Modular Evolution and the Origins of Symmetry: Reconstruction of a Three-Fold Symmetric Globular Protein

Aron Broom,[1,5] Andrew C. Doxey,[2,5,6] Yuri D. Lobsanov,[3] Lisa G. Berthin,[1] David R. Rose,[2] P. Lynne Howell,[3,4] Brendan J. McConkey,[2,*] and Elizabeth M. Meiering[1,*]
[1]Guelph-Waterloo Centre for Graduate Studies in Chemistry and Biochemistry
[2]Department of Biology
University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada
[3]Program in Molecular Structure and Function, The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada
[4]Department of Biochemistry, Faculty of Medicine, University of Toronto, Toronto, Ontario M5S 1A8, Canada
[5]These authors contributed equally to this work
[6]Present address: Department of Developmental Biology, Stanford University, Stanford, CA, 94305, USA
*Correspondence: meiering@uwaterloo.ca (E.M.M.), mcconkey@uwaterloo.ca (B.J.M.)
DOI 10.1016/j.str.2011.10.021

SUMMARY

The high frequency of internal structural symmetry in common protein folds is presumed to reflect their evolutionary origins from the repetition and fusion of ancient peptide modules, but little is known about the primary sequence and physical determinants of this process. Unexpectedly, a sequence and structural analysis of symmetric subdomain modules within an abundant and ancient globular fold, the β-trefoil, reveals that modular evolution is not simply a relic of the ancient past, but is an ongoing and recurring mechanism for regenerating symmetry, having occurred independently in numerous existing β-trefoil proteins. We performed a computational reconstruction of a β-trefoil subdomain module and repeated it to form a newly three-fold symmetric globular protein, ThreeFoil. In addition to its near perfect structural identity between symmetric modules, ThreeFoil is highly soluble, performs multivalent carbohydrate binding, and has remarkably high thermal stability. These findings have far-reaching implications for understanding the evolution and design of proteins via subdomain modules.

INTRODUCTION

Internal structural symmetry is observed very frequently in common protein folds (Orengo et al., 1994) and is thought to have arisen from the ancient evolution of these folds via the repetition and fusion of smaller peptide modules (Söding and Lupas, 2003; Lupas et al., 2001). The well-established occurrence of sequence duplication and fusion events in protein evolution (Orengo et al., 1994; Eisenbeis and Höcker, 2010) supports the structural evidence for this evolutionary mechanism. However, in modern globular proteins, symmetry at the primary sequence level is typically relatively low to undetectable, owing to sequence divergence (Söding and Lupas, 2003; Lupas et al., 2001). This represents a challenge for understanding the origins and molecular determinants of symmetric protein evolution. Elucidating how symmetric protein structures can be constructed from a set of basic “building blocks” or subdomain modules has far-reaching implications not only for understanding evolution, but also for rational protein design.

Seminal studies on $(\beta\alpha)_8$-barrel proteins have provided experimental proof of principle for the evolution of symmetric globular folds via the repetition of subdomain modules (Houbrechts et al., 1995; Richter et al., 2010; Höcker et al., 2009). Sterner, Höcker, and colleagues identified sequence and structural evidence for the evolution of this fold from a $(\beta\alpha)_4$-half-barrel ancestor (Lang et al., 2000). By fusing two identical copies of a half-barrel and stabilizing the resulting protein using a combination of rationally designed mutations and mutations selected from a library of variants, they obtained a stable and symmetric structure, although it was lacking in function (Höcker et al., 2009; Höcker et al., 2004). As protein design experiments often fail to produce the intended structure or properties (Houbrechts et al., 1995; Dantas et al., 2003; Hill et al., 2000), and data for other globular symmetric folds is limited, additional investigations are needed. The recent explosive growth in the availability of protein sequences and structures from genomics initiatives combined with new tools for reconstructing and designing proteins have set the stage for such investigations.

The focus of this study is the internally symmetric β-trefoil structure, an ancient fold adopted by many proteins with a great diversity of sequences and ligand-binding functions (Ponting and Russell, 2000; Murzin et al., 1992). β-trefoils currently include at least 14 families according to Pfam (Finn et al., 2010), such as the carbohydrate-binding ricin and agglutinin toxins, actin-bundling proteins, the fibroblast growth factor (FGF) and interleukin-1 cytokines, STI-like protease inhibitors, and LAG-1 DNA-binding proteins. The β-trefoil fold displays three-fold internal structural symmetry (Figure 1), and internal sequence similarities have been noted in some families, such as the multivalent sugar-binding ricins and actin-bundling

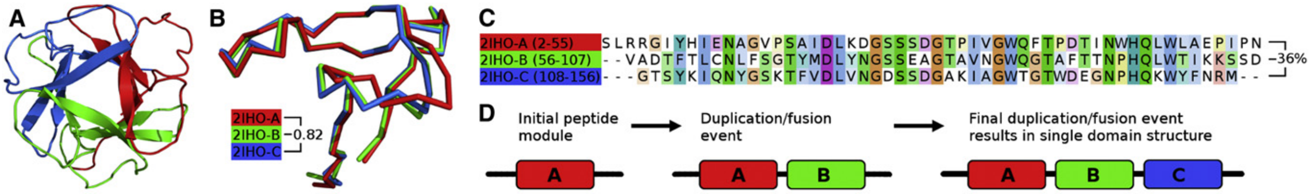
A
B
2IHO-A
2IHO-B –0.82
2IHO-C
C
2IHO-A (2-55) SLRRGIYHIENAGVPSAIDLKDGSSSDGTPIVGWQFTPDTINWHQLWLAEPIPN
2IHO-B (56-107) --VADTFTLCNLFSGTYMDLYNGSSEAGTAVNGWQGTAFTTNPHQLWTIKKSSD –36%
2IHO-C (108-156) ---GTSYKIQNYGSKTFVDLVNGDSSDGAKIAGWTGTWDEGNPHQKWYFNRM--
D
Initial peptide module
Duplication/fusion event
Final duplication/fusion event results in single domain structure
A
A
B
A
B
C

Figure 1. Internal Symmetry in the β-Trefoil Fold

(A) Structure of a typical β-trefoil domain (the *Marasmius oreades* mushroom lectin, PDB ID 2IHO [Grahn et al., 2007]), looking down the threefold symmetry axis. Subdomains are colored red, green, and blue from N- to C-terminus.

(B) Structural alignment of the C-α trace of subdomains, aligned using SSM (Krissinel and Henrick, 2004) and showing the average structural identity defined by SSM's Q-score parameter.

(C) The sequence alignment of 2IHO subdomains. The average sequence identity between pairs of aligned subdomains as determined using MUSCLE (Edgar, 2004) is 36%.

(D) One possible model of β-trefoil evolution from a single subdomain. All structure images were rendered using PyMol (http://www.pymol.org/).

proteins (Ponting and Russell, 2000). Each of the three subdomain modules is composed of four β-strands, with two strands from each module collectively forming a six-stranded β-barrel and the remaining two from each module together forming a β-hairpin triplet that caps one end of the barrel. Previous sequence analyses have suggested that modern β-trefoil proteins share a common homotrimer ancestor of identical subdomain modules (Ponting and Russell, 2000; Murzin et al., 1992; Mukhopadhyay, 2000; McLachlan, 1979). A recent parallel study to the work herein demonstrated the feasibility of this evolutionary model by constructing both homotrimer and fused three-fold symmetric β-trefoil structures, developed from a cytokine FGF template using rational design and library screening, in which protein function was lost (Lee and Blaber, 2011; Lee et al., 2011). This, together with a wealth of previous structural and folding studies on β-trefoil proteins (Murzin et al., 1992; Houliston et al., 2002; Dubey et al., 2007; Capraro et al., 2008), makes them an attractive candidate for examining modular evolution.

We report herein a large-scale sequence clustering analysis of β-trefoils. Based on previous analyses, we expected to find evidence for a single ancient homotrimer and triplication event (Ponting and Russell, 2000; Murzin et al., 1992; Mukhopadhyay, 2000; McLachlan, 1979). To our surprise, our analysis revealed that the repetition and fusion of subdomain modules to form new symmetric β-trefoils is an ongoing and recurring process that has occurred numerous times. We then reconstructed a completely three-fold symmetric β-trefoil sequence by using consensus sequence and protein design based on a carbohydrate-binding ricin sequence identified from the clustering results. The designed protein, ThreeFoil, forms a structure whose subdomain modules are essentially identical, and furthermore, it exhibits extremely high thermal stability, as well as functional multivalent carbohydrate-binding properties. The single-subdomain module, OneFoil, is poorly structured, however. Consequently, incorporating symmetry may be attractive both in evolution and in the design of multivalent binding proteins.

