frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 09 June 2022
doi: 10.3389/fmolb.2022.895496

New β-Propellers Are Continuously Amplified From Single Blades in all Major Lineages of the β-Propeller Superfamily

Joana Pereira[*†] *and Andrei N. Lupas*[*]

Department of Protein Evolution, Max Planck Institute for Biology, Tübingen, Germany

β-Propellers are toroidal folds, in which consecutive supersecondary structure units of four anti-parallel β-strands–called blades–are arranged radially around a central axis. Uniquely among toroidal folds, blades span the full range of sequence symmetry, from near identity to complete divergence, indicating an ongoing process of amplification and differentiation. We have proposed that the major lineages of β-propellers arose through this mechanism and that therefore their last common ancestor was a single blade, not a fully formed β-propeller. Here we show that this process of amplification and differentiation is also widespread within individual lineages, yielding β-propellers with blades of more than 60% pairwise sequence identity in most major β-propeller families. In some cases, the blades are nearly identical, indicating a very recent amplification event, but even in cases where such recently amplified β-propellers have more than 80% overall sequence identity to each other, comparison of their DNA sequence shows that the amplification occurred independently.

Keywords: protein evolution, β-propeller, repetition, amplification, differentiation

OPEN ACCESS

Edited by:
Suman Kundu,
University of Delhi, India

Reviewed by:
Liam Michael Longo,
Manatee–Sarasota, United States
Lionel Ballut,
Université de Lyon, France
Jeremy Tame,
Yokohama City University, Japan

***Correspondence:**
Joana Pereira
joana.pereira@unibas.ch
Andrei N. Lupas
andrei.lupas@tuebingen.mpg.de

***Present Address:**
Joana Pereira,
Biozentrum and SIB Swiss Institute of Bioinformatics, University of Basel, Basel, Switzerland

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

Received: *13 March 2022*
Accepted: *13 May 2022*
Published: *09 June 2022*

Citation:
Pereira J and Lupas AN (2022) New β-Propellers Are Continuously Amplified From Single Blades in all Major Lineages of the β-Propeller Superfamily. Front. Mol. Biosci. 9:895496. doi: 10.3389/fmolb.2022.895496

INTRODUCTION

Duplication is one of the most common mechanisms for the emergence of biomolecular novelties (Levasseur and Pontarotti, 2011). It can occur at different levels - from a small number of nucleotides up to whole genomes. When it happens at the scale of short nucleotide segments, it may lead to the amplification of subdomain-sized fragments and to the emergence of repetitive protein domains (Andrade et al., 2001; Söding and Lupas, 2003). Usually, the monotonous amplification of such fragments produces open-ended, solenoid folds, as in tetratricopeptide (TPR)-repeat (Blatch and Lässle, 1999) and leucine-rich repeat (LRR) proteins (Matsushima et al., 2005), but in some cases it may lead to closed, globular folds. One example is the β-propeller fold (Jawad and Paoli, 2002; Chaudhuri et al., 2008; Kopec and Lupas, 2013; Smock et al., 2016) adopted by a vast family of scaffolds involved in macromolecular interactions and catalysis. β-Propeller domains are found in all kingdoms of life (Pereira and Lupas, 2021) and are involved in a wide range of biological processes (Fülöp and Jones, 1999; Pons et al., 2003; Guruprasad and Dhamayanthi, 2004; Chen et al., 2011). Their repetitive unit, the 'blade', is an ancestral 4-stranded antiparallel β-meander of typically 40–50 residues (Alva et al., 2015), which is radially arranged around a central channel to form a toroidal fold, the 'propeller' (Andrade et al., 2001; Söding and Lupas, 2003). Currently, more than 300 structures have been experimentally obtained for a variety of β-propellers ranging from four up to 12 blades.

Despite the high structural similarity of their blades, β-propellers span a wide range of internal sequence symmetry, from near identity (>90% sequence identity) to almost full differentiation (<20% sequence identity) (Pereira and Lupas, 2021), a unique feature and a crucial factor for their wide functional diversification. The most symmetric β-propeller known to date is that of the WRAP domain in Npun_R6612 from the cyanobacterium Nostoc punctiforme PCC73102, with only 10 point mutations across 14 predicted blades (Chaudhuri et al., 2008; Dunin-Horkawicz et al., 2014). The near identity of its repetitive units and the absence of synonymous mutations in their encoding DNA suggests that this β-propeller was amplified very recently. The 14 blades fold into two separate 7-bladed β-propellers, but are able to produce a multitude of fold topologies when amplified to different copy numbers in the same chain (Afanasieva et al., 2019).

WRAP forms the C-terminal part of a protein with a STAND ATPase domain, whose gene is transcriptionally coupled to another STAND ATPase carrying an N-terminal Toll/Interleukin receptor (TIR) domain (Dunin-Horkawicz et al., 2014). This operon-like structure is conserved in other cyanobacteria, where the TIR-containing protein may also carry an additional N-terminal caspase domain and thus combines three core domains of eukaryotic innate immunity and apoptosis. The further discovery of similar proteins with recently amplified or disrupted (by frameshifts or in-frame stop codons) β-propellers, suggested a rapid turnover in the genes coding for these proteins likely due to a role in prokaryotic innate immunity (Dunin-Horkawicz et al., 2014).

WRAP β-propellers belong to the WD40 superfamily, characterised by blades of approximately 40 residues length, carrying a conserved Trp-Asp (WD) sequence motif. This constitutes one of the largest β-propeller superfamilies and corresponds to one of the three major hubs of the β-propeller sequence classification landscape (Chaudhuri et al., 2008; Kopec and Lupas, 2013; Pereira and Lupas, 2021). The other two are the Asp-Box, characterised by a SxDxGxTW motif (Quistgaard and Thirup, 2009), and the VCBS β-propellers, which contain a conserved cation binding DxDxDG motif and include the β-propellers in α-integrins, various archaeal toxins, and multiple eukaryotic and bacterial lectins (Rigden and Galperin, 2004; Rigden et al., 2011; Makarova et al., 2019; Pereira and Lupas, 2021). More than 60 unique β-propeller families are classified in ECOD (Cheng et al., 2014), most belong to one of these hubs and some form independent clusters that connect to them (Chaudhuri et al., 2008; Kopec and Lupas, 2013; Pereira and Lupas, 2021). Within these, high sequence and structure symmetry has been also reported for the β-propeller domain in tachylectin-2, a family of VCBS-like eukaryotic and bacterial lectins involved in the innate immune response of arthropods and extensively explored for protein design and the study of β-propeller folding dynamics (Beisel et al., 1999; Yadid and Tawfik, 2007, 2011; Vrancken et al., 2020; Pereira and Lupas, 2021).

