Computational design of a self-assembling symmetrical β-propeller protein

Arnout R. D. Voet,a,b Hiroki Noguchib, Christine Addyb, David Simoncini,a,c Daiki Teradaa, Satoru Unzai, Sam-Yong Parkb, Kam Y. J. Zhanga,d and Jeremy R. H. Tameb,2

*Structural Bioinformatics Team, Division of Structural and Synthetic Biology, Center for Life Science Technologies, RIKEN, 1-7-22 Suehiro, Yokohama, Kanagawa 230-0045, Japan; and *Drug Design Laboratory, Graduate School of Medical Life Science, Yokohama City University, 1-7-29 Suehiro, Yokohama, Kanagawa 230-0045, Japan

Edited* by David S. Eisenberg, University of California, Los Angeles, CA, and approved September 15, 2014 (received for review July 7, 2014)

The modular structure of many protein families, such as β-propeller proteins, strongly implies that duplication played an important role in their evolution, leading to highly symmetrical intermediate forms. Previous attempts to create perfectly symmetrical propeller proteins have failed, however. We have therefore developed a new and rapid computational approach to design such proteins. As a test case, we have created a sixfold symmetrical β-propeller protein and experimentally validated the structure using X-ray crystallography. Each blade consists of 42 residues. Proteins carrying 2–10 identical blades were also expressed and purified. Two or three tandem blades assemble to recreate the highly stable sixfold symmetrical architecture, consistent with the duplication and fusion theory. The other proteins produce different monodisperse complexes, up to 42 blades (180 kDa) in size, which self-assemble according to simple symmetry rules. Our procedure is suitable for creating nano-building blocks from different protein templates of desired symmetry.

protein evolution | computational protein design | self-assembly | β-propeller | protein crystallography

It is generally accepted that evolution is driven by duplications of genetic material. These events allow gene copies to develop independent regulation (1) and to express new proteins that inherit the stable architecture of the parent protein but possess a novel function (2, 3). Although this process largely explains the diversity of proteins with similar folds, it cannot account for the appearance of new protein folds. However, many proteins have a modular internal structure that most probably arose from duplication and fusion of structural elements. This type of process is most clearly demonstrated by proteins consisting of conserved domains repeated in tandem, giving a highly symmetrical tertiary structure (4, 5). Although symmetry remains a common feature of proteins (6), many present-day proteins show more limited symmetry than that of the ancestral intermediate forms suggested by the duplication theory of evolution (7–9). Since the group of Wilman demonstrated that a (β/α)8-barrel protein could be constructed out of two identical halves in 2000 (10), several other groups have also reported the artificial construction of symmetrical or modular proteins, providing evidence for duplication and fusion events in nature (11–15). In the case of β-trefoil proteins, a design procedure based on the Rosetta method proved much more efficient than directed evolution methods at producing a symmetrical structure (15). Structural plasticity and domain swapping (16, 17) allow such extended proteins to adopt novel tertiary and quaternary structures (18), but to date there is no report of a perfectly symmetrical β-propeller protein.

β-propeller proteins are composed of different numbers of repeats, each made from a single β-sheet, roughly 40 residues in length, that resembles the blade of a propeller (19, 20). β-propeller proteins are good examples of how proteins may have evolved from duplication and fusion events of simple peptide motifs (21). Examples are known of 4-, 5-, 6-, 7-, 8-, and 10-bladed proteins. These proteins have diverse functions, including varied enzymatic activities and protein–protein interactions, making them a highly interesting class to redesign both for synthetic biochemistry and as nano-building blocks. Previous attempts to create stable perfectly symmetrical β-propeller proteins have failed. Yadid and Tawfik (3, 22) screened genetic libraries encoding about 100 amino acid residues from a 236-residue five-bladed propeller (tachylectin-2) in attempts to create a fivefold symmetrical propeller. The initial proteins produced were poorly stable, but subsequent directed evolution to improve expression and folding led to domain-swapped structures through strand exchange (18). An artificial WD40-based repeat protein was designed by Nikkhah et al. using computational methods, but this protein failed to fold and adopted a molten globule state (23). Similarly, Figueroa et al. have recently described a putative artificial TIM barrel structure called “octarellin VI,” but this protein proved to be poorly soluble, and NMR indicated that it is not stably folded (24). It is widely believed that proteins with a perfectly repeated sequence motif experience “folding frustration,” the absence of a single strongly preferred tertiary structure, leading to unstable folds (11, 25). In fact, a search for identical sequence repeats within the same polypeptide chain failed to find any duplicated domains containing regular secondary structure in known natural proteins (26).