RESULTS

Sequence Analysis of β-Trefoil Subdomains Reveals Recurring Modular Evolution

We analyzed the evolutionary relationships among β-trefoil subdomain modules by constructing a dataset of subdomain modules and clustering these according to sequence similarity. First, a dataset of 1167 nonredundant sequences annotated as β-trefoils was obtained using the Conserved Domain Database from the National Center for Biotechnology Information (Marchler-Bauer et al., 2009). This set included members of 11 β-trefoil families, each with a representative of known structure (Table 1). Through alignment to representative structures, each β-trefoil sequence was subdivided into three β-β-β-loop-β subdomains, the putative building block of the β-trefoil fold (Murzin et al., 1992) (Figures 1B and 1C). In order to assess the evolutionary relationships between the subdomains, they were clustered by sequence similarity, where each subdomain pair with E < 1e-04 was connected.

Remarkably, we found a pattern of greater similarity between subdomains within a given β-trefoil sequence than between subdomains from different β-trefoils. Together these findings reveal ongoing evolution in which a distinct single-subdomain module was repeated to form a new symmetric protein. The predominant accepted model of protein domain evolution is the duplication and divergence of whole domains (Orengo and Thornton, 2005). According to this model, a given β-trefoil module should be most similar to the same module in a closely related sequence. Indeed, this mode of evolution is observed for the majority of subdomains, as is illustrated for a pair of proteins (Figure 2A), and for the entire dataset of sequences (representative clusters in Figure 3B; all clusters in Figure S1C available online). Strikingly, however, there are also multiple examples where each β-trefoil subdomain module is most similar to the other two modules within the same protein sequence, and less similar to the modules of other closely related β-trefoils. We illustrate this pattern for a pair of proteins (Figure 2B) and for the entire dataset of sequences, where we identified nine cases of subdomain module repetition through our clustering analysis, in the ricin, AbfB, and fascin families (representative clusters in Figure 3C; all clusters in Figure S1B; sequence alignments for representatives of each cluster in Table S1). To more sensitively detect subdomain-repetition events, including those occurring within a cluster, we performed a phylogenetic analysis of the subdomains showing the highest internal symmetry and identified nine additional (i.e., 18 total) distinct subdomain-repetition events (Figure S1D). These repetition events occurred most prominently within the ricin family, which included the eight sequences with greater similarity

Table 1. Dataset Construction and Calculated Sequence Symmetries

Family	CDD IDs[a]	Extracted Domain Sequences (No.)[b]	Domains after Filtering (No.)[c]	Average Sequence Symmetry[d]	Representative Structure Used for Alignment
AbfB	68828, pfam05270	24	15	17.7	1WD3
Agglutinin	70918, pfam07468	14	5	9.0	1JLX
CD Toxin	80015, pfam03498	69	28	9.2	1SR4
Fascin	29332, cd00257 87053, pfam06268	413	129	13.0	1DFC
FGF	28940, cd00058 47749, smart00442 84576, pfam00167	775	140	10.3	1NUN
IL1	28984, cd00100 64217, pfam00340	362	86	8.2	1MD6
STI/Kunitz	29140, cd00178 84601, pfam00197	452	89	7.7	1WBA
LAG1	72686, pfam09270	31	18	7.0	1TTU
MIR	86128, pfam02815	1267	65	10.3	1T9F
Ricin	29101, cd00161 47764, smart00458 84930, pfam00652	1604	518	14.3	1QXM
Toxin R Bind C	87408, pfam07951	89	15	7.3	3BTA

[a] See ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/cdd versions for conserved domain model accessions and version information.
[b] Sequences of “all related families” for each CDD ID were retrieved from the NCBI.
[c] Filtering involved removal of redundancy and partial sequences (see Experimental Procedures).
[d] The mean pairwise percent ID between the three repeats in each domain was calculated. The average sequence symmetry is the mean of this value for all domains in the family.

between subdomains than with any other subdomains (Figures S1D and S1E). This pattern of greater internal than external similarity demonstrates ongoing evolution of new β-trefoil folds by repetition of subdomain modules. The size of the clusters (Figure 3C) and the interrelationship of subdomains (Figure S1D) shows that such subdomain-repetition events may be preceded or followed by whole-domain duplication. A similar process of subdomain repetition has been postulated for the nonglobular β-propeller fold (Chaudhuri et al., 2008), and may apply for $(\beta\alpha)_8$-barrels (Richter et al., 2010) and many other internally symmetric protein folds (see Discussion).

Reconstructing a Progenitor Subdomain Module through Sequence Analyses and Computational Design

We tested the physical feasibility of evolution via subdomain repetition and fusion by using a combination of sequence analysis and computational design to reconstruct a β-trefoil consisting of three identical subdomain modules. We reasoned that the natural β-trefoil sequence with the highest internal sequence symmetry identified in the clustering analysis would be a good starting point. This was a member of the ricin family annotated as the carbohydrate binding module of a glycosidase from the halophilic red archaeon *Haloarcula marismortui* (NCBI accession no. AAV45265), which has 55% amino acid identity between all three modules.

In order to reconstruct a completely symmetrical β-trefoil, ThreeFoil, three steps were used to incorporate information from the template sequence, homologous sequences, and rational protein design. In the first step, the template sequence was split into its three constituent subdomain modules (Figure 4, Template), and those residues conserved in all three modules were fixed (Figure 4, Step 1); this left 21 of 47 positions undefined. In the second step, a small set of 13 highly homologous sequences were identified and aligned with the template sequence and split into their corresponding subdomain modules, and the residue frequency was calculated at each position (Figure 4, Homology; see Figure S2 for homologous subdomain module alignments). The residue frequency at each position was averaged between the homologous sequences and the template, and any residue with an average frequency >0.5 (50%) was incorporated into the reconstructed sequence (Figure 4, Step 2). This left 16 positions undefined.

The third step of reconstruction made use of computational design in the form of Rosetta Design (Dantas et al., 2003). Allowing only the 16 undefined positions to vary, Rosetta Design generated a set of 10,000 energetically favorable sequences, and the residue frequency at each position was calculated (Figure 4, Rosetta). Two points of concern were limitations in the successful design of all-β proteins using Rosetta Design (Dantas et al., 2003; Hu et al., 2008) and the low sequence conservation at some positions. To address this, the overall residue frequency at each position was calculated by equally weighting the frequency of residues in the template, in the homologous sequences, and from Rosetta Design, with the most frequent residue at a given position being incorporated into the final reconstructed sequence (Figure 4, Step 3). This approach allowed for inclusion of residues important for function and stability based on consensus information from the template and homologous sequences (Wetzel et al., 2008); at the same time, it allowed energetically favorable residues identified by