This led us to wonder about the extent of highly repetitive, recently amplified β-propellers in nature. Are WRAP blades especially prone to amplification or is this a more general, inherent property of β-propeller blades? In order to answer this question, we carried out a large-scale search for recently amplified β-propellers belonging to any of the families classified in ECOD.

MATERIALS AND METHODS

A summary of the workflow carried out to search, classify, and analyse highly repetitive, putative β-propellers in proteomes across all kingdoms of life is depicted in **Supplementary Figure S1**.

1 Seed Preparation

The searches for highly repetitive β-propellers were carried based on the sequences of individual blades from β-propellers and β-propeller-like folds of known structure. For that, we collected 426 structures of β-propeller-like, β-prisms and β-pinwheels domains from the ECOD database filtered to 70% sequence identity (ECOD version develop261) (Cheng et al., 2014). Their blades were extracted at the structure level by automated detection of internal structure symmetry and structural repeats with CE-Symm 2.1 (Bliven et al., 2019). The number of repeat units detected was then compared with the topology annotated in the ECOD database and manually curated.

tThese β-propeller-like domains were further classified at the sequence level with CLANS (Frickey and Lupas, 2004), which demonstrated that the set collected is overpopulated by the WD40 superfamily. To unskew our dataset and avoid a large number of redundant searches, clusters at a p-value of 10^{-4} were automatically detected using the linkage clustering method implemented in CLANS and, for each cluster, a maximum of three representatives selected. For that, and for each cluster, their constituent sequences were sorted in descending order of the number of blades and their total coverage of the full domain, and the top ranking three β-propellers selected. If a cluster was composed of less than three sequences, the full cluster was considered, and if individual sequences did not connect to any other cluster, they were considered a "singleton cluster" and the sequences of their blades saved too. This resulted in 103 clusters, corresponding to an unskewed set of 168 β-propeller-like domains and a total of 1,071 blade-like initial sequence seeds.

2 Sequence Searches

Sequence searches were carried out in parallel for each cluster and individually for each kingdom of life (bacteria, archaea, eukaryotes and virus). By doing so, we avoided that families that were over-represented in a given kingdom (e.g., VCBS sequences in bacteria and WD40 in eukaryotes) took over the PSSMs built and, thus, bias the searches towards that taxonomic group and that family, hindering the detection of under-represented groups. Each search step was composed of three stages: 1) PsiBlast searches for blade-like matches in a sequence database, 2) HMM searches for missed, degenerate blade-like sequences, and 3) selection of continuous, highly repetitive β-propeller-like sequence segments.

In stage 1), the corresponding blade seed sequences of the target cluster were searched with 15 rounds of PsiBlast (Altschul, 1997) against the set of sequences of a given kingdom of life in the NCBI non-redundant sequence database filtered to a maximum sequence identity of 90% (the nr_bac90, nr_arc90, nr_euk90 or nr_vir90 databases, respectively) (Zimmermann et al., 2018). Only matches at an E-value better than 10^{-4} were considered for PSSM building, and all matches at an E-value better than 1.0 and that covered at least 80% of the query sequence collected as putative blade matches.

At the end of stage 1), the matches collected over the corresponding sequence database were used to build an HMM sequence profile, which was used in stage 2) to search for missed blade-like sequences in unusually long matches, in long linkers between two consecutive matches in the same sequence, and N- and C-termini sequence segments. For that, only those matches whose sequence length was at maximum one median absolute deviation (MADe) away from the median sequence length in the corresponding set of matches were aligned with MUSCLE (Edgar, 2004) and the resulting alignment trimmed with TrimAl (Capella-Gutierrez et al., 2009), removing columns where >70% of the positions were a gap (gap score of 0.30) and sequences that only overlapped with less than 75% of the columns populated by 80% or more of the other sequences. The trimmed alignment was then used to build an HMM profile with HHmake (Söding, 2005), against which all the long matches that were not used to build it were searched. This allowed the detection and further break down of matches that included more than one blade. After that, the set of blades was updated and the HMM rebuilt (following the same procedure) to account for the newly detected matches. Long linkers between two consecutive blade-like matches and with a size compatible with at least one blade were then searched against the updated HMM, the set of blades updated, and the procedure repeated for the N-termini and C-termini of the parental full-length sequence.

In stage 3), the individual blade-like sequence matches detected were grouped into β-propeller-like domains based on the length of the linkers between them. For that, all linkers between two consecutive blade-like matches in the corresponding set of sequences were collected and the median and MADe of their lengths computed. With this, two consecutive blade-like matches were considered to belong to the same β-propeller if the linker between them was less than three MADe away from the median linker length. For each putative β-propeller domain with at least two consecutive blade matches, their constituent blades were aligned with MUSCLE and the median sequence identity between individual pairs of blades computed. Only those with a median blade sequence identity higher than 60%, totalling 14,176 preliminary matches across 11,807 protein sequences, were considered for further analysis.

3 Validation of Detected Highly Repetitive β-Propeller-Like Sequences

In order to validate the boundaries of the blades and make sure that no blade was left behind after stage 3), we further validated our β-propellers with HHrepID (Biegert and Söding, 2008). For that, the full-length sequences encompassing the selected β-propellers were analysed, searching for significant repeat units at a *p*-value of 10^{-1}. For a given target full-length sequence, if the repeats detected with HHrepID covered more of it than the putative blades already annotated in the previous steps, this region was further compared with the previous annotations.

As the host full-length protein sequence may contain other repeat patterns in addition to those in β-propeller regions, and HHrepID is able to differentiate between different types of repeat units, only those repeat types in regions matching already annotated blades were considered. However, even if a given repeat type can be assigned to correspond to a β-propeller region, we cannot assume that the repeat unit corresponds to a blade; it can also correspond to a larger unit or even to a full-length β-propeller depending on the evolutionary history of the target protein. Thus, and to detect such cases, the number of blades within a given repeat unit was always predicted based on the median length of the blades of the corresponding annotated β-propeller. If a given repeat unit could accommodate at least two blades, it was further analysed with HHrepID iteratively until no more repeats with a blade-like size were detected. On the other hand, if a repeat unit could accommodate one blade only, it was considered a putative blade.

