ORIGINAL RESEARCH
published: 19 April 2022
doi: 10.3389/fmolb.2022.889943

Variable and Conserved Regions of Secondary Structure in the β-Trefoil Fold: Structure Versus Function

*Michael Blaber**

Department of Biomedical Sciences, College of Medicine, Florida State University, Tallahassee, FL, United States

β-trefoil proteins exhibit an approximate C_3 rotational symmetry. An analysis of the secondary structure for members of this diverse superfamily of proteins indicates that it is comprised of remarkably conserved β-strands and highly-divergent turn regions. A fundamental "minimal" architecture can be identified that is devoid of heterogenous and extended turn regions, and is conserved among all family members. Conversely, the different functional families of β-trefoils can potentially be identified by their unique turn patterns (or turn "signature"). Such analyses provide clues as to the evolution of the β-trefoil family, suggesting a folding/stability role for the β-strands and a functional role for turn regions. This viewpoint can also guide *de novo* protein design of β-trefoil proteins having novel functionality.

Keywords: protein symmetry, *de novo* design, hydrophobic patterning, ligand, folding nucleus

Edited by:
Delia Picone,
University of Naples Federico II, Italy

Reviewed by:
Takeshi Kikuchi,
Ritsumeikan University, Japan
Serena Leone,
Zoological Station Anton Dohrn, Italy

Correspondence:
Michael Blaber
michael.blaber@med.fsu.edu

Specialty section:
This article was submitted to
Structural Biology,
a section of the journal
Frontiers in Molecular Biosciences

Received: *04 March 2022*
Accepted: *01 April 2022*
Published: *19 April 2022*

Citation:
Blaber M (2022) Variable and Conserved Regions of Secondary Structure in the β-Trefoil Fold: Structure Versus Function.
Front. Mol. Biosci. 9:889943.
doi: 10.3389/fmolb.2022.889943

INTRODUCTION

The β-trefoil is a common protein architecture, with 10 different superfamilies, and constituting approximately 1% of the proteome (Andreeva et al., 2013) (**Table 1**). A notable feature of the β-trefoil is a discernable C_3 rotational symmetry where the repeating "trefoil" motif is approximately 40–50 amino acids in length and contains four anti-parallel β-strands connected by turn/loop regions (Sweet et al., 1974; McLachlan, 1979; Murzin et al., 1992) (**Figure 1**). β-trefoil proteins encompass diverse ligand-type functionalities, including toxins, protease inhibitors, cytokines, growth factors, agglutinins, lectins, and other types of ligands [SCOP database (Andreeva et al., 2019)], although no known enzymatic functionality. These ligand functionalities are associated with specific turn/loop regions that may define certain β-trefoil families (Blow et al., 1974; Veerapandian et al., 1992; Notenboom et al., 2002; Bovi et al., 2012; Blaber, 2020).

Symmetry in a subset of common protein folds has been evident from the earliest days of protein structure determination, and has stimulated hypotheses of gene duplication and fusion in their evolutionary emergence from simpler peptide motifs (Eck and Dayhoff, 1966; Ohno, 1970; McLachlan, 1972). Alternative hypotheses for such evolution of the β-trefoil have been proposed, including "emergent architecture" and "conserved architecture" models, where the simple peptide motif comprises two anti-parallel β-hairpins known as a "trefoil" (Mukhopadhyay, 2000; Ponting and Russell, 2000; Blaber and Lee, 2012; Balaji, 2015). In the emergent architecture model the structural complexity increases with each gene duplication and fusion event, such that the overall β-trefoil architecture only emerges upon a final triplet repeat of the trefoil motif. In the conserved architecture model, the trefoil peptide has the property of oligomerizing as a trimer, thereby generating an intact β-trefoil architecture. A tandem repeat also oligomerizes as a domain-swapped trimer that generates two intact β-trefoils. A triplet repeat of the trefoil motif yields a single polypeptide that folds into β-trefoil. Experimental studies lend greater

TABLE 1 | β-trefoil superfamily and structures utilized in characterization of secondary structure heterogeneity. The overlay statistics with Symfoil-4T (RCSB 3O4B) are also provided.

Superfamily	Family	Domain	RCSB	Res. (Å)	#Cα Ovl	Ovl rmsd (Å)
Ricin B-like lectin	Ricin B-like	β-zylanase	1XYF	1.90	93	1.13
		β-galactoside-specific lectin 1	1SZ6	2.05	96	1.36
		Hemolytic lectin CEL-III	1VCL	1.70	94	1.33
		29-kDa galactose-binding lectin	2ZQO	1.80	86	1.31
		Main hemagglutinin component type C	3AH2	1.70	102	1.24
		Agglutinin	5D61	1.60	98	1.01
		Endo-1,4-β-xylanase A	1KNL	1.20	90	1.14
		Cytolethal distending toxin	1SR4	2.00	104	1.28
		Abrin-A	1ABR	2.14	95	1.22
	Cysteine rich domain	Cysteine rich domain	1FWV	1.90	88	1.19
	GlcNAc-alpha-1,4-Gal-releasing endo-β-galactosidase	GlcNAc-alpha-1,4-Gal-releasing endo-β-galactosidase	1UPS	1.82	104	1.16
	HylA β-trefoil domain-like	HylA β-trefoil domain-like	1XEZ	2.30	88	1.55
	Kunitz (STI) inhibitors	Chymotrypsin inhibitor 3	1EYL	1.90	79	1.41
		Trypsin inhibitor A	1AVW	1.75	75	1.37
		Alpha-amylase/subtilisin inhibitor	3BX1	1.85	80	1.34
		Kunitz-type serine proteinase inhibitor DrTI	1R8N	1.75	78	1.42
		Albumin-1	1WBA	1.80	74	1.34
	Clostridium neurotoxins, C-terminal domain	Botulinum neurotoxin type B	1EPW	1.90	85	1.41
		Botulinum neurotoxin type A	5MK6	1.45	79	1.19
		Tetanus toxin	1A8D	1.57	80	1.25
	Clitocypin-like	Clitocypin-5	3H6S	2.22	87	1.18
		Clitocypin-2	3H6R	1.95	89	1.26
Cytokine	Fibroblast growth factors	FGF-1	1RG8	1.10	115	1.06
		FGF-2	1BFG	1.60	113	0.98
		FGF-4	1IJT	1.80	115	1.23
		FGF-8	2FDB	2.28	110	1.23
		FGF-9	1IHK	2.20	113	1.19
		FGF-12	1Q1U	1.70	113	1.39
		FGF-19	1PWA	1.30	93	1.18
	Interleukin-1 (IL-1)	Interleukin-1 β	5R7W	1.27	95	1.34
		Interleukin-18	3WO2	2.33	89	1.33
		Interleukin-36 receptor agonist protein	1MD6	1.60	81	1.30
Actin-crosslinking proteins	Fascin	Fascin-1	3LLP	1.80	104	1.24
DNA-binding protein LAG-1 (CSL)	DNA-binding protein LAG-1 (CSL)	Lin-12 and Glp-1 phenotype	3BRD	2.21	83	1.04
AbfB domain	AbfB domain	Alpha-L-arabinofuranosidase B	1WD3	1.75	96	1.22
Agglutinin	Agglutinin	Agglutinin	1JLY	2.20	98	1.38
MIR domain	MIR domain	Inositol 1,4,5-trisphosphate receptor type 1	1N4K	2.20	101	1.12
		Uncharacterized protein (*C. elegans*)	1T9F	2.00	105	0.90
30 K Lipoprotein C-terminal domain-like	30 K Lipoprotein C-terminal domain-like	30 K protein 2	4EFP	1.33	107	1.12
		Low molecular mass 30 kDa lipoprotein 19G1	4IY9	2.10	107	1.10
		30 K lipoprotein	4PC4	1.80	104	1.10
Proteinase inhibitor 1-like	Proteinase inhibitor 1-like	Serine protease inhibitor 1	3VWC	1.50	95	1.22
de novo Symmetric	*de novo* Symmetric	Symfoil (Symfoil-4T variant)	3O4B	1.80	126 (Ref)	N/A (Ref)
		Threefoil	3PG0	1.62	105	1.00
		Mitsuba-1	5XG5	1.54	103	1.03

support to the conserved architecture model (Lee and Blaber, 2011; Lee et al., 2011), indicating that an appropriate trefoil motif peptide can spontaneously oligomerize as a trimer to form an intact β-trefoil. Sequence and structure analyses suggest that extant β-trefoil proteins are unlikely to share a common ancestor, but are more likely to have evolved independently from simpler peptide motifs many times, and indeed, this may be a reoccurring and ongoing evolutionary process (Broom et al., 2012).

