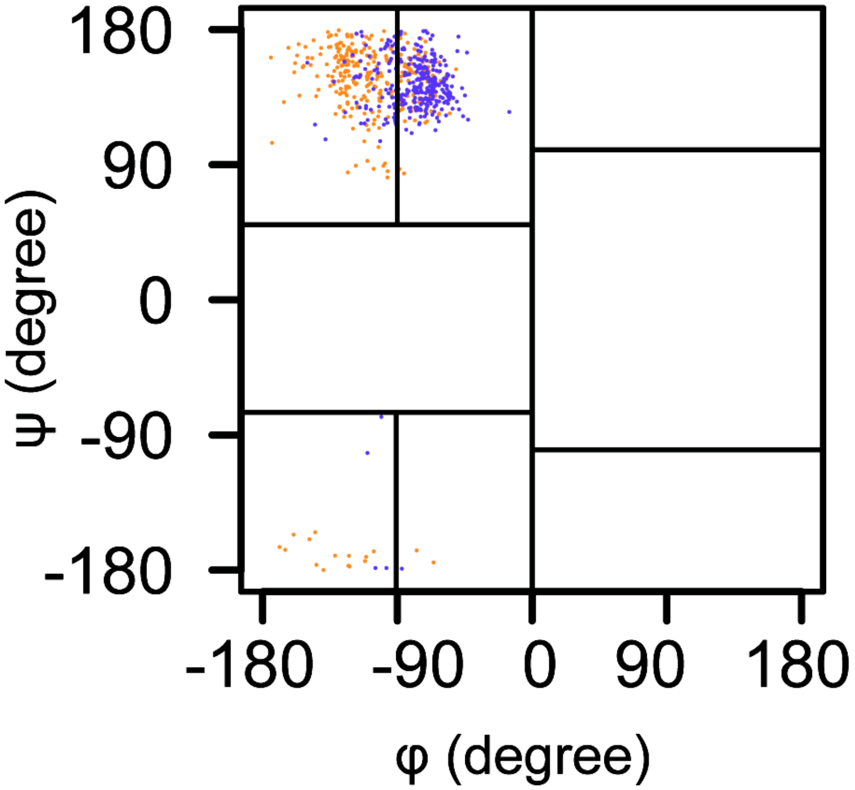


Figure S3-5: New border line dividing region B into two subclasses shown alongside the dihedral angle data of the first B site of GBB. Orange dots represent the data of corners and purple dots represent the data of hairpins. This figure is to show how the new border divides the populations into two classes.