The intervals of putative blade-like repeat units and sequence-based blade like matches were compared and merged to maximize the length of the detected β-propeller. If a given repeat type annotated with HHrepID encompassed more than one of the β-propellers annotated before, these were merged into one β-propeller, whose boundaries were defined to maximise the coverage of the full-length protein. On the other hand, if only one β-propeller would fit into the repeated region, the interval with the largest number of repeats/blades was considered the β-propeller. This procedure allowed for the detection of further degenerate blades and the merging of seemingly short β-propeller-like segments, into longer putative β-propeller domains. These β-propellers were again analysed for the median sequence identity of their individual blades, resulting in 10,480 putative highly symmetric β-propeller domain sequences with a median blade sequence identity higher than 60%, from a total of 9,919 protein sequences.

4 Sequence Annotation

The putative highly-repetitive β-propeller domains identified were assigned an ECOD family with HHsearch (Söding, 2005). For that, each β-propeller sequence was searched against a database of HMM profiles built for the ECOD database filtered to 70% maximum sequence identity [the HHpred ECOD70 database as of November 2021 (Zimmermann et al., 2018)] and assigned the best match at a probability better than 50%.

To analyse the domain environments of each β-propeller, i.e., annotate the domain composition of the proteins bearing the highly repetitive β-propellers to classify them into "globally" or "locally" repetitive, full-length sequences were annotated similarly (using the HHpred database as of May 2020), but iteratively, so that sequence regions not yet mapped to a

domain in the previous iteration were searched individually again. A maximum of five iterations were carried out and only the best matches outside already annotated β-propeller regions at a probability better than 50%, larger than 40 residues and that covered at least 30% of the target profile considered.

5 Three-Dimensional Structure Prediction and Model Quality Estimation

Three-dimensional structural models of selected repetitive β-propeller examples were produced with AlphaFold v2.1.2 monomer (Jumper et al., 2021). Modelling was carried out with af2@scicore v2.2.0 (https://git.scicore.unibas.ch/schwede/af2-at-scicore) using default settings. For each example, five models were generated. Model quality was estimated based on the predicted LDDT scores reported by AlphaFold but also independently, with QMEANDisCo, through the SWISS-MODEL "Structure Assessment" tool (Waterhouse et al., 2018; Studer et al., 2020).

6 Identification of Non-Coding β-Propeller Fragments

Given that most of the highly repetitive β-propeller regions were found at protein extremities (N- or C-terminal regions), we analysed the immediately adjacent non-coding genomic regions that flank their corresponding genes for nucleotide segments with β-propeller-encoding potential. For that, the genome assembly corresponding to each protein was mapped using Biopython. Entrez (Cock et al., 2009) by first querying for each unique EntrezID the corresponding NCBI Identical Protein Group (IPG), collecting the accession codes of all genome assemblies matched and then filtering out repeated assemblies (i.e., GCA assemblies for which there is an identical, GCF, entry on RefSeq).

For all protein coding regions mapped, the nucleotide sequence of the immediate 5′ and 3' non-coding intergenic regions (i.e., the genomic regions between the current open reading frame (ORF) and the two closest, flanking ORFs excluding pseudogenes, in the same direction) were collected and translated in all three frames possible in the same direction of the target ORF. Stops resulting from stop codons were replaced by X, emulating an unknown amino acid. These protein-like sequences were then iteratively annotated against the ECOD70 database using the same approach described above; for this case, a maximum of five iterations were carried out and only the best matches to β-propeller-like folds at a probability better than 1%, larger than 20 residues and with a maximum 15% content of stop codons were considered.

RESULTS

We started with the blades of 168 non-redundant β-propellers and β-propeller-like domains of known structure, and used them as seeds for deep sequence similarity searches in the non-redundant sequence database at NCBI (nr_bac, nr_arc, nr_euk and nr_vir; **Supplementary Figure S1**) (Zimmermann et al., 2018). This initial set comprised not only the β-propeller domains classified in ECOD but also β-pinwheels and type II β-prisms, which are repetitive, symmetric folds based on the repetition of a 4-stranded blade-like unit distantly related to β-propeller blades (Kopec and Lupas, 2013). β-pinwheels and type II β-prisms are represented by the DNA-binding domains of bacterial type IIA topoisomerases (Corbett et al., 2004) and Galanthus nivalis agglutinin-related lectins (GNA-related lectins) (Hester and Wright, 1996; Kaus et al., 2019), respectively.

In our searches, we specifically probed for those sequences with at least two consecutive non-overlapping matches to the seed blade and a median internal blade-to-blade sequence identity higher than 60%. We selected this threshold as, apart from WRAP, no natural β-propeller was known with such a high internal sequence symmetry (Pereira and Lupas, 2021). Our approach resulted in 10,474 putative β-propeller domains across 9,919 host protein sequences from all kingdoms of life.

1 β-Propeller Families With Highly Repetitive β-Propellers

HHsearches against the ECOD database allowed us to map each of these domains to 32 out of 65 unique ECOD β-propeller families (**Figure 1A**). While a 50% probability threshold was used for annotation, most matches have a probability better than 90% (**Figure 1B**). Most of these correspond to those of β-propeller domains involved in macromolecular interactions, but mappings to at least eight commonly associated with a catalytic activity were also found. Still, only 17 sequences did not find significant matches to any family and represent remote homologs of ECOD β-propeller families that fall below the probability threshold (UNKNOWN label), while three found their best match in a non-propeller fold (NO_PROPELLER label); these 3 cases were excluded from further analysis.

The sequences in our dataset span a wide range of median sequence identities of their blades, reaching up to full identity in half the lineages of the dataset (**Supplementary Figure S2A**).

Half of the sequences collected are of the WD40 superfamily, to which WRAP belongs. The WRAP domain in Npun_R6612 of Nostoc punctiforme PCC 73102 was the β-propeller with the highest internal sequence symmetry we had found up to that point and contains no synonymous substitutions at the DNA level (**Figure 2**), pointing to a very recent amplification event (Dunin-Horkawicz et al., 2014). Keeping in mind our hypothesis that new β-propellers continuously arise through cycles of amplification and differentiation, we attempted to track the process of differentiation in this very recently arisen β-propeller through the emergence of synonymous mutations in close homologs. Indeed, these homologs show a proportion of synonymous mutations, along with a higher number of non-synonymous ones, but to our surprise appeared to have resulted from independent amplification events, as judged by the presence of positions at which all the blades of one β-propeller contained a different nucleotide than all the others (**Figure 2**, **Supplementary Figure S3**). Thus, even at pairwise sequence identities above 80%, ungapped, these β-propellers are only homologous at the level of

