molecules

MDPI

Article

The Structural Rule Distinguishing a Superfold: A Case Study of Ferredoxin Fold and the Reverse Ferredoxin Fold

Takumi Nishina [1], Megumi Nakajima [1], Masaki Sasai [1,2,3,*] and George Chikenji [1,*]

[1] Department of Applied Physics, Nagoya University, Nagoya 464-8601, Japan;
nishina@tbp.ap.pse.nagoya-u.ac.jp (T.N.); nakajima@tbp.cse.nagoya-u.ac.jp (M.N.)
[2] Department of Complex Systems Science, Nagoya University, Nagoya 464-8601, Japan
[3] Fukui Institute for Fundamental Chemistry, Kyoto University, Kyoto 606-8501, Japan
[*] Correspondence: masakisasai@nagoya-u.jp (M.S.); chikenji@tbp.ap.pse.nagoya-u.ac.jp (G.C.)

Abstract: Superfolds are folds commonly observed among evolutionarily unrelated multiple superfamilies of proteins. Since discovering superfolds almost two decades ago, structural rules distinguishing superfolds from the other ordinary folds have been explored but remained elusive. Here, we analyzed a typical superfold, the ferredoxin fold, and the fold which reverses the N to C terminus direction from the ferredoxin fold as a case study to find the rule to distinguish superfolds from the other folds. Though all the known structural characteristics for superfolds apply to both the ferredoxin fold and the reverse ferredoxin fold, the reverse fold has been found only in a single superfamily. The database analyses in the present study revealed the structural preferences of $\alpha\beta$- and $\beta\alpha$-units; the preferences separate two α-helices in the ferredoxin fold, preventing their collision and stabilizing the fold. In contrast, in the reverse ferredoxin fold, the preferences bring two helices near each other, inducing structural conflict. The Rosetta folding simulations suggested that the ferredoxin fold is physically much more realizable than the reverse ferredoxin fold. Therefore, we propose that minimal structural conflict or minimal frustration among secondary structures is the rule to distinguish a superfold from ordinary folds. Intriguingly, the database analyses revealed that a most stringent structural rule in proteins, the right-handedness of the $\beta\alpha\beta$-unit, is broken in a set of structures to prevent the frustration, suggesting the proposed rule of minimum frustration among secondary structural units is comparably strong as the right-handedness rule of the $\beta\alpha\beta$-unit.

Keywords: protein design; reverse fold; minimum frustration

check for updates

Citation: Nishina, T.; Nakajima, M.; Sasai, M.; Chikenji, G. The Structural Rule Distinguishing a Superfold: A Case Study of Ferredoxin Fold and the Reverse Ferredoxin Fold. *Molecules* **2022**, *27*, 3547. https://doi.org/10.3390/molecules27113547

Academic Editors: Kunihiro Kuwajima, Yuko Okamoto, Tuomas Knowles and Michele Vendruscolo

Received: 28 March 2022
Accepted: 28 May 2022
Published: 31 May 2022

1. Introduction

A principal goal of protein science is to elucidate the relationship among sequences, structures, and functions [1,2]. Toward such a goal, remarkable progress has been achieved in structure prediction from the knowledge of amino-acid sequences [3,4]. Also, in protein design, which is a reverse problem of structure prediction, elucidation of design principles [5–7] led to an increasing number of successful examples to find amino-acid sequences that can fold into the designed structures [5,6,8–12]. Here, for further advancing the design technology, it is crucial to develop a systematic method to distinguish less designable structures and highly designable ones into each of which a large number of different sequences can fold [13]. Investigating the occurrence of structural folds among natural proteins provides a clue to this problem [14–18]. An ordinary fold appears in only one or a few superfamilies, but a particular fold is shared by a large number of superfamilies; such a particular fold was called a superfold [19]. Here, a superfamily is defined as the largest group of proteins for which common ancestry can be inferred [20]. Superfolds are rare in the entire fold categories but are robust against mutations, suggesting superfolds represent highly designable structures. Each superfold corresponds to many different functions, in sharp contrast to the ordinary folds showing the nearly one-to-one correspondence between fold and function.

Since the discovery of superfolds [19], features distinguishing superfolds from the other ordinary folds have been explored, leading to the several empirical rules that characterize the superfolds, some of which are (1) frequent appearance of super secondary structures [21], (2) avoidance of mixing parallel and anti-parallel β-sheets [14], (3) infrequent jumps between β-strands [16], and (4) high structural symmetry [22]. However, examples of ordinary folds satisfy the rules from (1) through (4), showing the need for further rules to distinguish superfolds. The reverse ferredoxin fold is such an example. The ferredoxin fold, a typical superfold, comprises four β-strands connected in the order and directions as designated in Figure 1A. The reverse ferredoxin fold reverses the N to C terminus direction from the ferredoxin fold (Figure 1B). According to the SCOPe classification [23,24], the ferredoxin fold is found in 62 superfamilies, whereas the reverse ferredoxin fold is found only in one superfamily. Therefore, the reverse ferredoxin fold is not a superfold, but both the ferredoxin fold and the reverse ferredoxin fold satisfy the rules (1) through (4). Other examples show the significant difference between the fold and the reverse fold in the number of occurrences in the spectrum of families [15]. The reason for this difference between folds and reverse folds remains elusive; there have been arguments suggesting physical or functional necessities to avoid the reverse folds [15] and those suggesting the bias occasionally acquired in evolutionary history [25].

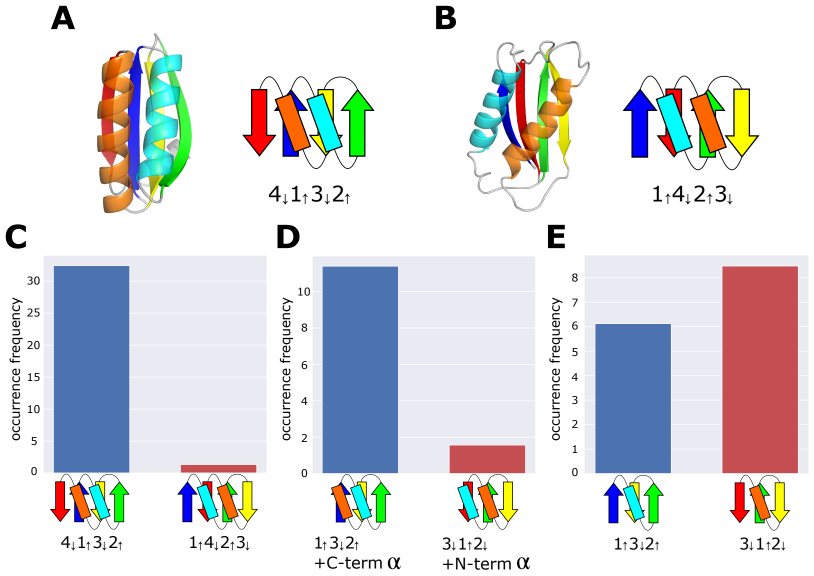

Figure 1. Topology and occurrence frequency of the ferredoxin fold and the reverse ferredoxin fold. (**A**) An example structure (a microcompartment protein, PDB ID: 4QIV) and the topology $4_{\downarrow}1_{\uparrow}3_{\downarrow}2_{\uparrow}$ of the ferredoxin fold. (**B**) An example structure (the catalytic core of human DNA polymerase kappa, PDB ID: 1T94) and the topology $1_{\uparrow}4_{\downarrow}2_{\uparrow}3_{\downarrow}$ of the reverse ferredoxin fold. (**C**) Occurrence frequency of the ferredoxin topology $4_{\downarrow}1_{\uparrow}3_{\downarrow}2_{\uparrow}$ and the reverse ferredoxin topology $1_{\uparrow}4_{\downarrow}2_{\uparrow}3_{\downarrow}$. (**D**) Occurrence frequency of the topology $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ + C-term α and the topology $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α. (**E**) Occurrence frequency of the topology $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ and the topology $3_{\downarrow}1_{\uparrow}2_{\downarrow}$. In (**C**–**E**), the dataset of the 99% sequence identity representatives derived from ECOD was used. Chains are colored from blue (N-terminus) to red (C-terminus). In the topology diagram, β-strands are represented with arrows and α-helices are rectangles.

Here, we explored the factor to distinguish superfolds from the ordinary folds by comparing the ferredoxin fold and the reverse ferredoxin fold as a case study. By analyzing the database, we found the structural tendency shown by the $\alpha\beta$-unit and $\beta\alpha$-unit, suggesting that the structure comprises multiple $\alpha\beta$- and $\beta\alpha$-units should satisfy a rule to minimize the conflict between structural tendencies of these units. We show that the ferredoxin fold satisfies this rule for minimal conflict or frustration, whereas the reverse ferredoxin fold does not. We also performed the Rosetta folding simulations to test the foldability of structures [5]; the test results suggested that the ferredoxin fold is physically much more realizable than the reverse ferredoxin fold. Thus, we propose that the minimum frustration rule to consistently satisfy the structural preference of multiple parts of the protein is a rule to distinguish superfolds from ordinary folds.

