Vol. 18, pp. 159–167 (2021)
doi: 10.2142/biophysico.bppb-v18.017

Biophysics and Physicobiology
https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

Limitations of the ABEGO-based backbone design: ambiguity between αα-corner and αα-hairpin

Koya Sakuma[1,2]

[1] *SOKENDAI, The Graduate University for Advanced Studies, Okazaki, Aichi 444-8585, Japan*
[2] *Institute for Molecular Science, Okazaki, Aichi 444-8585, Japan*

Received April 15, 2021; accepted May 25, 2021; Released online in J-STAGE as advance publication May 28, 2021

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, I show a nontrivial example of two local motifs that could not be distinguished by their ABEGO representations. I 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. I 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, I 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.

Key words: protein design, fragment assembly simulation

Corresponding author: Koya Sakuma, SOKENDAI, The Graduate University for Advanced Studies, 38 Nishigonaka, Myodaiji, Okazaki, Aichi 444-8585, Japan. e-mail: sakuma@ims.ac.jp

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

◀ ***Significance*** ▶

ABEGO is a coarse-grained representation for polypeptide backbone dihedral angles. I show ABEGO representation is unable to distinguish certain type of helix-loop-helix fragments, which causes the loss of efficiency in the fragment-assembly simulations for construction of backbone model in de novo protein design. Understanding the limitation of commonly used coarse-grained representations is important for improvement of backbone-building strategies in de novo protein design.

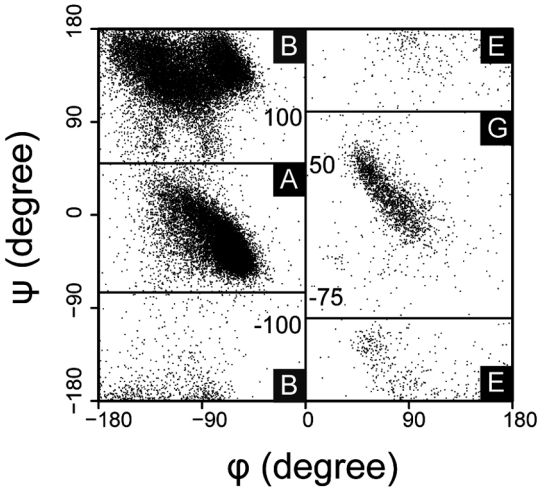

Figure 1 Definition of ABEGO. Horizontal axis represents φ and vertical axis represents ψ angle of polypeptide backbone structure. Ramachandran plot is divided into four sections named A, B, E, and G. The values of phi and psi angles for boarder line are indicated on the left or right of the boarder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

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 (Fig. 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.

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, I show a non-trivial example of two famous local motifs that are indistinguishable by their ABEGO representation and point 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

I 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 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 2, 3 and Supplementary Figures S1, S4, and S5. 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 S8. Next, I identified the helix–helix crossing angle (Supplementary Figure S1) using the helix orientation vector defined by Krissinel *et al.* [18] and confirmed that the clustering can clearly separate αα-corners and αα-hairpins (Fig. 2).

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.