Current knowledge regarding symmetric protein architecture suggests that utilization of symmetry is an efficient and practical strategy for simplifying the *de novo* design problem (Hocker et al., 2004; Nikkhah et al., 2006; Yadid and Tawfik, 2007; Richter et al., 2010; Kopec and Lupas, 2013; Voet et al., 2014; Broom et al., 2015; Brunette et al., 2015; Huang et al., 2016; Terada et al., 2017; Afanasieva et al., 2019; Kimura et al., 2020). Furthermore, it may be practical to divide the design problem into two parts: 1) the initial design of a stable, foldable but functionless "scaffold",

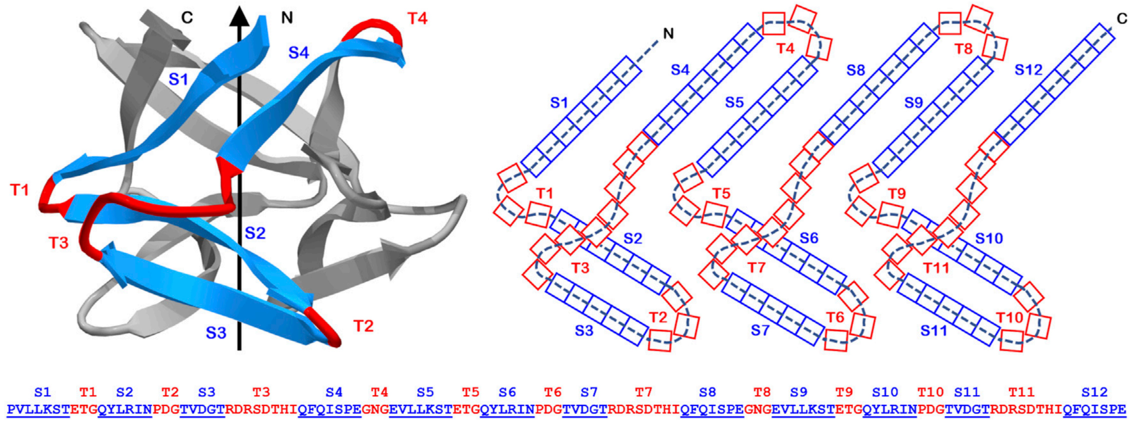
C
N
T4
S4
S1
T1
T3
S2
S3
T2
N
S1
S4
T4
S5
S8
T8
S9
S12
C
T1
T5
T9
S2
S6
S10
T3
T7
T11
T2
T6
T10
S3
S7
S11
S1 T1 S2 T2 S3 T3 S4 T4 S5 T5 S6 T6 S7 T7 S8 T8 S9 T9 S10 T10 S11 T11 S12
PVLLKSTETGQYLRINPDGTVDGTRDRSDTHIQFQISPEGNGEVLLKSTETGQYLRINPDGTVDGTRDRSDTHIQFQISPEGNGEVLLKSTETGQYLRINPDGTVDGTRDRSDTHIQFQISPE

FIGURE 1 | The primary, secondary, and tertiary structure of the Symfoil ("Symfoil-4T") reference β-trefoil protein. Upper panel: A "ribbon" diagram of the Symfoil protein (RCSB 3O4B). The colored region (blue: β-strand; red: turn) identifies the first of three repeating "trefoil" motifs in the structure (the other two colored in gray). Middle panel: A two-dimensional representation of the overall β-trefoil architecture and indicating the strand and turn numbering and the number of residues in each type of secondary structure (referencing Symfoil). Lower panel: the primary structure of the Symfoil protein indicating the secondary structure positions (β-strands underlined and indicated by "S", and turns indicated by "T").

followed by 2) specific functionalization (Bolon et al., 2002; Dwyer et al., 2004; Claren et al., 2009). In the case of the β-trefoil (and perhaps also the β-propeller architecture), this strategy appears especially appropriate for the design of proteins having novel ligand functionalities. It would therefore be extremely useful to elucidate the structural parameters that dictate stable, foldable architecture, from parameters that generate specific functionality.

In this report we examine the hypothesis that the structural determinants of stability and folding for the β-trefoil are principally the β-strand secondary structure (and that this is an essentially conserved structural feature in this superfamily), while specific functionality is provided by turn/loop regions (and that this is a divergent, and unique feature, among functionally-distinct β-trefoil proteins). The analysis suggests an efficient *de novo* protein design pathway that leverages symmetric principles of protein architecture.

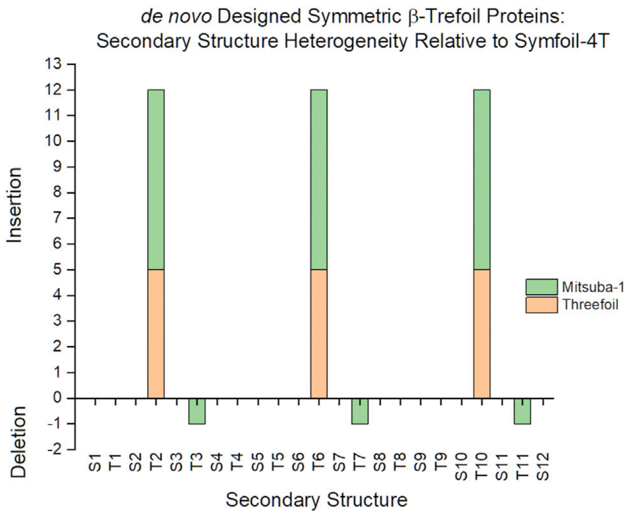

FIGURE 2 | A comparison of secondary structure insertions/deletions for three symmetric designed β-trefoil proteins. The Symfoil, Threefoil, and Mitsuba-1 proteins are three independently *de novo* designed, purely-symmetric β-trefoil proteins. The most compact of these three is Symfoil, primarily due to ligand-binding turns T2, T6, and T10, engineered into both Threefoil and Misuba-1.

MATERIALS AND METHODS

Selection of Reference β-Trefoil Structure

The identification of insertions or deletions of secondary structure within a protein architecture depends upon the reference protein used for such comparison. The reference protein should ideally comprise the essential structural architecture, with no extraneous insertions or deletions beyond the basic folding and stability requirements. In the case of the β-trefoil, where extant naturally evolved proteins exhibit varying degree of C_3 rotational symmetry, the reference protein would ideally constitute a purely-symmetric architecture so that any asymmetric features in an evaluated protein can readily be identified. There are several *de novo* designed β-trefoil proteins having an exact threefold symmetric primary structure; including Threefoil (Broom et al., 2015), Mitsuba-1 (Terada et al., 2017), Phifoil (Longo et al., 2014) and the Symfoil family of proteins (Lee and Blaber, 2011; Lee et al., 2011). Threefoil was designed to have carbohydrate binding function and contains specific turn/loop secondary structure for this purpose. Similarly, Mitsuba-1 was designed to have a galactose binding site afforded by specific surface turn/loop secondary structure. In contrast, Symfoil was designed exclusively from the standpoint of optimized folding kinetics and thermodynamics, and is notably devoid of any specific functionality. Symfoil (using the Symfoil-4T variant) as a reference structure identifies five residue insertions within turns T2, T6 and T10 in Threefoil, and seven residue insertions of the same turns in Mitsuba-1 (**Figure 2**). Thus, the Symfoil protein was considered as the most appropriate reference protein with which to quantify secondary structure heterogeneity among β-trefoil proteins.