2. Results

2.1. Occurrence Frequency of Topologies

Previous analyses showed that the ferredoxin fold is frequently found, whereas the reverse ferredoxin fold is rare among protein families [17,25]. We confirmed this imbalance in the most recent version of a semi-manually curated database, ECOD (version 20210511: develop280), which hierarchically classifies protein domains according to homology, reflecting their evolutionary relationship [26]. ECOD has been frequently updated, suited to estimating the most recent number of homology groups having a topology on which we focus. The ECOD database classifies homologous protein domains according to categories of family and homology. The family (F) group consists of evolutionarily related protein domains with substantial sequence similarity, and the homology (H) group comprises multiple F-groups having functional and structural similarities. The H-group corresponds to the superfamily in the other structural databases, SCOP [27] and CATH [28]. The X-group in ECOD comprises multiple H-groups that share similar features in the structure but lack a convincing evidence for homology. In this study, we used the 99% sequence identity representatives in ECOD as the dataset for the analyses.

We detected secondary structures and hydrogen bonds in protein domains recorded in the dataset using STRIDE [29]. Then, based on the thus found hydrogen-bond pattern among β-strands, we defined the β-sheet topology as in Ref. [15]; we describe the β-sheet topology by representing the strand directions with up and down arrows with the sequential number from the N- to C-termini (4132, for example). Then, topology T of the ferredoxin fold is $T = 4_{\downarrow}1_{\uparrow}3_{\downarrow}2_{\uparrow}$ (Figure 1A) and topology T of the reverse ferredoxin fold is $T = 1_{\uparrow}4_{\downarrow}2_{\uparrow}3_{\downarrow}$ (Figure 1B).

We estimated the occurrence frequency $OF(T)$ of a given topology T by summing the occupation ratio $OR(T,i)$ of protein domains having T in the ith H-group as

$$OF(T) = \sum_{i=1}^{N_{\text{homology}}} OR(T,i), \tag{1}$$

where N_{homology} is the total number of H groups in the dataset, and

$$OR(T,i) = \frac{1}{N_{\text{family}}(i)} \sum_{j=1}^{N_{\text{family}}(i)} \frac{N_{\text{domain}}(T,i,j)}{N_{\text{domain}}(i,j)}. \tag{2}$$

Here, $N_{\text{domain}}(T,i,j)$ is the number of protein domains having topology T in the jth F-group, which belongs to the ith H-group in the dataset. $N_{\text{domain}}(i,j) = \sum_T N_{\text{domain}}(T,j)$ is the total number of protein domains in the jth F-group, and $N_{\text{family}}(i)$ is the number of F-groups in the ith H-group. Figure 1C shows that the occurrence frequency of the ferredoxin topology, $OF(4_{\downarrow}1_{\uparrow}3_{\downarrow}2_{\uparrow})$, is more than 10 times larger than the occurrence frequency of the reverse ferredoxin topology, $OF(1_{\uparrow}4_{\downarrow}2_{\uparrow}3_{\downarrow})$, confirming the previously reported ubiquity of the ferredoxin fold and the rareness of the reverse ferredoxin fold [17,25].

Here, we should note that topology has often been classified with ECOD in terms of X-groups; for example, an X-group called "alpha-beta plaits" has been regarded as the group representing the ferredoxin topology. However, we used STRIDE for a more precise topological classification instead of the X-group classification. Therefore, the $OF(T)$ defined in Equation (1) does not precisely correlate with the number of H-groups in the X-group. Tetracycline resistance protein, tetM (PDB ID: 3J25), for example, belongs to the X-group of alpha-beta plaits, but we did not count tetM as a ferredoxin-topology protein because STRIDE identifies only two β-strands in tetM. Similarly, surface-layer (S-layer) protein

(PDB ID: 3CVZ) belongs to the reverse ferredoxin X-group in ECOD, but we did not count S-layer protein as a protein with the reverse-ferredoxin fold because STRIDE identifies a topology $1_\uparrow 5_\uparrow 4_\downarrow 2_\uparrow 3_\downarrow$ for S-layer protein instead of $1_\uparrow 4_\downarrow 2_\uparrow 3_\downarrow$. See Supplementary Figure S1 for the structure of tetM and S-layer protein.

We examine the minimal structural units that induce the difference between $4_\downarrow 1_\uparrow 3_\downarrow 2_\uparrow$ and $1_\uparrow 4_\downarrow 2_\uparrow 3_\downarrow$. We consider the topology in which the C-terminal strand (β-strand 4) is deleted from the ferredoxin topology by retaining the α-helix connecting β-strands 4 and 3 in the structure, and write the thus obtained topology as $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$. We also consider the topology in which the C-term α is further deleted from $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$ and write such a topology as $1_\uparrow 3_\downarrow 2_\uparrow$. Similarly, we consider the topology in which the N-terminal strand (β-strand 1) is deleted from the reverse ferredoxin topology by retaining the α-helix connecting β-strands 1 and 2 in the structure. Then, we renumber the strands as $4, 2, 3 \to 3, 1, 2$, and write the thus-obtained topology as $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$, which is the reverse of $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$. We also consider the topology in which the N-term α is further deleted from $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$ and write such a topology as $3_\downarrow 1_\uparrow 2_\downarrow$, which is the reverse of $1_\uparrow 3_\downarrow 2_\uparrow$.

We consider protein domains whose entire (not the partial) structure has the topology $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$ or $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$, and calculated occurrence frequencies, $OF(1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha)$ and $OF(3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha)$ (Figure 1D). We should note that with the topology of $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$, the C-term α can lie on either side of the β-sheet plane. However, in the ferredoxin fold, this helix is always on the same side of the plane as the α-helix of the $\beta\alpha\beta$-unit consisting of β-strands 1 and 2; therefore, we here calculated $OF(1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha)$ for the structures in which the C-term α is on the same side of the plane as the α-helix of the $\beta\alpha\beta$-unit. Similarly, we calculated $OF(3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha)$ for structures in which the N-term α is on the same side of the β-sheet plane as the α-helix of the $\beta\alpha\beta$-unit consisting of β-strands 2 and 3. See the Materials and Methods section for the way to judge which side of the plane the terminal helix lies in a given structure in calculating OFs. Figure 1D shows that $OF(1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha)$ is significantly larger than $OF(3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha)$, suggesting that the determining structural factor distinguishing the ferredoxin fold and the reverse ferredoxin fold exists in the difference between $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$ and $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$. The population of the structures with two helices lying on the opposite side of the β-sheet plane is small in the $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$ topology and in the $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$ topology, and there is no significant difference between occurrence frequencies of two topologies for those structures with helices lying on the opposite side of the plane. The large difference between two topologies only appear for structures in which two helices lie on the same side of the plane (Supplementary Figures S2 and S3).

Similarly, we calculated occurrence frequencies, $OF(1_\uparrow 3_\downarrow 2_\uparrow)$ and $OF(3_\downarrow 1_\uparrow 2_\downarrow)$ (Figure 1E), showing that $OF(3_\downarrow 1_\uparrow 2_\downarrow)$ is mildly larger than $OF(1_\uparrow 3_\downarrow 2_\uparrow)$. These results suggest that the determinant structural factor that induces the difference between $4_\downarrow 1_\uparrow 3_\downarrow 2_\uparrow$ and $1_\uparrow 4_\downarrow 2_\uparrow 3_\downarrow$ is in the difference between $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$ and $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$. Addition of the C-term α-helix to $1_\uparrow 3_\downarrow 2_\uparrow$ and addition of the N-term α-helix to $3_\downarrow 1_\uparrow 2_\downarrow$ bring about the difference in the occurrence frequency between the ferredoxin topology and the reverse ferredoxin topology. Hereafter, the ferreoxin fold and the $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$ topology are referred to collectively as the ferredoxin-type topology, and the reverse ferredoxin fold and the $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$ topology are referred to collectively as the reverse ferredoxin-type topology.

2.2. Conflict between Structural Preferences of $\alpha\beta$- and $\beta\alpha$-Units

Because positions of the $\alpha\beta$- and $\beta\alpha$-units are different in $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$ and $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$ (Figure 1A,B), analyses on these structural units should give critical insights on the difference between $1_\uparrow 3_\downarrow 2_\uparrow + \text{C-term } \alpha$ and $3_\downarrow 1_\uparrow 2_\downarrow + \text{N-term } \alpha$. For the structural analyses of these units, we defined the distance x between the plane of the β-pleats in the strand and the α-helix (Figure 2A). See the Materials and Methods section for the precise definition of x. We derived the distribution of x by analyzing the dataset culled from PDB with constraints of the sequence identity $< 30\%$, the finer resolution than 2.0 Å, and the R-factor < 0.25 [30]. For the statistical analyses, we selected typical $\alpha\beta$- and $\beta\alpha$-units following the criterion of Ref. [31]; we used the structural units satisfying the conditions that the linker loop between α-helix and β-strand is shorter than five-residue length and the angle between α-helix and β-strand is less than $60\,^\circ$.