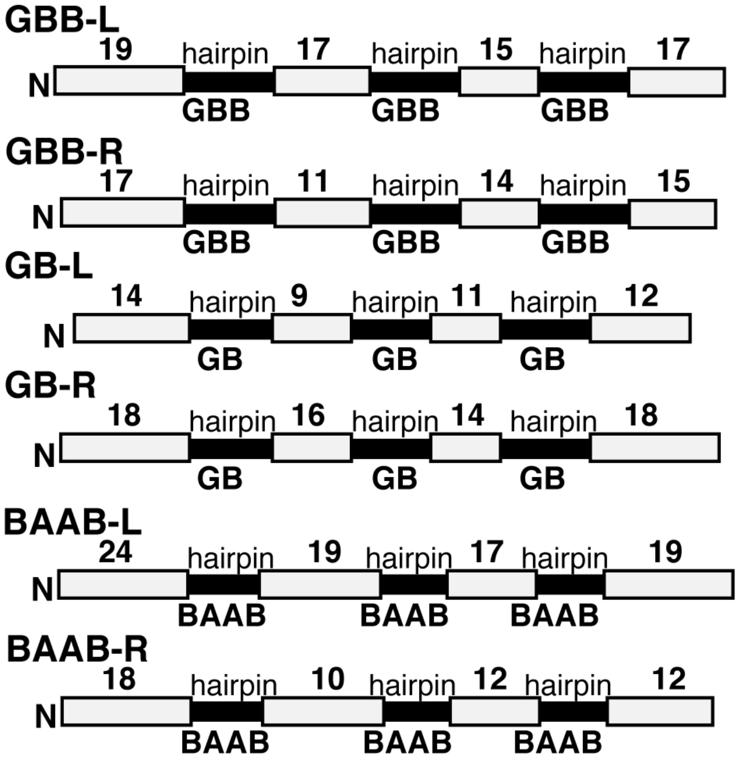


Figure S3-6: Blueprints for GBB-L, GBB-R, GB-L, GB-R, BAAB-L, and BAAB-R four-helix up-down bundles. The gray bars represent the α-helix and black bars represent loop regions. The ABEGO of the loop is indicated beneath the loop region. The numbers indicate the residue number of α-helix.

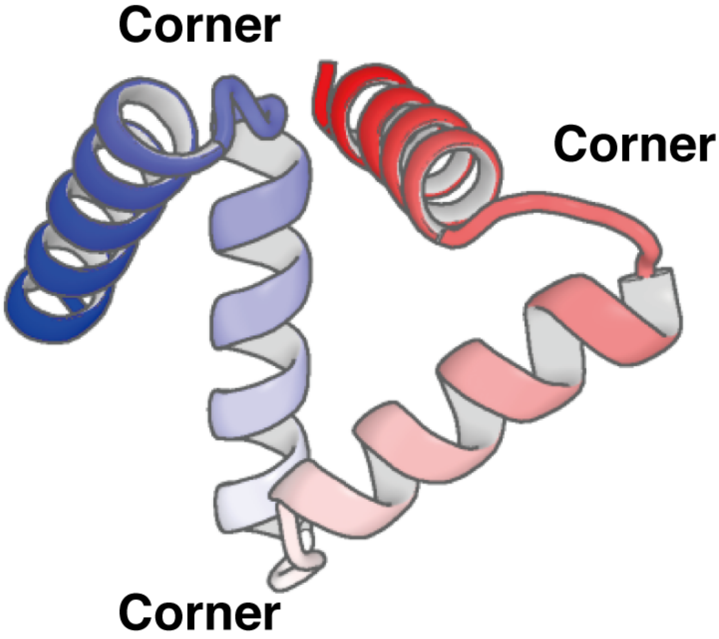


Figure S3-7: An example of misfolded structure observed in backbone building simulations for GBB up-down bundle. The loop regions take αα-corner conformations instead of αα-hairpins. This confusion between corner and hairpins makes it difficult to construct a simple up-down bundle structure by ABEGO-based backbone building simulations.

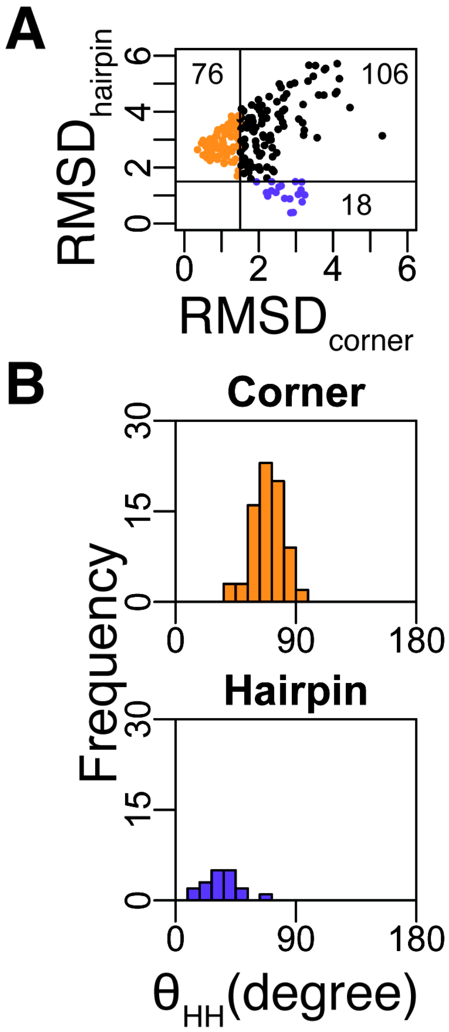
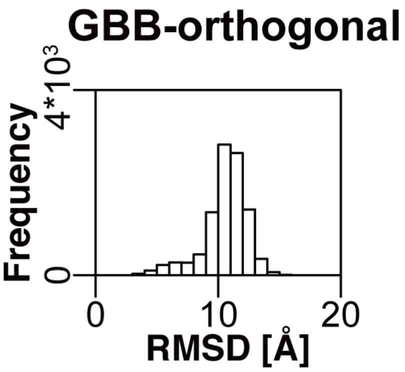