In this study, we have designed and experimentally validated, to our knowledge, the first perfectly symmetrical β-propeller protein. Our results provide insight not only into protein evolution through duplication events, but also into methods for creating designer proteins that self-assemble according to simple mathematical rules. Such proteins may have very wide uses in bionano-technology. Furthermore our design approach is both rapid and applicable to many different protein templates. Our novel propeller protein consists of six identical domains known as “blades.” Using a variety of biophysical techniques, we show it to be highly stable and report several high-resolution crystal structures of different forms of the protein. Domain swapping allows us to generate related oligomeric forms with fixed numbers of blades per complex.

Author contributions: A.R.D.V., K.Y.J.Z., and J.R.H.T. designed research; A.R.D.V., H.N., C.A., D.T., S.U., and J.R.H.T. performed research; A.R.D.V. and D.S. contributed new reagents/analytic tools; A.R.D.V., H.N., S.U., S.Y.P., and J.R.H.T. analyzed data; and A.R.D.V., C.A., D.T., S.U., and J.R.H.T. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data Bank (PDB), www.pdb.org [PDB ID codes 3WW7 (Pfizz2), 3WW8 (Pfizz3), 3WW9 (Pfizz6), 3WWA (Pfizz7), 3WWB (Pfizz8a), 3WWF (Pfizz8b) Form A], and 3WWF (Pfizz8b) Form B).

1Present address: Mathématiques et Informatique Appliquées de Toulouse, Unité de Recherche 875, Institut National de la Recherche Agronomique, F-31320 Castanet-Tolosan, France.

2To whom correspondence should be addressed. Email: kamzhang@riken.jp or jame@tsurumi.yokohama-cu.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1412768111/-/DCSupplemental.
We have applied a novel computational approach to the problem of creating a stable, perfectly symmetrical propeller by reverse engineering the supposed evolutionary pathway. Specifically, we wanted to address the question whether we can construct from a nonsymmetrical protein a symmetrical one that could have originated from smaller protein fragments. Ancestral sequence reconstruction was used to derive likely parent sequences assuming evolution through duplication, and then these sequences were computationally evaluated for protein stability (Fig. S1). We chose a six-bladed protein, given its additional two- and threefold pseudosymmetry. To agree with the duplication and fusion theory, such a protein should be divisible into a self-assembling unit consisting of 2 or 3 domains. Additionally, we created polypeptides carrying up to 10 identical blades and showed that these molecules also fold to give stable structures.

Results

One hundred seventy-four models of six-bladed β-propellers were identified in the PDB and examined by eye for suitability as templates for protein design. The NHL repeat structure PDB entry 1RWL (27), the sensor domain of a protein kinase from Mycobacterium tuberculosis, was found to be the most visually appealing due to the apparent symmetry of the Cα trace and was selected on this basis alone. In common with almost all β-propeller proteins, this structure shows a so-called “Velcro” strap, the last β-strand completing the first domain (Fig. S2) (19). The sequence of the blades were considered as separate genes, aligned and used for ancestral sequence reconstruction of the parent blade (Fig. 1A and B) (28). Three out of the six blades have an identical number of amino acids per blade, and insertions or deletions with respect to these blades were not allowed. Sequence comparison indicated that likely ancestral sequences are most closely related to the third blade of the template protein, residues 107–148, and so this blade was used to construct a symmetrical template with the help of RosettaDock (29). The glycine residues G106 and G148, which sandwich blade 3, were not included, in the expectation that these missing residues would be compensated by N- to C-terminal salt bridges between neighboring subunits. The best scoring sixfold symmetrical structure after the docking step did indeed show these bonds (Fig. 1C and D). The six chains were fused into a single polypeptide by reintroducing the glycine residues, and these putative ancestral sequences were mapped onto the protein structure using a Rosetta-based algorithm (Fig. 1E). Our procedure was similar to that used by Broom et al. (15) to create a symmetrical β-trefoil protein. However, contrary to their approach using many known protein sequences from a given family, our method employed ancestor reconstruction, reverse engineering the evolutionary process for a single protein template. No β-propeller sequences other than the chosen template were used at any step.