Representative β-Trefoil Proteins

The RCSB structural databank (www.rcsb.org) was queried for β-trefoil proteins solved to better than 2.5 Å resolution. A total of 45 proteins were identified, representing 10 superfamilies, 17 families, and 45 domains, and with an average resolution of 1. 81 ± 0.31 Å (**Table 1**). Only the *de novo* designed β-trefoil proteins exhibit an exact threefold rotational symmetry; all naturally-evolved β-trefoil proteins exhibit varying degrees of primary, secondary and tertiary structure symmetry.

Structural Overlay

Structural overlays of individual β-trefoil proteins onto the Symfoil protein coordinates (using the Symfoil-4T variant, RCSB 3O4B) were performed using the Swiss PDB Viewer software (Guex and Peitsch, 1997) and selecting for Cα atoms. An iterative fitting process was used to optimize the overlay. The number of matching Cα atoms was noted, as well as the rmsd for the fit (**Table 1**). This overlay was then examined for insertions or deletions in specific secondary structure elements as defined in the Symfoil structure (**Figure 1**). The percent of Cα matches per secondary structure element was also determined.

Sequence Logo Plots

Sequence logo plots are a graphical representation of an amino acid (or nucleic acid) multiple sequence alignment (Schneider and Stephens, 1990; Crooks et al., 2004). Each logo consists of stacks of symbols, one stack for each position in the sequence. The height of symbols within a stack indicates the relative frequency of each amino at that position. A sequence logo plot was generated for β-strands S1, S5, and S9 as a group; similarly, S2, S6, and S10 as a group; S3, S7, and S11 as a group; and S4, S8, and S12 as a group (i.e., all sets of C_3 symmetry related strands, $n = 126$), for all representative β-trefoil proteins in **Table 1** and using structural overlays as described above. Image generation utilized the web logo server at https://weblogo.berkeley.edu/ with colors based on chemical properties: polar amino acids (G,S,T,Y,C,Q,N) are green, basic (K,R,H) blue, acidic (D,E) red and hydrophobic (A,V,L,I,P,W,F,M) amino acids are black.

RESULTS

Secondary Structure Length and Conformational Heterogeneity

An analysis of the secondary structure length heterogeneity for the β-trefoil superfamily of proteins, compared to the Symfoil reference, shows that the heterogeneity is localized almost exclusively to turn secondary structure; indeed, all β-strands show a remarkable absence of relative insertion or deletion (i.e., all β-strands show a marked conservation of length (**Figure 3**). Furthermore, the heterogeneity in the turn regions principally involves insertions, as opposed to deletions, compared to the Symfoil reference protein. However, there are two notable exceptions to this general rule at turns T4 and T8, where some β-trefoils have limited deletions of up to three amino acids.

An analysis of the Cα structural conservation for regions of secondary structure in β-trefoil proteins, compared to the Symfoil-4T reference, shows that not only do β-strand regions show highly-conserved lengths, but that their overall conformation as β-strands is also highly-conserved (**Figure 4**). It can be seen that for the entire superfamily of β-trefoils a >90% structural conservation (i.e., <1.5 Å rmsd) is present with the symmetry-related sets of β-strands S1/S5/S9, S3/S7/S11, and S4/S8/S12. The S2/S6/S10 set exhibits 76–84% Cα structural conservation. Among turn secondary structure, turns T4 and T8 (which are symmetry-related) exhibit the least Cα structural conservation.

The Ricin B-like, Cytokine, and 30 K Lipoprotein superfamilies have the greatest number of members, with 22, 10, and 3 members, respectively (**Table 1**). The secondary structure length heterogeneity for these individual families is shown in **Figure 5**. This graph suggests that the general turn heterogeneity observed in the overall superfamily graph (**Figure 3**) is a composite of patterns of turn heterogeneity unique to the individual superfamilies or families. Thus, the Ricin B-like lectin superfamily exhibits the greatest turn heterogeneity (i.e., extensions) at T2, T3, T4, T6, and T10; while the Cytokine superfamily exhibits turn extensions principally at T3, T4, T7, T9, and T11; and the 30 K Lipoprotein superfamily exhibits turn extensions principally at T2, T6, and T10. Thus, each different superfamily exhibits characteristically different turn heterogeneity (i.e., extensions).

Sequence Logo Plots

The sequence logo plots for the β-strand secondary structure exhibit characteristic patterns of hydrophobic residues (**Figure 6**). In β-strands S1/S5/S9 position #4 is principally hydrophobic: Ile and Leu account for 80% of all amino acids at this position, with the other residues being Phe, Tyr, Val and Met. There is some indication of hydrophobic preference at position #2, with Val and Phe accounting for approximately 40% of positions (and if Y is considered hydrophobic, then ~50% of residues at position #2 are hydrophobic). In β-strands S2/S6/S10 positions #3 and #5 show a clear hydrophobic preference. Leu accounts for ~50% of residues at position #3, with the majority of other residues being either Val, Ile, Phe or Trp. At position #5 Leu, Val, Ile, Ala and Met account for ~66% of residues. In β-strands S3/S7/S11 Val, Leu and Ile account for ~75% of residues at position #2. Ala, Leu, Val and Ile account for ~50% of residues at position #4, with Gly another major residue at this position. In β-strands S4/S8/S12 there is a remarkable ~70% preference of aromatic residues W or F at position #2 (with Leu, Val and Ile comprising the majority of the remainder). Hydrophobic residues are also preferred at position #4, with Ile, Leu, Phe, and Val comprising ~60% of residues. Thus, in all β-strands there is a hydrophobic (P)/hydrophilic (H) pattern of H-P-H-P-H. Binary patterning of hydrophobic/hydrophilic amino acids is a key determinant of protein secondary

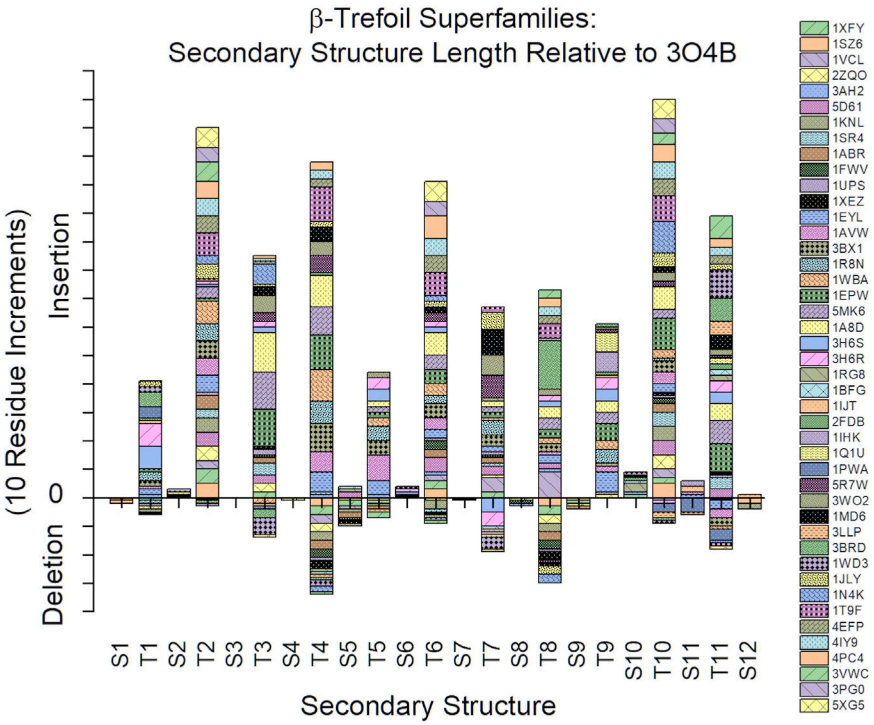

FIGURE 3 | Relative insertions or deletions in secondary structure elements among the β-trefoil superfamily of proteins. The reference protein is the Symfoil protein—a *de novo* designed, purely-symmetric, minimalist, and functionless β-trefoil protein (see text).

