PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Evolution of the β-propeller fold

Indronil Chaudhuri, Johannes Söding, and Andrei N. Lupas*

Department for Protein Evolution, Max Planck Institute for Developmental Biology, 72076 Tuebingen, Germany

ABSTRACT

β-Propellers are toroidal folds, in which repeated, four-stranded β-meanders are arranged in a circular and slightly tilted fashion, like the blades of a propeller. They are found in all domains of life, with a strong preponderance among eukaryotes. Propellers show considerable sequence diversity and are classified into six separate structural groups by the SCOP and CATH databases. Despite this diversity, they often show similarities across groups, not only in structure but also in sequence, raising the possibility of a common origin. In agreement with this hypothesis, most propellers group together in a cluster map of all-β folds generated by sequence similarity, because of numerous pairwise matches, many of which are individually nonsignificant. In total, 45 of 60 propellers in the SCOP25 database, covering four SCOP folds, are clustered in this group and analysis with sensitive sequence comparison methods shows that they are similar at a level indicative of homology. Two mechanisms appear to contribute to the evolution of β-propellers: amplification from single blades and subsequent functional differentiation. The observation of propellers with nearly identical blades in genomic sequences show that these mechanisms are still operating today.

Proteins 2008; 71:795–803.

Key words: cluster analysis; evolution; homology; profile hidden Markov model; protein.

INTRODUCTION

Repetition is a widespread feature of natural proteins. The repeated units vary considerably in size, from single residues to entire domains. Repeats that fold into single domains yield a hierarchy of structural complexity, from fibrous domains made by the repetition of patterns only a few amino acids long (collagen, coiled coils, β-helices), to solenoid domains formed by the repetition of simple supersecondary structures (αα-hairpins: tetratricopeptide and HEAT repeats; ββ-hairpins: choline-binding domains; βα-hairpins: leucine-rich repeats), to globular domains formed by the repetition, frequently in interleaved form, of complex supersecondary structure units (ββαβ: cradle-loop barrels; βαβ: ferredoxins). Toroids are intermediate in complexity between solenoids and globular proteins, as they are usually formed by simple, noninterleaved supersecondary structure units, but fold into a closed, rather than open structure (αα-hairpins: protein prenyltransferases; ββ-hairpins: porins; βα-hairpins: TIM barrels).

β-Propellers are circular folds with 4–8 repeats. This variability is unusual among toroids, which are generally formed by a specific number of repeats, differing in special cases by at most one unit (as, e.g., in the six-hairpin versus seven-hairpin glycosidases). From an evolutionary point of view, β-propellers are also remarkable as their degree of internal sequence symmetry ranges over the full evolutionary spectrum, from binding proteins with nearly identical repeats to fully differentiated enzymes whose origin from repetition can only be seen in their structures. The repeated unit is a four-stranded β-meander, whose strands are labeled A to D from N- to C-terminus. The meanders are arranged radially and in slightly tilted fashion around a central pore, with strand A innermost. They resemble the blades of a propeller, hence the name of the structure. Although their detailed interaction varies somewhat between propellers with different numbers of blades, the blades themselves are structurally very similar, irrespective of the size of the propeller from which they originate, and their inner three β-strands can generally be superimposed with an RMSD of less than 1 Å (Fig. 1). The central pore they enclose is funnel-shaped rather than cylindrical, with the narrow end often being the binding site for ions or substrates.

In many β-propellers, up to three strands of the last blade are circularly permuted to the N-terminus of the protein. This places the termini of the protein into adjacent, hydrogen-bonded strands ("velcro" closure), instead of between blades, and presumably confers additional structural stability to the propeller. Some propellers also have N-terminal extensions that add a fifth strand to the first blade.

Grant sponsor: Max Planck Society.
*Correspondence to: Andrei N. Lupas, Department for Protein Evolution, Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tuebingen, Germany. E-mail: andrei.lupas@tuebingen.mpg.de
Received 29 May 2007; Revised 8 July 2007; Accepted 27 July 2007
Published online 2 November 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21764

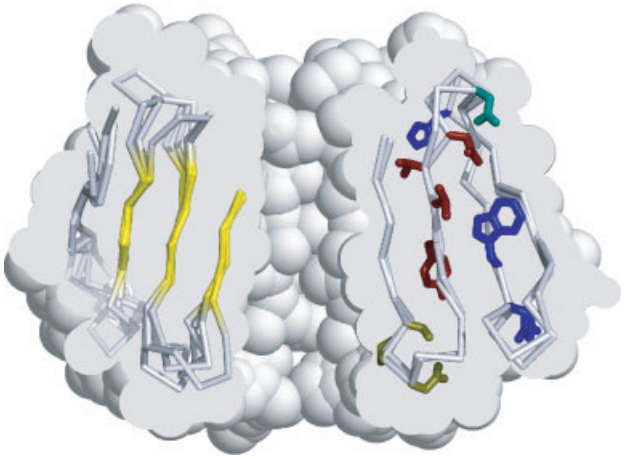


Figure 1

Structural features of β-propellers. On the left is shown a superposition of propeller blades from 4-, 5-, 6-, 7-, and 8-bladed β-propellers (blade 1 of 1hxn, blade 2 of 1gyd, blade 4 of 1ijq, blade 4 of 2trc, and blade 7 of 1n90, respectively). The three inner β-strands superimpose very well, irrespective of the blade number of the full propeller. The spacefilling slab view of a propeller in the background illustrates the position of the blades in the protein. On the right is shown the spatial location of characteristic sequence motifs: SPDG at the end of β-strand A (yellow), YVTN as the hydrophobic core of strand B (red), DG in the turn between strands B and C (green), WD at the end of strand C and GH at the end of strand D (blue). For the location of these motifs in sequence see Figure 2. As no single protein combines all motifs, we superimposed blades from two proteins (blade 4 of 2trc and blade 3 of 1l0q). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Despite their similar structures, β-propellers have remarkably diverse sequences and several families have been described, usually based on the occurrence of the following specific sequence patterns.