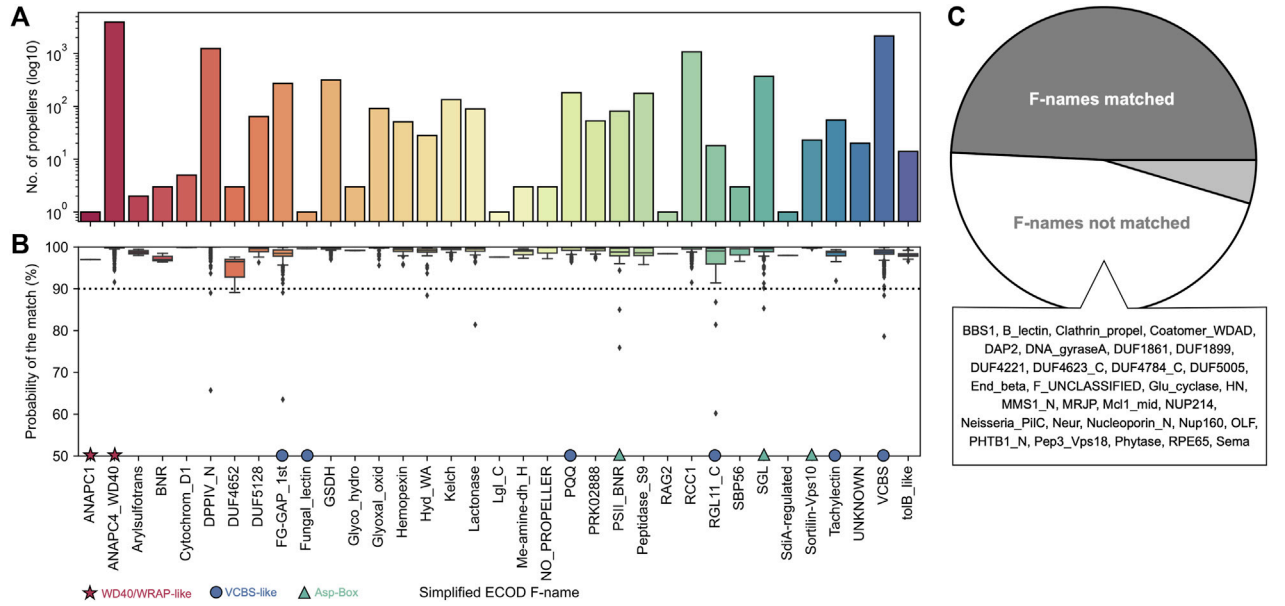

FIGURE 1 | The ECOD families matched. **(A)** Frequency of the different ECOD F-names (in their simplified form) matched. **(B)** Boxplots of the HHsearch probability of the match, per family. The most abundant families belonging to one of the three main groups of β-propellers, as described in the text, are marked: stars mark WD40, WRAP-like propellers, circles mark VCBS and VCBS-like (including PQQ) propellers, and triangles the three main families making the Asp-Box group. **(C)** Proportion of ECOD F-names mapped, with a list of those without any match to β-propeller-like sequences in our dataset. The UNKNOWN family corresponds to the sequences without significant matches to any of the 65 ECOD F-names.

the blade that had been amplified, but analogous in the fully amplified form.

The VCBS superfamily is the second most sampled, encompassing about 20% of the β-propeller-like sequences in the dataset. We previously demonstrated that VCBS β-propellers form a large and diverse hub in the β-propeller classification landscape that is tightly connected to the FG-GAP/α-integrin, tachylectin-2 and Family 11 rhamnogalacturonan lyase families and is bridged to the WD40 supercluster by the PQQ family (Pereira and Lupas, 2021). In line with this, we also found highly repetitive β-propeller-like sequences assigned to the FG-GAP/α-integrin (2%), tachylectin-2 (0.7%), Family 11 rhamnogalacturonan lyase (RGL11, 1%) and PQQ (1.7%) families. While VCBS-like β-propellers are typically associated with the recognition of different carbohydrates (but also protein-protein interactions) (Pereira and Lupas, 2021), RGL11 β-propellers are catalytic and involved in the cleavage of the rhamnogalacturonan type-I region of plant cell wall pectin (Ochiai et al., 2007, 2009), and PQQ β-propellers in protein binding in various biological processes, including outer membrane protein assembly, alcohol metabolism and unfolded protein response (Ghosh et al., 1995; Kim and Paetzel, 2011; Kopec and Lupas, 2013). In contrast to recently amplified propellers in the WD40 superfamily, where all blades of one domain appear to have been amplified in one event, repetitive VCBS propellers may show a further, protein-specific amplification of an internal blade, leading to a diversity of blade numbers within the same lineage, even at pairwise sequence identity of >80% between individual members. An example of this is found in a group of closely related cyanobacterial chitinases, in whose N-terminal propeller domains we observe forms with 3, 4, 6, and 9 blades (**Supplementary Figure S4**). Particularly conspicuous are two paralogs of the protein encoded in the genome of Mastigocoleus testarum BC008, one with 6 blades and the other with 9, which are otherwise 100% sequence identical.

The third most populated group, making up 10% of the set, is RCC1/BLIPII. Well-known members of this group are the β-propeller domains in the eukaryotic Regulator of Chromosome Condensation 1 (RCC1) protein, a nuclear interactor involved in the regulation of spindle formation and nuclear assembly during mitosis (Hadjebi et al., 2008; Stevens and Paoli, 2008), and the bacterial β-lactamase inhibitor protein II (BLIP-II), a secreted binder produced by soil bacteria as a potent β-lactamase inhibitor (Park and Lee, 1998; Brown et al., 2013). In our dataset, members of this group are mostly of bacterial origin, but a few eukaryotic and viral members were also collected (**Figure 3B**).

The fourth largest set of sequences maps to the Asp-Box superfamily (e.g., Sortillin-, PSII- and SGL-like β-propellers), making up to 5% of the dataset. The remaining sequences belong to, among others, 1) the DPPIV and Peptidase S9 families (Pfam: PF00930 and PF02897), found in multiple bacterial and eukaryotic serine, prolyl oligopeptidases where the β-propeller domain controls the access to the active site (Fülöp et al., 1998; Polgár, 2002; Hiramatsu et al., 2003), 2) the Hemopexin family (Pfam: PF00045), which encompasses the domains in various exported eukaryotic proteins involved in haem transport and protein-protein interactions (Das et al.,