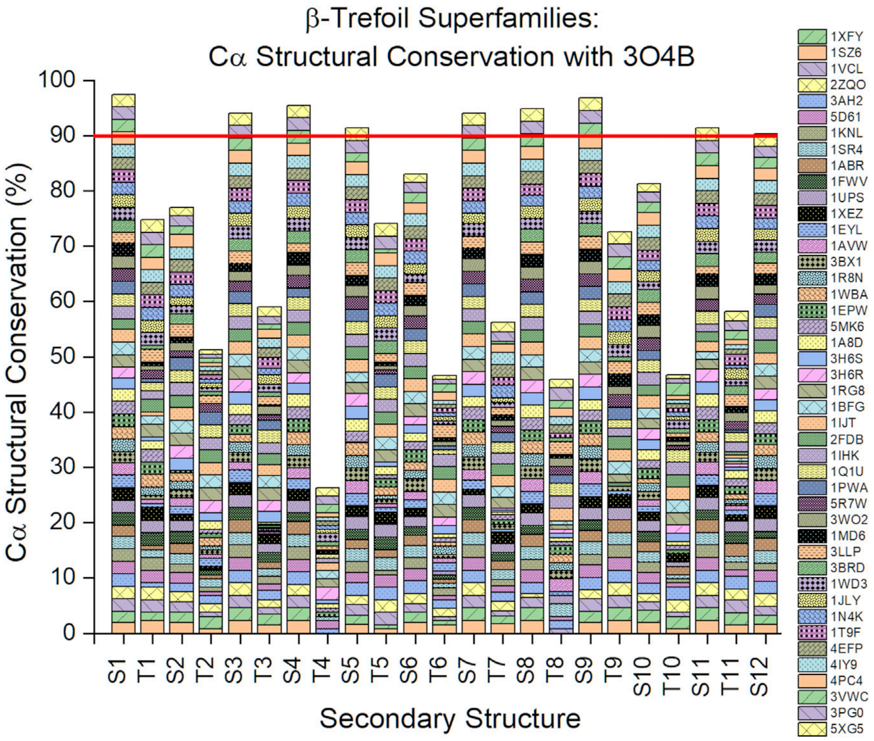

FIGURE 4 | Cα structural conservation (<1.5 Å rmsd) within secondary structure elements for the β-trefoil family of proteins. The reference protein is the Symfoil protein (RCSB 3O4B)—a *de novo* designed, purely-symmetric, minimalist, and functionless β-trefoil protein (see **Figure 1**).

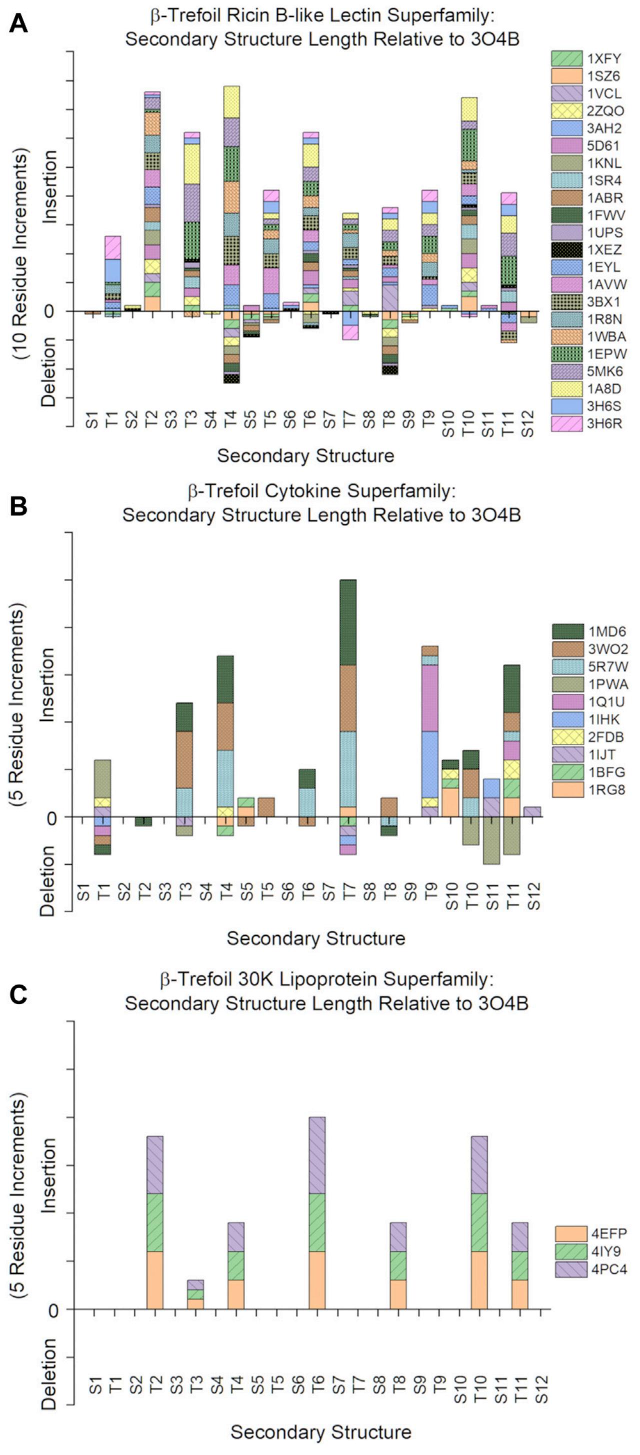

FIGURE 5 | Relative secondary structure insertions or deletions of individual β-trefoil superfamilies. **(A)**: Ricin B-like lectin superfamily (*n* = 22). **(B)**: Cytokine superfamily (*n* = 10). **(C)**: 30 K Lipoprotein superfamily (*n* = 3).

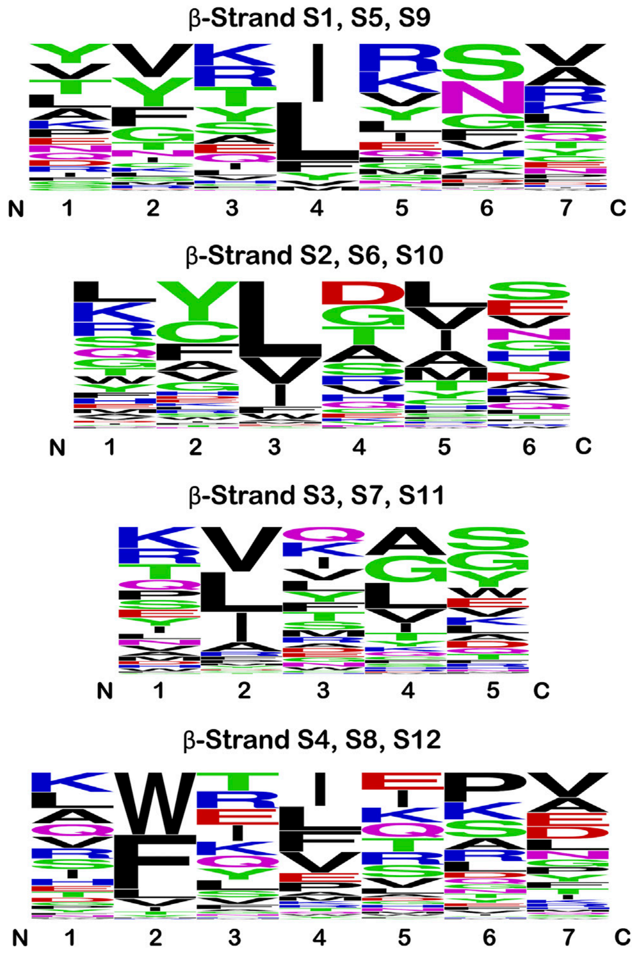

FIGURE 6 | Sequence logo plots for the β-strand secondary structure in the β-trefoil superfamily. Equivalent β-strands are grouped by the C_3 rotational symmetry of the β-trefoil for all members of the superfamily in **Table 1**. Thus, strands S2, S6, and S10 are grouped together in this analysis, and similarly for the other symmetry-related β-strands (therefore, *n* = 126 at each position). The single letter amino acid code is utilized, and the height indicates the relative prevalence of a particular amino acid at each position. The amino acids are colored according to chemical properties (see text); however, hydrophobic is indicated by black.

structure, with an alternating hydrophobic/hydrophilic pattern favoring the formation of amphipathic β-strand secondary structure (West and Hecht, 1995; Xiong et al., 1995). These hydrophobic residues within the H-P-H-P-H patterning of the β-trefoil β-strands contribute to a highly-cooperative core packing group in the β-trefoil structure (Blaber, 2021).