WD40: This family of 7-bladed propellers, discovered in the β-subunits of heterotrimeric G proteins, was the first to be described and represents by far the biggest family of propellers.[1,2] The repeat was named for its length of 40 residues and for a prominent Trp-Asp (WD) motif. The name was meant to evoke a popular brand of motor oil (WD-40) because of a proposed functional analogy. The WD motif was originally thought to occur at the C-terminus of the repeat, but is in fact located at the end of β-strand C. WD40 proteins are involved in a wide range of processes, including signal transduction, transcriptional regulation, and apoptosis.[3] So far, no dominant mode of action has been identified, but it seems that most act as mediators of protein–protein interactions.

RCC1: The regulator of chromosome condensation 1 (RCC1) and related proteins are involved in transport between the nucleus and the cytoplasm, as well as in cell cycle control.[4] Like WD40, they form 7-bladed propellers with velcro closure, but their repeat unit is more than 11 residues longer due to longer connecting loops in the β-meander. RCC1 repeats do not have strongly conserved sequence motifs. The most visible is an Asp-Gly (DG) motif at the beginning of strand C.

KELCH: This repeat was named for the Drosophila Kelch protein, where it was first observed. Proteins in this group fold into 6- as well as 7-bladed propellers. In several cases, KELCH proteins were found to outcompete Gβ subunits from binding to G protein alpha.[5] KELCH resembles WD40 in containing a Tyr-Asp (YD) motif at the end of strand C.

YWTD: This repeat was discovered in proteins of the extracellular matrix and in extracellular domains of receptors,[6,7] such as the LDL receptor.[8] The name stems from the characteristic amino acid sequence Tyr-Trp-Thr-Asp (YWTD), which forms the end of strand B.

NHL: This repeat was named for the three proteins in which it was first found, NCL-1, HT2A, and LIN-41.[9] Reported at the same time as YWTD, its core motif, also at the end of strand B, is a variant thereof: YVTD.

YVTN: This repeat was described as a divergent variant of YWTD in archaeal and metazoan surface proteins, but is in fact closer to NHL.[10] A representative YVTN structure from the surface layer protein of methanosarcina displays an unusually high level of structural symmetry.[11] In addition to the name-giving motif, the repeat also contains a Ser-Pro-Asp-Gly (SPDG) motif in the loop between strands A and B, and an Ile-Asp (ID) motif at the end of strand C, in the same location as the WD motif in WD40 and the YD motif in KELCH. Neither motif appears in NHL proteins.

Most families listed above were described independently and their subsequently discovered structural similarity to other β-propellers was often seen as the result of convergent evolution[5,12] due to the low levels of sequence similarity (but note the proposed homology of YVTN, NHL, and YWTD[11]). We were, however, intrigued by the tendency of iterated sequence profile searches to connect proteins from different families and by the tendency of motifs characteristic for one family to show up, sometimes in variant form, in one or more of the other families. Both observations generally indicate homologous relationships. We therefore undertook to investigate the issue of homology versus analogy using advanced bioinformatic tools and found that most, if not all, β-propellers are likely to have originated from a common ancestor. The evolutionary mechanism appears to operate by the amplification of single blades and subsequent differentiation.

METHODS

Structural superposition of β-propeller blades

To generate structural superpositions, proteins were split into their individual blades and these were interac-