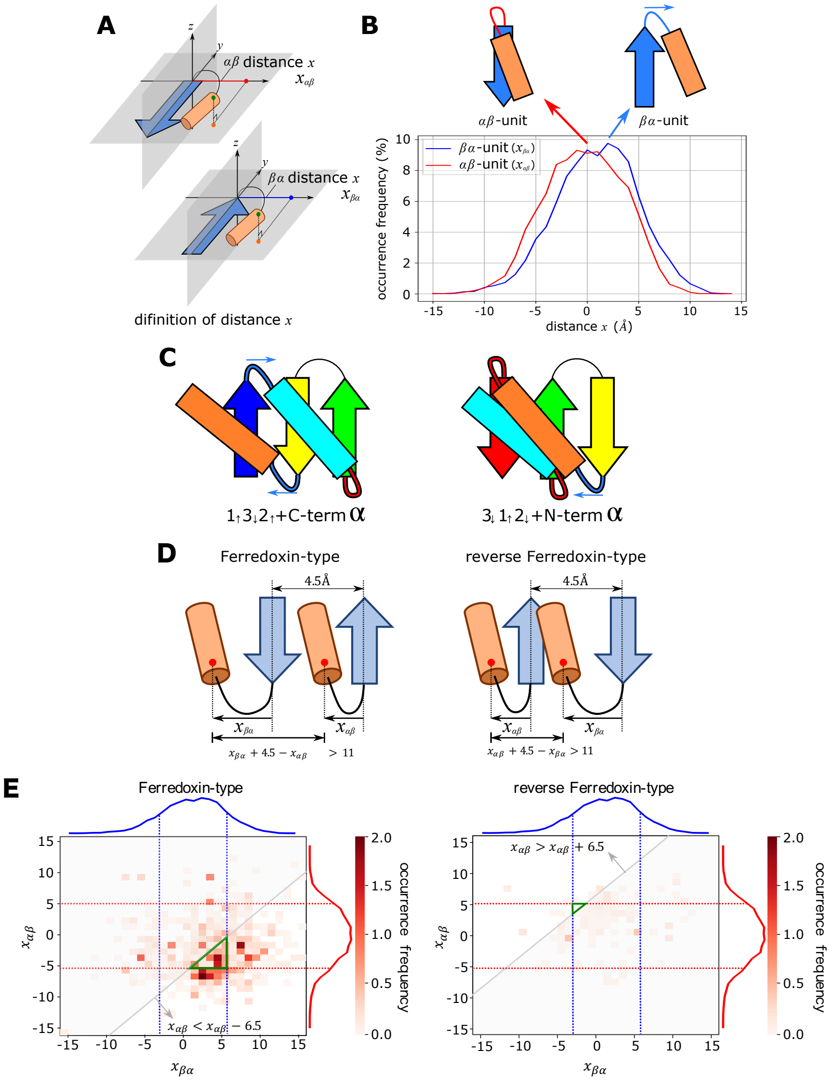

Figure 2. Absence or presence of the structural conflict between α-helices. (**A**) Definition of the distance x between the pleated plane of the β-strand and the α-helix in the $\alpha\beta$-unit (top) and the $\beta\alpha$-unit (bottom). (**B**) Distribution of x in the $\alpha\beta$-unit (red) and the $\beta\alpha$-unit (blue). The distribution was found in the culled PDB dataset with the parameters of the sequence identity $< 30\%$, the finer resolution than 2.0 Å, and the R-factor < 0.25. (**C**) Structural preferences of the the $\alpha\beta$-unit (connected by a red linker) and the $\beta\alpha$-unit (connected by a blue linker) prevent collision between the terminal helix and the helix in the $\beta\alpha\beta$ structure in the $1_\uparrow 3_\downarrow 2_\uparrow$ + C-term α topology (left), while they induce a collision in the $3_\downarrow 1_\uparrow 2_\downarrow$ + N-term α topology (right). Blue arrows show the shift of α-helix induced by the $x > 0$ preference of the $\beta\alpha$-unit. (**D**) The necessary condition to avoid the collision of two helices. $x_{\beta\alpha} - x_{\alpha\beta} + 4.5\,\text{Å} > 11\text{Å}$ for the ferredoxin-type topology and $x_{\alpha\beta} - x_{\beta\alpha} + 4.5\,\text{Å} > 11\text{Å}$ for the reverse ferredoxin-type topology. (**E**) The realizable area to avoid the collision and the occurrence frequency of $(x_{\beta\alpha}, x_{\alpha\beta})$ in the ECOD database. The realizable area satisfying the three conditions; the necessary condition to avoid the collision, the condition of the frequency $> 5\%$ in the $x_{\beta\alpha}$ distribution, and the condition of the frequency $> 5\%$ in the $x_{\alpha\beta}$ distribution; is shown with a green triangle on the $(x_{\beta\alpha}, x_{\alpha\beta})$ plane. The occurrence frequency shown with the gray-scale is superposed. Blue and red curves are distributions in (**B**).

Figure 2B shows the distribution of x obtained by the dataset analyses. The distribution of x in the $\beta\alpha$-unit peaked at $2{\sim}4$ Å, whereas the distribution of x in the $\alpha\beta$-unit peaked at ~ 0 Å, showing a distinct tendency of positive x in the $\beta\alpha$-unit. This positive x distribution implies the tendency of shifting the α-helix toward the direction of blue arrows in Figure 2C. In the $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ + C-term α structure, this shift separates the C-term α-helix from the helix in the $\beta\alpha\beta$ structure, while in the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α structure, the shift induces collision of the N-term α-helix against the helix in the $\beta\alpha\beta$ structure when two helices are on the same side of the β-sheet surface. Therefore, the structural conflict arising between two helices destabilizes the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α structure; and hence, destabilizes the reverse ferredoxin fold.

We can quantitatively assess how the difference in the distribution of the distance x in Figure 2B determines the absence/presence of the structural conflict. We write x in the $\beta\alpha$-unit and the $\alpha\beta$-unit as $x_{\beta\alpha}$ and $x_{\alpha\beta}$, respectively. Considering that a typical distance between two adjacent β-strands in a β-sheet is 4.5 Å [32], the distance between two helices in the ferredoxin-type topology is $x_{\beta\alpha} - x_{\alpha\beta} + 4.5$ Å. Similarly, the distance between two helices in the reverse ferredoxin-type topology is $x_{\alpha\beta} - x_{\beta\alpha} + 4.5$ Å (Figure 2D). Because the helix diameter is approximately 11.0 Å [33], the necessary condition to avoid the collision of two helices is $x_{\beta\alpha} - x_{\alpha\beta} + 4.5\,\text{Å} > 11\text{Å}$ for the ferredoxin-type topology and $x_{\alpha\beta} - x_{\beta\alpha} + 4.5\,\text{Å} > 11\text{Å}$ for the reverse ferredoxin-type topology. In Figure 2E, the region satisfying three conditions at the same time is designated by a green triangle on a two-dimensional plane of $x_{\beta\alpha}$ and $x_{\alpha\beta}$: (i) the necessary condition to avoid the collision, (ii) the condition of frequency $> 5\,\%$ in the frequency distribution of $x_{\beta\alpha}$ in Figure 2B, and (iii) the condition of frequency $> 5\,\%$ in the frequency distribution of $x_{\alpha\beta}$ in Figure 2B. The thus-defined green triangle, i.e., the realizable area to avoid the collision, is extremely narrow in the reverse ferredoxin-type topology, whereas it is wide in the ferredoxin-type topology. Figure 2E shows that the occurrence frequency of $(x_{\beta\alpha}, x_{\alpha\beta})$ in the ECOD database is large around the green triangle in the ferredoxin-type fold, while the frequency is small everywhere on the plane of $(x_{\beta\alpha}, x_{\alpha\beta})$ in the reverse ferredoxin-type fold. Thus, the shift of $2{\sim}4$ Å in distributions in Figure 2B is a determining factor for the realizability of the structure. In the reverse ferredoxin-type topology, the structures are realized by breaking at least one of three conditions (i)–(iii). Different ways of breaking the conditions in the reverse ferredoxin-type topology make the distribution scattered on the $(x_{\beta\alpha}, x_{\alpha\beta})$ plane in Figure 2E. Supplementary Figure S4 shows example proteins with the reverse ferredoxin topology showing uncommon configuration of the $\beta\alpha$- or $\alpha\beta$-unit.

We should note that the results shown in Figure 2B,E are the plots for proteins with loops shorter than five-residue length. The longer loops allow the structural variety to obscure the realizability conditions in Figure 2B,E. However, the stability of native structures inversely correlates to the loop length [34,35], making the proteins having the longer loops rare. See Supplementary Figure S5 for the distribution of the loop length found in the ECOD database. Here, it is sufficient to consider non-rare proteins with short enough loops for clarifying how the ferredoxin-type topology is much more realizable than the reverse ferredoxin-type topology.