DISCUSSION

Is Symfoil-4T a "Minimal" β-Trefoil?

Among the *de novo* designed symmetric β-trefoil proteins Symfoil is the most compact, principally due to the absence of specific functional surface turns/loops. Analyses of structural variations (i.e., insertions or deletions) of other β-trefoil proteins indicate that the vast majority of structural heterogeneity is associated with insertions in surface turn/loop regions in comparison to Symfoil. However, there is evidence of some β-trefoil proteins having relative truncations in the T4 and

T8 regions (**Figures 3**, **5A**). Specifically, 1FWV, 1ABR, 1KNL, 2ZQO, 1SZ6, 1XYF, and 1XEZ (all members of the Ricin B-like lectin superfamily, **Table 1**) have three amino acid deletions in both the T4 and T8 regions. These deletions effectively eliminate the hydrophobic residue at the #2 position in the S5 and S9 β-strands (which participate in the cooperative central core); thus, these truncations of the T4 and T8 turns may result in a less stable, or less cooperatively-folding, protein. The Symfoil protein therefore represents a "minimal" or "essential" β-trefoil architecture—one that is highly-conserved in the family of β-trefoil proteins—and is therefore a useful reference structure by which to characterize secondary structure heterogeneity in β-trefoil proteins.

Is There a Segregation of β-Strand and Turn Secondary Structure as Regards Protein Structure and Function?

The highly-conserved β-strands, and highly-divergent turn/loop regions, when comparing members of the β-trefoil superfamily, strongly suggests that functionality has its principle basis in turn/loop structure. For example, the specific heparin-binding functionality of FGF-1 (Cytokine superfamily) has been localized principally to an extension within the T11 region (Brych et al., 2004) while interaction with FGF receptor involves the T1, T4, and T8 regions (Olsen et al., 2004). Lectin functionality in the shellfish lectin MytiLec-1 and *M. oreades* mushroom lectin is localized to regions T2, T6, and T10 (Broom et al., 2015; Terada et al., 2017). The inhibitory function of Kunitz (STI) protease inhibitors is due to active site binding of an extended T4 loop region (Song and Suh, 1998). Ricin B-like lectin interactions involve the T2/T3 and T10/T11 regions (Suzuki et al., 2009). The Pmt2-MIR domain (superfamily MIR domain) interaction with tetraethylene glycol ligand involves regions T4 and T7 (Chiapparino et al., 2020). The interaction between LAG-1 (CSL) DNA-binding protein and DNA ligand principally involves the T1 region (Friedmann et al., 2008). The interaction between Agglutinin and T-disaccharide involves the T6 and T10 region (Transue et al., 1997). The interaction between clitocypin and cathepsin V involves the T1 and T3 regions (Renko et al., 2010). This representative summary of binding interactions provides strong support for a primary assignment of functionality to specific and structurally-heterogenous turn/loop regions in β-trefoil proteins.

Can Turn Structure Provide Evidence of Evolutionary Gene Duplication/Fusion Processes?

Symmetric relationships among turn/loop structures in β-trefoils appears most apparent within the symmetry-related set of T2/T6/T10 turn positions. There are β-trefoil proteins having relative insertions of $n = +1$ (1UPS), $n = +5$ (3PG0), $n = +6$ (4IY9), $n = +7$ (5XG5), and $n = +8$ (1T9F) amino acids, relative to the Symfoil (i.e., 3O4B) reference structure. Additionally, similar examples exist having no relative insertions (i.e., $n = 0$; 1Q1U/1IHK/2FDB/1IJT/1BFG/1RG8) as well as $n = -1$ deletions (1WD3) (**Figure 7**). The most parsimonious explanation for such structural conservation of these symmetry-related turns is for duplication/fusion events to occur subsequent to trefoil motif structural evolution. This implies the likelihood of multiple independent instances of the evolution of β-trefoil proteins from simpler (i.e., trefoil-fold) motifs, and supports the evolutionary hypothesis put forth by Meiering (Broom et al., 2012) that the emergence of β-trefoil proteins is a recurring and ongoing evolutionary mechanism.

In the simplest example of duplication and fusion of individual trefoil-motifs leading ultimately to formation of a β-trefoil protein, the junction of gene fusion is the T4 turn region (Ponting and Russell, 2000; Lee and Blaber, 2011; Lee et al., 2011). Thus, the β-trefoil architecture contains two symmetry-related turns T4 and T8, with the "third" member of this symmetrically-related set defined by the adjacent (but discontinuous) N- and C- termini (see **Figure 1**). As with the T2/T6/T10 turns, a number of β-trefoil proteins exhibit a unique structural symmetry when comparing the T4 and T8 turns (e.g., 1FWV, 1ABR, 1KNL, 2ZQO, 1SZ6, 1XYF, 1XEZ; as described above). This implies that this turn formed prior to the duplication and fusion event that yielded the mature β-trefoil architecture. However, this results in a structural conundrum. The existence of a T4 region results from the fusion of two trefoil motifs. Two such turns (i.e., T4 and T8) would be generated by a subsequent tandem duplication of such a construct; however, this would yield a total of four sequential trefoil motifs. The apparent solution to the presence of an "extra" trefoil motif is for the latter fusion to include a truncation event affecting one trefoil motif (Jeltsch, 1999; Peisajovich et al., 2006; Longo et al., 2013).

Turns and the Folding Nucleus

In addition to providing a potential functional role, turns also serve to connect adjacent β-strand secondary structure (forming a β-hairpin), minimizing the entropic penalty of association, and thereby influencing stability and folding (Nagi et al., 1999; Thompson and Eisenberg, 1999; Lindberg et al., 2006). The reaction coordinate of cooperative protein folding typically describes a highly-polarized transition state or folding nucleus (Abkevich et al., 1994; Went and Jackson, 2005; Faísca, 2009). Establishment of this folding nucleus is the rate limiting step in folding, and once formed, serves to rapidly condense formation of the overall native structure. An isolated 42-mer trefoil motif (i.e., "Monofoil") derived from the Symfoil protein spontaneously oligomerizes to yield an intact β-trefoil architecture (Lee and Blaber, 2011; Lee et al., 2011); thus, a serviceable folding nucleus resides within each repeating motif in the Symfoil protein (Blaber, 2020; Parker et al., 2021). However, phi-value analysis (Fersht and Sato, 2004) indicates that the effective folding nucleus in the Symfoil protein, and the related fibroblast-growth factor-1 β-trefoil protein, while not identical, are both centrally-located and more expansive than an individual trefoil motif (Longo et al., 2014; Xia et al., 2016). This more expansive central definition includes turns T4 and T8, which are novel turn structures generated by the fusion of trefoil motif repeats. These novel turns are postulated to promote local

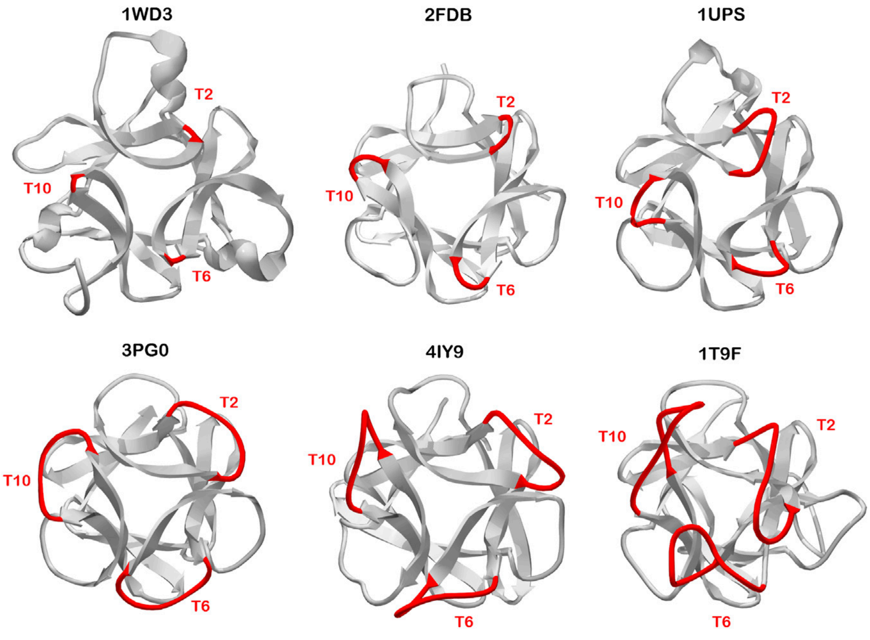

FIGURE 7 | Examples of β-trefoil proteins having distinct C_3 symmetry at the T2/T6/T10 turn region. The turn length in reference to the Symfoil protein is -1 (3PG0), 0 (2FDB), +1 (1UPS), +5 (3PG0), +6 (4IY9), and +8 (1T9F). The view is down the C_3 axis of rotational symmetry. Such symmetric relationships in turn structure suggests divergence of this turn structure occurred prior to duplication/fusion/truncation events leading to the extant β-trefoil architecture.