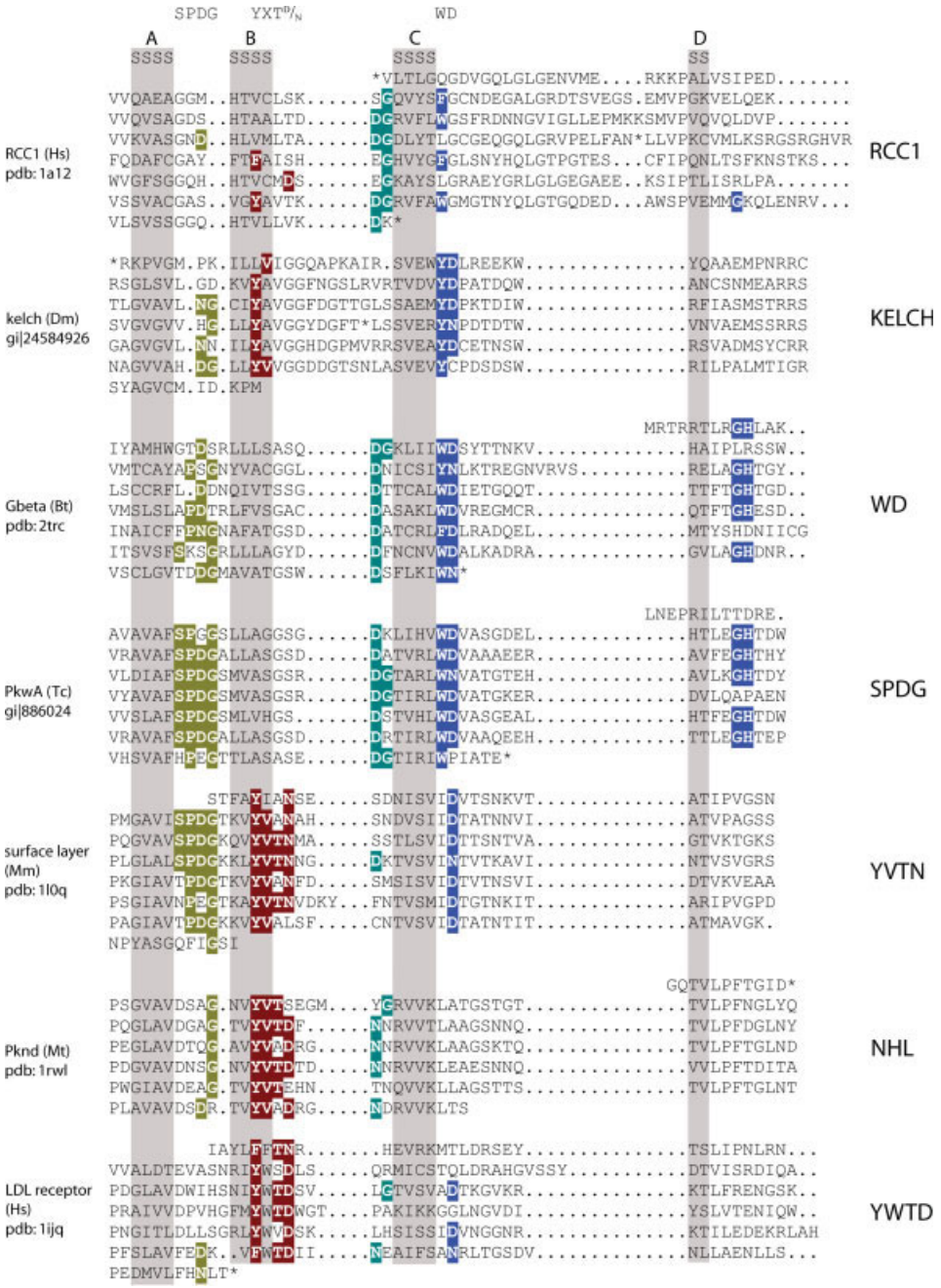


Figure 2

Structure-based sequence alignment of β-propeller protein families. The location of key motifs within the blades is shown with the same color coding as in Figure 1. (Hs: Homo sapiens; Dm: Drosophila melanogaster; Bt: Bos taurus; Tc: Thermomonospora curvata; Mm: Methanosarcina mazei; Mt: Mycobacterium tuberculosis). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

tively superimposed with Swiss PDB Viewer.[13] The positions used for superposition correspond to the residues of strands A, B, and C that are highlighted in grey in Figure 2. For Figure 1 (left), one blade each was chosen from propellers with 4, 5, 6, 7, and 8 blades (as listed in the figure legend), such that the connecting loops would not be too extended and obstruct the view onto the central β-meander. For Figure 1 (right), the blades were chosen to contain exemplars of the sequence motifs highlighted in Figure 2. For Figure 2, we used the proteins from which the main propeller families were originally defined. Five of these have known structures and one

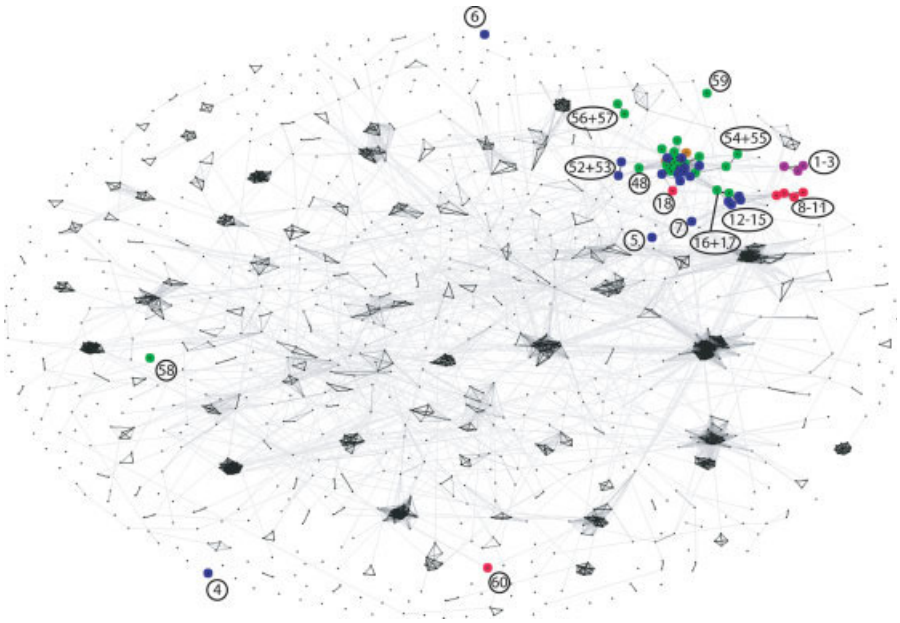


Figure 3

Cluster map of proteins classified in SCOP25 as all-β. The colored dots mark the location of propellers (violet = 4-bladed; red = 5-bladed; blue = 6-bladed; green = 7-bladed; brown = 8-bladed). The color code and sequence numbering are used consistently with Figure 4. For clarity, only proteins outside the central cluster are labeled with their number in the matrix of pairwise connections shown in Figure 4. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

(*Drosophila* kelch) has a close homolog of known structure. We superimposed all the blades of one structure to generate a structural alignment for that protein. We then took one blade from each protein to generate a second structural superposition, which was afterwards used to adjust the alignments of the individual proteins to generate one structural alignment of all selected blades. Attention was paid to precisely align the β-strands of the first three blades.