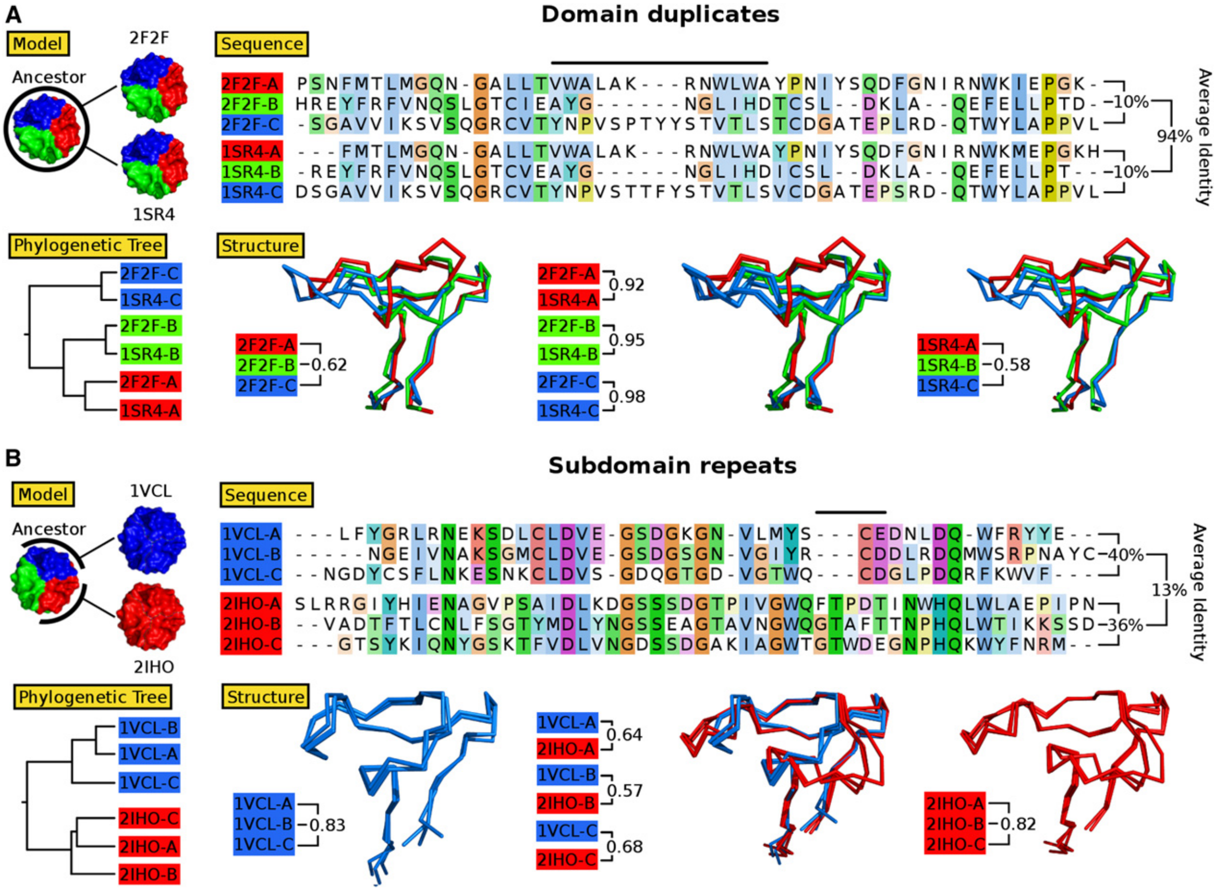
A
Domain duplicates
Model
2F2F
Ancestor
1SR4
Sequence
2F2F-A PSNFMTLMGQN-GALLTVWALAK---RNWLWAYPNIYSQDFGNIRNWKIEPGK-
2F2F-B HREYFRFVNQSLGTCIEAYG------NGLIHDTCSL--DKLA--QEFELLPTD-
2F2F-C -SGAVVIKSVSQGRCVTYNPVSPTYYSTVTLSTCDGATEPLRD-QTWYLAPPVL
1SR4-A ---FMTLMGQN-GALLTVWALAK---RNWLWAYPNIYSQDFGNIRNWKMEPGKH
1SR4-B -REYFRFVNQSLGTCVEAYG------NGLIHDICSL--DKLA--QEFELLPT--
1SR4-C DSGAVVIKSVSQGRCVTYNPVSTTFYSTVTLSVCDGATEPSRD-QTWYLAPPVL
10%
10%
94%
Average Identity
Phylogenetic Tree
2F2F-C
1SR4-C
2F2F-B
1SR4-B
2F2F-A
1SR4-A
Structure
2F2F-A
2F2F-B
2F2F-C
0.62
2F2F-A
1SR4-A
0.92
2F2F-B
1SR4-B
0.95
2F2F-C
1SR4-C
0.98
1SR4-A
1SR4-B
1SR4-C
0.58
B
Subdomain repeats
Model
1VCL
Ancestor
2IHO
Sequence
1VCL-A ---LFYGRLRNEKSDLCLDVE-GSDGKGN-VLMYS---CEDNLDQ-WFRYYE--
1VCL-B -----NGEIVNAKSGMCLDVE-GSDGSGN-VGIYR---CDDLRDQMWSRPNAYC
1VCL-C --NGDYCSFLNKESNKCLDVS-GDQGTGD-VGTWQ---CDGLPDQRFKWVF---
2IHO-A SLRRGIYHIENAGVPSAIDLKDGSSSDGTPIVGWQFTPDTINWHQLWLAEPIPN
2IHO-B --VADTFTLCNLFSGTYMDLYNGSSEAGTAVNGWQGTAFTTNPHQLWTIKKSSD
2IHO-C ---GTSYKIQNYGSKTFVDLVNGDSSDGAKIAGWTGTWDEGNPHQKWYFNRM--
40%
36%
13%
Average Identity
Phylogenetic Tree
1VCL-B
1VCL-A
1VCL-C
2IHO-C
2IHO-A
2IHO-B
Structure
1VCL-A
1VCL-B
1VCL-C
0.83
1VCL-A
2IHO-A
0.64
1VCL-B
2IHO-B
0.57
1VCL-C
2IHO-C
0.68
2IHO-A
2IHO-B
2IHO-C
0.82

Figure 2. Evolution by Whole-Domain Duplication versus Subdomain Repetition

β-trefoils are labeled with their PDB IDs, and different subdomains are colored as in Figure 1. Evolution by whole-domain duplication (A) and subdomain module repetition mechanisms (B). For each mechanism, a set of representative amino acid sequences given in single letter code illustrates the respective mode of evolution. In (A) and (B), a structural representation of each evolutionary mode is shown with space-filled β-trefoil structures (Model), illustrating the putative evolutionary path, which is supported by the phylogenetic tree inferred from the sequence alignment (Sequence; black bars highlight key regions) using Phylip (http://evolution.genetics.washington.edu/phylip.html) and MUSCLE (Edgar, 2004), respectively. In addition, the same pattern is seen in structure alignments (Structure) made using SSM (Krissinel and Henrick, 2004). Sequence identities were given by MUSCLE and structural similarities were defined by the Q-Score parameter in SSM (with a score of 1.0 representing identical C-α traces).

Rosetta Design to be incorporated. In this respect, it is reassuring to note that the residues most frequently identified by Rosetta Design were well represented in the homologous sequences.

The final reconstructed subdomain module sequence was expressed as a single module, OneFoil, and as a fused three-fold repeat, ThreeFoil. The proteins were characterized as described below.

ThreeFoil Has Near-Perfect Three-Fold Structural and Ligand-Binding Symmetry

The structure of ThreeFoil was determined by X-ray crystallography to a resolution of 1.62 Å, and refined to high quality (see Table 2 for refinement statistics; structure deposited as PDB code 3PG0). The structure exhibits exceptionally high symmetry, as evidenced by a very low backbone RMSD of only 0.2 Å between subdomain modules (Figures 5A–5C). The symmetry of ThreeFoil is also apparent in its binding of ligands, including galactose, a metal ion, and ordered water molecules. The template sequence from *Haloarcula marismortui* is annotated by BLAST (Altschul et al., 1997) as a member of the ricin family of β-trefoils, which bind carbohydrates (often terminating in galactose) in a shallow pocket formed by the second and third β-strands and the long loop between strands 3 and 4 (Hazes, 1996). This pocket in ThreeFoil contains bound bis-tris from the crystallization buffer in all three symmetric units (Figure 5D). Bis-tris has been shown previously to occupy expected active or binding sites within a protein (Stenmark et al., 2004), and given its many hydroxyl groups, it likely mimics the natural carbohydrate ligand. The binding of D-galactose to ThreeFoil was measured via changes in the intrinsic protein fluorescence upon sugar binding (Figure 6A), giving a dissociation constant (K_d) of ~1 mM, which is very similar to the measured K_d for D-galactose binding to one of the proteins used in the homology modeling of ThreeFoil (Winter et al., 2002). In addition, the binding of ThreeFoil to a series of glycans was measured using a glycan array. The results clearly show that ThreeFoil's symmetry allows for multivalent binding, as seen in the pronounced improvement

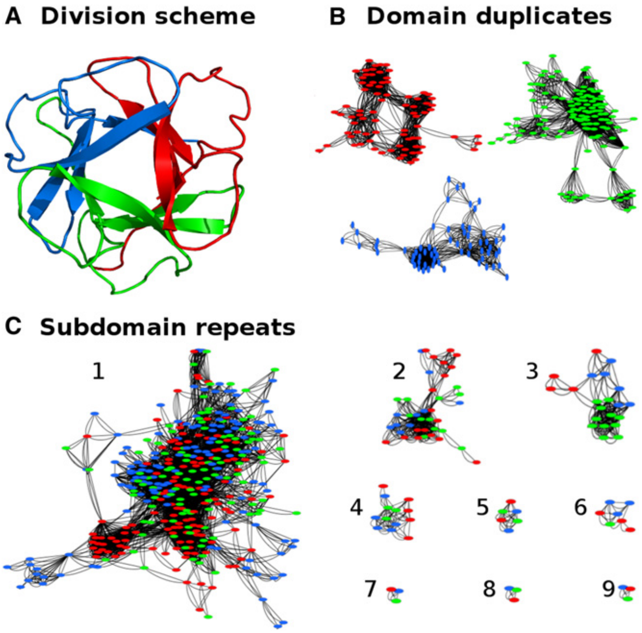

Figure 3. Representative Sequence Clusters of β-Trefoil Subdomain Modules Show Occurrences of Both Whole-Domain Duplication and Subdomain Repetition

Individual subdomain modules represented by an oval are colored as in Figure 1 and clustered according to sequence similarity, as described in the Experimental Procedures and Results sections.