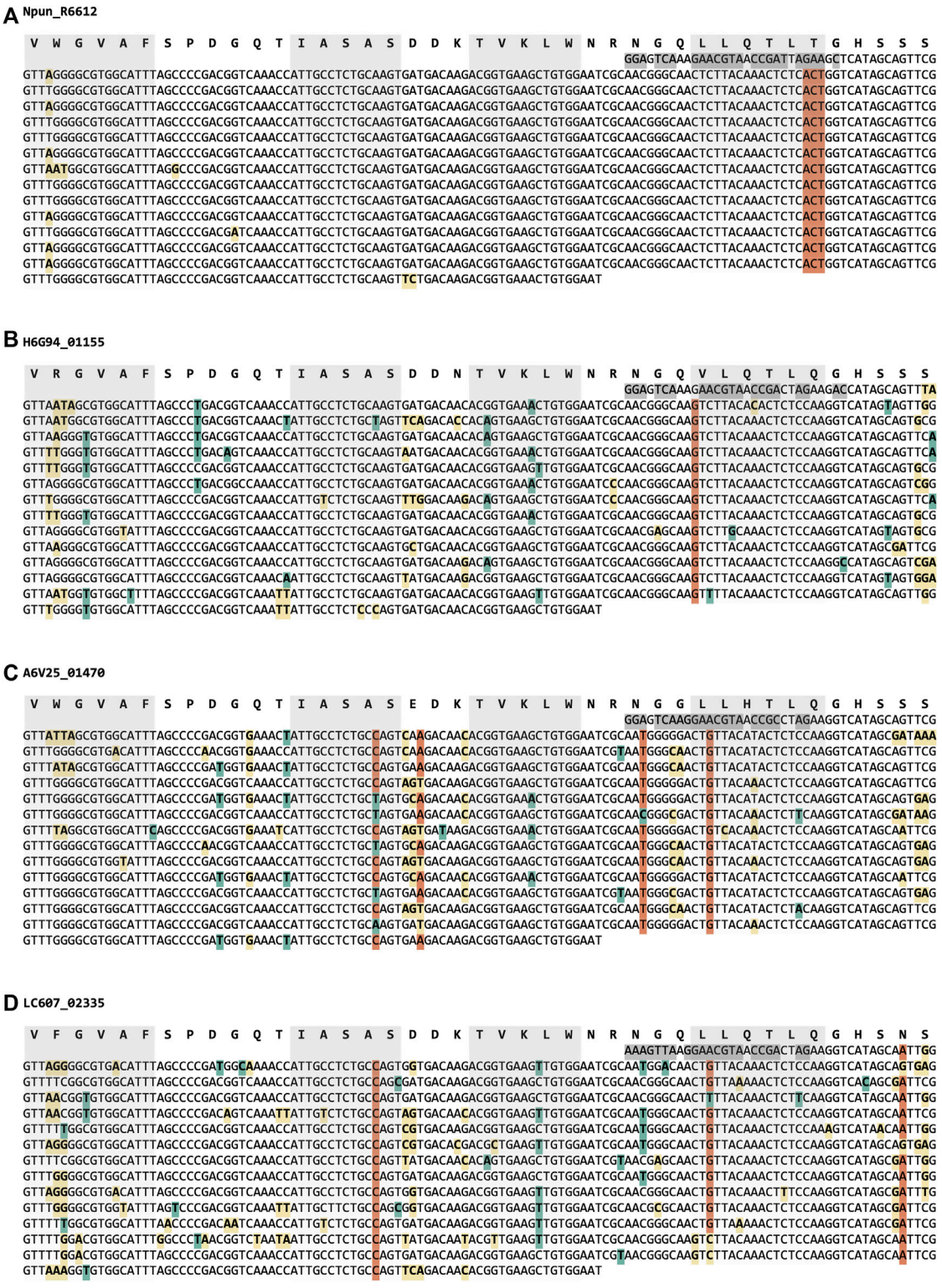

FIGURE 2 | Synonymous and non-synonymous mutations in the nucleotide sequence of four closely related WRAP domains (≥80% pairwise sequence identity, **Supplementary Figure S3**). Synonymous mutations relative to the majority rule consensus of the respective β-propeller are highlighted in teal, and non-synonymous ones in yellow. Positions where a nucleotide is present in at least two thirds of the blades of one β-propeller, but different from the equivalent nucleotides in at least two of

(*Continued*)

FIGURE 2 | the other β-propellers are highlighted in orange. The nucleotide sequence encoding the first β-strand of the structure is different from that of the equivalent strand in the amplified region and highly conserved between homologs; it is highlighted in dark grey. The individual strands in each β-propeller blade (based on the experimental structure of the Npun_R6612 WRAP domain, Afanasieva et al., 2019) are highlighted in grey. **(A)** Npun_R6612 of Nostoc punctiforme PCC 73102 (ACC84870.1), **(B)** H6G94_01,155 of Nostoc punctiforme FACHB-252 (MBD2609895.1), **(C)** A6V25_01,470 of Nostoc sp. ATCC 53789 (RCJ36011.1), and **(D)** LC607_02,335 of Nostoc sp. CHAB 5824 (MCC5641816.1).

2003), 3) the Glucose/Sorbosone dehydrogenase (GSDH) family (Pfam: PF07995), which are bacterial pyrroloquinoline quinone (PQQ)-dependent enzymes (Oubrie et al., 1999), 4) two distinct Lactonase-like families (Pfam: PF10282), that bind the co-factors of various enzymes involved in the breakage of lactone rings (Amara et al., 2011) and include the protein-binding domain of the archaeal Lp49 surface antigen (Giuseppe et al., 2008), and 5) the Kelch superfamily (Pfam: PF01344), whose members are functionally diverse and may be involved in protein binding (Gupta and Beggs, 2014).

On the other hand, and although they were included in the initial set of seed sequences, no highly repetitive members were found for families with a β-prism or a β-pinwheel fold, nor for members of the Ph1500 family, which form, by oligomerization, the largest, 12-bladed β-propellers known to date (Varnay, 2009).