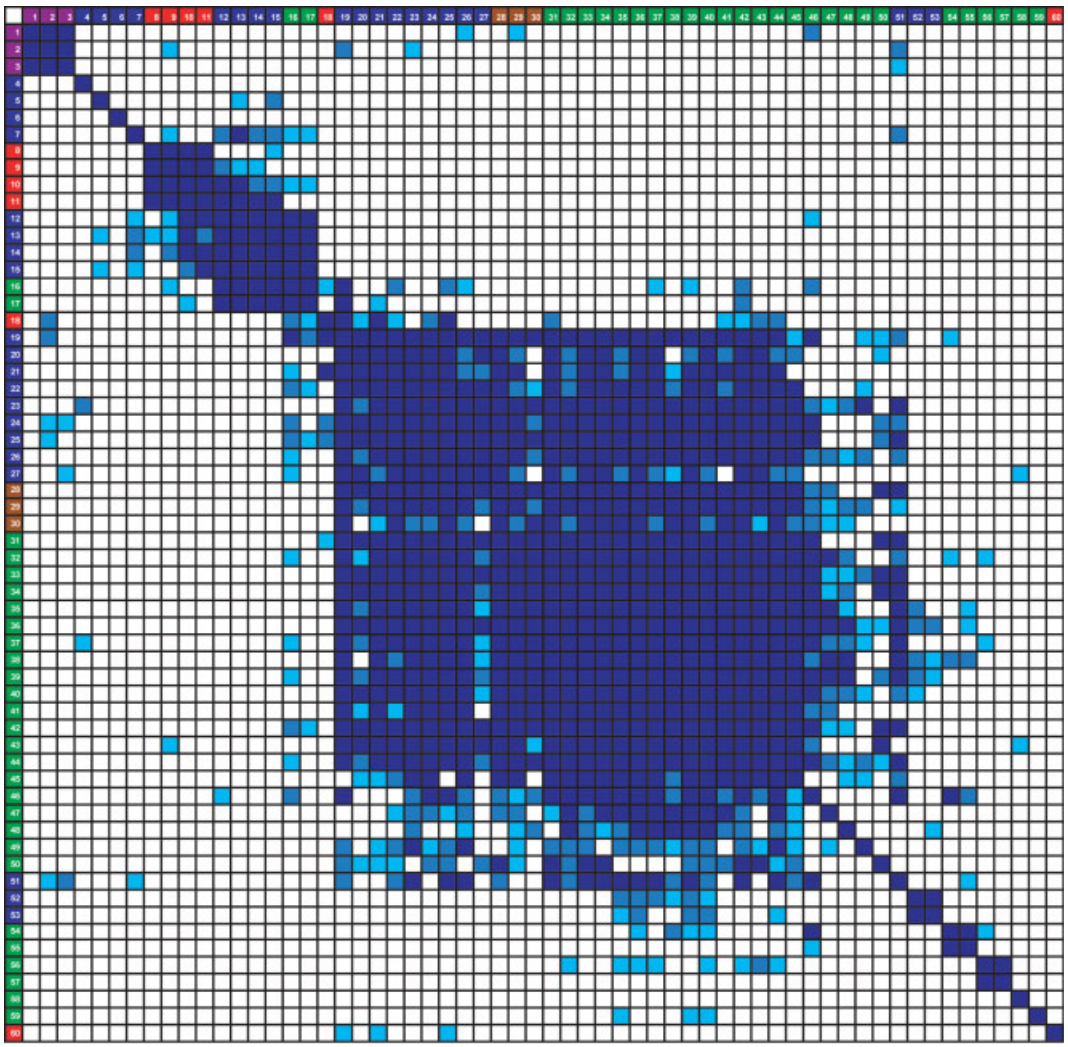
Cluster map of all-β folds

We used the SCOP database,[14] version 1.71, filtered to 25% sequence identity (SCOP25), downloadable from the HHpred web server,[15] and extracted all proteins with an all-beta fold description (SCOP identifier b.). For the construction of the HHpred alignment and HMM databases, a modified PSI-BLAST[16] procedure (buildali.pl) has been developed that largely suppresses the corruption of alignments by preventing the inclusion of nonhomologous sequence-stretches at the ends of PSI-BLAST high-scoring segment pairs (J. Söding, unpublished). For the all beta domains, we computed an all-against-all comparison using HHsearch (version 1.5.0).[17] HHsearch is a highly sensitive homology search tool based on the pairwise comparison of hidden Markov models (HMMs).[15] Importantly, version 1.5.0 employs an amino acid bias correction method, which suppresses a positive score contribution arising merely from the similarity of the amino acid composition of two proteins.

We clustered the proteins by their pairwise P-values using CLANS,[18] an implementation of the Fruchterman-Reingold algorithm that scales log-*P*-values into attractive forces in a force field (Fig. 3). During the clustering step, pairwise connections up to a *P*-value of 1e-3 were used and the CLANS parameters were set to: attraction and repulsion = 10 and the attraction exponent = 2. Clustering was done in three dimensions until convergence. The reproducibility of the map was ascertained by multiple independent cluster runs from random start conditions.