The output models showed significantly lower energy than the sequences corresponding to the blades of the template (1RWL), but the best-scoring solution had a ring of adjacent symmetry-related arginine residues. Therefore, the second-ranked solution was selected, with histidine at this position in the sequence and an additional serine-to-asparagine mutation. The sixfold repeating amino acid code was back-translated into a degenerate nucleotide sequence that was synthesized and cloned for protein expression. From its shape, the six-bladed designer protein was named Pizza6.

Pizza6 protein expressed to very high levels (roughly 100 mg/mL) in Escherichia coli BL21(DE3) cells using pET vectors and was purified by a very simple procedure. The protein was shown to be monodisperse by size-exclusion chromatography (SEC), electrospray ionization mass spectrometry (ESI-MS), and analytical ultracentrifugation (AUC) (Fig. 2 and Table 1). Crystals were obtained under a wide variety of conditions, mostly in fewer than 24 h. After optimization, X-ray data were collected to 1.33 Å resolution, and the structure was rapidly solved by molecular replacement using the predicted structure as the search model (Fig. 3) (see Table S3). The backbone-rmsd of 0.68 Å between the final and expected structures validate our design strategy for a fully symmetrical protein made from a minimal, nonnatural domain.

Multimeric versions of Pizza6 were created by truncating the protein after two or three repeats. Both Pizza2 and Pizza3 express as monodisperse proteins with the same molecular weight in solution as Pizza6. Their crystal structures are essentially identical to that of Pizza6, demonstrating that propeller proteins could have evolved by gene duplication and fusion. A Pizza2 mutant, corresponding to the top-scoring design, was also created with two identical blades, each carrying two internal mutations, N16S and H31R. This Pizza2-SR protein also proved stable, despite carrying the ring of neighboring arginine residues.

Fig. 1. Computational design of a fully symmetric β-propeller. From the nonsymmetrical six-bladed 1RWL template protein (A), the sequences of each blade were aligned (B) and used for ancestral sequence reconstruction. For comparison, the final Pizza sequence is also shown on the bottom line of B. Blade 3 was identified as closest to the most probable ancestral sequence and was used for the generation of a sixfold symmetrical template protein using RosettaDock with C6 symmetry. From the scatter plot (C) of the docking scores versus the rmsds between the different solutions and the best scoring solution (D), it is clear that the higher the deviation from the six-bladed propeller fold the worse the docking score becomes. The ancestral sequences and three WT sequences were mapped onto the fully symmetrical template and scored using Rosetta (E). The green bars indicate the 1RWL sequence scores (blades 3, 4, and 5). The red bar indicates the top-scoring sequence (Pizza2-SR). The orange bar corresponds to the selected Pizza sequence, which is also depicted as a Cα trace in F, colored blue to red from the N to C terminus. The differences between the Pizza and the Pizza2-SR sequences are annotated in red in B and F.
To investigate evolution by partial duplication and the self-assembling behavior of the Pizza proteins, we created a range of proteins carrying 4–10 copies of the Pizza blade. All of them seem to be folded and can be purified as monodisperse species, which were characterized using analytical gel filtration, AUC, dynamic light scattering (DLS), and ESI-MS (Fig. 2 and Table 1). Pizza4 has a molecular weight that corresponds to a trimeric state in solution, consistent with two six-bladed structures linked by domain swapping. Pizza5, Pizza8, Pizza9, and Pizza10 form complexes with the mass of five, four, three, and five six-bladed units respectively, also indicating a strong preference for the six-bladed structure. Each complex has a size determined by the lowest common multiple (LCM) of six and the number of blades per polypeptide chain (Fig. 2 and Table 1). Pizza7 is also a highly soluble protein, but gel filtration indicates the presence of two different complexes, with sizes corresponding to a hexamer, presumably composed of seven six-bladed units (referred to as Pizza7.1) and a 7-bladed monomer (referred to as Pizza7.2).