β-hairpin interactions, thereby generating a more efficient folding nucleus compared to an isolated trefoil motif. However, destabilizing mutations targeting the folding nucleus region of Symfoil indicate that the C_3 symmetry provides for alternative folding nuclei in other regions of the protein able to salvage foldability (Longo et al., 2013; Tenorio et al., 2020). The survey of turn region lengths in the β-trefoil superfamily indicates that the central region comprises turns having generally the shortest lengths (**Figure 3**). Thus, central turns may be somewhat "privileged" regions of secondary structure where considerations of efficient folding nucleus formation impact the optimal turn length and sequence design.

Implications and Suitability of β-Trefoil Proteins for *de novo* Design

The secondary structure elements of the fundamental β-trefoil are limited to β-strand and reverse turn, and thus describe a comparatively simple protein architecture. Knowledge essential for the *de novo* design of β-trefoil proteins is extensive: 1) The β-strand secondary structure is the key determinant of the conserved basic architecture for this protein superfamily; 2) Conserved β-strand characteristics have been elucidated as regards length and hydrophobic patterning; and 3) The role of β-strand hydrophobic residues in cooperative core-packing interactions has been well-characterized. In this regard, it is interesting to note the different independent solutions for the set of hydrophobic core-packing residues (referencing **Figure 6**) utilized by the *de novo* designed symmetric β-trefoil proteins Symfoil [3O4B; generated through top-down symmetric deconstruction of FGF-1 (Lee and Blaber, 2011; Lee et al., 2011)], Phifoil [4O4W; generated by folding nucleus symmetric expansion of FGF-1 (Longo et al., 2014)], Threefoil [3PG0; generated by consensus sequence of a carbohydrate-binding ricin sequence (Broom et al., 2012)], and Mitsuba-1 [5XG5; generated by computational sequence constraint of the shellfish lectin MytiLec-1 (Terada et al., 2017)]. The sequence logo plot for this set of core-packing residues (**Figure 8**) suggests that, as long as the appropriate hydrophobic patterning and compatible van der Waals interactions are satisfied, a variety of alternative core-packing arrangements are permissible, thereby indicating a lowered threshold for successful design.

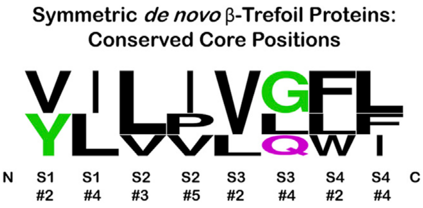

FIGURE 8 | Sequence logo plot of the set of symmetric core-packing residues (see **Figure 6**) present in *de novo* designed symmetric β-trefoil proteins Symfoil-4T (3O4B), Phifoil (4OW4), Threefoil (3PG0), and Misuba-1 (5XG5). The positions within a single trefoil motif are shown, but these are replicated exactly for the other two trefoil motifs in each protein. Position #4 in S1, and position #3 in S2, have the highest neighbor contacts among the set of core residues (Blaber, 2021).

The general attributes of the folding nucleus for Symfoil have been identified, and the potential for redundant folding nuclei demonstrated. Evolutionary considerations indicate highly-permissive design pathways of foldability involving diverse fusion/truncation of trefoil motifs. Turn regions have been identified as the key regions of structural variability, and are the principle determinants of ligand functionality characteristic of this superfamily. As connectors of adjacent β-strand secondary structure, turn regions also influence the entropic penalty for the assembly of local β-hairpin structure, and this plays an important role in the formation of the folding nucleus.

Protein design must simultaneously solve at least three different problems: 1) protein foldability (i.e., folding kinetics requirements), 2) protein stability (i.e., thermodynamic requirements), and 3) the accommodation of specific function (with potential structural dynamics requirements). Analysis of the β-trefoil architecture suggests that it is readily amenable to a two-step design process, with the initial step focusing upon the design of a foldable, stable "scaffold" (and many avenues appear possible); subsequently followed by a second step of functional mutation. The present analysis indicates that the first step involves β-strand secondary structure and key hydrophobic patterning design (building upon current extensive knowledge in this area). The C_3 symmetry substantially reduces the combinatorial search of appropriate primary structure solutions. The second step focuses upon turn/loop regions and their mutation to generate desired functionality (the β-trefoil architecture perhaps best suited to ligand functionality). This second step is less-well characterized and therefore open to expansive and novel opportunities. The C_3 symmetry provides for monovalent or multivalent ligand binding opportunities. In an alternative approach, if specific loop regions are associated with unique functional properties, and the β-strands as structural elements, then diverse chimeras with novel combined structure/function attributes might be constructed using computational approaches (Ferruz et al., 2021). Overall, the β-trefoil architecture has many attractive features for *de novo* protein design, applied especially to ligand functionality. The adoption of heparin-binding functionality into a benign β-trefoil scaffold using the principles described herein has recently been demonstrated (Tenorio et al., Forthcoming 2022).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MB is responsible for planning, data analysis, and writing of this report.

FUNDING

This work was supported by research grant RF02551 from Trefoil Therapeutics Inc. Additional support was provided by the department of Biomedical Sciences, FSU College of Medicine.

ACKNOWLEDGMENTS

The work is dedicated to Brian Matthews on the occasion of his retirement. A truly outstanding scientist and mentor who led naturally and effectively by example.

REFERENCES

Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1994). Specific Nucleus as the Transition State for Protein Folding: Evidence from the Lattice Model. *Biochemistry* 33 (33), 10026–10036. doi:10.1021/bi00199a029

Afanasieva, E., Chaudhuri, I., Martin, J., Hertle, E., Ursinus, A., Alva, V., et al. (2019). Structural Diversity of Oligomeric β-propellers with Different Numbers of Identical Blades. *eLife* 8, e49853. doi:10.7554/eLife.49853

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2013). SCOP2 Prototype: a New Approach to Protein Structure Mining. *Nucl. Acids Res.* 42 (D1), D310–D314. doi:10.1093/nar/gkt1242

Andreeva, A., Kulesha, E., Gough, J., and Murzin, A. G. (2019). The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures. *Nucleic Acids Res.* 48 (D1), D376–D382. doi:10.1093/nar/gkz1064

Balaji, S. (2015). Internal Symmetry in Protein Structures: Prevalence, Functional Relevance and Evolution. *Curr. Opin. Struct. Biol.* 32, 156–166. doi:10.1016/j.sbi.2015.05.004

Blaber, M. (2020). Conserved Buried Water Molecules Enable the β-trefoil Architecture. *Protein Sci.* 29 (8), 1794–1802. doi:10.1002/pro.3899

Blaber, M. (2021). Cooperative Hydrophobic Core Interactions in the β-trefoil Architecture. *Protein Sci.* 30 (5), 956–965. doi:10.1002/pro.4059

Blaber, M., and Lee, J. (2012). Designing Proteins from Simple Motifs: Opportunities in Top-Down Symmetric Deconstruction. *Curr. Opin. Struct. Biol.* 22 (4), 442–450. doi:10.1016/j.sbi.2012.05.008