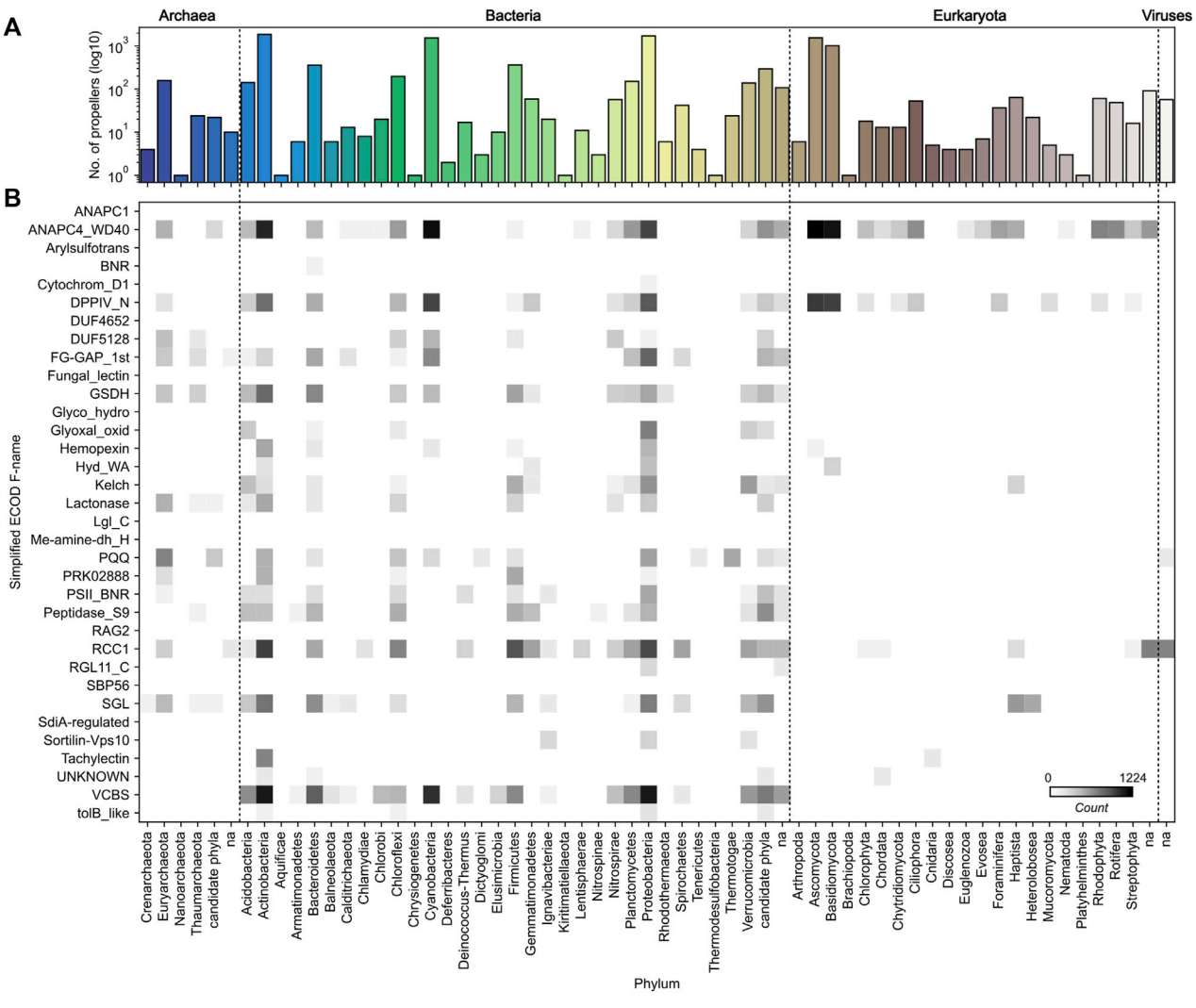

FIGURE 3 | Taxonomic distribution of highly repetitive β-propeller sequences. **(A)** The frequency, in logarithmic scale, of highly repetitive β-propeller sequences from different phyla in the collected dataset. **(B)** Heatmap of the frequency of highly repetitive β-propeller sequences of different ECOD families in different phyla. Colors are in logarithmic scale.

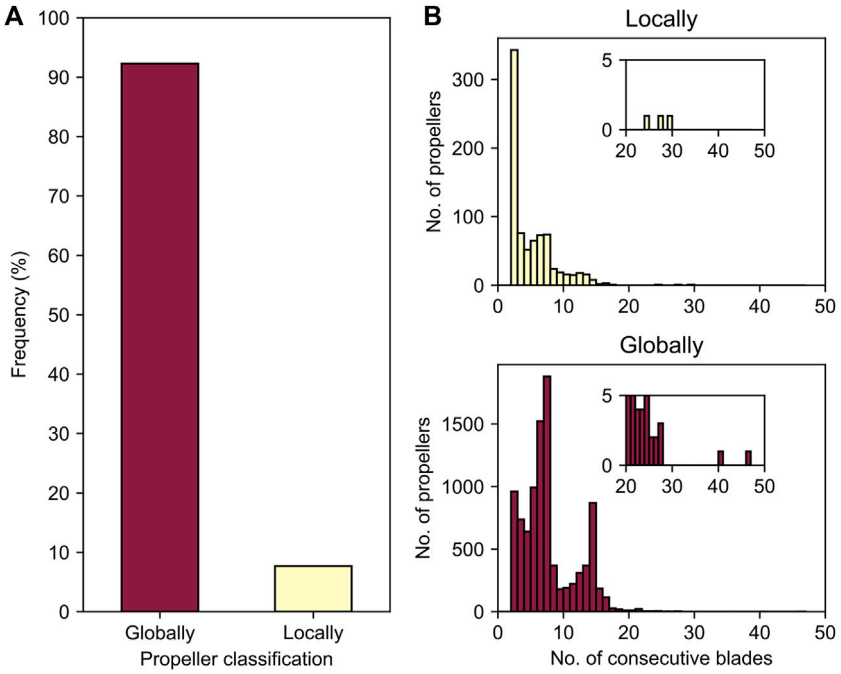

FIGURE 4 | Globally versus locally highly repetitive β-propellers. **(A)** The proportion of β-propellers in the dataset categorised into "globally" or "locally" based on the absence or presence of non-repetitive β-propeller matches in their immediate domain environment. **(B)** Histograms for the absolute number of consecutive highly identical blades, separated by highly repetitive β-propeller type.

2 Taxonomic Distribution of Proteins With Highly Repetitive β-Propellers

Within the β-propeller-like sequences in our dataset, we found representatives from all kingdoms of life, including viruses (**Figure 3**). While examples for most phyla could be found, the most prevalent are Actinobacteria, Cyanobacteria, Proteobacteria, Ascomycota and Basidiomycota, followed by Bacteroidetes, Chloroflexi, Firmicutes and Euryarchaeota (**Figure 3A**). Indeed, ~70% of the β-propellers collected (and the same proportion of full-length host protein sequences) belong to bacterial species, and it is in bacteria that the largest variety is found, followed by archaea, eukaryotes, and viruses (**Figure 3B**). Sequences from Ascomycota and Basidiomycota, two phyla of fungi, make ~25% of the dataset and fully 84% of the eukaryotic sequences.