2.3. Minimum Frustration Rule

The dataset analyses showed that the structural preference of $\alpha\beta$- and $\beta\alpha$-units leads to the structural conflict in the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α structure, while the conflict is avoided in the $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ + C-term α structure. We examined the effect of presence/absence of the structural conflict by performing the Rosetta folding simulations. In these simulations, we substituted all the residues in the model to Valine, and assembled the fragments of one-, three-, or nine-residue length, which have the compatible main-chain dihedral angles with the secondary structures in the blueprints designated in Figure 3. We used the all-Valine sequence to focus on the role of structural consistency among the assembled fragments instead of the effects of the residue-specific interactions. We regard structures generated through the simulations as compatible structures when they have low energy and the same topology as the blueprint. For each blueprint, we performed the fragment-assembly simulation 10,000 times and counted how many compatible structures were obtained through simulations. Koga et al. showed that the topology designated by the blueprint is physically realizable by avoiding the structural conflict when the number of the obtained compatible structures is large, while it is physically unrealizable with the structural inconsistency when the number is small [5]. See the Materials and Methods section for the details of the simulations.

Figure 3A shows the number of structures compatible with the $3_{\downarrow}1_{\uparrow}2_{\downarrow} +$ N-term α topology and the number of structures compatible with the $1_{\uparrow}3_{\downarrow}2_{\uparrow} +$ C-term α topology. The compatible structures were 229 and 10 for the $1_{\uparrow}3_{\downarrow}2_{\uparrow} +$ C-term α topology and the $3_{\downarrow}1_{\uparrow}2_{\downarrow} +$ N-term α topology, respectively, showing the $1_{\uparrow}3_{\downarrow}2_{\uparrow} +$ C-term α topology is much more realizable than the $3_{\downarrow}1_{\uparrow}2_{\downarrow} +$ N-term α topology. We performed the same test for the $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ topology and the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ topology. Figure 3B shows that the number of compatible structures for the $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ topology is almost same as the number of compatible structures for the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ topology, indicating that there is no significant difference between the realizability of these topologies. Figure 3A,B are qualitatively same as Figure 1D,E, showing that the difference in the realizability of the $1_{\uparrow}3_{\downarrow}2_{\uparrow} +$ C-term α topology and the $3_{\downarrow}1_{\uparrow}2_{\downarrow} +$ N-term α topology arises from absence/presence of the conflict between local structural units.

Combined analyses of databases and Rosetta folding simulations showed that the structural conflict or frustration is minimized in the largely realizable topology, which characterizes the superfold; therefore, we propose that the minimum frustration among local preferences of secondary structures is the rule to distinguish a superfold from the ordinary folds.

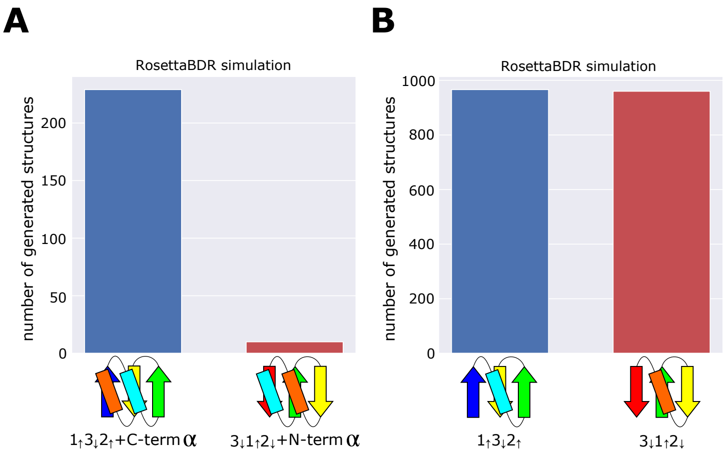

Figure 3. The number of simulated structures compatible with the blueprint. We repeated the Rosetta folding simulations 10,000 times and counted the number of compatible structures generated. (**A**) Comparison between the $1_{\uparrow}3_{\downarrow}2_{\uparrow} +$ C-term α topology and the $3_{\downarrow}1_{\uparrow}2_{\downarrow} +$ N-term α topology. In simulations, the number of structures in which two helices lie on the same side of the β-sheet surface was counted. (**B**) Comparison between the $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ topology and the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ topology.

3. Discussion

In this study, we proposed a rule that the minimum frustration among local structural preferences of secondary structures is the necessary condition for superfolds. In this section, we discuss the meaning of this rule by explaining how the rule predicts occurrence frequency of other structures, the relation of the rule with the other design rule, and the relation with protein function.

3.1. Occurrence Frequency of Other Structures

The present analyses of the ferredoxin fold and the reverse ferredoxin fold showed that the frequently occurring topology is designed to minimize frustration among multiple secondary-structure units that lie near each other on the same side the β-sheet plane. We can examine whether this rule predicts the occurrence frequency of other structures in the dataset. Figure 4A–D are four examples of pairs of topologies; in each pair, one is the topology minimizing frustration, and the other is its reverse topology exhibiting frustration. We should note that pairs in Figure 4B–D have the same arrangement of β-strands but have different connections of terminal α-helices showing different topologies. Our rule of minimum frustration predicts that the topology shown on the left side in each pair in Figure 4 is more realizable than the topology on the right side. We counted the occurrence frequency of these topologies in the dataset and found a significant difference as expected. In particular, we found the zero occurrence frequency of the frustrated topology in Figure 4D. The absence of this topology is reasonable because the frustrated topology of Figure 4D has two positions of structural collisions between helices, whereas the other frustrated topologies in Figure 4A–C show only a single collision in each. These results support our proposal that the minimum frustration among secondary structures is the requirement for the frequently occurring topologies; therefore, the necessary condition for the superfolds.

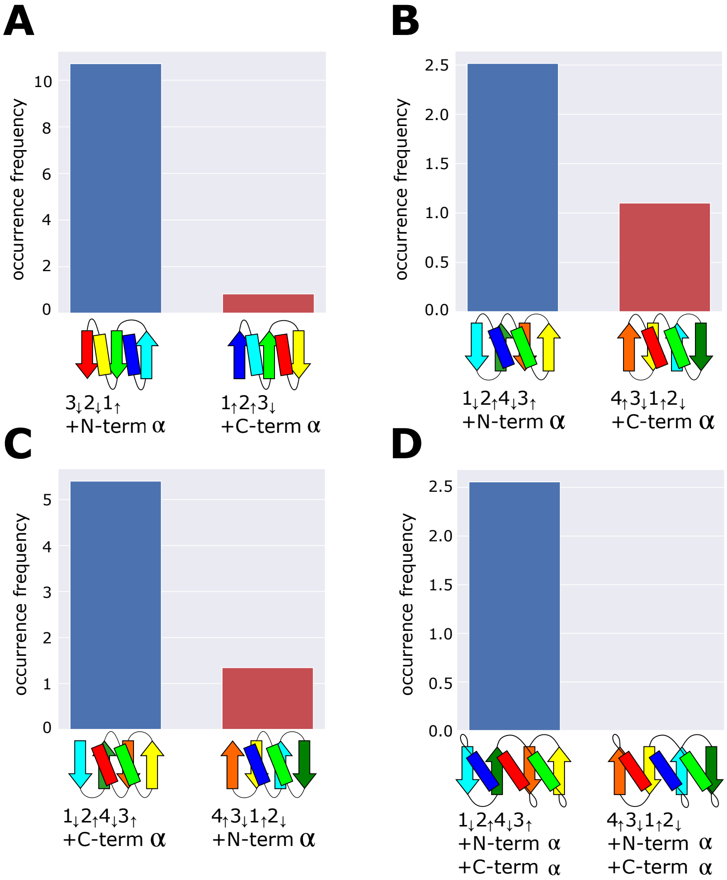

Figure 4. Comparisons of occurrence frequency between topologies minimizing frustration and their reverse topologies exhibiting frustration. (**A**) $3_{\downarrow}2_{\downarrow}1_{\uparrow}$ + N-term α and $1_{\uparrow}2_{\uparrow}3_{\downarrow}$ + C-term α, (**B**) $1_{\downarrow}2_{\uparrow}4_{\downarrow}3_{\uparrow}$ + N-term α and $4_{\uparrow}3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + C-term α, (**C**) $1_{\downarrow}2_{\uparrow}4_{\downarrow}3_{\uparrow}$ + C-term α and $4_{\uparrow}3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α, and (**D**) $1_{\downarrow}2_{\uparrow}4_{\downarrow}3_{\uparrow}$ + N-term α + C-term α and $4_{\uparrow}3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α + C-term α. The dataset was the 99% sequence identity representatives derived from the ECOD database.

3.2. The Left-Handed $\beta\alpha\beta$-Unit Is Selectively Found in the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-termα Structures

We showed that the collision between two helices arising from the structural preference of nearby $\alpha\beta$- and $\beta\alpha$-units decreases the occurrence frequency of the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α topology. However, this collision disappears when the two helices lie on the opposite side of the β-sheet surface. Such configurations are possible in two different ways. One is the configuration that the $\beta\alpha\beta$-unit consisting of β-strands 2 and 3 is right-handed and the terminal helix is on the opposite side; we have a small number of such examples in the dataset as shown in Supplementary Figure S3. The other is the configuration that the $\beta\alpha\beta$-unit is left-handed with the terminal helix in the position similar to that in the reverse ferredoxin fold. Here, we cannot expect the frequent occurrence of the latter structure because more than 98% of the known $\beta\alpha\beta$-unit structures are right-handed [14,36–38]. Indeed, in our dataset derived from ECOD, there is no left-handed $\beta\alpha\beta$-unit in protein domains with the $1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha$ or the $3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha$ topology.