(A) Division scheme for splitting of β-trefoil domains into three constituent symmetrical subdomain modules.

(B) Representative clusters demonstrating whole-domain duplication and divergence resulting in subdomains that are most closely related to the corresponding subdomain in homologous sequences.

(C) Clusters demonstrating evolution via subdomain repetition. Internal subdomains are more closely related to one another than to extant subdomains. Clusters are numbered as in Table S1, and all clusters can be seen in Figure S1C. In addition, a phylogenetic tree and heat map of the most internally symmetric subdomains are shown in Figure S1D, with a boxplot of internal symmetry by sequence family in Figure S1E.

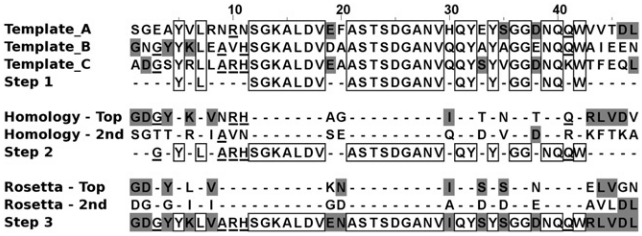
10
20
30
40
Template_A SGEAYVLRNRNSGKALDVEFASTSDGANVHQYEYSGGDNQQWVVTDL
Template_B GNGYYKLEAVHSGKALDVDAASTSDGANVQQYAYAGGENQQWAIEEN
Template_C ADGSYRLLARHSGKALDVEAASTSDGANVQQYSYVGGDNQKWTFEQL
Step 1 ----Y-L----SGKALDV--ASTSDGANV-QY-Y-GG-NQ-W-----
Homology - Top GDGY-K-VNRH--------AG----------I--T-N--T--Q-RLVDV
Homology - 2nd SGTT-R-IAVN--------SE----------Q--D-V--D--R-KFTKA
Step 2 --G-Y-L-ARHSGKALDV--ASTSDGANV-QY-Y-GG-NQQW-----
Rosetta - Top GD-Y-L-V-----------KN----------I--S-S--N----ELVGN
Rosetta - 2nd DG-G-I-I-----------GD----------A--D-D--E----AVLDL
Step 3 GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL

Figure 4. Reconstruction of a Three-Fold Symmetric Sequence, ThreeFoil

The template sequence is shown split into its three subdomain modules (Template A–Template C), with conserved residues used to reconstruct the partial identity of the putative progenitor subdomain module (Step 1, boxed residues). The frequency of amino acids at each remaining position in a set of homologous sequences (see Figure S2) (most frequently seen amino acids in Homology [top], second most frequent in Homology [2nd]) were used together with the template sequence for further reconstruction (Step 2, underlined residues). The frequency of amino acids at the remaining positions, as predicted in a set of low-energy Rosetta Design models (most frequently predicted amino acids in Rosetta [Top], second most frequent in Rosetta [2nd]), was used with the template and homologous sequence data to complete reconstruction (Step 3, gray highlighting). For more details, see Results and Experimental Procedures.

in binding from a single glycan chain to a multiantennary one (Figure 6B).

In addition to binding carbohydrates, many β-trefoil structures have structurally conserved buried water molecules in each symmetrical unit (Murzin et al., 1992). ThreeFoil also binds three symmetrical buried water molecules, which make important bridging hydrogen bonds between strands 1, 2, and 4 of each symmetrical unit (Figure 5E). Finally, ThreeFoil binds a single metal ion along the three-fold axis of symmetry, which coordinates one backbone oxygen atom and one side-chain asparagine oxygen atom from each symmetric unit (Figure 5F) in an octahedral manner. The presence of a metal ion located on the axis of symmetry is very common for cyclically symmetric protein structures (Goodsell and Olson, 2000), and may point to a primordial role for metal ions in stabilizing symmetric structures.

ThreeFoil Is Well Behaved in Solution: Monomeric, Highly Soluble, and Extremely Stable

Since many designed proteins have a tendency to misfold or aggregate (Houbrechts et al., 1995; Dantas et al., 2003; Hill et al., 2000), we further tested the success of the ThreeFoil design using a battery of biophysical measurements. These showed that ThreeFoil is a highly soluble monomer with high thermal stability. Static light scattering (SLS) (Figure 7A), dynamic light scattering (DLS), and size exclusion chromatography (SEC) (Figures S3A and S3B) showed that ThreeFoil is a highly soluble monomer in solution. ^{1}H-NMR spectroscopy showed that ThreeFoil is well folded and has a well defined structure (Figure 7B), with numerous downfield amide resonances, as expected for β-sheet structure, and upfield methyl resonances indicating a well packed hydrophobic core. In addition, ThreeFoil unfolds at a high temperature of ~94°C by differential scanning calorimetry (DSC) (Figure 7C), further demonstrating its stability.

It is interesting that the single-peptide module used to generate ThreeFoil, termed OneFoil, is sufficiently stable for expression; however, it appears to be unfolded, based on NMR (Figure 7D) and fluorescence (Figure S3C). In contrast, fluorescence spectroscopy of ThreeFoil showed that aromatic residues undergo a very pronounced blue shift upon folding (Figure S3D), characteristic of burial in a solvent-inaccessible hydrophobic core (Vivian and Callis, 2001). This indicates a significant energetic penalty for forming the β-trefoil fold from multiple smaller chains, as has also been reported for other proteins (Lee and Blaber, 2011; Lee et al., 2011; Akanuma et al., 2010). This suggests the possibility that while the first symmetry-forming event for β-trefoils may have proceeded from a homotrimer (Ponting and Russell, 2000; Murzin et al., 1992; Mukhopadhyay, 2000; McLachlan, 1979), the recurring symmetry-forming events highlighted by our analysis may proceed from a subdomain module within an existing whole domain, thereby avoiding the energetically penalized homotrimeric form and also suggesting an explanation for why no isolated subdomain sequences have been reported.

Table 2. Data Collection and Refinement Statistics

Data Collection	
Space group	$P4_32_12$
Cell dimensions	
a, *b*, *c* (Å)	45.0, 45.0, 113.4
α, β, γ (°)	90.0, 90.0, 90.0
Resolution (Å)	1.62 (1.68–1.62)[a]
R_{merge}[b] (%)	0.068 (0.318)
Average *I*/σ*I*	13.7 (3.5)
Completeness (%)	99.5 (96.0)
Redundancy	6.25 (4.29)
Refinement	
Resolution (Å)	1.62
No. measured reflections	97142
No. unique reflections	15533
R_{cryst}/R_{free}[c]	16.7/18.5
No. atoms	
Protein	1151
Ligand/ion	61
Water	115
Average *B*-factors (Å^2)	
Protein	14.7
Ligands[d]	
BTB 1,2,3;	10.8, 22.6, 34.4;
Glycerol 1,2;	38.5, 38.5;
Na^+ ion	10.6
Water	28.5
Rmsds	
Bond lengths (Å)	0.006
Bond angles (°)	1.07
Coordinate error (ML-based, Å)(8)	0.22
Ramachandran plot (%)	
Most favored	86
Allowed	14

[a] Values in parentheses are for last resolution shell.

[b] $R_{merge} = \sum\sum | I(k) - \langle I \rangle | / \sum I(k)$, where I(k) is the measured intensity for each symmetry related reflection and <I> is the mean intensity for the unique reflection. The summation is over all unique reflections.

[c] $R_{cryst} = \sum |F_o| - |F_c| / \sum |F_o|$ and $R_{free} = \sum |F_{os}| - |F_{cs}|) / \sum |F_{os}|$, where "s" refers to a subset of data not used in the refinement, representing 7% of the total number of observations.

[d] Ligand atoms BTB 1,2,3 and glycerol 1,2 refer to three Bis-Tris methane molecules of Bis-Tris buffer (BTB) and two glycerol molecules of the cryoprotectant that were identified in the electron density and built into the structure.

DISCUSSION

Our protein sequence analyses and design results provide exciting experimental support for ancient as well as ongoing evolution of globular proteins via the repetition of subdomain modules. The recurring symmetry-forming events in globular folds are a surprising discovery, challenging the common view that symmetric globular folds were generated only in the ancient past and subsequently diverged, losing internal sequence symmetry (Söding and Lupas, 2003; Lupas et al., 2001; Ponting and Russell, 2000; Lee and Blaber, 2011). Several groups have undertaken to make fully symmetric versions of common globular folds from different structural classes: two-fold symmetric four-helix bundles, two- and four-fold symmetric $(\beta\alpha)_8$-barrels (Houbrechts et al., 1995; Richter et al., 2010; Höcker et al., 2009; Akanuma et al., 2010; Osterhout et al., 1992), and now, in this study and a parallel study on FGF by the Blaber group (Lee and Blaber, 2011; Lee et al., 2011), two very different three-fold symmetric β-trefoils. Among these designed proteins, demonstration of successful design by biophysical and structural analyses has been reported only for a $(\beta\alpha)_8$-barrel (Höcker et al., 2009) and for β-trefoils (Lee and Blaber, 2011; Lee et al., 2011).