Figure S3-8: Estimating the ratio of αα-corners against αα-hairpins in the Rosetta-derived GBB fragment library. (A) The populations were classified into three classes, corners (orange), hairpins (purple), and others (black) by the RMSDs from reference structures. The inset numbers represent the number of class members. The reference structures for RMSD calculations were the same as the structure shown in figure 2. (B) The distribution helix-helix crossing angle in corner and hairpin class. The corner-type fragments dominate in the fragment library compared to the hairpins.



S3-9: Distribution of RMSDs from the sequence-independent ABEGO-guided backbone building simulations of four-helix orthogonal bundle. Similar to the four-helix up-down bundle composed of GBB-hairpins, the distribution of GBB-orthogonal bundles also showed a large peak around 10 Å. This indicates the GBB orthogonal bundle is difficult to build by ABEGO-guided backbone building simulations.

Chapter 07: **Conclusion**

Recently, de novo designs of proteins have made large progress [1]. However, the structure of designed proteins remains quite undiversified, compared to naturally occurring protein. Especially, the structure of designed all-α proteins until today remains simple and alike compared to naturally occuring diverse proteins, which is problematic because the structures of proteins define their functions. This lack of diversity possibly would bottleneck the diversification of functions of de novo designed proteins. Therefore there's a strong need to expand the repertoire of desigable protein structures.

When compared to designed proteins, naturally occurring proteins are complicated: John Kendrew described the first structure of myoglobin as “difficult to describe in the simple terms”, and most naturally occurring proteins similarly have complicated appearances. On the other hand, artificial proteins the humans have designed or created so far look simpler than naturally occuring proteins. As for all-α proteins previously designed, most of them can be classified into up-down helical bundles. The author’s motivation in this thesis was to fill the gap between them; the author aimed to diversify the structure of design proteins and make them more complicated so as to be described as “difficult-to-describe”.

In this thesis, the author reported 5 different studies concerning the design of α-helix structures to address this problem: diversification of de novo designed protein structures.The all-α class structures were selected as design targets because the class seemed to have rich potential to yield diverse topologies. The author investigated why the previous all-α designs remain simple, proposed novel strategies to diversify them, and seeked for applications of the new method.

In the second section of the thesis, the authors investigated how the simplest class of all-α protein structures can be built from scratch. FIrst, the author started from the statistical analysis of helix-loop-helix motifs and identified three typical αα-hairpins that are specifically related to the left- or right-handedness of helix-helix packing. Using these typical αα-hairpins, the author constructed various types of three-helix bundle structures. By enumerating the possible combinations of those building blocks, the author found that the lengths of the second alpha-helices play a significant role in the compaction of the three-helix bundle structures. The enumeration of possible combinations resulted in a comprehensive set of compact three-helix bundles. This research indicated that it is not always predictable beforehand which combinations of the length and types of secondary structure and loops are appropriate to yield compact and globular protein-like conformations, and protein designers need to exhaustively explore the conformational space to discover which combinations give desired complicated topologies.

In the third section of this research, the author investigated how a four-helix orthogonal bundle can be built from scratch. Four-helix orthogonal bundle is the simplest example of “non-up-down” types of all-α structures, and therefore was the best target to start with to investigate how to compose such complicated or “difficult-to-describe” topologies. The author found incorporation of that GBB type of loop can cause severe efficiency loss in the backbone-building simulations. This was because the GBB can refer to both αα-hairpins and αα-corners, and therefore introduce structural disturbance in the backbone building simulations. This seems to have been lowering the efficiency of the all-α protein design by fragment assembly simulation and bottlenecking the diversity of designed proteins. Nevertheless, the author also found that such GBB loops can be firmly designed with amino-acid sequence design. This suggested that “difficult-to-describe” structures are “difficult-to-build”, but are not always “difficult-to-design”. These results motivated the author to step into the design of much more complicated topologies.