Construction of target structures

I 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

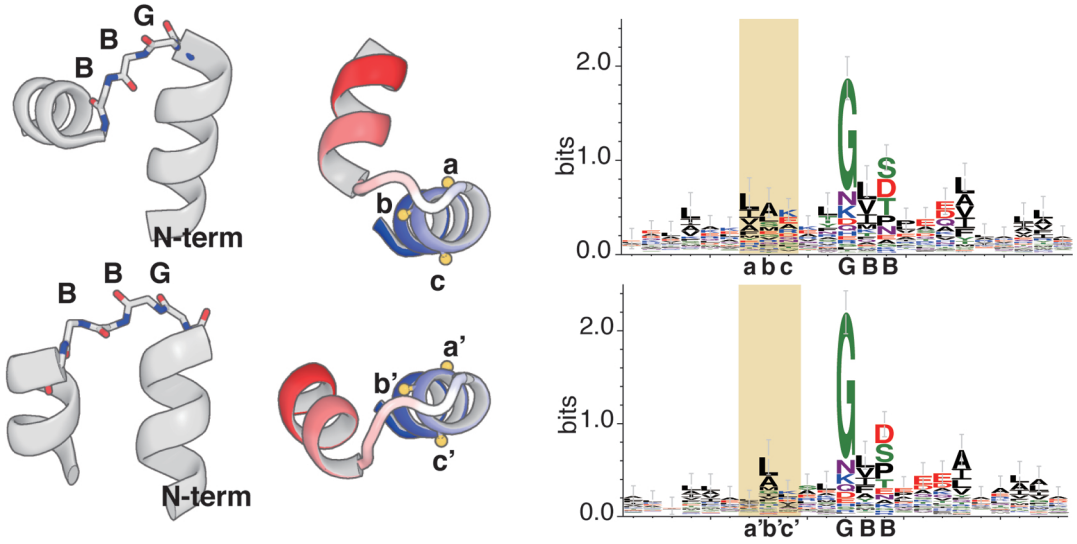

Figure 2 Comparison of αα-corner and αα-hairpins. They have similar backbone torsions but provide distinct contact patterns between two flanking α-helices. (Left) The overall structures of αα-corner and αα-hairpins. The loop regions are shown as sticks and colored in CPK-scheme. The α-helices are shown 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.

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

I 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. I 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. I 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, I investigated a nontrivial example in which ABEGO representation could not distinguish two structurally different local motifs. Using structural informatics analysis, I identified two distinct types of helix–loop–helix fragments that were indistinguishable based on their

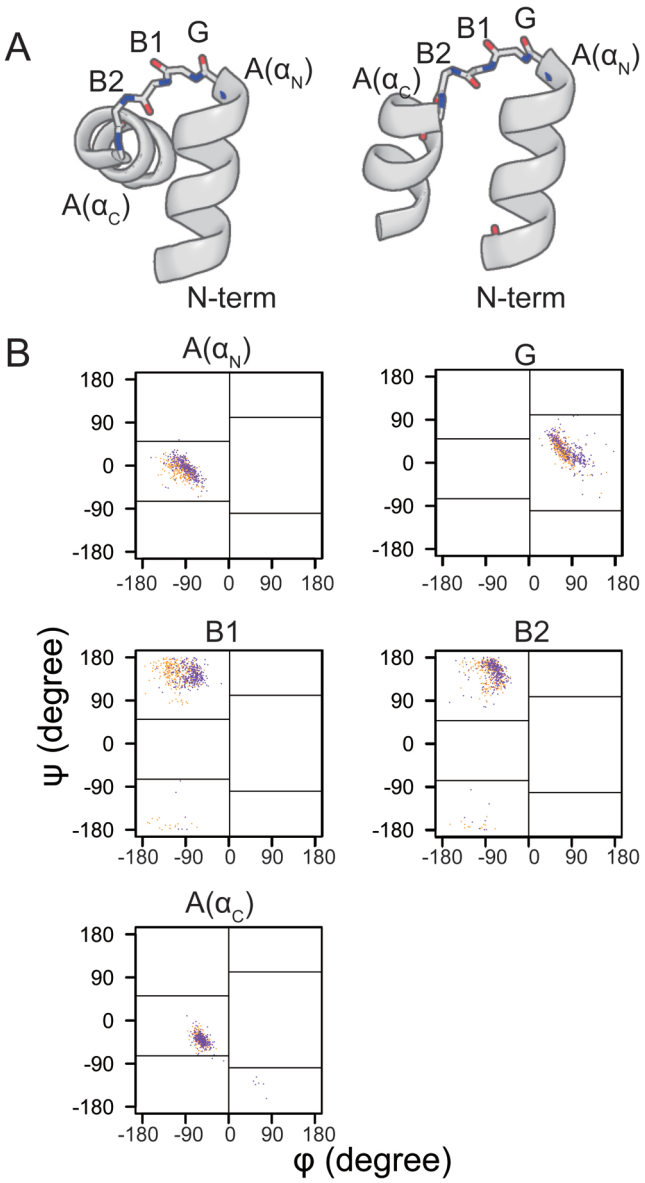

Figure 3 Identification of the residues responsible for the diversification between αα-corners and αα-hairpins. (A) Structure of αα-corner and αα-hairpin and assignment of site names. The loop regions are shown in sticks. (B) The Ramachandran plots for site $A(\alpha_N)$, G, B1, B2, and $A(\alpha_C)$. 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.