The rational design and selection approach used by the Blaber group differs from the bioinformatics and rational design approach herein in that it uses multiple rounds of incorporating a few selected mutations to gradually increase symmetry, followed by screening for stability, ultimately resulting in a highly stable but nonfunctional protein. The primary sequences of the FGF design and ThreeFoil, which are based on proteins from different β-trefoil superfamilies (cytokines and ricin toxins, respectively), are well below the twilight zone and into the midnight zone of similarity (only ~15% identity) (Rost, 1999). Thus, the results reported here may illustrate how common folds can persist in evolution due to their compatibility with highly diverse sequences (Orengo and Thornton, 2005). Furthermore, common symmetric folds may be highly populated because they are generated repeatedly.

Our results for the β-trefoil fold have broad implications for understanding evolution not only of symmetric globular proteins but of other types of protein structures containing repeated subdomain modules, in particular, elongated repeat proteins and toroidal proteins (such as β-propellers), as well as oligomeric proteins. Consideration of the types of interfaces between modules in these proteins reveals key structural relationships (Figure 8). In proteins containing two repeated modules, or homodimers, a single, typically symmetric, interface is formed between the repeated structural elements. In all proteins with more than two repeats, at least two distinct interface surfaces are formed (Goodsell and Olson, 2000). For example, in β-trefoils each subdomain module packs front to back against the other two modules. Similarly, front-to-back packing of multiple ~20- to 40-amino-acid modules is the basis for the structure of elongated repeat proteins and toroidal proteins. It is important to note, though, that repeat and toroidal proteins are fundamentally different from globular proteins, because their hydrophobic core lacks interactions between residues that are distant in the primary sequence (Main et al., 2005; Yadid et al., 2010). The front-to-back packing and globular arrangement of modules in β-trefoils is also similar to that frequently observed in compact homooligomers with more than two subunits and cyclic symmetry (Goodsell and Olson, 2000). Thus, ThreeFoil shares structural characteristics with many other protein folds.

Common structural characteristics may underlie intriguing similarities in the stability and folding of β-trefoils, toroidal proteins, repeat proteins, and homooligomers. Consensus-sequence designed repeat proteins containing identical repeats

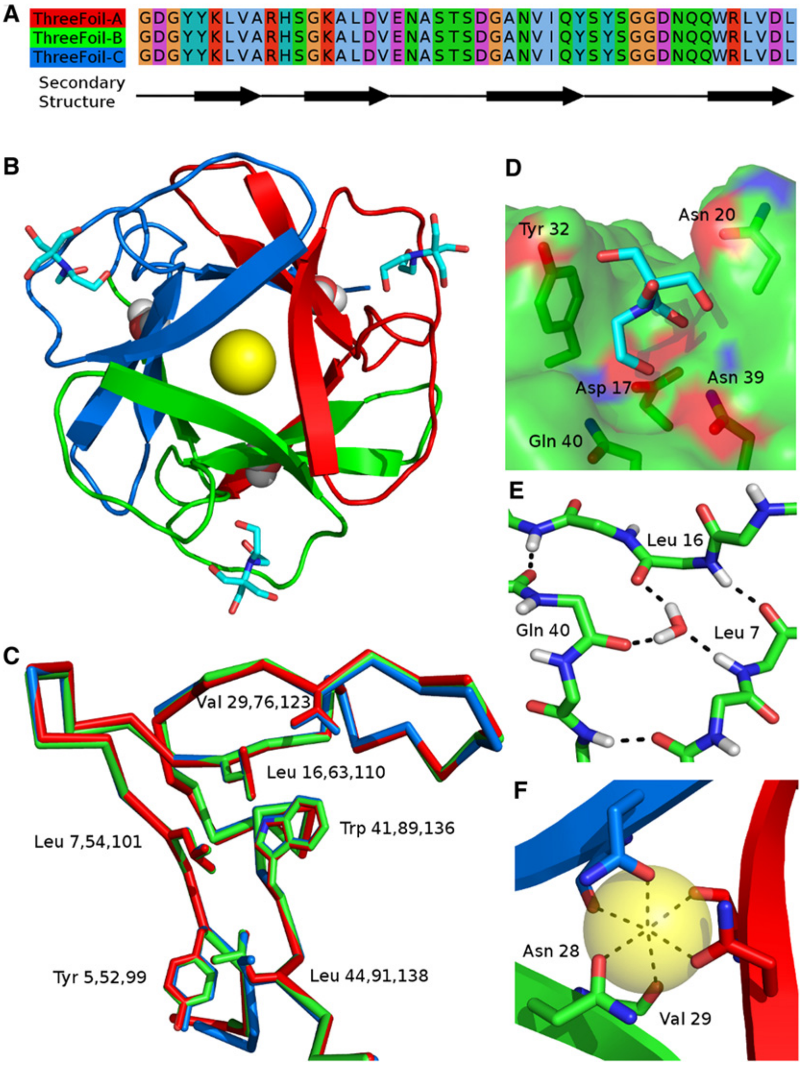
A
ThreeFoil-A GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL
ThreeFoil-B GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL
ThreeFoil-C GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL
Secondary Structure
B
C
Val 29,76,123
Leu 16,63,110
Trp 41,89,136
Leu 7,54,101
Tyr 5,52,99
Leu 44,91,138
D
Tyr 32
Asn 20
Asp 17
Asn 39
Gln 40
E
Leu 16
Gln 40
Leu 7
F
Asn 28
Val 29

Figure 5. Symmetric Structure in ThreeFoil

(A) The three subdomain modules of ThreeFoil aligned with the secondary structure shown below.

(B) A view of ThreeFoil along its three-fold symmetry axis, with subdomains indicated using the same colors as in Figures 1–3, bound bis-tris carbon atoms in cyan, bound waters as red (oxygen) and white (hydrogen) spheres, and the bound sodium as a yellow sphere.

(C) Each subdomain module of ThreeFoil structurally aligned by C-alpha using SSM (Krissinel and Henrick, 2004) shown as a C-alpha trace, with core hydrophobic sidechains shown as sticks.

(D) Bis-tris bound to ThreeFoil in the shallow pocket that forms the carbohydrate binding site in related ricins.

(E) The buried water molecule in each subdomain forms hydrogen bonds with three different β-strands.

(F) Sodium binding site, showing the symmetric backbone and side-chain oxygen atoms involved in the octahedral coordination.

have been found to have very high stability (Main et al., 2005; Wetzel et al., 2008). For repeat proteins, stability increases with increasing number of identical repeats (Main et al., 2005; Barrick et al., 2008). Similarly, additional interfaces may also result in oligomers being generally more stable than monomers (André et al., 2008). Additional entropic stabilization of symmetric globular proteins may be obtained by combining subdomain modules into a single chain. Such stabilization is suggested by the well structured ThreeFoil compared with the unstructured OneFoil, and by similar results for a symmetric FGF β-trefoil (Lee and Blaber, 2011; Lee et al., 2011) and a four-helix bundle protein (Akanuma et al., 2010). Thus, combining identical modules into a single chain could favor folding in multiple ways. In general, the roles of structural symmetry in protein folding are not yet well understood, and proteins with completely symmetric tertiary structure and sequence, like ThreeFoil, are intriguing models for examining these roles further.

Internally symmetric structures may provide significant benefits for protein function. For instance, β-trefoils and toroidal, repeat, and oligomeric proteins appear to be particularly well suited for a wide range of binding functions, and they often use multiple repeats for multivalent binding of ligands, as seen with ThreeFoil and reported in many other cases (Murzin et al., 1992; Hazes, 1996; Main et al., 2005; Beisel et al., 1999). Thus, these protein structures may provide stable scaffolds for displaying a wide variety of loop structures for binding diverse ligands. It is now widely accepted that functional features such as binding sites can be a significant source of instability in proteins (Tokuriki and Tawfik, 2009; Meiering et al., 1992), and symmetrical structures may be more stable (André et al., 2008), as well as more robust to mutations and therefore more designable (André et al., 2008; Li et al., 1996). Together, this suggests that the repetition of structural modules in proteins may confer sufficient stability to accommodate destabilizing functional features. In addition, selection for multivalent binding functions may give rise to symmetry in globular, repeat, toroidal, and oligomeric proteins (Goodsell and Olson, 2000). In this respect, it is noteworthy that the ricins, AbfB, and fascin β-trefoil families most prominently exhibit subdomain repetition, and these families are involved in the multivalent binding of ligands: carbohydrates in the case of the ricins (Hazes, 1996) and AbfBs (Miyanaga et al., 2004) and actin in the case of the fascins (Sedeh et al., 2010). Ricin β-trefoils are involved in host-pathogen interactions, which often require multivalent binding achieved through symmetry (Collins and Paulson, 2004). Also, they exist as domains within rapidly evolving toxins (Doxey et al., 2008), which may provide an increased opportunity to observe repetition events.