In the fourth section of this research, the author aimed to design all-α structures with more complicated topologies. As the author had revealed that some helix-loop-helix building blocks cause the severe efficiency loss in fragment-assembly simulations for backbone structure building, the author developed a different approach to efficiently sample complicated all-α topologies. First, the author performed statistical analysis of naturally occuring helix-loop-helix loop motifs, and found there’s dominant and typical loop conformations. The author classified them into 18 clusters, and composed a minimal set of building blocks for all-α protein structures. Then the author performed a literally combinatorial generation of these typical fragments, and obtained a vast number of all-α backbone structures. After discarding extended or severely clashing structural models, the author constructed a library of globular all-α protein structure models. Of course the native combinatorial computation resulted in a large number of extended or clashing structure models. Nevertheless, the author obtained more than 300,000 globular protein-like conformations from the naive combinations. To demonstrate the designability of these generated structural models, the author selected 5 topologies from the library and performed amino-acid sequence design. The designed protein showed high-solubility, α-rich spectra in CD experiments, and monomeric mono-disperse distributions of molecular weights in SEC-MALS experiments. The 13C-15N 2D HSQC NMR suggested that these proteins have well-folded native states. Therefore, the author asked the NMR-expert and crystallography-expert collaborators to solve the three-dimensional structure of these design proteins, and they revealed that their structures agreed so well with the design models, which tells that the such “difficult-to-describe” structures are indeed designable.

In the fifth section of this research, the author constructed a de-novo designed protein library that encodes 294 topologies by 7,350 amino-acid sequences. Though the experimental validation of this library remains to be done in future, the whole library is designed to be encoded by state-of-art DNA oligo pool as the library is composed of the 70-residue structures. The structural diversity of the library would be advantageous to design functional proteins or specific binders, and such “many-fold” libraries would be standard approaches for library design coupled with high-throughput screenings.

In the sixth section of this research, the author composed two types of globin-like topologies. The author designed 8 amino-acid sequences for each of these globin-like folds. The designed sequences are predicted to fold into the target structures by fragment assembly simulations. Though the experimental validation for these design proteins are yet to be done, it is surprising that one of the most complicated topologies, globin fold, can be built up from the simple set of local motifs. This suggests that the complicated topology can be reduced into their simplified version of structures using appropriate sets of typical local motifs.

To summarize, the author concluded that the lower diversity of previously designed all-α protein structures can be attributed to two factors. First factor is a technical problem; ABEGO-based fragment assembly simulations, which have routinely been utilized in the de novo backbone design, have low efficiency when two similar fragments are mixed up in the fragment library as the author has revealed in the second chapter of this thesis. The second factor is related to human bias; the complicated structures are difficult for humans to draft out so that protein designers tend to design easier structures. To overcome these problems, the author developed the novel strategy to model the backbone all-α structures that can yield massive structural diversity of design templates. The experimental efforts have confirmed that the author’s approach can indeed be able to generate designable structures, and the designed protein structure showed complicatedness comparable to the naturally occuring globin structures. Supported by the experimental results, the approach can encourage the protein designers to break down the possible prejudice that the design proteins have to be simpler than naturally occuring proteins and certainly extend the repertoire of de novo designed proteins.

Finally the author summarizes three lessons from this series of studies, which may help designers to efficiently and confidently perform de novo designs:

(1) Use typical structural motifs rather than atypical ones. Quality of backbone structures largely determines the fate of the design. Low-quality backbone structures cannot produce high-quality amino-acid sequences that preferentially fold into the target structures, and lead to failure of designs, or at least cannot be predicted to fold into the target structure in silico. The starting point, the stage to construct the template backbone structures, largely determines the overall fate of the design process, which is difficult to compensate for in the later stages.