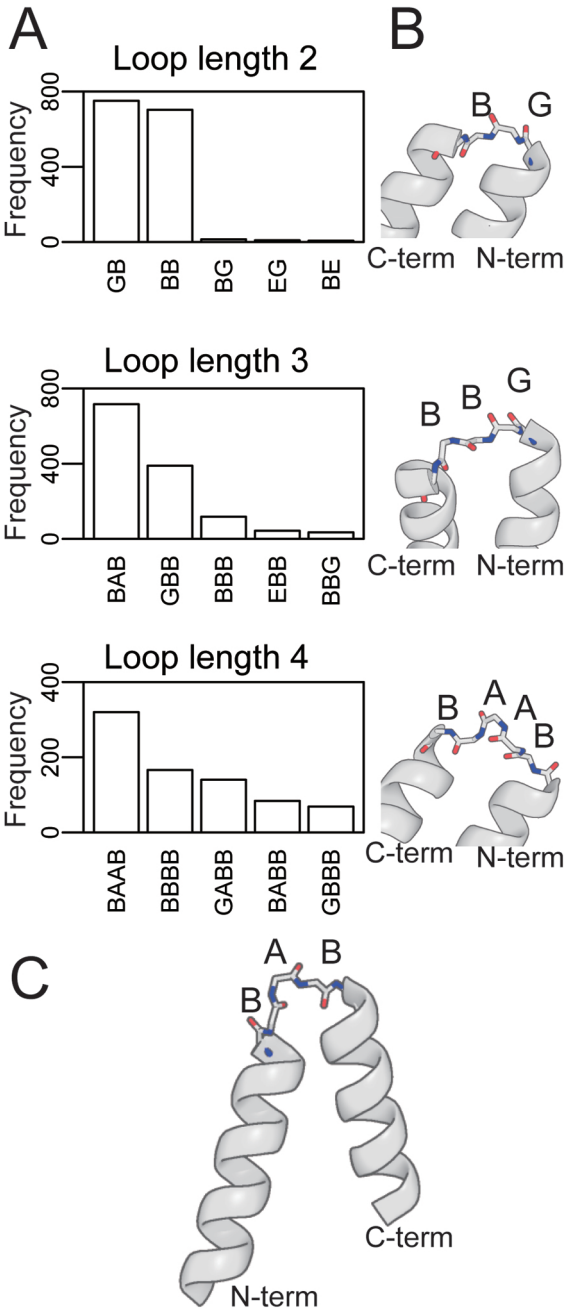

Figure 4 Statistical analysis of helix-loop-helix fragments revealed GB, GBB, and BAAB loops are most frequent αα-hairpins. (A) Histogram of ABEGO types for length 2, 3, and 4 loops. (B) Structures of GB, GBB, and BAAB αα-hairpins. (C) Although BAB-loop is the most frequent loop types in the statistics 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.

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 (Figs. 2 and Supplementary Fig. S1). 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, I found that the first B site (B1) showed most divergent torsion angles between αα-corners and αα-hairpins (Fig. 3). I also confirmed that αα-hairpin can