The repetition of subdomain modules to form internally symmetric structures may provide an inherent benefit for functional plasticity, as compared with oligomerization in a repeat protein. In an oligomeric protein, any mutations are necessarily present in all subunits of the oligomer, and this may limit the opportunity to acquire new or improved functions that require

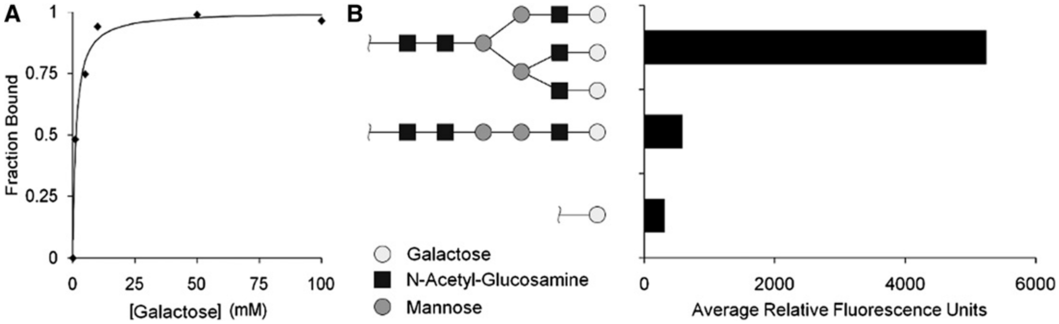

Figure 6. Multivalent Glycan Binding by ThreeFoil

(A) Galactose binding curve for ThreeFoil measured by fluorescence showing a dissociation constant (K_d) of 1 mM.

(B) Binding results from a glycan array demonstrating that ThreeFoil has considerably improved binding to multiantennary glycan structures (top structure as compared with bottom two).

a combination of mutations, if any of the individual mutations along the way are functionally deleterious. By contrast, repetition of a subdomain module into a single larger protein with multiple initially identical functional sites means that at least one of these functional sites may be free to accumulate mutations that are initially deleterious to function but may eventually lead to novel or improved function (Yadid et al., 2010), and this modified subdomain may then itself repeat to form a newly symmetric protein with amplified function.

Symmetric protein structures (due to internal repetition or oligomerization) are the rule rather than the exception in nature. The reasons for this have been the subject of much speculation (Goodsell and Olson, 2000; André et al., 2008). Both physical factors, as described above, and evolutionary mechanisms at the level of DNA replication may play important roles (Goodsell and Olson, 2000). Using sequence analysis and rational design tools we have successfully reconstructed a fully internally symmetric protein sequence with the intended monomeric structure, which has the attractive features of being well folded, highly soluble, and functional, and exhibiting high thermal stability. The structure itself and the design strategy of combining bioinformatics and protein modeling should be useful for future studies of the origins and determinants of symmetric protein folds, as well as for designing proteins with desirable properties such as multivalent binding and high stability. In particular, the application of modular evolution or modular protein design may prove to be an elegant solution to incorporating functional properties into a scaffold while retaining sufficient stability, and recent work in addition to this study highlights the promise of such approaches (Richter et al., 2010; Lee and Blaber, 2011; Akanuma et al., 2010; Fernandez-Fuentes et al., 2010; Yadid and Tawfik, 2011). Although there is still relatively little experimental data on the sequence repetition of subdomain modules in globular proteins, it seems likely that closer examination of the vast and ever-expanding protein sequence and structure databases will continue to provide further evidence for these processes in other symmetric folds.

EXPERIMENTAL PROCEDURES

Sequence Dataset Construction and Analysis

All annotated β-trefoil domain sequences were retrieved from the National Center for Biotechnology Information (NCBI) using the Conserved Domain Database (CDD). All families annotated as β-trefoils by SCOP (Murzin et al., 1995) and Pfam (Finn et al., 2010) with an available structure in the Protein Data Bank (http://www.pdb.org) (Berman et al., 2000) were included. See Table 1 for statistics on construction of the dataset. Sequences were parsed and their β-trefoil regions extracted according to the CDD information included in the NCBI's GenPept file. All β-trefoil domains in each protein chain were extracted, which resulted in an initial dataset of 5287 domain sequences. To remove redundancy, all domain sequences were grouped into clusters of highly similar sequences using the BLASTCLUST algorithm from the BLAST package (Altschul et al., 1997) with default parameters (length coverage threshold = 0.9; score coverage threshold = 1.75). The longest sequence from each cluster was selected as a representative, and the remaining sequences were removed from the dataset. β-trefoil sequences were then parsed into their individual subdomain modules by aligning all sequences to their corresponding β-trefoil family HMM using the program HMMalign (http://hmmer.janelia.org), and dividing the sequences into three parts according to the repeat pattern evident within a representative structure. The representative structures used in subdomain module parsing are listed in Table 1. Sequences that were truncated and/or contained insufficient data were excluded by only including sequences containing three subdomain modules with lengths longer than 20 residues. The final dataset consisted of 3501 subdomain modules from 1167 β-trefoil domains. Subdomain modules were then clustered using a graph-based approach. First, an all-by-all BLAST search was performed, and any two repeats with E < 1e-04 were connected. The results were visualized with the program Cytoscape (http://www.cytoscape.org). The choice of clustering parameter will change the evolutionary resolution of the analysis. A lower BLAST E-value will result in a larger number of clusters and require the detected subdomain-repetition events to be higher in similarity and thus more recent. Conversely, a higher E-value threshold will result in fewer clusters but identify potentially more ancestral subdomain-repetition events. The cutoff of 1e-04 was chosen as a reasonable middle-ground. We also constructed a phylogenetic tree via neighbor joining (see Supplemental Experimental Procedure), from which subdomain-repetition events can be inferred (Figure S1D).

Design of ThreeFoil

An overview of the design methodology is given in the Results section, with additional details below. The template sequence was divided into its three subdomain modules, each of length 47 amino acids, using the method employed for sequence dataset construction and analysis. The homologous sequences used in reconstruction were the 13 most closely related nonredundant sequences identified using BLAST (Altschul et al., 1997). These were split into their three constituent subdomain modules after multiple sequence alignment with the template using MUSCLE (Edgar, 2004), giving 39 homologous subdomain modules (Figure S2). For Rosetta Design, an initial structure was needed. Three structures were generated through homology modeling using MODELER (Sali and Blundell, 1993) (http://www.salilab.org/modeller/) and the three most closely related structures, PDB IDs 1KNM (Notenboom et al., 2002), 2IHO (Grahn et al., 2007), and 1YBI (Arndt et al., 2005). The

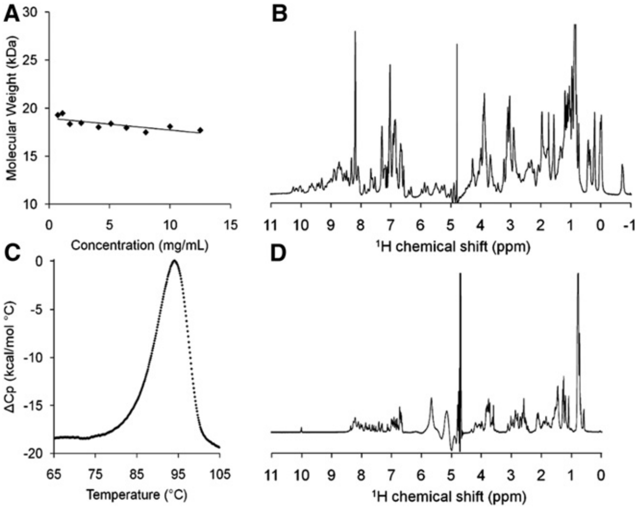

Figure 7. Biophysical Characterization of ThreeFoil

(A) Molecular weight Debye plot (Zimm, 1948) of SLS measurements, consistent with expected size of a ThreeFoil monomer (see also Figure S3).

(B) ^{1}H-NMR spectrum of ThreeFoil in H_2O (containing 7% D_2O). The relatively sharp and well dispersed lines are indicative of a well folded monomeric structure.

(C) DSC of ThreeFoil showing a large endothermic peak is typical of a cooperative thermal unfolding transition with a midpoint of ~94°C.