To analyze their stability, melting experiments were performed with each Pizza protein, as well as the original template protein, using a thermofluor-based assay (Fig. 4 and Fig. S3) (30). Most of the Pizza proteins have a high melting temperature, close to 80 °C, similar to the template protein, 1RWL. For proteins with fewer than six blades, no melting could be observed below 99 °C (Fig. S3). The monomeric fraction of Pizza7 (Pizza7.2), however, has a melting temperature of only 57 °C. In an attempt to improve the purification of the Pizza7.1 multimer, the cell lysates were heated to 70 °C in the hope of removing the unstable fraction, but this heat-treated Pizza7 protein (referred to as Pizza7H) then behaved only as a monomer. Pizza7H readily crystallizes, and X-ray analysis showed it to have the six-bladed structure of Pizza6 plus one disordered domain. Pizza7.2 is therefore unlikely to include a sixfold symmetrical structure despite being monomeric; it is not heat-stable and does not readily crystallize.

Circular dichroism (CD) indicates that Pizza7.2 is folded (Fig. 4). Both light-scattering and analytical ultracentrifugation showed that it forms a compact monomer in solution, but different from Pizza7H. These results suggest that Pizza7.2 may adopt a strained sevenfold symmetrical form, analogous to evolution by partial duplication and fusion. This symmetry

| Table 1. Biophysical characterization of the Pizza complexes |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Protein  | Repeats | LCM(6)* | n‡ | MW, kDa† | LCM MW§ | SEC, ml{| | SEC, kDa | AUC, S|$ | ESI, kDa | DLS, nm | DSF, °C |
| Pizza2   | 2       | 6      | 3   | 8.9   | 26.7   | 84.1           | 20  | 2.6           | 26.7           | 5.5             | ND*             |
| Pizza3   | 3       | 6      | 2   | 13.2  | 26.3   | 84.5           | 20  | 2.6           | 26.2           | 5.4             | ND*             |
| Pizza4   | 4       | 12     | 3   | 17.4  | 52.2   | 77.8           | 50  | 3.8           | 52.2           | 8.8             | ND*             |
| Pizza5   | 5       | 30     | 6   | 21.6  | 129.6  | 63.9           | 125 | 6.3           | 129.7          | 11.8            | ND*             |
| Pizza6   | 6       | 6      | 1   | 25.9  | 25.4   | 85.1           | 20  | 2.6           | 25.9           | 5.9             | 77.4            |
| Pizza7.1 | 7       | 42     | 6   | 30.1  | 180.6  | 57.3           | 175 | 8.1           | 182.7          | 14.0            | 78.2            |
| Pizza7.2 | 7       | n/a    | 1   | 30.1  | n/a    | 78.8           | 30  | 2.7           | 30.1           | 7.5             | 57.0            |
| Pizza7H  | 7       | n/a    | 1   | 30.1  | n/a    | 81.3           | 25  | 2.6           | 30.1           | 5.5             | 77.4            |
| Pizza8   | 8       | 24     | 3   | 34.3  | 103.0  | 65.2           | 125 | 5.6           | 103.2          | 10.2            | 77.8            |
| Pizza9   | 9       | 18     | 2   | 38.6  | 77.2   | 69.7           | 100 | 4.7           | 77.2           | 9.0             | 75.4            |
| Pizza10  | 10      | 30     | 3   | 42.8  | 128.4  | 63.4           | 130 | 6.5           | 128.6          | 10.0            | 76.2            |
| 1RWL     |         |        |     | 28.1  | 83.7   | 25             | 26  | 2.6           | 28.1           | 5.4             | 77.8            |

*The lowest common multiple (LCM) of the number of repeats and six, the number of blades in Pizza6.
†The degree of oligomerization in solution.
‡Theoretical molecular mass of the monomer.
§Theoretical molecular mass of the LCM complex (kDa).
{| The elution volume used to derive estimated molecular mass, shown in the next column.
| Not determined; no melting was observed up to 99 °C.
^Not applicable. Pizza7.2 and Pizza7H are monomeric species.
$Sedimentation coefficient (Svedberg).
change implies considerable adaptability of the tertiary structure to accommodate the extra blade.