A matrix of β-propeller connections based on HHsearch

We extracted all β-propellers contained in SCOP25 (folds b.66 to b.70, 60 proteins total), and submitted them to the HHpred webserver (http://toolkit.tuebingen.mpg.de/hhpred) in April 2007, using the SCOP70 1.71 database of profile HMMs. We turned off secondary structure scoring to reduce the chance of matches resulting primarily from agreements in local secondary structure. The pairwise probabilities obtained are discussed in the text and summarized in Figure 4. As expected for pairwise comparisons, the matrix is almost symmetrical, the slight asymmetry in the probabilities (as opposed to



1	1hxn	Hemopexin
2	1qhu	Hemopexin
3	1itv	Gelatinase
4	1ofz	Fucose-specific lectin
5	1v3e	Hemagglutinin-neuraminidase
6	1f8e	Influenza neuraminidase
7	1v0e	Endo-alpha-sialidase
8	1oyg	Levansucrase
9	1gyh	Alpha-L-arabinanase
10	1y4w	Exo-inulinase
11	1vkd	Hypothetical protein TM1225
12	1w0p	Vibrio cholerae sialidase
13	1so7	Sialidase 2
14	2sli	Leech trans-sialidase
15	3sil	Salmonella sialidase
16	1sqja1	Oligoxyloglucan reducing CBH
17	1sqja2	Oligoxyloglucan reducing CBH
18	1tl2	Tachylectin-2
19	1rwi	Ser/Thr-protein kinase PknD
20	1v04	Paraoxonase/arylesterase 1
21	1cru	Soluble quinoprotein glucose DH
22	1ijq	Low density lipoprotein receptor
23	1crz	TolB
24	1pjx	Diisopropylfluorophosphatase
25	1q7f	Brain tumor cg10719-pa
26	1k32a2	Tricorn protease N-domain
27	1h6l	Thermostable phytase
28	1qks	Cytochrome nitrite reductase
29	1orv	Dipeptidyl peptidase IV/CD26
30	1w6s	Methanol DH
31	2bbk	Methylamine DH
32	1yfq	Cell cycle arrest protein BUB3
33	1l0q	Surface layer protein
34	1pby	Quinohemoprotein amine DH
35	1tbg	Beta-subunit G protein
36	1gxr	Groucho/tle1
37	1k8k	Arp2/3 complex
38	1nex	Cdc4 propeller domain
39	1nr0a1	Actin interacting protein 1
40	1nr0a2	Actin interacting protein 1
41	1sq9	Antiviral protein Ski8
42	1ri6	Putative isomerase YbhE
43	1jof	Mucoante lactonizing enzyme
44	1k32a3	Tricorn protease domain 2
45	1qfm	Prolyl oligopeptidase
46	1k3i	Galactose oxidase
47	1xip	Nucleoporin NUP159
48	1utc	Clathrin heavy-chain
49	1ve6	Acylamino-acid-release enzyme
50	1fwx	Nitrous oxide reductase
51	1u6d	Kelch-like ECH-associated
52	1wp5	Topoisomerase IV
53	1suu	DNA gyrase A
54	1a12	RCC 1
55	1jtd	Beta-lactamase inhibitor protein
56	1olz	Semaphorin 4d
57	1shy	Hepatocyte growth factor recept
58	1txv	Integrin alpha
59	1xks	Nuclear pore complex Nup133
60	1s1d	Ca-activated nucleotidase

Figure 4

Pairwise sequence comparison of β-propellers in SCOP25. The query structures are listed in the top row. Each column indicates which other propellers from the dataset are detected by HMM-HMM comparison at probabilities >90% (dark blue), 50–90% (blue), and 20–50% (light blue). The probabilities correspond to the likelihood that two proteins at the respective E-value are classified as homologous in the SCOP database. The colors on the edges indicate the size of the propeller, as in Figure 3. The five proteins of known structure from Figure 2, as well as the human ortholog of a sixth (Kelch), are highlighted in yellow.

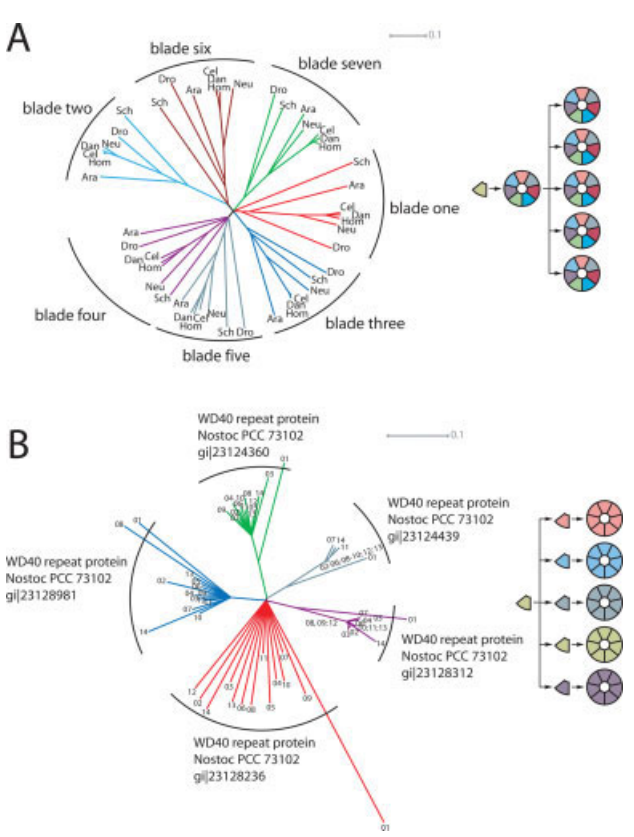


Figure 5

*Two evolutionary scenarios for β-propellers, illustrated by neighbor-joining phylogenetic trees. (**A**) Divergent evolution of a fully formed propeller, as seen in the β-subunits of heterotrimeric G proteins (Ara: Arabidopsis thaliana; Cel: Caenorhabditis elegans; Dan: Danio rerio; Dro: Drosophila melanogaster; Hom: Homo sapiens; Neu: Neurospora crassa; Sch: Schizosaccharomyces pombe). (**B**) Amplification from single blades, as seen in 14-bladed WD40 proteins from Nostoc. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

the raw scores) being mainly caused by the dependence on the calibration of individual HMMs.