When looking at the individual taxonomic distribution of specific families, there are families that are widespread while others have highly repetitive members in just a few taxonomic groups (**Figure 3B**). For example, among highly repetitive β-propellers, WD40 members are widespread, but especially prevalent in fungi, Actinobacteria, Cyanobacteria and Proteobacteria, VCBS members are only found in bacteria, FG-GAP/α-integrin members are found in both archaea and bacteria, and tachylectin-like members are unique to Actinobacteria and, surprisingly, Cnidaria.

3 The Diverse Blade Numbers of Highly Repetitive β-Propellers

Sequences in our dataset may contain as few as two or as many as 47 highly similar consecutive blades (**Figure 4**). Their number depends primarily on the β-propeller family (**Supplementary Figures S2B, S4**). For example, highly repetitive WD40-like β-propellers have a median number of seven blades, while VCBS-like have 6 and DPPIV-like 13. Looking at the distribution of blade numbers in individual lineages, we observe that there is a continuum of blade numbers, which are however biased towards the median number in their family. For example, for β-propellers whose best match in ECOD is a WD40 β-propeller, the histogram of the number of consecutive highly identical blades has peaks at 7 and 14, with a continuum of less frequent intermediate blade numbers.

However, HHsearches with the full-length sequences of all 9,919 proteins containing highly repetitive β-propellers indicated that in 8% of the cases these regions have adjacent, non-highly repetitive β-propeller matches, suggesting they correspond to parts of larger, degenerate β-propellers in which only some consecutive blades are highly similar. Such cases we denote as "locally repetitive β-propellers". The remaining 92% correspond to continuous regions of consecutively amplified, highly similar blades. These correspond to "globally repetitive β-propellers" (**Figure 4A**). There are no family or taxonomic differences between globally and locally repetitive β-propellers, yet their preferences for the number of highly similar blades differ (**Figure 4B**). The labels "locally" and "globally repetitive" are not always mutually exclusive, as seen for the aforementioned VCBS-like propellers (**Supplementary Figure S4**), where a recent local amplification led to internal blades of up to 100% sequence identity, embedded within a global amplification of blades with >60% sequence identity, which gave rise to the domain.

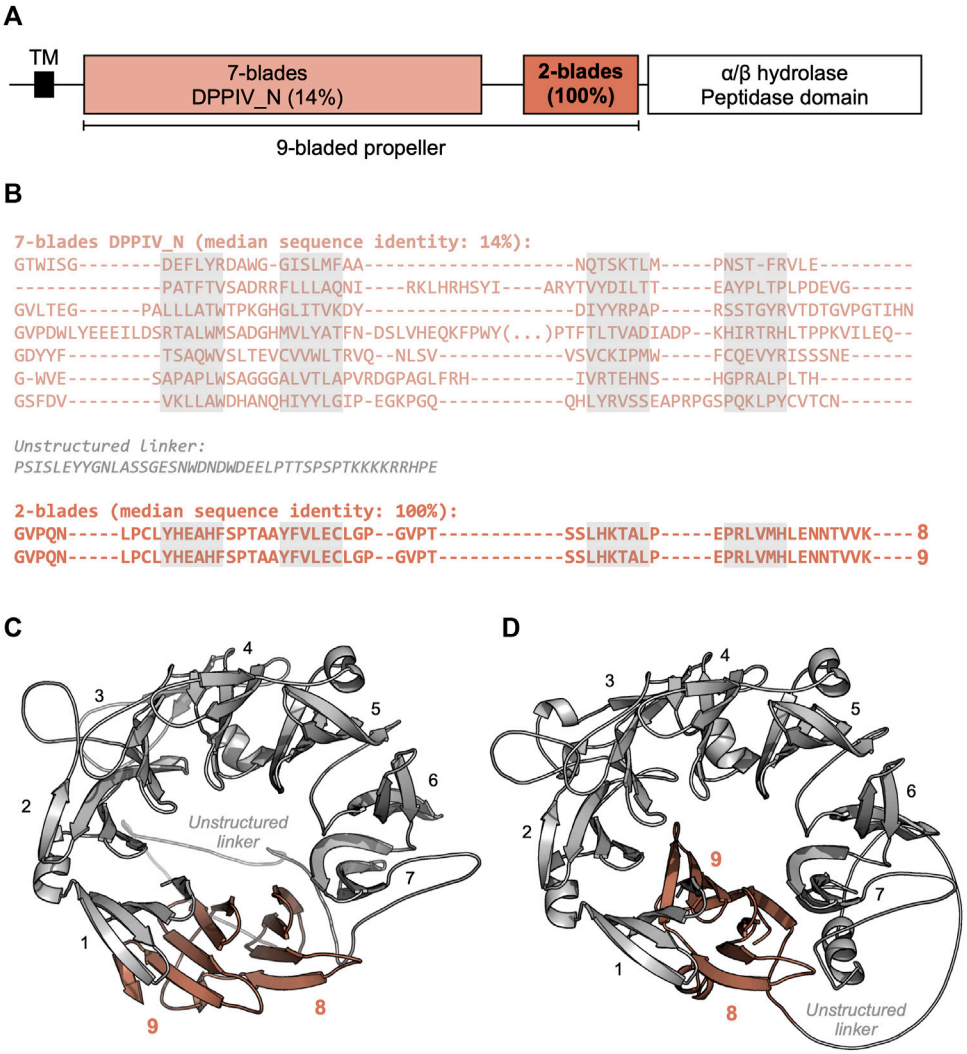

FIGURE 5 | The locally highly repetitive β-propeller of *Bombyx* mandarina inactive dipeptidyl peptidase 10 (DPPY) (XP_028029463.1). **(A)** Predicted domain composition of the full-length DPPY sequence. **(B)** Annotated β-propeller sequence, highlighting the two highly identical, consecutive blades. β-propeller blade sequences were aligned with Promals3D (Pei and Grishin, 2014) based on their predicted structure. The individual strands in each β-propeller blade are highlighted in grey. **(C)** Three-dimensional model of the β-propeller region, predicted with ColabFold MMseqs protocol as of August 2021 (Mirdita et al., 2021) (QMEANDisCo Global (Studer et al., 2020) score of 0.59 ± 0.05, average pLDDT (Jumper et al., 2021) of 82.9). **(D)** Three-dimensional model of the β-propeller region, predicted with AlphaFold v2.1.2 (Jumper et al., 2021) (QMEANDisCo Global (Studer et al., 2020) score of 0.62 ± 0.05, average pLDDT (Jumper et al., 2021) of 85.3).