(2) Do not stick to routinely used backbone-building methods. For some classes of protein structures, ABEGO-based fragment assembly simulations fail to generate desired structures. There should be more efforts to develop diverse backbone-building methods to free the designers from the technical limitations of currently available methods. It is possible that if the rule (1) is satisfied, any means to build backbone structures can offer the starting point for the de novo protein designs.

(3) Topology may not matter as long as it’s globular. Tertiary structures of the proteins the author designed and validated in the chapter #4 were originally generated by random combinations of typical local motifs. Therefore, there were less intentions to “design” the topology than previous design researches did, and the author can conclude that the overall topology of the target does not affect the successfulness of the design as long as the topologies look globular. In other words, what designers should pay attention to is the typicality of the local motifs, which is what the rule (1) says. As long as the local quality of the fragments are guaranteed and local/non-local backbone hydrogen bonds are satisfied, the overall topologies can be any.

Reference

1. X. Pan, T. Kortemme, Recent advances in de novo protein design: Principles, methods, and applications. J. Biol. Chem. 296, 100558 (2021).

Appendix

Template modeling score (TM-score)

TM-score is one of the structure-similarity measures to compare two protein conformations, which plays central roles in the series of studies described in this thesis. TM-score is defined by the following two equations:

$$\text{TM-score} = \max\left[\frac{1}{L_N} \sum_{i}^{L_T} \frac{1}{1+\left(\frac{d_i}{d_0}\right)^2} \right] \qquad (1)$$

$$d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8 \qquad (2)$$

In the equation (1), the L_N denotes the length of native protein structure, L_T denotes the length of aligned segments, d_i denotes the distance between the i-th pair of aligned residues, and max means "take the maximum of the function" after the optimal structural superposition between two structures. The equation (2) defines the d_0, whose value appears about 0.17 independently from the number of residues. TM-score takes the values between 0 and 1, where higher value means higher structural similarity between two structures [1]. Practically, TM-score is computed by the program named TM-align, which utilizes TM-score rotation matrix and dynamic programming to approximately search the optimal superposition to maximize TM-score [2]. When a pair of structures shows TM-score higher than 0.5, the pair of structures belong to the same fold [3].

References

1. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins Struct Funct Genet. 2004;57:702–10.

2. Zhang Y, Skolnick J. TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33:2302–9.

3. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010;26:889–95.

Acknowledgements

The author would like to thank Dr. Naohiro Kobayashi and Dr. Toshihiko Sugiki for NMR structural determination of the five design proteins Chantal, Rei, Gogy, Nomur, and Mussoc, which clarified that these designed proteins indeed fold into the target conformations. The author would like to thank Dr. Kano Suzuki and Prof. Takeshi Murata for structural determination of the design protein Elsa by X-ray crystallography, which provided the first experimental validation for the author's design. The author is grateful that these structure-biology experts permitted the use of the structures that they solved in this thesis. The author would like to thank Dr. Shintaro Minami for his mentorship in the composition of my first publication, daily discussions on protein design and informatics, and the technical support in construction of the mini-protein library studied in chapter 5. The author also would like to thank Dr. Naoya Kobayashi for discussion on protein design and synthetic biology, experimental support in chapter 4, and inspiring the author to reach the concept of the "many-fold" library studied in chapter 5. The author also would like to thank Dr. Nobu C. Shirai for daily discussion and advice on the manuscripts: his unique insights based on lattice protein models were helpful in critically treating combinatorial enumeration of all-α protein structures. The author is grateful to members of Koga laboratory for helping the author pursue the study for the design of complicated all-α topologies for chapter 4, offering experimental equipment, and providing advice on the protein design. The author would like to thank Prof. Masahiro Ehara for providing comments on the manuscripts and supportive supervision during the final year of the doctoral course. The author would like to thank the Japan Society for the Promotion of Science (JSPS) for financial support as DC1 and support by a Grant-in-Aid for JSPS Fellows Grant Number 15J02427, which encouraged me to jump into the field of protein design. The author would like to thank the Institute for Molecular Science for financial support as a research assistant: without this support, I would not have continued this research in 2018 and 2020. Finally, I would like to thank my families for supporting me in every aspect of my life during my doctoral course.