Phylogenies of β-propeller proteins

We chose a phylogenetically representative set of heterotrimeric G protein β subunits and five propellers from Nostoc PCC 73102, an organism rich in highly repetitive WD40 proteins. We aligned the sequences of individual blades in ClustalX,[19] adjusted the gapping locally, and computed distance-based neighbour-joining trees using the default parameters of ClustalX, the tree was visualized using SplitsTree[20] (Fig. 5).

RESULTS AND DISCUSSION

Because of the limited number of structural solutions available to a folded polypeptide chain, unrelated proteins tend to converge upon similar local solutions, termed supersecondary structures, the most prominent of which are αα- and ββ-hairpins, and βαβ-elements.[21] In contrast, the combinatorial sequence space is gigantic and many sequences are compatible with a particular local structure, so that sequence convergence is rare. For this reason, sequence similarity is viewed as the hallmark for homology. We therefore focussed on sequence analyses to evaluate the possibility of homologous relationships among β-propellers.

A continuum of sequence motifs

Our starting point was a comparison of sequence motifs across the main described families of β-propellers, whose spatial arrangement in a blade is shown in Figure 1. To this end, we built a multiple alignment from representatives of each family, broken down to the level of single blades (Fig. 2). This was guided, where available, by structural information. For Figure 2, we selected as representative sequences the proteins from which the respective family was originally defined. In addition to the families listed in the introduction, we generated a group exemplified by the protein kinase PkwA from the actinobacterium *Thermomonospora curvata*, which we call SPDG for the very prominent pattern of residues connecting strands A and B. This is a subgroup of WD40, whose proteins have particularly well-conserved sequence repeats. In profile sequence searches with β-propellers, these proteins are generally the first ones identified outside the family of the query sequence. As can be seen from Figure 2, the main motifs described as characteristic for a family almost invariably span several families, forming a continuous motif space across propellers. Only the GH motif at the end of strand D appears to be specific for the extended WD40 family. The most prominent motif consists of a large hydrophobic residue, followed by Asp or Asn, at the end of strand C; this motif is present, at least partially, in most β-propellers.

β-propellers group together in a cluster map of all-β folds

In an attempt to evaluate whether these sequence motifs are convergent and in fact mainly reflect structural constraints of ββ-hairpins in general and β-meanders in particular, we combined a recently developed method for comparing HMMs, HHsearch,[15] with a clustering method built on the Fruchterman-Reingold algorithm, CLANS,[18] to produce a cluster map of all-β folds from the structural classification of proteins (SCOP) database, filtered for a maximum pairwise identity of 25% (SCOP25). This map was produced using sequence information alone (Fig. 3). We reasoned that all-β folds, being built largely of ββ-hairpins and containing many instances of β-meanders, would provide a suitable background

of convergent sequence motifs to disrupt any grouping of convergently arisen propellers. The map, however, showed that, with the exception of four proteins, all propellers grouped in the same area, regardless of their SCOP fold number. Six of the seven proteins aligned in Figure 2 either have known structures or, in the case of Kelch, orthologs of known structure and appear in this map. Five cluster in the central group of β-propellers, while RCC1 (No. 54) is still proximate and connected to the central group.

HMM-HMM comparisons of β-propellers

To systematically explore the sequence relationships between β-propellers, we took the 60 propellers in the SCOP25 database, version 1.71 (folds b.66 to b.70), and compared profile HMMs of their sequences to all β-propeller profile HMMs in SCOP70, using the webserver HHpred.[15] We excluded the partial propeller 1n7v, which shows no sequence similarity to other propellers and is classified by itself in SCOP (b.126). All propellers had their top match to another propeller and, of the 2597 matches with E-values less than 1e-5, all were to other β-propellers.

The best matches to proteins with a different fold were between *Vibrio cholerae* sialidase (1w0p) and *Serratia marcescens* chitobiase (1qba) at $E = 3.9e\text{-}3$, and between leech sialidase (2sli) and the same chitobiase at $E = 4.4e\text{-}3$. This relationship has been described previously[22,23] and has been discussed as a possible instance of remote homology between proteins with different folds.[22,24] It is based on a conserved β-hairpin with a SxDxGxxW turn motif (the "Asp-box"), which occurs between strands C and D in β-propellers. Asp-boxes are found in immunoglobulin-like beta-sandwiches (SCOP b.1.18), microbial ribonucleases (d.1.1), sialidases (6-bladed β-propellers; b.68.1), and cellobiohydrolases (7-bladed β-propellers; b.69.13); and, in a variant form that is one residue shorter (SxDxxxW), in the arabinanase/levansucrase/invertase group of 5-bladed β-propellers (b.67.2), and in carbohydrate binding domains (b.18.1). It is striking to note in this context that all Asp-box-containing propellers act on carbohydrates; a direct role of this motif in carbohydrate binding remains, however, unclear.

The pairwise relationships between propellers are summarized in Figure 4. The relationships are shown as HHsearch probabilities, which correspond to the likelihood that matches at the respective E-value connect homologous proteins in SCOP (i.e., proteins of the same family or superfamily).[17]

More than half of all propellers form a tightly knit group (Nos. 18–51), networked by many, statistically highly significant pairwise connections. This central group is structurally and functionally diverse, containing propellers with 6, 7, and 8 blades, as well as peripherally one 5-bladed propeller, tachylectin (1tl2; No. 18). Of the seven proteins in Figure 2, the five that clustered together in Figure 3 (Nos. 19, 22, 33, 35, and 51) are also present in this central group.