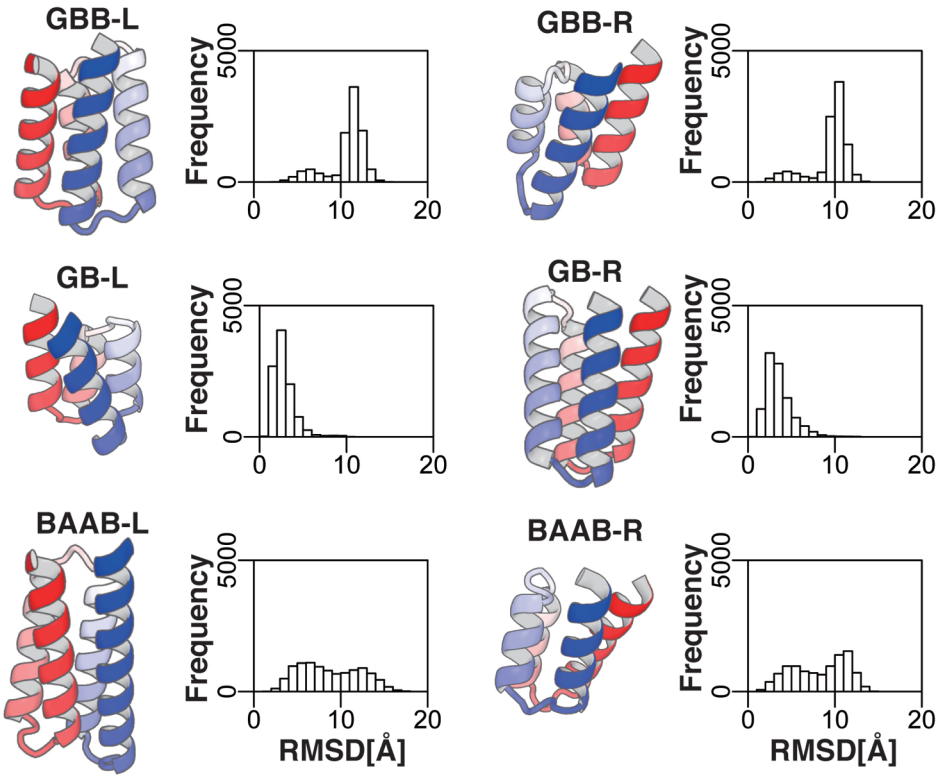

Figure 5 The foldability of four-helix up-down bundles. The structure four-helix up-down bundles are shown on the left of each column. The ABEGO of hairpins and handedness of bundles are indicated above each structure. The distributions of Cα RMSDs from 10^5 trajectories of backbone building simulations 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.

be transformed into αα-corner by systematically changing the value of dihedral angle φ at the site B1 from −70° to −150° (Supplementary Fig. S2). From these observations, I 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 Figs. S3, S4, and S5). 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, I 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. I 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 (Fig. 4). I 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 (Fig. 5). Based on these decoy structures, the backbone dihedral angles was roughly specified by the ABEGO representations (Supplementary Fig. S6) 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 (Fig. 5). The result 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 Fig. S7), which is undesirable for building the up-down bundles.

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, I 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 (Fig. 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, I 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 Fig. S8). 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 Fig. S1), 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.

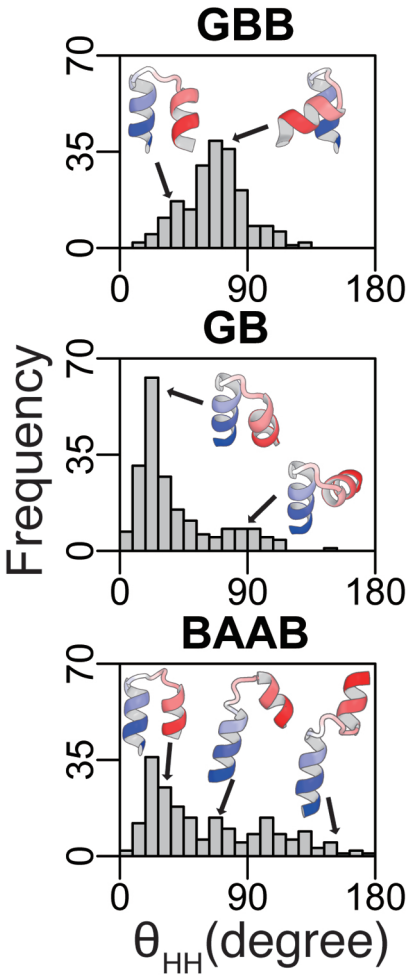

Figure 6 Distributions of helix-helix crossing angles in Rosetta-derived fragment library. GBB library shows a large peak at 90°, which corresponds to αα-corners. GB and BAAB libraries have the largest peak around 30°, which corresponds to the αα-hairpins. GBB fragment library is largely biased to the 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, I sought to identify whether they can be designed when their amino acid sequences are completely specified. Are they difficult to build again? I 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 (Figs. 7A and 7B). Similar to the ABEGO-based backbone-building simulations for the GBB up-down bundle, those for the