However, in the dataset, we found a small number of left-handed $\beta\alpha\beta$-units in protein domains having the extended structures including $1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha$ or $3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha$ as a partial structure (Figure 5B,C). See the Materials and Methods section for the method to detect the left-handed $\beta\alpha\beta$-unit in the dataset. Figure 5A shows occurrence frequencies of domains in the dataset having more than four β-strands and include the $1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha$ or the $3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha$ topology as their partial structure. For these extended domains, we counted occurrence frequencies separately for those having a left-handed $\beta\alpha\beta$-unit, $OF(\text{Extended-}1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha; \text{Left})$ and $OF(\text{Extended-}3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha; \text{Left})$, and for those having the right-handed $\beta\alpha\beta$-unit, $OF(\text{Extended-}1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha; \text{Right})$ and $OF(\text{Extended-}3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha; \text{Right})$. We found $OF(\text{Extended-}1_{\uparrow}3_{\downarrow}2_{\uparrow}\text{+C-term } \alpha;$ Right) = 73.8, $OF(\text{Extended-}1_{\uparrow}3_{\downarrow}2_{\uparrow}\text{+C-term } \alpha;$ Left) = 0.5, $OF(\text{Extended-}3_{\downarrow}1_{\uparrow}2_{\downarrow}\text{+N-term } \alpha;$ Right) = 16.0, and $OF(\text{Extended-}3_{\downarrow}1_{\uparrow}2_{\downarrow}\text{+N-term } \alpha;$ Left) = 2.5, leading to the ratios,

$$\begin{aligned} \frac{OF(\text{Extended-}1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha; \text{Left})}{OF(\text{Extended-}1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha; \text{Right})} &\approx 0.0068, \\ \frac{OF(\text{Extended-}3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha; \text{Left})}{OF(\text{Extended-}3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha; \text{Right})} &\approx 0.156, \end{aligned} \tag{3}$$

suggesting that some mechanism exists for enhancing the occurrence of the left-handed $\beta\alpha\beta$-unit in the $3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha$ structure. A plausible explanation is that the left-handed $\beta\alpha\beta$-unit was chosen in these domains to avoid the collision between two helices lying on the same side of the β-sheet in the $\text{Extended-}3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha$ structures. This mechanism suggests that the rule for minimizing frustration between the structural preferences of secondary structures lying nearby on the same side of the β-sheet is comparably strong as the rule of the right-handedness of the $\beta\alpha\beta$-unit.

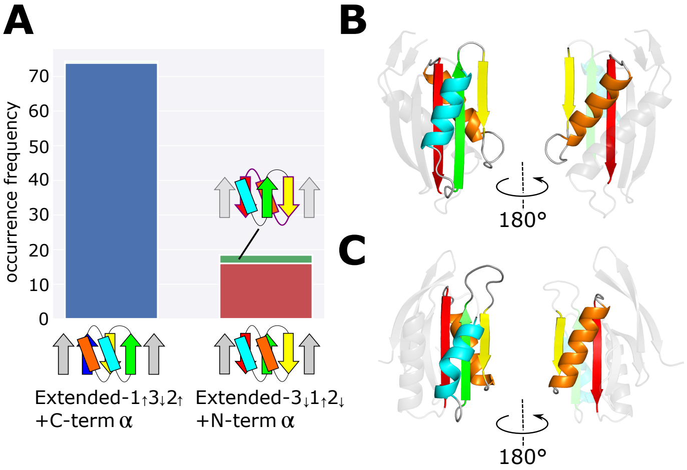

Figure 5. Occurrence of the left-handed and right-handed $\beta\alpha\beta$-units in the extended domains which include the $1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha$ or $3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha$ structure. (**A**) Comparison between occurrence frequencies of extended domains that include the $1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha$ or $3_{\downarrow}1_{\uparrow}2_{\downarrow} + \text{N-term } \alpha$ as the partial structure. The occurrence frequency of the extended $1_{\uparrow}3_{\downarrow}2_{\uparrow} + \text{C-term } \alpha$ is 74.3 among which the occurrence frequency of structures having the left-handed $\beta\alpha\beta$-unit is 0.5 (invisible in the figure).

The occurrence frequency of the extended $3_{\downarrow}1_{\uparrow}2_{\downarrow} +$ N-term α structure is 18.5 among which the occurrence frequency of structures having the left-handed $\beta\alpha\beta$-unit is 2.5 (green). (**B**,**C**) Examples of the extended $3_{\downarrow}1_{\uparrow}2_{\downarrow} +$ N-term α domains having the left-handed $\beta\alpha\beta$-unit. (**B**) PDB ID: 2CVE. (**C**) PDB ID: 1RLH.

3.3. Frustration and Function

A remaining critical question is the reason for the existence of protein domains having the reverse ferredoxin topology. Because proteins have evolved not for their stability but their functions, a possible explanation is that frustrated structures are necessary for their functions. Roles of frustration in functions have been discussed with theoretical methods by inferring the local degree of frustration using the coarse-grained energy function of protein conformation [39]. By computationally perturbing the sequence or configuration of a local part of the protein, the local part was regarded as less frustrated when most of the perturbations increase the calculated free energy significantly, while the local part was regarded as frustrated when the free energy change upon perturbations is insignificant [40]. It was shown that the local frustration can guide thermal motions [41] and specific associations [42], suggesting the positive role of frustration in protein functioning.

In this study, we proposed a new definition of frustration as the conflict between structural preferences of local parts of the protein. This definition of frustration should shed further light on the role of frustration. The frustrating interaction between helices in the reverse ferredoxin fold destabilizes the structure. This tendency may be compensated for by a specific residue design to stabilize the fold, or the protein may utilize the tendency to enhance the fluctuation and facilitate the structural change, which is needed for its functioning. An example shown in Figure 1B was the catalytic core of human DNA polymerase kappa. Because the sizeable structural change is necessary for activating a molecular motor motion of DNA polymerase, we can expect that the frustration in this structure helps function DNA polymerase.

The definition of frustration introduced in this study, the structural conflict among the local parts' structural preferences, provides a new perspective to the frustration-function relationship. In particular, the hypothesis proposed in this subsection suggests an intriguing possibility that the designed incorporation of frustration in the structure helps design the protein whose function is related to mobility with the significant structural change. To test this hypothesis, the dynamics and stability of the frustrated proteins and the specific design of sequences to fold the frustrated structures should be examined with further direct and systematic methods.

4. Materials and Methods

4.1. Detecting the Position of the C/N Terminal α-Helix

We explain in Figure 6 the method to judge on which side of the β-sheet plane the C or N-terminal α-helix lies in protein domains. We defined three vectors, **a**, **b**, and **c** in the $1_{\uparrow}3_{\downarrow}2_{\uparrow} +$ C-term α (Figure 6A) and $3_{\downarrow}1_{\uparrow}2_{\downarrow} +$ N-term α (Figure 6B) structures. The terminal α-helix is on the upper side of the β-sheet plane of Figure 6 if $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} > 0$ and the helix is on the lower side of the plane if $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} < 0$.

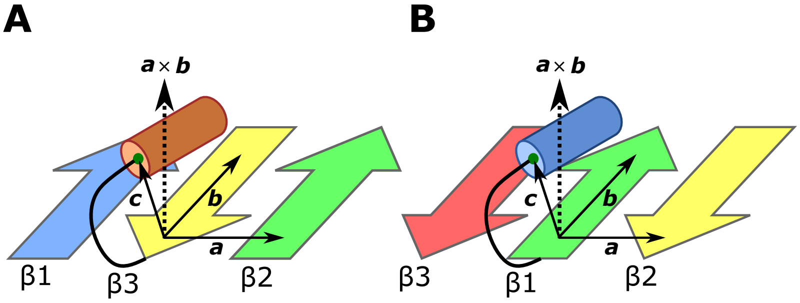

Figure 6. The method to judge on which side of the β-sheet the C or N-terminal α-helix lies. We defined three vectors, **a**, **b**, and **c**. The helix lies on the upper side of the β-sheet plane if $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} > 0$ and the helix lies on the lower side of the plane if $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} < 0$. (**A**) In the $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ + C-term α structure, the vector **a** is a vector extending from the Cα atom of the C-terminal residue of the β-strand 3 (yellow arrow) to the Cα atom of the N-terminal residue of the β-strand 2 (green arrow). The vector **b** is a vector extending from the Cα atom of the C-terminal residue of the β-strand 3 to the Cα atom of the second residue before the C-terminal residue of the β-strand 3. The vector **c** is a vector extending from the Cα atom of the C-terminal residue of the β-strand 3 to the center of mass (green dot) of Cα atoms of four N-terminal residues of the α-helix (orange cylinder). (**B**) In the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α structure, the vector **a** is a vector extending from the Cα atom of the N-terminal residue of the β-strand 1 (green arrow) to the Cα atom of the C-terminal residue of the β-strand 2 (yellow arrow). The vector **b** is a vector extending from the Cα atom of the N-terminal residue of the β-strand 1 to the Cα atom of the second residue after the N-terminal residue of the β-strand 1. The vector **c** is a vector extending from the Cα atom of the N-terminal residue of the β-strand 1 to the center of mass (green dot) of Cα atoms of four C-terminal residues of the α-helix (blue cylinder).