The second largest group, consisting of 5, 6, and 7-bladed propellers, contains the aforementioned carbohydrate-modifying propellers, in which the Asp-boxes form the dominant sequence feature (Nos. 7–17). Connections within this group are strengthened by intermediate structures that contain both canonical and variant Asp-boxes (*Salmonella sialidase*, 3sil, and *Aspergillus inulinase*, 1y4w).

The two main groups of propellers show multiple connections at probabilities >50%, as well as several at >90%, bringing together 45 of the 60 propellers. These connections are primarily due to the two propeller domains in cellobiohydrolase (1sqj), which, in addition to multiple Asp-boxes, also show several motifs typical of other propellers, including a clear WD motif in the first blade of the first propeller. The cellobiohydrolase HMM also includes further bridging proteins, such as a surface protein from *Methanoculleus marisnigri* (gil126179850), whose propeller has a high degree of internal sequence symmetry and combines canonical Asp-boxes with an xPDG motif at the same location as the SPDG motif of the extended WD40 family.

The remaining 15 propellers in Figure 4, including all three 4-bladed propellers, show fewer connections to other propellers, but several, such as RCC1 (No. 54), still make some connections at better than 90% probability. Only one propeller remains entirely unconnected at a 10% probability cutoff. This is influenza neuraminidase (1f8e), which is thought to be a divergent member of the sialidase family (SCOP b.68.1) and belong to the group of carbohydrate-modifying propellers, even though it lacks Asp-boxes.

The number and strength of sequence matches between most propellers in Figure 4 is indicative of homology, judging by the calibration of HHsearch on the SCOP database. Thus, the grouping of propellers into five different SCOP folds, based on their blade number, is not consistent with the levels of sequence similarity at which other proteins are typically grouped into a single superfamily. Apparently, in producing the classification, blade number outweighed sequence considerations to cause a division of propellers into different folds, even though in other toroids (TIM barrels, glycosidases), variations in the number of repeat elements—admittedly not as large as in propellers—did not prevent their grouping into single folds. We propose that, at a minimum, the 45 well-connected propellers in Figure 4 are better understood as members of the same superfamily. As the absence of evidence is not evidence of absence, we would further argue that the weak or missing connections for the remaining 15 propellers is not evidence of their analogous origin. Given the critical role that bridging proteins have in connecting groups of propellers in

sequence space, such as, for example, cellobiohydrolase and tachylectin in Figure 4, we think that future additions to the structure database may well provide currently missing connections.

Evolutionary scenarios

Many globular protein families are structurally repetitive. For the most part, the events that led to the repetition are ancient and predated the radiation of the family; thus, each repeat is more similar in sequence to the equivalent repeat in the other members of the family than to other repeats of the same protein. This situation is certainly also observed in some β-propellers; for example, in phylogenies of G protein β-subunits, equivalent blades branch together at the exclusion of other blades [Fig. 5(A)]. In general, however, β-propeller sequences show features that set them aside from most other repetitive folds:

i. Their sequences span the full range of internal symmetry, from highly repetitive to fully diverged. This suggests that the propeller fold has arisen repeatedly by amplification from single blades at different times in evolution. Unless constrained by a symmetrical substrate or by the need for a high degree of structural stability, repeats are expected to diverge by adaptive differentiation and neutral drift. Thus, propellers with high internal sequence symmetry should, on balance, be of more recent evolutionary origin than fully differentiated propellers. The existence of propellers with nearly identical blades, such as for example ORF Npun02007829 of Nostoc punctiforme (gi|23124439), where all 14 blades recognizable in the sequence are identical in all but one position, shows that this is an ongoing process. Indeed we found that Nostoc is an organism with a particularly high number of recently amplified propellers [Fig. 5(B)].
ii. Propellers do not display any symmetry other than that based on the repetition of single blades. Thus, for example, we have not found six-bladed propellers with two- or threefold symmetry. If a propeller is repetitive in sequence, the repeat is always a single blade. There are no instances known to us that would parallel the HisA and HisF situation in TIM barrels, where a twofold sequence symmetry overlies the basic eightfold structural symmetry of the fold.[25,26] We conclude that propellers are always amplified from single blades and not by duplication from units containing more than one blade.
iii. The blades of a propeller usually display the same sequence motifs. Blades in propellers of the WD40 family adhere to the WD consensus and blades of YTWD propellers adhere to the YWTD consensus, although we observed a small number of blades in the structure database that seemed closer to the blades of another family (the best example being the already mentioned first blade in cellobiohydrolase, which carries a WD motif in place of the Asp-box). We conclude that the recombination between propellers from different families is rare, and that the common ancestor at the branching nodes between major propeller families was a single blade, not a fully formed propeller.

This analysis suggests that the single, ancestral blades may have had the capacity to fold as oligomers. It is, therefore, striking that there is only one oligomeric propeller found in the entire current structure database and this is a trimer of double-bladed subunits (*Ralstonia* fucose-specific lectin, 2bt9, which is not yet included in SCOP and is linked by HHpred to fungal fucose-specific lectin, 1ofz, at >90%). One possible explanation for this absence is that oligomeric propellers are less effective than their single-chain counterparts (e.g., because of lower stability) and therefore only appear as evolutionary intermediates, being rapidly displaced by their fully amplified versions.