Discussion

To our knowledge, Pizza6 is the first successfully designed, completely symmetrical propeller protein. The theory that the β-propeller protein family originated from duplication and fusion of ancestral fragments suggests that this evolutionary process can be reverse-engineered, and Pizza6 recapitulates a possible intermediate structure, shared by Pizza2 and Pizza3, in the evolution of a natural propeller protein. These highly stable artificial proteins indicate that it is indeed possible that propeller proteins may have arisen from symmetric multimers. The *Ralstonia solanacearum* lectin is one example of a natural β-propeller assembly that consists of a trimeric two-bladed protein, also indicative of this process (31).

Although it is clear that duplication of an entire gene can lead to symmetrical structures with an even number of repeats, the origin of protein structures with odd symmetry is more obscure. Previous experiments have indicated that structural plasticity allows for duplication or loss of repeats by domain swapping to create oligomeric assemblies. For example, symfoil is an artificial protein with perfect internal threefold symmetry; when a polypeptide carrying 2 repeats instead of 3 repeats was expressed, it assembled into a trimer with two trefoil domains, each with threefold symmetry (2). Similarly, tachylectin-2 is a protein with 5 repeats. Expressing two tandem copies of a designed tachylectin-2 domain led to a complex with 10 repeats in total (18).

This mechanism, however, does not explain the diversity of symmetry in the β-propeller family. Insertion or deletion of a single domain could have created odd-numbered symmetry from an evenly symmetrical precursor. At first, such a protein would be relatively unstable but would evolve by subsequent, less drastic changes into a stable form. This process is reflected by the creation of Pizza7.2 from Pizza6. Although Pizza7 predominantly folds as a hexameric complex, with a total number of 42 repeats (Pizza7.1), there is also a smaller fraction of the isolated protein that is monomeric (Pizza7.2). From the melting experiments and CD, it can be observed that this protein is folded, but it is less stable and no crystals have been obtained despite considerable effort. It is different from the heat-treated protein, Pizza7H, a six-bladed protein with one free domain. The data are therefore consistent with Pizza7.2 possessing a sevenfold symmetrical shape, with lower stability than the sixfold. We have previously demonstrated a similar change of rotational symmetry with the 11mer ring protein TRAP. TRAP is able to switch to 12-fold symmetry by simple tandem duplication of two, three, or four copies of the protein (32, 33).

Not only do the crystal structures presented here support the possibility that ancestral β-propeller proteins were symmetrical multimers, they also validate our design strategy inspired by ancestral reconstruction. The prediction of probable ancestor sequences, and then selection of those most compatible with a perfectly symmetrical structure template, are two critical elements to our successful design. This strategy is very rapid, and the very first expression experiments yielded the desired protein. It may be extended to the design of perfectly symmetrical proteins generally. In the case of the Pizza proteins, the differences between the predicted and experimental structures are close to experimental error, and the sixfold structure is remarkably stable. Pizza2-SR was also crystallized, showing the fold is stable to considerable surface variation, including the uncompensated charges of six arginine side chains in contact. Arginine stacking is not unprecedented however (34) and, contrary to expectations, does not destabilize the protein.

The crystal structures of the Pizza proteins show that they are highly symmetric and assemble into a zig-zag pattern, which is found in all of the structures solved (Figs. S3 and S4).

![Fig. 3. Crystallographic structures of the Pizza proteins. X-ray crystallographic analysis of five Pizza proteins confirmed the expected quaternary structure in each case, showing a six-bladed propeller. One blade of Pizza7H is not visible in the electron density maps. Superposition of the expected and experimental structures (bottom row) demonstrates close agreement with the backbone-rmsd as shown. The mutated residues in Pizza2-SR are depicted as spheres.](image)

---

**A** Differential Scanning Fluorimetry

**B** Circular Dichroism

---

Voet et al.  
PNAS | October 21, 2014 | vol. 111 | no. 42 | 15105
Well-ordered crystals were obtained very rapidly in a wide variety of conditions tested. These structures show very similar crystal packing contacts (Fig. S5), despite having very different cell dimensions. This packing suggests that the Pizza proteins may be suitable as crystallization tags for mono-, di-, and trimeric proteins and that their self-assembly properties may have other applications.