Blow, D. M., Janin, J., and Sweet, R. M. (1974). Mode of Action of Soybean Trypsin Inhibitor (Kunitz) as a Model for Specific Protein-Protein Interactions. *Nature* 249, 54–57. doi:10.1038/249054a0

Bolon, D. N., Voigt, C. A., and Mayo, S. L. (2002). De Novo design of Biocatalysts. *Curr. Opin. Chem. Biol.* 6 (2), 125–129. doi:10.1016/S1367-5931(02)00303-4

Bovi, M., Cenci, L., Perduca, M., Capaldi, S., Carrizo, M. E., Civiero, L., et al. (2012). BEL β-trefoil: A Novel Lectin with Antineoplastic Properties in king Bolete (Boletus Edulis) Mushrooms. *Glycobiology* 23 (5), 578–592. doi:10.1093/glycob/cws164

Broom, A., Doxey, A. C., Lobsanov, Y. D., Berthin, L. G., Rose, D. R., Howell, P. L., et al. (2012). Modular Evolution and the Origins of Symmetry: Reconstruction of a Three-fold Symmetric Globular Protein. *Structure* 20, 161–171. doi:10.1016/j.str.2011.10.021

Broom, A., Ma, S. M., Xia, K., Rafalia, H., Trainor, K., Colón, W., et al. (2015). Designed Protein Reveals Structural Determinants of Extreme Kinetic Stability. *Proc. Natl. Acad. Sci. U.S.A.* 112, 14605–14610. doi:10.1073/pnas.1510748112

Brunette, T., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., et al. (2015). Exploring the Repeat Protein Universe through Computational Protein Design. *Nature* 528, 580–584. doi:10.1038/nature16162

Brych, S. R., Dubey, V. K., Bienkiewicz, E., Lee, J., Logan, T. M., and Blaber, M. (2004). Symmetric Primary and Tertiary Structure Mutations within a Symmetric Superfold: a Solution, Not a Constraint, to Achieve a Foldable Polypeptide. *J. Mol. Biol.* 344 (3), 769–780. doi:10.1016/j.jmb.2004.09.060

Chiapparino, A., Grbavac, A., Jonker, H. R., Hackmann, Y., Mortensen, S., Zatorska, E., et al. (2020). Functional Implications of MIR Domains in Protein O-Mannosylation. *eLife* 9, e61189. doi:10.7554/eLife.61189

Claren, J., Malisi, C., Höcker, B., and Sterner, R. (2009). Establishing Wild-type Levels of Catalytic Activity on Natural and Artificial $(\beta\alpha)_8$-barrel Protein Scaffolds. *Proc. Natl. Acad. Sci. U.S.A.* 106 (10), 3704–3709. doi:10.1073/pnas.0810342106

Crooks, G. E., Wolfe, J., and Brenner, S. E. (2004). Measurements of Protein Sequence-Structure Correlations. *Proteins* 57 (4), 804–810. doi:10.1002/prot.20262

Dwyer, M. A., Looger, L. L., and Hellinga, H. W. (2004). Computational Design of a Biologically Active Enzyme. *Science* 304 (5679), 1967–1971. doi:10.1126/science.1098432

Eck, R. V., and Dayhoff, M. O. (1966). Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences. *Science* 152 (April 15), 363–366. doi:10.1126/science.152.3720.363

Faísca, P. F. N. (2009). The Nucleation Mechanism of Protein Folding: a Survey of Computer Simulation Studies. *J. Phys. Condens. Matter* 21 (37), 373102. doi:10.1088/0953-8984/21/37/373102

Ferruz, N., Noske, J., and Höcker, B. (2021). Protlego: A Python Package for the Analysis and Design of Chimeric Proteins. *Bioinformatics* 37 (19), 3182–3189. doi:10.1093/bioinformatics/btab253

Fersht, A. R., and Sato, S. (2004). Φ-Value Analysis and the Nature of Protein-Folding Transition States. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7976–7981. doi:10.1073/pnas.0402684101

Friedmann, D. R., Wilson, J. J., and Kovall, R. A. (2008). RAM-induced Allostery Facilitates Assembly of a Notch Pathway Active Transcription Complex. *J. Biol. Chem.* 283 (21), 14781–14791. doi:10.1074/jbc.M709501200

Guex, N., and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-Pdb Viewer: An Environment for Comparative Protein Modeling. *Electrophoresis* 18, 2714–2723. doi:10.1002/elps.1150181505

Höcker, B., Claren, J., and Sterner, R. (2004). Mimicking Enzyme Evolution by Generating New (βα) 8 -barrels from (βα) 4 -Half-Barrels. *Proc. Natl. Acad. Sci. U.S.A.* 101, 16448–16453. doi:10.1073/pnas.0405832101

Huang, P.-S., Feldmeier, K., Parmeggiani, F., Fernandez Velasco, D. A., Höcker, B., and Baker, D. (2016). De Novo design of a Four-fold Symmetric TIM-Barrel Protein with Atomic-Level Accuracy. *Nat. Chem. Biol.* 12 (1), 29–34. doi:10.1038/nchembio.1966

Jeltsch, A. (1999). Circular Permutations in the Molecular Evolution of DNA Methyltransferases. *J. Mol. Evol.* 49, 161–164. doi:10.1007/pl00006529

Kimura, R., Aumpuchin, P., Hamaue, S., Shimomura, T., and Kikuchi, T. (2020). Analyses of the Folding Sites of Irregular β-trefoil Fold Proteins through Sequence-Based Techniques and Gō-Model Simulations. *BMC Mol. Cel Biol* 21 (1), 1–17. doi:10.1186/s12860-020-00271-4

Kopec, K. O., and Lupas, A. N. (2013). β-Propeller Blades as Ancestral Peptides in Protein Evolution. *PLoS ONE* 8 (10), e77074. doi:10.1371/journal.pone.0077074

Lee, J., and Blaber, M. (2011). Experimental Support for the Evolution of Symmetric Protein Architecture from a Simple Peptide Motif. *Proc. Natl. Acad. Sci. U.S.A.* 108, 126–130. doi:10.1073/pnas.1015032108

Lee, J., Blaber, S. I., Dubey, V. K., and Blaber, M. (2011). A Polypeptide "Building Block" for the β-Trefoil Fold Identified by "Top-Down Symmetric Deconstruction". *J. Mol. Biol.* 407, 744–763. doi:10.1016/j.jmb.2011.02.002

Lindberg, M. O., Haglund, E., Hubner, I. A., Shakhnovich, E. I., and Oliveberg, M. (2006). Identification of the Minimal Protein-Folding Nucleus through Loop-Entropy Perturbations. *Proc. Natl. Acad. Sci. U.S.A.* 103 (11), 4083–4088. doi:10.1073/pnas.0508863103

Longo, L. M., Lee, J., Tenorio, C. A., and Blaber, M. (2013). Alternative Folding Nuclei Definitions Facilitate the Evolution of a Symmetric Protein Fold from a Smaller Peptide Motif. *Structure* 21, 2042–2050. doi:10.1016/j.str.2013.09.003

Longo, L. M., Kumru, O. S., Middaugh, C. R., and Blaber, M. (2014). Evolution and Design of Protein Structure by Folding Nucleus Symmetric Expansion. *Structure* 22, 1377–1384. doi:10.1016/j.str.2014.08.008

McLachlan, A. D. (1972). Repeating Sequences and Gene Duplication in Proteins. *J. Mol. Biol.* 64 (2), 417–437. doi:10.1016/0022-2836(72)90508-6

McLachlan, A. D. (1979). Three-fold Structural Pattern in the Soybean Trypsin Inhibitor (Kunitz). *J. Mol. Biol.* 133, 557–563. doi:10.1016/0022-2836(79)90408-x

Mukhopadhyay, D. (2000). The Molecular Evolutionary History of a Winged Bean α-Chymotrypsin Inhibitor and Modeling of its Mutations through Structural Analyses. *J. Mol. Evol.* 50, 214–223. doi:10.1007/s002399910024