CONCLUSIONS

We have analyzed the sequences of β-propellers and have found that at least a core group of these, comprising about 80% of propellers of known structure, are similar at a level indicative of homology. Despite their large sequence diversity, several lines of evidence point to their monophyletic origin: (i) shared sequence motifs across families; (ii) cluster formation in a map of all-beta folds; and (iii) statistically significant similarity in their profile HMMs. We cannot judge on the basis of this analysis, whether propellers not connected to the core group by these metrics are analogous developments, or homologs in which the evidence for common ancestry has been lost through sequence divergence. We note, however, that many of these proteins have few homologs in current databases; with the rapid progress of sequencing projects, their profile HMMs are likely to increase in depth, and bridging sequences may be determined that connect them to the core group. In addition, sequence analysis tools will probably continue to increase in sensitivity, allowing more distant connections to be made reliably. This situation resembles that of the TIM barrel fold, a toroidal fold which comprises a comparably large group of protein families with marginal sequence similarity, for which a monophyletic origin is now also being discussed.[27–29]

Even though we propose a common origin for all β-propellers, we do not imply that they diverged from an ancestral propeller. Rather, we find evidence that the major families of propellers were amplified independently from single blades and that this is an ongoing process.

Our scenario is based on the hypothesis that folded proteins arose from an ancestral set of peptides with the potential to form supersecondary structures.[24,30] One of these was a β-meander capable of forming toroidal structures by repetition, similar in size and complexity to the βαβ peptide that gave rise to cradle-loop barrels[31,32] or the αβα peptide that yielded domains of AAA+ proteins and histones.[33] The differentiation of the first proto-propellers yielded a population of new β-meanders, which could serve as starting points for further amplification. Since then, the world of propellers has continuously expanded through successive rounds of amplification and differentiation.

ACKNOWLEDGEMENTS

We thank Vikram Alva for the analysis of Asp-box occurrence.

REFERENCES

1. Neer EJ, Schmidt CJ, Nambudripad R, Smith TF. The ancient regulatory-protein family of WD-repeat proteins. Nature 1994;371:297–300.
2. Smith TF, Gaitatzes C, Saxena K, Neer EJ. The WD repeat: a common architecture for diverse functions. Trends Biochem Sci 1999;24:181–185.
3. Li D, Roberts R. WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. Cell Mol Life Sci 2001;58:2085–2097.
4. Renault L, Nassar N, Vetter I, Becker J, Klebe C, Roth M, Wittinghofer A. The 1.7 A crystal structure of the regulator of chromosome condensation (RCC1) reveals a seven-bladed propeller. Nature 1998;392:97–101.
5. Gettemans J, Meerschaert K, Vandekerckhove J, De Corte V. A kelch beta propeller featuring as a G beta structural mimic: reinventing the wheel? Sci STKE 2003;2003:PE27.
6. Bork P, Downing AK, Kieffer B, Campbell ID. Structure and distribution of modules in extracellular proteins. Q Rev Biophys 1996;29:119–167.
7. Springer TA. An extracellular beta-propeller module predicted in lipoprotein and scavenger receptors, tyrosine kinases, epidermal growth factor precursor, and extracellular matrix components. J Mol Biol 1998;283:837–862.
8. Jeon H, Meng W, Takagi J, Eck MJ, Springer TA, Blacklow SC. Implications for familial hypercholesterolemia from the structure of the LDL receptor YWTD-EGF domain pair. Nat Struct Biol 2001;8:499–504.
9. Slack FJ, Ruvkun G. A novel repeat domain that is often associated with RING finger and B-box motifs. Trends Biochem Sci 1998;23:474–475.
10. Edwards TA, Wilkinson BD, Wharton RP, Aggarwal AK. Model of the brain tumor-Pumilio translation repressor complex. Genes Dev 2003;17:2508–2513.
11. Jing H, Takagi J, Liu JH, Lindgren S, Zhang RG, Joachimiak A, Wang JH, Springer TA. Archaeal surface layer proteins contain beta propeller, PKD, and beta helix domains and are related to metazoan cell surface proteins. Structure 2002;10:1453–1464.
12. Andrade MA, Ponting CP, Gibson TJ, Bork P. Homology-based method for identification of protein repeats using statistical significance estimates. J Mol Biol 2000;298:521–537.
13. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
14. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–D229.
15. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33:W244–W248.
16. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
17. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.
18. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 2004;20:3702–3704.
19. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876–4882.
20. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 2006;23:254–267.
21. Orengo CA, Thornton JM. Protein families and their evolution—a structural perspective. Annu Rev Biochem 2005;74:867–900.
22. Copley RR, Russell RB, Ponting CP. Sialidase-like Asp-boxes: sequence-similar structures within different protein folds. Protein Sci 2001;10:285–292.
23. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1227.
24. Lupas AN, Ponting CP, Russell RB. 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 2001;134:191–203.
25. Hocker B, Claren J, Sterner R. Mimicking enzyme evolution by generating new (beta alpha)(8)-barrels from (beta alpha)(4)-half-barrels. Proc Natl Acad Sci USA 2004;101:16448–16453.
26. Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M. Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. Science 2000;289:1546–1550.
27. Copley RR, Bork P. Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. J Mol Biol 2000;303:627–641.
28. Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321:741–765.
29. Soding J, Remmert M, Biegert A. HHrep: de novo protein repeat detection and the origin of TIM barrels. Nucleic Acids Res 2006;34(Web Server issue):W137–W142.
30. Soding J, Lupas AN. More than the sum of their parts: on the evolution of proteins from peptides. Bioessays 2003;25:837–846.
31. Coles M, Diercks T, Liermann J, Groger A, Rockel B, Baumeister W, Koretke KK, Lupas A, Peters J, Kessler H. The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple betaalphabetabeta element. Curr Biol 1999;9:1158–1168.
32. Coles M, Hulko M, Djuranovic S, Truffault V, Koretke K, Martin J, Lupas AN. Common evolutionary origin of swapped-hairpin and double-psi beta barrels. Structure 2006;14:1489–1498.
33. Alva V, Ammelburg M, Soding J, Lupas AN. On the origin of the histone fold. BMC Struct Biol 2007;7:17.