4.2. Definition of the Distance x between the Plane of β-Pleats and the α-Helix in the αβ- or βα-Unit

We measured the distance x between the plane of β-pleats and the α-helix in the $\alpha\beta$- and $\beta\alpha$-units by introducing a xyz-coordinate system in each unit (Figure 7). For defining the coordinate system, we set the direction of the y-axis parallel to the β-strand axis, and set the y-z plane parallel to the plane defined by the terminal three Cα atoms of the β-strand. We set the direction of the z-axis so as to place the helix on the $z > 0$ side. This idea of the coordinate system can be described in a precise way by defining the basis vectors, $\overrightarrow{e_x}$, $\overrightarrow{e_y}$, and $\overrightarrow{e_z}$, of the xyz-coordinate system with $\overrightarrow{e_z}$ being $\overrightarrow{e_z} = \overrightarrow{e_x} \times \overrightarrow{e_y}$.

We defined $\overrightarrow{e_x}$ and $\overrightarrow{e_y}$ as in the following way. Let i be the number of the terminal residue of the β-strand (the C-terminal residue in the $\beta\alpha$-unit and the N-terminal residue in the $\alpha\beta$-unit) and $C\alpha_i$ be the position of the ith Cα atom. We defined $\overrightarrow{e_x}$ by categorizing the $\beta\alpha$- or $\alpha\beta$-unit into two types, the parallel and antiparallel unit (Figure 7A,B). Then, we defined $\overrightarrow{e_x}$ as a normalized vector having the direction, which places both the starting and ending points of the α-helix on the coordinate of $x > 0$;

$$\overrightarrow{e_x} \parallel \begin{cases} \overrightarrow{C\alpha_{i-2}C\alpha_{i-1}} \times \overrightarrow{C\alpha_{i-1}C\alpha_i} & \text{(parallel } \beta\alpha\text{-unit)}, \\ \overrightarrow{C\alpha_i C\alpha_{i-1}} \times \overrightarrow{C\alpha_{i-1}C\alpha_{i-2}} & \text{(antiparallel } \beta\alpha\text{-unit)}, \\ \overrightarrow{C\alpha_i C\alpha_{i+1}} \times \overrightarrow{C\alpha_{i+1}C\alpha_{i+2}} & \text{(parallel } \alpha\beta\text{-unit)}, \\ \overrightarrow{C\alpha_{i+2}C\alpha_{i+1}} \times \overrightarrow{C\alpha_{i+1}C\alpha_i} & \text{(antiparallel } \alpha\beta\text{-unit)}, \end{cases} \tag{4}$$

and $\overrightarrow{e_y}$ is a normalized vector, whose direction is

$$\overrightarrow{e_y} \parallel \begin{cases} \overrightarrow{C\alpha_{i-2}C\alpha_i} & (\beta\alpha\text{-unit}), \\ \overrightarrow{C\alpha_{i+2}C\alpha_i} & (\alpha\beta\text{-unit}). \end{cases} \tag{5}$$

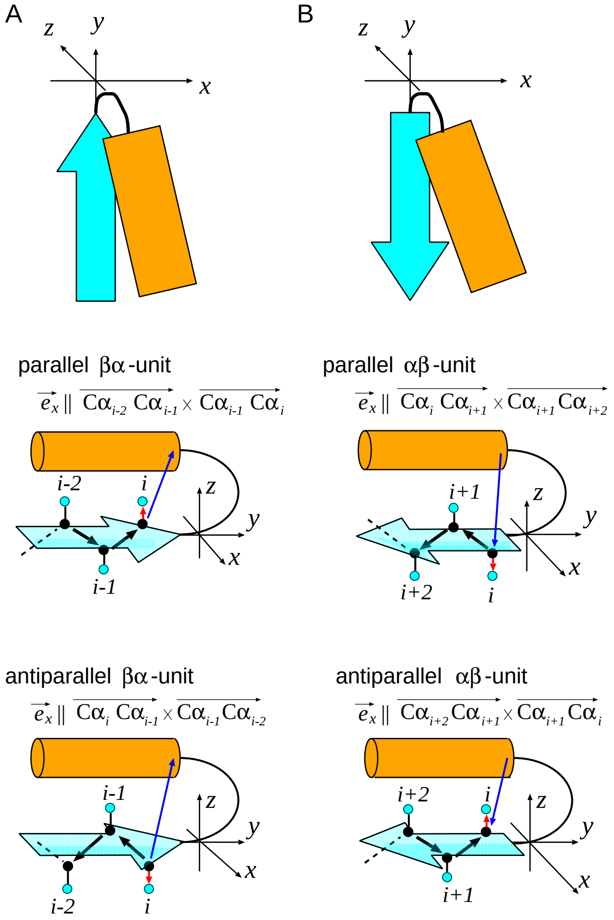
A
z
y
x
B
z
y
x
parallel βα-unit
$\overrightarrow{e_x} \parallel \overrightarrow{C\alpha_{i-2}\,C\alpha_{i-1}} \times \overrightarrow{C\alpha_{i-1}\,C\alpha_{i}}$
i-2
i
z
y
x
i-1
parallel αβ-unit
$\overrightarrow{e_x} \parallel \overrightarrow{C\alpha_{i}\,C\alpha_{i+1}} \times \overrightarrow{C\alpha_{i+1}C\alpha_{i+2}}$
i+1
z
y
x
i+2
i
antiparallel βα-unit
$\overrightarrow{e_x} \parallel \overrightarrow{C\alpha_{i}\,C\alpha_{i-1}} \times \overrightarrow{C\alpha_{i-1}C\alpha_{i-2}}$
i-1
z
y
x
i-2
i
antiparallel αβ-unit
$\overrightarrow{e_x} \parallel \overrightarrow{C\alpha_{i+2}C\alpha_{i+1}} \times \overrightarrow{C\alpha_{i+1}\,C\alpha_{i}}$
i+2
i
z
y
x
i+1

Figure 7. The xyz-coordinate system to define the distance x between the plane of β-pleats and the α-helix. (**A**) The $\beta\alpha$-unit and (**B**) the $\alpha\beta$-unit. These units consist of a β-strand (cyan arrow) and an α-helix (orange rectangle). Top panels represent the rough sketch of the coordinate system. Middle and bottom panels show Cα atoms (black dots), Cβ atoms (cyan dots), a vector spanning from the Cα to the Cβ of the terminal residue of the β-strand (i.e., the residue in the strand nearest to the helix) in each unit (red arrow), and a vector spanning from the Cα of the terminal residue of the β-strand to the center of mass of terminal four residues of the α-helix (i.e., four residues in the helix nearest to the strand) in each unit. Unit is referred to as "parallel" when the inner product of red and blue arrows is positive, and as "antiparallel" when the inner product is negative.

4.3. Rosetta Folding Simulations

We performed the Rosetta folding simulations to test the realizability of the blueprint structures. Here, Rosetta is a software suite that includes algorithms for macromolecular modeling, docking, protein design, etc [43]. Among the many algorithms included in the Rosetta software, we used the Rosetta BluePrintBDR protocol [43] for folding simulations. With this protocol, we performed the folding simulations by assembling one, three, or nine-residue length fragments so as to make the assembled structure compatible with a "blueprint", which describes the length of the secondary structure elements, strand pairings, and backbone torsion ranges for each residue. In thsese simulations, the main chain was represented by N, NH, Cα, C, and CO, and the side chain was represented by a sphere using the centroid model of Rosetta. We used the simulated annealing method to search for low-energy structures, and recorded the last structure of each simulated annealing run as a compatible structure only when the structure met the conditions specified in the blueprint.

As in models of Ref. [44], we represented all the residues as Valine, and used the same energy parameters as in Ref. [44]. The use of the poly-Valine sequence is because our purpose is to determine whether the phenomena observed in the database are explained by backbone properties rather than by the sequence-specific properties. Valine is the smallest and strongest hydrophobic amino acid, which suits this purpose, as shown in Ref. [5]. Figure 8 shows the blueprints we used in the BluePrintBDR protocol. In these blueprints, we used the same length of secondary structures and loops as optimized in Ref. [44]. The purpose of the present Rosetta simulations is to analyze the statistical tendency among different topologies. Because loops in each topology are shorter than five-residue length in most folds, and their distribution is peaked at around the two- to three-residue length (Figure S5), it is sufficient to use the short loops in the blueprints. Here, for the computational efficiency, we restricted ourselves to the loops with two- to three residue length for $\beta\alpha$- and $\alpha\beta$-loops. For β-hairpin loops, we assumed that loop consists of two, four, or five residues in the blueprints because the two or five-residue length is necessary for keeping the chirality rule of the hairpin loop [5] (Figure 8).