Murzin, A. G., Lesk, A. M., and Chothia, C. (1992). β-Trefoil Fold. Patterns Of Structure And Sequence In The Kunitz Inhibitors Interleukins-1β and 1α And Fibroblast Growth Factors. *J. Mol. Biol.* 223, 531–543. doi:10.1016/0022-2836(92)90668-a

Nagi, A. D., Anderson, K. S., and Regan, L. (1999). Using Loop Length Variants to Dissect the Folding Pathway of a four-helix-bundle Protein. *J. Mol. Biol.* 286 (1), 257–265. doi:10.1006/jmbi.1998.2474

Nikkhah, M., Jawad-Alami, Z., Demydchuk, M., Ribbons, D., and Paoli, M. (2006). Engineering of β-propeller Protein Scaffolds by Multiple Gene Duplication and Fusion of an Idealized WD Repeat. *Biomol. Eng.* 23, 185–194. doi:10.1016/j.bioeng.2006.02.002

Notenboom, V., Boraston, A. B., Williams, S. J., Kilburn, D. G., and Rose, D. R. (2002). High-Resolution Crystal Structures of the Lectin-like Xylan Binding Domain from Streptomyces Lividans Xylanase 10A with Bound Substrates Reveal a Novel Mode of Xylan Binding, *Biochemistry* 41 (13), 4246–4254. doi:10.1021/bi015865j

Ohno, S. (1970). *Evolution by Gene Duplication*. New York: Allen & Unwin.

Olsen, S. K., Ibrahimi, O. A., Raucci, A., Zhang, F., Eliseenkova, A. V., Yayon, A., et al. (2004). Insights into the Molecular Basis for Fibroblast Growth Factor Receptor Autoinhibition and Ligand-Binding Promiscuity. *Proc. Natl. Acad. Sci. U.S.A.* 101, 935–940. doi:10.1073/pnas.0307287101

Parker, J. B., Tenorio, C. A., and Blaber, M. (2021). The Ubiquitous Buried Water in the Beta-Trefoil Architecture Contributes to the Folding Nucleus and ~20% of the Folding Enthalpy. *Protein Sci.* 30 (11), 2287–2297. doi:10.1002/pro.4192

Peisajovich, S. G., Rockah, L., and Tawfik, D. S. (2006). Evolution of New Protein Topologies through Multistep Gene Rearrangements. *Nat. Genet.* 38, 168–174. doi:10.1038/ng1717

Ponting, C. P., and Russell, R. B. (2000). Identification of Distant Homologues of Fibroblast Growth Factors Suggests a Common Ancestor for All β-trefoil Proteins 1 1Edited by J. Thornton. *J. Mol. Biol.* 302, 1041–1047. doi:10.1006/jmbi.2000.4087

Renko, M., Sabotič, J., Mihelič, M., Brzin, J., Kos, J., and Turk, D. (2010). Versatile Loops in Mycocypins Inhibit Three Protease Families. *J. Biol. Chem.* 285 (1), 308–316. doi:10.1074/jbc.M109.043331

Richter, M., Bosnali, M., Carstensen, L., Seitz, T., Durchschlag, H., Blanquart, S., et al. (2010). Computational and Experimental Evidence for the Evolution of a $(\beta\alpha)_8$-Barrel Protein from an Ancestral Quarter-Barrel Stabilised by Disulfide Bonds. *J. Mol. Biol.* 398, 763–773. doi:10.1016/j.jmb.2010.03.057

Schneider, T. D., and Stephens, R. M. (1990). Sequence Logos: a New Way to Display Consensus Sequences. *Nucl. Acids Res.* 18 (20), 6097–6100. doi:10.1093/nar/18.20.6097

Song, H. K., and Suh, S. W. (1998). Kunitz-type Soybean Trypsin Inhibitor Revisited: Refined Structure of its Complex with Porcine Trypsin Reveals an Insight into the Interaction between a Homologous Inhibitor from Erythrina Caffra and Tissue-type Plasminogen Activator. *J. Mol. Biol.* 275, 347–363. doi:10.1006/jmbi.1997.1469

Suzuki, R., Kuno, A., Hasegawa, T., Hirabayashi, J., Kasai, K.-i., Momma, M., et al. (2009). Sugar-complex Structures of the C-Half Domain of the Galactose-Binding Lectin EW29 from the earthwormLumbricus Terrestris. *Acta Crystallogr. D Biol. Cryst.* 65**,** 49–57. doi:10.1107/S0907444908037451

Sweet, R. M., Wright, H. T., Janin, J., Chothia, C. H., and Blow, D. M. (1974). Crystal Structure of the Complex of Porcine Trypsin with Soybean Trypsin Inhibitor (Kunitz) at 2.6 Angstrom Resolution. *Biochemistry* 13, 4212–4228. doi:10.1021/bi00717a024

Tenorio, C. A., Parker, J. B., and Blaber, M. (Forthcoming 2022). Functionalization of a Symmetric Protein Scaffold: Redundant Folding Nuclei and Alternative Oligomeric Folding Pathways. *Protein Sci.* 31. doi:10.1002/pro.4301

Tenorio, C. A., Parker, J. B., and Blaber, M. (2020). Oligomerization of a Symmetric β-trefoil Protein in Response to Folding Nucleus Perturbation. *Protein Sci.* 29 (7), 1629–1640. doi:10.1002/pro.3877

Terada, D., Voet, A. R. D., Noguchi, H., Kamata, K., Ohki, M., Addy, C., et al. (2017). Computational Design of a Symmetrical β-trefoil Lectin with Cancer Cell Binding Activity. *Sci. Rep.* 7 (1), 5943. doi:10.1038/s41598-017-06332-7

Thompson, M. J., and Eisenberg, D. (1999). Transproteomic Evidence of a Loop-Deletion Mechanism for Enhancing Protein Thermostability. *J. Mol. Biol.* 290, 595–604. doi:10.1006/jmbi.1999.2889

Transue, T. R., Smith, A. K., Mo, H., Goldstein, I. J., and Saper, M. A. (1997). Structure of Benzyl T-Antigen Disaccharide Bound to Amaranthus Caudatus Agglutinin. *Nat. Struct. Mol. Biol.* 4 (10), 779–783. doi:10.1038/nsb1097-779

Veerapandian, B., Gilliland, G. L., Raag, R., Svensson, A. L., Masui, Y., Hirai, Y., et al. (1992). Functional Implications of Interleukin-1β Based on the Three-Dimensional Structure. *Proteins* 12, 10–23. doi:10.1002/prot.340120103

Voet, A. R. D., Noguchi, H., Addy, C., Simoncini, D., Terada, D., Unzai, S., et al. (2014). Computational Design of a Self-Assembling Symmetrical β-propeller Protein. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15102–15107. doi:10.1073/pnas.1412768111

Went, H. M., and Jackson, S. E. (2005). Ubiquitin Folds through a Highly Polarized Transition State. *Protein Eng. Des. Selec.* 18 (5), 229–237. doi:10.1093/protein/gzi025

West, M. W., and Hecht, M. H. (1995). Binary Patterning of Polar and Nonpolar Amino Acids in the Sequences and Structures of Native Proteins. *Protein Sci.* 4 (10), 2032–2039. doi:10.1002/pro.5560041008

Xia, X., Longo, L. M., Sutherland, M. A., and Blaber, M. (2016). Evolution of a Protein Folding Nucleus. *Protein Sci.* 25 (7), 1227–1240. doi:10.1002/pro.2848

Xiong, H., Buckwalter, B. L., Shieh, H. M., and Hecht, M. H. (1995). Periodicity of Polar and Nonpolar Amino Acids Is the Major Determinant of Secondary Structure in Self-Assembling Oligomeric Peptides. *Proc. Natl. Acad. Sci. U.S.A.* 92 (14), 6349–6353. doi:10.1073/pnas.92.14.6349

Yadid, I., and Tawfik, D. S. (2007). Reconstruction of Functional β-Propeller Lectins via Homo-Oligomeric Assembly of Shorter Fragments. *J. Mol. Biol.* 365, 10–17. doi:10.1016/j.jmb.2006.09.055

Conflict of Interest: MB is a cofounder and has equity ownership in Trefoil Therapeutics Inc.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.