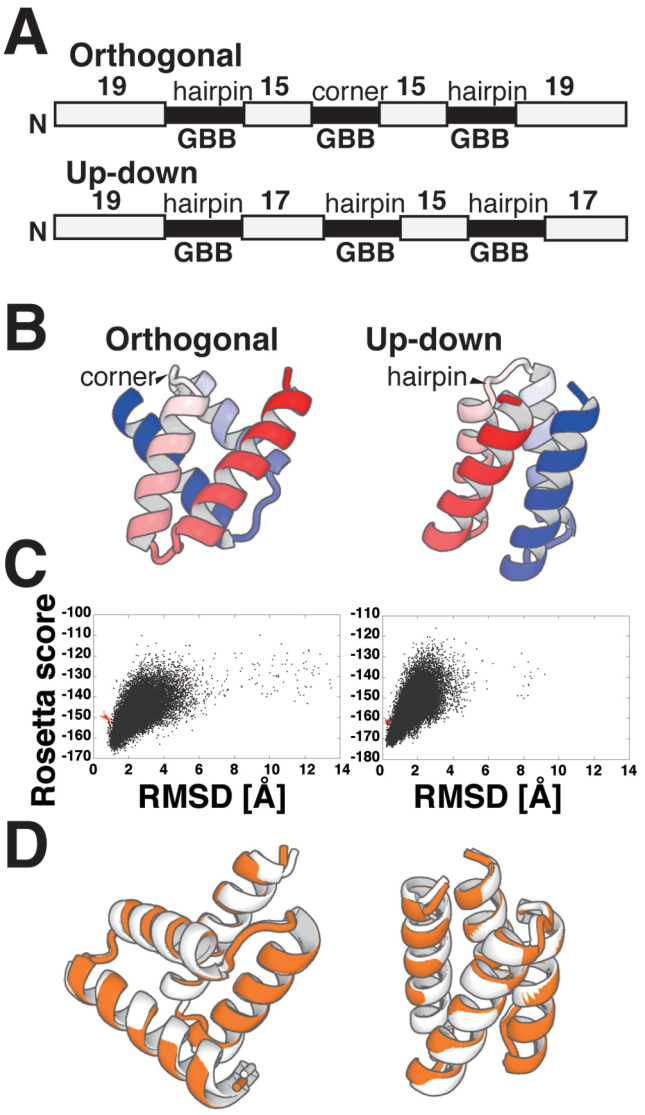

Figure 7 Design and sequence-dependent folding simulations of the four-helix orthogonal bundle and up-down bundles. (A) (B) Blueprints and structures of the GBB orthogonal bundles (left) and up-down bundles (right). The gray bars represent the α-helix and black bars represent loop regions. As all the 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.

GBB orthogonal bundle were also trapped in a misfolded state and showed low efficiency for achieving the target conformation (Supplementary Fig. S9), 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 (Figs. 7C and 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.

Conclusion

In this study, I 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 I 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 subsection 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; I 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.

A preliminary version of this work, DOI: 10.1101/2021.04.13.439694, was deposited in the bioRxiv on April 14, 2021.

Acknowledgments

K.S. would like to thank Dr. Shintaro Minami for providing a curated domain structure dataset and suggestion on structural-informatics analysis. K.S would like to thank the Koga laboratory for offering computational resources. Most of the computational analysis was performed using the facilities at the Research Center for Computational Science, Okazaki, Japan. K.S. was supported by a Grant-in-Aid for JSPS Fellows (grant number 15J02427). Additionally, K.S. would like to thank the Institute for Molecular Science for the financial support received as a research assistant during the doctoral course.

Conflicts of Interest

The author declares no conflicts of interest.

Author Contributions

K.S. designed the research, performed numerical experiments, analyzed data, and wrote manuscripts.

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. USA* **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. USA* **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. USA* **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 β barrels. *Science* **371**, eabc8182 (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. USA* **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**, e1003926 (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**, e20161 (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

(*Edited by Motonori Ota*)