All Pizza proteins self-assemble into complexes corresponding to the lowest common multiple (LCM) of the number of repeats and six. Other designed proteins with repeated domains [based on a trefoil architecture (2) or tachylectin-2 (18)] have been shown to dimerize to maintain either three- or fivefold rotational symmetry, but the Pizza proteins demonstrate a much larger range of multimeric forms generated by a simple arithmetical rule. This behavior may be facilitated by the Velcro strap, which is suitable for domain swapping. Our results show that the Pizza proteins’ tendency to reassemble into a sixfold propeller can drive association of many protein chains. This strong preference for a particular symmetry can therefore be used to direct self-assembly according to simple rules and can be exploited for the rational design of novel protein building blocks for bionanotechnology, to develop "crystaline"-like materials or other shapes such as capsids (35–39).

Conclusion
We have designed a novel, symmetrical protein to study the role of domain duplication in protein evolution. Starting from a natural nonsymmetrical template protein, our rapid computational procedure yielded on the first attempt a 42-residue repeat sequence capable of assembling into a sixfold symmetric propeller. High-resolution X-ray crystallographic analysis confirmed the expected structure, showing that evolution of modern natural propeller proteins may have occurred via such an intermediate. The patterns of resolution X-ray crystallographic analysis confirmed the expected procedure yielded on the first attempt a 42-residue repeat sequence nonsymmetrical template protein, our rapid computational procedure may be generally useful for readily creating, proteins may have occurred via such an intermediate. The patterns of resolution X-ray crystallographic analysis confirmed the expected procedure yielded on the first attempt a 42-residue repeat sequence.
using an An-50 Ti rotor. For sedimentation velocity experiments, cells with a standard Epon two-channel centrepiece and sapphire windows were used. Four hundred microlitres of protein (1.0 mg/mL) and 420 μL of reference buffer (20 mM Hepes, pH 8, 100 mM NaCl) were used in each experiment. The rotor temperature was equilibrated at 20 °C in the vacuum chamber for 2 h before starting each measurement. Absorbance (280 nm) scans were collected at 5-min intervals during sedimentation at 50,000 rpm (182,000 × g). The resulting scans were analyzed using the continuous distribution c(s) analysis module in the program SEDFIT (47). Sedimentation coefficient increments of 200 were used in the calculation to give a separate range for each sample. The frictional coefficient was allowed to float during fitting. Partial specific volume of the proteins, solvent density, and solvent viscosity were calculated using the program SEDNTERP (48).

Electrospray Ionization Mass Spectrometry. Samples for Nanoflow ESI were prepared by extensive dialysis against 20 mM ammonium acetate. The protein concentration was adjusted to 10 μM by dilution with 20 mM ammonium acetate. The mass spectra were obtained by Synapt G2 HDMS mass spectrometer (Waters) with a nanoESI source. The mass spectra were calibrated with CsCl, Cs2SO4 ions from m/z 1,000 to m/z 10,000. MassLynx version 4.1 software (Waters) was used for data processing and peak integration. The temperature of the ion source was set to 70 °C. An aliquot of 3 μL of the sample solution was placed in a nanospray tip (HUMANIX) and electrosprayed at 0.8–1.0 kV.

Differential Scanning Fluorimetry. Differential scanning fluorimetry (DSF) was performed using a Roche LightCycler 480 to determine the stability of the protein using RT-PCR. Then, 15-μL samples of each Pizza protein (0.5 mg/mL, in 20 mM Hepes, pH 8) mixed with Sypro-orange (5x, final) (Sigma) were incubated at an increasing temperature from 25 °C to 90 °C with a temperature gradient of 3 °C per min. The fluorescence was monitored using standard excitation/emission wavelengths, and the protein Tm was determined using the manufacturer’s software.

ACKNOWLEDGMENTS. We thank Prof. Satoko Akashi for electrospray mass spectrometry measurements and Dr. Shiniichiro Egashira for help with RT-PCR. A.R.D.V. acknowledges the RIKEN Foreign Postdoctoral Research program for a postdoctoral grant and the Japan Society for the Promotion of Science for a grant-in-aid, as well as the RIKEN Integrated Cluster of Clusters for computational time.