(D) ^{1}H-NMR spectrum of OneFoil in H_2O (containing 7% D_2O) with features typical of an unfolded protein (lack of amide resonances >8.5 ppm, and methyl resonances <1 ppm).

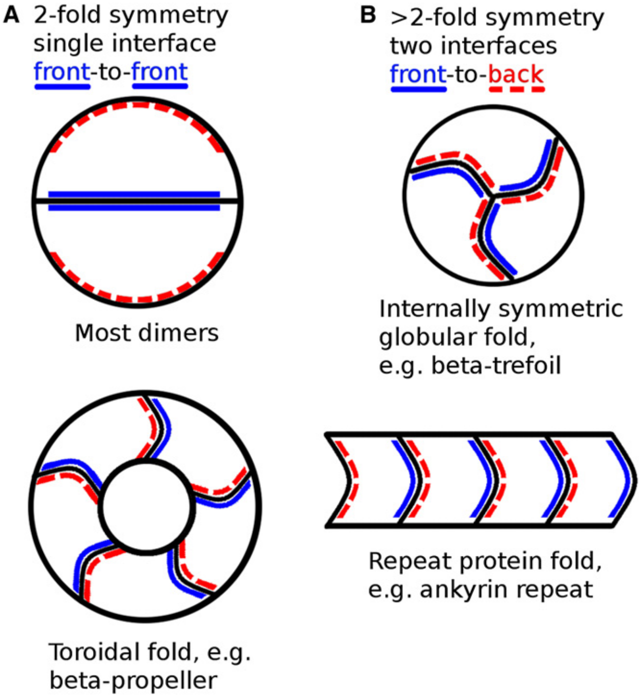

Figure 8. Packing of Symmetric, Repeat, and Oligomeric Proteins

Proteins with two-fold internal symmetry and dimers both tend to interact through a single interface, packing front to front (A). Proteins with more than two-fold symmetry or oligomers require two different interfaces and pack front to back (B).

sequence used to generate the homology models for step 3 of the reconstruction was from step 2 (Figure 4, Step 2), with the 16 gaps filled in by the most frequent residue based on the step 2 frequency calculation (see Results).

Expression and Purification of ThreeFoil and OneFoil

The nucleotide sequence of ThreeFoil was synthesized and supplied in a pUC57 vector (GenScript). The sequence was subcloned into a modified pET-28a vector containing an N-terminal deca-histidine tag (modified from the original hexa-histidine tag). OneFoil was generated by annealing six overlapping oligonucleotides (Sigma Aldrich) followed by direct ligation of the oligos into linearized modified pET-28a. Both ThreeFoil and OneFoil were expressed in *E. coli* after induction with IPTG (1 mM). Cells were harvested after 48 and 24 hr of growth at 37°C and 25°C for ThreeFoil and OneFoil, respectively. Both proteins were isolated as inclusion bodies and solubilized in buffered urea (6 M urea, 100 mM phosphate, and 10 mM tris, pH 8.1), bound to a Ni-NTA column, and eluted at pH 4.5. The purified protein was then refolded by dialysis (SpectraPor10) against 300 mM NaCl and 100 mM phosphate, pH 6.6 (the standard buffer used for all subsequent experiments) at a concentration of 0.15 mg/ml, and then concentrated to 12.5 and 1.0 mg/ml for ThreeFoil and OneFoil, respectively, using ultrafiltration (YM10 membranes, Amicon). Due to solubility limits, OneFoil could not be concentrated to the same levels as ThreeFoil. Molar extinction coefficients of 33,600 and 11,200 Lmol^{-1} cm^{-1} for ThreeFoil and OneFoil, respectively, were determined using the method of Pace and co-workers (Pace et al., 1995) and used for determination of protein concentrations.

Structure Determination and Refinement

ThreeFoil was screened for crystallization conditions using the Index HT screen (Hampton Research). Crystals of ThreeFoil appeared after one month from sitting drops (2.4 M ammonium sulfate and 100 mM bis-tris, pH 6.5) at a protein concentration of 7 mg/ml and were soaked in the aforementioned solution with the addition of 25% (v/v) glycerol as a cryoprotectant before being flash-frozen in a −170°C N_2 stream. Data were collected in-house at The Hospital for Sick Children in Toronto and processed using d*TREK (Pflugrath, 1999). The structure was determined using molecular replacement techniques and refined to 1.62 Å resolution (deposited as PDB code 3PG0). For details of molecular replacement, see Supplemental Experimental Procedure.

Biophysical Characterization of ThreeFoil and OneFoil

All samples of ThreeFoil and OneFoil were prepared in 300 mM NaCl and 100 mM phosphate, pH 6.6, and analyzed at ambient temperature unless otherwise noted. DLS measurements were made using a protein concentration of 12.5 mg/ml, with a 0.4-cm-pathlength cuvette and a NanoSZ particle sizer (Malvern). SLS measurements were obtained at the same time as DLS for protein concentrations ranging from 0.7 to 12.5 mg/ml. SEC was performed using a Superose 12 HR 10/30 column (GE Healthcare), with a 0.5 ml/min flow-rate, with buffer supplemented with D-galactose (1.5 M). Fluorescence measurements were performed using a Flourolog322 (Spex) with excitation and emission wavelengths of 280 nm and 313 nm, respectively, and excitation and emission slit widths of 1 nm and 5 nm, respectively. DSC was performed at a ThreeFoil concentration of 0.6 mg/ml and a scan rate of 1°C/min. One-dimensional ^{1}H-NMR spectra were acquired at 25°C using a Bruker AVANCE 600 MHz spectrometer with a TSI probe and excitation sculpting for water suppression (Hwang and Olson, 1995), with a ThreeFoil concentration of 12.5 mg/ml and a OneFoil concentration of 1.0 mg/ml. Glycan array analysis was performed as reported by the Consortium for Functional Glycomics (Blixt et al., 2004) (http://www.functionalglycomics.org/fg/), using a ThreeFoil concentration of 0.2 mg/ml.

ACCESSION NUMBERS

The PDB accession code for ThreeFoil reported in this paper is 3PG0.

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures, one table, and Supplemental Experimental Procedures and is available with this article online at doi:10.1016/j.str.2011.10.021.

ACKNOWLEDGMENTS

This work was supported by research grants from the National Science and Engineering Council (NSERC; to E.M.M. and B.J.M.) and the Canadian Institutes of Health Research (CIHR; to P.L.H [#43998]). A.C.D. is the recipient of an NSERC Canada Graduate Scholarship. P.L.H. is the recipient of a Canada Research Chair. The authors would like to acknowledge the Protein-Glycan Interaction Core (H) of the Consortium for Functional Glycomics funded by the National Institute of General Medical Sciences (GM62116) for the glycan array analysis.

Received: August 22, 2011
Revised: October 9, 2011
Accepted: October 24, 2011
Published online: December 15, 2011

REFERENCES

Akanuma, S., Matsuba, T., Ueno, E., Umeda, N., and Yamagishi, A. (2010). Mimicking the evolution of a thermally stable monomeric four-helix bundle by fusion of four identical single-helix peptides. J. Biochem. *147*, 371–379.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

André, I., Strauss, C.E.M., Kaplan, D.B., Bradley, P., and Baker, D. (2008). Emergence of symmetry in homooligomeric biological assemblies. Proc. Natl. Acad. Sci. USA *105*, 16148–16152.

Arndt, J.W., Gu, J., Jaroszewski, L., Schwarzenbacher, R., Hanson, M.A., Lebeda, F.J., and Stevens, R.C. (2005). The structure of the neurotoxin-associated protein HA33/A from Clostridium botulinum suggests a reoccurring β-trefoil fold in the progenitor toxin complex. J. Mol. Biol. *346*, 1083–1093.

Barrick, D., Ferreiro, D.U., and Komives, E.A. (2008). Folding landscapes of ankyrin repeat proteins: experiments meet theory. Curr. Opin. Struct. Biol. *18*, 27–34.

Beisel, H.G., Kawabata, S., Iwanaga, S., Huber, R., and Bode, W. (1999). Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab Tachypleus tridentatus. EMBO J. *18*, 2313–2322.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Blixt, O., Head, S., Mondala, T., Scanlan, C., Huflejt, M.E., Alvarez, R., Bryan, M.C., Fazio, F., Calarese, D., Stevens, J., et al. (2004). Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. Proc. Natl. Acad. Sci. USA *101*, 17033–17038.

Capraro, D.T., Roy, M., Onuchic, J.N., and Jennings, P.A. (2008). Backtracking on the folding landscape of the β-trefoil protein interleukin-1β? Proc. Natl. Acad. Sci. USA *105*, 14844–14848.