In the folding simulations, we did not impose the ABEGO constraint on the loop regions, but we imposed the constraint on the secondary structure regions by making the dihedral angles of the main chain in these regions fall into the ABEGO classes compatible with the secondary structures designated by the bluprint. Here, the ABEGO classification is a coarse-grained representation of the dihedral angles, specifying the regions in a Ramachandran plot with the alphabetic symbols: A, B, E, G, and O denote the right-handed α-helix region, right-handed β-strand region, left-handed β-strand region, left-handed helix region, and the cis peptide conformation, respectively [45].

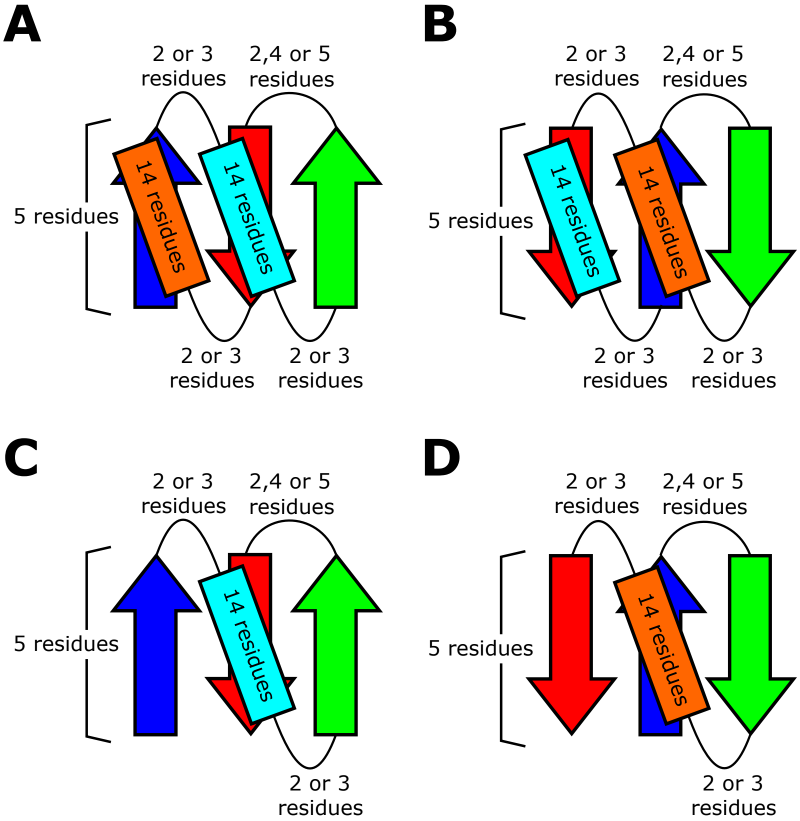

Figure 8. Blueprints used in the Rosetta folding simulations. The blueprints are represented by β-strands (arrows), α-helices (rectangles), and loops (curved lines). Blueprints of (**A**) the $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ + C-term α topology, (**B**) the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ + N-term α topology, (**C**) the $1_{\uparrow}3_{\downarrow}2_{\uparrow}$ topology, and (**D**) the $3_{\downarrow}1_{\uparrow}2_{\downarrow}$ topology.

4.4. Score to Detect the Left-Handed $\beta\alpha\beta$-Unit

We detected protein domains having the left-handed $\beta\alpha\beta$-unit by calculating the score of the left-handedness (*L-score*) . Here, for defining the *L-score*, we consider a $\beta\alpha\beta$-unit exemplified in Figure 9A. We refer to the N-terminal β-strand in the $\beta\alpha\beta$-unit as $\beta1$, and the C-terminal β-strand as $\beta 2$. We should note that the following *L-score* is applicable to evaluating the left-handedness of structures in which $\beta 1$ and $\beta 2$ are not connected directly to each other by hydrogen bonds, but multiple β-strands intervene between $\beta 1$ and $\beta 2$. We write the residue length of $\beta 1$, $\beta 2$, and the linker part connecting $\beta 1$ and $\beta 2$ as n, m, and l, respectively. We label the residues in those parts as $(N_1, N_2, \cdots, N_n)$, $(C_1, C_2, \cdots, C_m)$, and $(L_1, L_2, \cdots, L_l)$.

We define the residue number $C_{\max}(N_i, N_{i+1})$ so as to maximize the peak angle in Figure 9B when the residues N_i and N_{i+1} are given. Similarly, we define the residue number $N_{\max}(C_j, C_{j+1})$ to maximize the peak angle;

$$\begin{aligned} C_{\max}(N_i, N_{i+1}) &= \arg\max_{C_j}\left[\angle C\alpha_{N_i} C\alpha_{C_j} C\alpha_{N_{i+1}}\right], \\ N_{\max}(C_j, C_{j+1}) &= \arg\max_{N_i}\left[\angle C\alpha_{C_j} C\alpha_{N_i} C\alpha_{C_{j+1}}\right]. \end{aligned} \tag{6}$$

Then, using the Heaviside function, $H[x] = 1$ for $x > 0$ and $H[x] = 0$ for $x \leq 0$, the *L-score* is defined as

$$\begin{aligned} L\text{-}score &= \frac{1}{[(n-1)+(m-1)]\cdot l}\sum_{k=1}^{l}\left[\sum_{i=1}^{n-1} H\left[\left(\overrightarrow{C\alpha_{N_i} C\alpha_{N_{i+1}}} \times \overrightarrow{C\alpha_{N_i} C\alpha_{C_{\max}(N_i,N_{i+1})}}\right)\cdot \overrightarrow{C\alpha_{N_i} C\alpha_{L_k}}\right]\right. \\ &+ \left.\sum_{j=1}^{m-1} H\left[\left(\overrightarrow{C\alpha_{C_{j+1}} C\alpha_{C_j}} \times \overrightarrow{C\alpha_{C_j} C\alpha_{N_{\max}(C_j,C_{j+1})}}\right)\cdot \overrightarrow{C\alpha_{C_j} C\alpha_{L_k}}\right]\right]. \end{aligned} \tag{7}$$

The *L-score* ranges from 0 to 1 (Figure 9C). The higher the score, the more left-handed the $\beta\alpha\beta$-unit becomes. We judged the unit is left-handed when *L-score* ≥ 0.6.

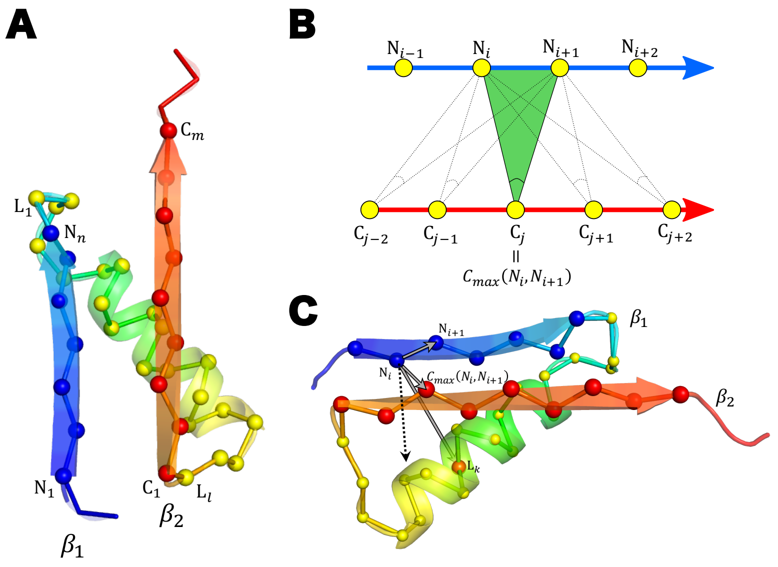

Figure 9. Calculation of the left-handedness score, *L-score*. (**A**) An example left-handed $\beta\alpha\beta$-unit. The cartoon representation and the backbone representation of the main chain are superposed. Cα atoms are drawn with spheres in the backbone representation. The first and the last residue numbers of $\beta 1$, $\beta 2$, and the linker part are labeled on the chain. (**B**) Determination of $C_{\max}(N_i, N_{i+1})$. (**C**) Calculation of a term in *L-score*. The vector connecting $C\alpha_{N_i} C\alpha_{N_{i+1}}$, the one connecting $C\alpha_{N_i} C\alpha_{C_{\max}(N_i,N_{i+1})}$, and the one connecting $C\alpha_{N_i} C\alpha_{L_k}$ in Equation (7) are drawn with gray arrows and the vector product of the first two vectors are drawn with a dashed arrow. The calculated score of this example $\beta\alpha\beta$-unit is *L-score* $= 0.86$.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules27113547/s1, Supplementary Figures S1–S5.