Chaudhuri, I., Söding, J., and Lupas, A.N. (2008). Evolution of the β-propeller fold. Proteins *71*, 795–803.

Collins, B.E., and Paulson, J.C. (2004). Cell surface biology mediated by low affinity multivalent protein-glycan interactions. Curr. Opin. Chem. Biol. *8*, 617–625.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J. Mol. Biol. *332*, 449–460.

Doxey, A.C., Lynch, M.D.J., Müller, K.M., Meiering, E.M., and McConkey, B.J. (2008). Insights into the evolutionary origins of clostridial neurotoxins from analysis of the Clostridium botulinum strain A neurotoxin gene cluster. BMC Evol. Biol. *8*, 316.

Dubey, V.K., Lee, J., Somasundaram, T., Blaber, S., and Blaber, M. (2007). Spackling the crack: stabilizing human fibroblast growth factor-1 by targeting the N and C terminus β-strand interactions. J. Mol. Biol. *371*, 256–268.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Eisenbeis, S., and Höcker, B. (2010). Evolutionary mechanism as a template for protein engineering. J. Pept. Sci. *16*, 538–544.

Fernandez-Fuentes, N., Dybas, J.M., and Fiser, A. (2010). Structural characteristics of novel protein folds. PLoS Comput. Biol. *6*, e1000750.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The Pfam protein families database. Nucleic Acids Res. *38* (Database issue), D211–D222.

Goodsell, D.S., and Olson, A.J. (2000). Structural symmetry and protein function. Annu. Rev. Biophys. Biomol. Struct. *29*, 105–153.

Grahn, E., Askarieh, G., Holmner, Å., Tateno, H., Winter, H.C., Goldstein, I.J., and Krengel, U. (2007). Crystal structure of the Marasmius oreades mushroom lectin in complex with a xenotransplantation epitope. J. Mol. Biol. *369*, 710–721.

Hazes, B. (1996). The (QxW)3 domain: a flexible lectin scaffold. Protein Sci. *5*, 1490–1501.

Hill, R.B., Raleigh, D.P., Lombardi, A., and DeGrado, W.F. (2000). De novo design of helical bundles as models for understanding protein folding and function. Acc. Chem. Res. *33*, 745–754.

Höcker, B., Claren, J., and Sterner, R. (2004). Mimicking enzyme evolution by generating new $\beta\alpha_8$-barrels from $\beta\alpha_4$-half-barrels. Proc. Natl. Acad. Sci. USA *101*, 16448–16453.

Höcker, B., Lochner, A., Seitz, T., Claren, J., and Sterner, R. (2009). High-resolution crystal structure of an artificial $\beta\alpha_8$-barrel protein designed from identical half-barrels. Biochemistry *48*, 1145–1147.

Houbrechts, A., Moreau, B., Abagyan, R., Mainfroid, V., Préaux, G., Lamproye, A., Poncin, A., Goormaghtigh, E., Ruysschaert, J.-M., Martial, J.A., et al. (1995). Second-generation octarellins: two new de novo (β/α)8 polypeptides designed for investigating the influence of β-residue packing on the α/β-barrel structure stability. Protein Eng. *8*, 249–259.

Houliston, R.S., Liu, C., Singh, L.M.R., and Meiering, E.M. (2002). pH and urea dependence of amide hydrogen-deuterium exchange rates in the β-trefoil protein hisactophilin. Biochemistry *41*, 1182–1194.

Hu, X., Wang, H., Ke, H., and Kuhlman, B. (2008). Computer-based redesign of a β sandwich protein suggests that extensive negative design is not required for de novo β sheet design. Structure *16*, 1799–1805.

Hwang, T., and Olson, A. (1995). Water suppression that works: excitation sculpting using arbitrary wave-forms and pulsed-field gradients. J. Magn. Reson. A *112*, 275–279.

Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. *60*, 2256–2268.

Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., and Wilmanns, M. (2000). Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science *289*, 1546–1550.

Lee, J., and Blaber, M. (2011). Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. Proc. Natl. Acad. Sci. USA *108*, 126–130.

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.

Li, H., Helling, R., Tang, C., and Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. Science *273*, 666–669.

Lupas, A.N., Ponting, C.P., and Russell, R.B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J. Struct. Biol. *134*, 191–203.

Main, E.R., Lowe, A.R., Mochrie, S.G., Jackson, S.E., and Regan, L. (2005). A recurring theme in protein engineering: the design, stability and folding of repeat proteins. Curr. Opin. Struct. Biol. *15*, 464–471.

Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., et al. (2009). CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. *37* (Database issue), D205–D210.

McLachlan, A.D. (1979). Three-fold structural pattern in the soybean trypsin inhibitor (Kunitz). J. Mol. Biol. *133*, 557–563.

Meiering, E.M., Serrano, L., and Fersht, A.R. (1992). Effect of active site residues in barnase on activity and stability. J. Mol. Biol. *225*, 585–589.

Miyanaga, A., Koseki, T., Matsuzawa, H., Wakagi, T., Shoun, H., and Fushinobu, S. (2004). Crystal structure of a family 54 α-L-arabinofuranosidase reveals a novel carbohydrate-binding module that can bind arabinose. J. Biol. Chem. *279*, 44907–44914.

Mukhopadhyay, D. (2000). The molecular evolutionary history of a winged bean alpha-chymotrypsin inhibitor and modeling of its mutations through structural analyses. J. Mol. Evol. *50*, 214–223.

Murzin, A.G., Lesk, A.M., and Chothia, C. (1992). β-Trefoil fold. Patterns of structure and sequence in the Kunitz inhibitors interleukins-1 β and 1 α and fibroblast growth factors. J. Mol. Biol. *223*, 531–543.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

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*, 4246–4254.

Orengo, C.A., and Thornton, J.M. (2005). Protein families and their evolution-a structural perspective. Annu. Rev. Biochem. *74*, 867–900.

Orengo, C.A., Jones, D.T., and Thornton, J.M. (1994). Protein superfamilies and domain superfolds. Nature *372*, 631–634.

Osterhout, J., Handel, T., Na, G., Toumadje, A., Long, R., Connolly, P., Hoch, J., Johnson, W., Live, D., and Degrado, W. (1992). Characterization of the structural-properties of alpha-1B, a peptide designed to form a 4-helix bundle. J. Am. Chem. Soc. *114*, 331–337.

Pace, C.N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995). How to measure and predict the molar absorption coefficient of a protein. Protein Sci. *4*, 2411–2423.

Pflugrath, J.W. (1999). The finer things in X-ray diffraction data collection. Acta Crystallogr. D Biol. Crystallogr. *55*, 1718–1725.

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. J. Mol. Biol. *302*, 1041–1047.

Richter, M., Bosnali, M., Carstensen, L., Seitz, T., Durchschlag, H., Blanquart, S., Merkl, R., and Sterner, R. (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.

Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng. *12*, 85–94.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Sedeh, R.S., Fedorov, A.A., Fedorov, E.V., Ono, S., Matsumura, F., Almo, S.C., and Bathe, M. (2010). Structure, evolutionary conservation, and conformational dynamics of Homo sapiens fascin-1, an F-actin crosslinking protein. J. Mol. Biol. *400*, 589–604.

Söding, J., and Lupas, A.N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. Bioessays *25*, 837–846.

Stenmark, P., Gurmu, D., and Nordlund, P. (2004). Crystal structure of CaiB, a type-III CoA transferase in carnitine metabolism. Biochemistry *43*, 13996–14003.

Tokuriki, N., and Tawfik, D.S. (2009). Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. *19*, 596–604.

Vivian, J.T., and Callis, P.R. (2001). Mechanisms of tryptophan fluorescence shifts in proteins. Biophys. J. *80*, 2093–2109.

Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K., and Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. J. Mol. Biol. *376*, 241–257.

Winter, H.C., Mostafapour, K., and Goldstein, I.J. (2002). The mushroom Marasmius oreades lectin is a blood group type B agglutinin that recognizes the Galα 1,3Gal and Galα 1,3Galβ 1,4GlcNAc porcine xenotransplantation epitopes with high affinity. J. Biol. Chem. *277*, 14996–15001.

Yadid, I., and Tawfik, D.S. (2011). Functional β-propeller lectins by tandem duplications of repetitive units. Protein Eng. Des. Sel. *24*, 185–195.

Yadid, I., Kirshenbaum, N., Sharon, M., Dym, O., and Tawfik, D.S. (2010). Metamorphic proteins mediate evolutionary transitions of structure. Proc. Natl. Acad. Sci. USA *107*, 7287–7292.

Zimm, B.H. (1948). The scattering of light and the radial distribution function of high polymer solutions. J. Chem. Phys. *16*, 1093–1099.