Author Contributions: Conceptualization, G.C.; methodology, T.N., M.N. and G.C.; software, T.N. and M.N.; investigation, T.N., M.N. and G.C.; writing—original draft preparation, M.S. and G.C.; writing—review and editing, M.S. and G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the KAKENHI Grants, 20H05530, 21H00248, and 22H00406 of Japan Society for the Promotion of Science for M.S. and 19H03166 for G.C. and by Platform Project for Supporting Drug Discovery and Life Science Research (JP21am0101111) from Japan Agency for Medical Research and Development for G.C.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SCOP	Structural Classification of Proteins
ECOD	Evolutionary Classification of protein Domains
PDB	Protein Data Bank

References

1. Tramontano, A.; Cozzetto, D. *Supramolecular Structure and Function 8*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 15–29.
2. Sadowski, M.I.; Jones, D.T. The sequence-structure relationship and protein function prediction. *Curr. Opin. Str. Biol.* **2019**, *19*, 357–362. [CrossRef] [PubMed]
3. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710. [CrossRef] [PubMed]
4. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
5. Koga, N.; Tatsumi-Koga, R.; Liu, G.; Xiao, R.; Acton, T.B.; Montelione, G.T.; Baker, D. Principles for designing ideal protein structures. *Nature* **2012**, *491*, 222–227. [CrossRef] [PubMed]
6. Marcos, E.; Chidyausiku, T.K.; McShan, A.C.; Evangelidis, T.; Nerli, S.; Carter, L.; Nivón, L.G.G.; Davis, A.; Oberdorfer, G.; Tripsianes, K.; et al. De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **2018**, *25*, 1028–1034. [CrossRef]
7. Murata, H.; Imakawa, H.; Koga, N.; Chikenji, G. The register shift rules for $\beta\alpha\beta$-motifs for de novo protein design. *PLoS ONE* **2021**, *16*, e0256895. [CrossRef]
8. Minami, S.; Kobayashi, N.; Sugiki, T.; Nagashima, T.; Fujiwara, T.; Koga, R.; Chikenji, G.; Koga, N. Exploration of novel $\alpha\beta$-protein folds through de novo design. *bioRxiv* **2021**. [CrossRef]
9. Huang, P.S.; Feldmeier, K.; Parmeggiani, F.; Velasco, D.A.F.; Höcker, B.; Baker, D. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **2016**, *12*, 29–34. [CrossRef]
10. Dou, J.; Vorobieva, A.A.; Sheffler, W.; Doyle, L.A.; Park, H.; Bick, M.J.; Mao, B.; Foight, G.W.; Lee, M.Y.; Gagnon, L.A.; et al. De novo design of a fluorescence-activating β-barrel. *Nature* **2018**, *561*, 485–491. [CrossRef]
11. Kuhlman, B.; Dantas, G.; Ireton, G.C.; Varani, G.; Stoddard, B.L.; Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364–1368. [CrossRef]
12. Doyle, L.; Hallinan, J.; Bolduc, J.; Parmeggiani, F.; Baker, D.; Stoddard, B.L.; Bradley, P. Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **2015**, *528*, 585–588. [CrossRef] [PubMed]
13. Pan, F.; Zhang, Y.; Liu, X.; Zhang, J. Estimating the designability of protein structures. *bioRxiv* **2021**. 11.03.467111. [CrossRef]
14. Richardson, J.S. beta-Sheet topology and the relatedness of proteins. *Nature* **1977**, *268*, 495–500. [CrossRef] [PubMed]
15. Richardson, J.S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **1981**, *34*, 167–339.
16. Ruczinski, I.; Kooperberg, C.; Bonneau, R.; Baker, D. Distributions of beta sheets in proteins with application to structure prediction. *Proteins* **2002**, *48*, 85–97. [CrossRef]
17. Chitturi, B.; Shi, S.; Kinch, L.N.; Grishin, N.V. Compact Structure Patterns in Proteins. *J. Mol. Biol.* **2016**, *428*, 4392–4412. [CrossRef]

18. Minami, S.; Chikenji, G.; Ota, M. Rules for connectivity of secondary structure elements in protein: Two-layer $\alpha\beta$ sandwiches. *Protein Sci.* **2017**, *26*, 2257–2267. [CrossRef]
19. Orengo, C.A.; Jones, D.T.; Thornton, J.M. Protein superfamilles and domain superfolds. *Nature* **1994**, *372*, 631–634. [CrossRef]
20. Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540. [CrossRef]
21. Salem, G.M.; Hutchinson, E.G.; Orengo, C.A.; Thornton, J.M. Correlation of observed fold frequency with the occurrence of local structural motifs. *J. Mol. Biol.* **1999**, *287*, 969–981. [CrossRef]
22. Kinoshita, K.; Kidera, A.; Go, N. Diversity of functions of proteins with internal symmetry in spatial arrangement of secondary structural elements. *Protein Sci.* **1999**, *8*, 1210–1217. [CrossRef] [PubMed]
23. Fox, N.K.; Brenner, S.E.; Chandonia, J.M. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **2014**, *42*, D304–D309. [CrossRef] [PubMed]
24. Chandonia, J.M.; Fox, N.K.; Brenner, S.E. SCOPe: Classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res.* **2019**, *47*, D475–D481. [CrossRef] [PubMed]
25. Zhang, C.; Kim, S.H. The anatomy of protein beta-sheet topology. *J. Mol. Biol.* **2000**, *299*, 1075–1089. [CrossRef] [PubMed]
26. Cheng, H.; Schaeffer, R.D.; Liao, Y.; Kinch, L.N.; Pei, J.; Shi, S.; Kim, B.H.; Grishin, N.V. ECOD: An evolutionary classification of protein domains. *PLoS Comput. Biol.* **2014**, *10*, e1003926. [CrossRef]
27. Andreeva, A.; Howorth, D.; Chandonia, J.M.; Brenner, S.E.; Hubbard, T.J.P.; Chothia, C.; Murzin, A.G. Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res.* **2008**, *36*, D419–D425. [CrossRef]
28. Sillitoe, I.; Bordin, N.; Dawson, N.; Waman, V.P.; Ashford, P.; Scholes, H.M.; Pang, C.S.M.; Woodridge, L.; Rauer, C.; Sen, N.; et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **2021**, *49*, D266–D273. [CrossRef]
29. Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **1995**, *23*, 566–579. [CrossRef]
30. Wang, G.; Dunbrack, R.L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589–1591. [CrossRef]
31. Street, T.O.; Fitzkee, N.C.; Perskie, L.L.; Rose, G.D. Physical-chemical determinants of turn conformations in globular proteins. *Protein Sci.* **2007**, *16*, 1720–1727. [CrossRef]
32. Lesk, A.M.; Brändén, C.I.; Chothia, C. Structural principles of α/β barrel proteins: The packing of the interior of the sheet. *Proteins Str. Funct. Bioinform.* **1989**, *5*, 139–148. [CrossRef] [PubMed]
33. Murzin, A.G.; Finkelstein, A.V. General architecture of the α-helical globule. *J. Mol. Biol.* **1988**, *204*, 749–769. [CrossRef]
34. Nagi, A.D.; Regan, L. An inverse correlation between loop length and stability in a four-helix-bundle protein. *Fold. Des.* **1997**, *2*, 67–75. [CrossRef]
35. Linse, S.; Thulin, E.; Nilsson, H.; Stigler, J. Benefits and constrains of covalency: the role of loop length in protein stability and ligand binding. *Sci. Rep.* **2020**, *10*, 20108. [CrossRef] [PubMed]
36. Richardson, J.S. Handedness of crossover connections in beta sheets. *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 2619–2623. [CrossRef] [PubMed]
37. Sternberg, M.J.E.; Thornton, J.M. On the conformation of proteins: The handedness of the connection between parallel β-strands. *J. Mol. Biol.* **1976**, *110*, 269–283. [CrossRef]
38. Cole, B.J.; Bystroff, C. Alpha helical crossovers favor right-handed supersecondary structures by kinetic trapping: The phone cord effect in protein folding. *Protein Sci.* **2009**, *18*, 1602–1608. [CrossRef]
39. Ferreiro, D.U.; Komives, E.A.; Wolynes, P.G. Frustration, function and folding. *Curr. Opin. Struct. Biol.* **2018**, *48*, 68–73. [CrossRef]
40. Parra, R.G.; Schafer, N.P.; Radusky, L.G.; Tsai, M.Y.; Guzovsky, A.B.; Wolynes, P.G.; Ferreiro, D.U. Protein frustratometer 2: A tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res.* **2016**, *44*, W356–W360. [CrossRef]
41. Ferreiro, D.U.; Hegler, J.A.; Komives, E.A.; Wolynes, P.G. On the role of frustration in the energy landscapes of allosteric proteins. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3499–3503. [CrossRef]
42. Ferreiro, D.U.; Hegler, J.A.; Komives, E.A.; Wolynes, P.G. Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19819–19824. [CrossRef] [PubMed]
43. Fleishman, S.; Leaver-Fay, A.; Corn, J.E.; Strauch, E.M.; Khare, S.D.; Koga, N.; Ashworth, J.; Murphy, P.; Richter, F.; Lemmon, G.; et al. RosettaScripts: A scripting language interface to the Rosetta macromolecular modeling suite. *PLoS ONE* **2011**, *6*, e20161. [CrossRef] [PubMed]
44. Lin, Y.R.; Koga, N.; Tatsumi-Koga, R.; Liu, G.; Clouser, A.F.; Montelione, G.T.; Baker, D. Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E5478–E5485. [CrossRef] [PubMed]
45. Wintjens, R.T.; Rooman, M.J.; Wodak, S.J. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J. Mol. Biol.* **1996**, *255*, 235–253. [CrossRef] [PubMed]