3.1 The Blade Numbers of Locally Repetitive β-Propellers

The most frequent number of highly similar blades in locally repetitive β-propellers is 2, but, in some instances, it may be as high as 29 (**Figure 4**). A representative example for a locally repetitive β-propeller is the β-propeller domain of the inactive dipeptidyl peptidase 10 (DPPY) of a wild silk moth (*Bombyx* mandarina) (XP_028029463.1, **Figure 5**). This 929-residue membrane protein contains seven divergent β-propeller blades of the DPPIV family with a median sequence identity of 14%, followed by two blades with a pairwise sequence identity of 100% (**Figure 5A**) that resulted from the amplification of a single exon. While DPPY β-propellers usually adopt an 8-bladed fold (Bezerra et al., 2015), we obtained two possible models for the 9 blades of the *Bombyx* protein (**Figures 5C,D**). While an early version of ColabFold (Mirdita et al., 2021) produced a 9-bladed β-propeller (**Figure 5C**), recent versions of ColabFold and AlphaFold (Jumper et al., 2021) (as of February 2022) always produced 8-bladed β-propellers, where the first identical blade is part of the domain and the second identical blade is modelled as a single, unpacked blade (**Figure 5D**). As eight- and 9-bladed propellers exist for DPPIV propellers of known structure, and no striking differences are observed in the interfaces between blade one and blades 9 and 8 as well as in their quality estimation metrics, it is not clear which of these two is the best model.

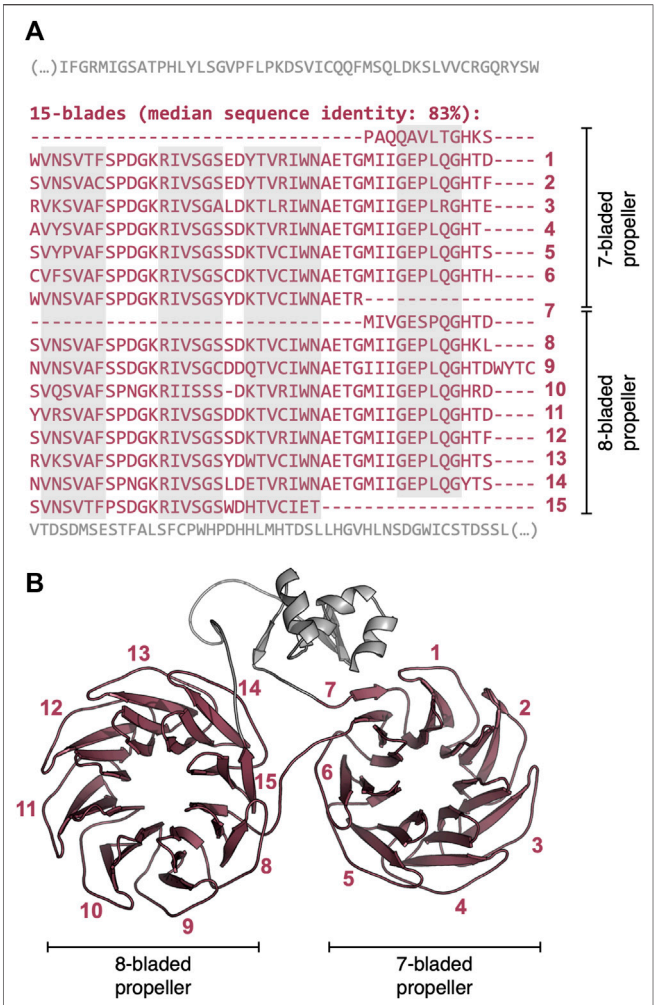

FIGURE 6 | The globally highly repetitive β-propeller of WD40 repeat-like protein KAE9403162.1 from Gymnopus androsaceus JB14. **(A)** Annotated β-propeller sequence, highlighting the 15 nearly identical, consecutive blades and the flanking non-propeller sequences. The individual strands in each β-propeller blade are highlighted in grey. **(B)** Three-dimensional model of the β-propeller and immediate flanking regions predicted with AlphaFold v2.1.2 (Jumper et al., 2021) (QMEANDisCo Global (Studer et al., 2020) score of 0.65 ± 0.05, average pLDDT (Jumper et al., 2021) of 88.9). Modelling was carried out with the full-length protein sequence but only the region whose sequence is shown in **(A)** is depicted.

3.2 The Blade Numbers of Globally Repetitive β-Propellers

In contrast to the locally repetitive β-propellers, globally repetitive ones strongly prefer larger numbers of blades (**Figure 4B**, **Supplementary Figure S4**). In fact, they often have a greater number of blades than the median number for their lineage. For example, the β-propeller region in the WD40 repeat-like protein KAE9403162.1 from Gymnopus androsaceus JB14 corresponds to 15 WD40-like consecutive blades with a median sequence identity of 83% (**Figure 6A**). AlphaFold predicts, at high confidence (average pLDDT of 89%), that this region folds into two individual β-propeller domains, one with seven blades and the other with 8 (**Figure 6B**).

However, modelling globally repetitive β-propellers is not always so straightforward. For example, the highest number of blades in our set is 47, found in a hypothetical NACHT ATPase from Penicillium camemberti (CRL31138.1) (**Supplementary Figure S6A**). The 47 blades are compatible with multiple combinations of 6 to 9-bladed β-propellers, but no high-quality three-dimensional model could be obtained through AlphaFold (**Supplementary Figure S6B**). In all five models generated, β-propeller silhouettes can be seen but they all collapse into each other. A reason may lie in their extreme level of sequence symmetry: while the median blade identity is 81%, consecutive fragments of two blades have a median sequence identity of 100% at both the protein and the nucleic acid levels. The unit of repetition is thus 2-bladed and the sequence is extremely symmetric (**Supplementary Figure S6C**), representing the only case of amplification from a 2-bladed fragment we have observed so far. This makes it difficult to generate meaningful multiple sequence alignments for protein structure prediction based on co-evolutionary potentials and impossible, so far, to predict how many β-propellers it may be able to fold into. Its complete internal sequence identity, even at the nucleotide level, also makes it difficult to judge whether there are indeed 47 blades given the problems of assembling different fragments with identical overlaps. Even though not as pronounced, this may be a recurrent problem in our highly repetitive β-propellers.

We also found globally repetitive β-propellers with fewer blades than the median number for their lineage and interpret these as fragments of full domains (**Supplementary Figure S5**). One example is provided by the two blades found in the hypothetical protein DL7770_005,219 from Menosporascus sp. CRB-9-2 (**Figure 7**). This protein carries two WD40-like tandem blades with a sequence identity of 92%, which are flanked by an N-terminal α-helical region and a C-terminal region rich in β-strands. We obtained different models for this part of the protein (**Figures 7B–D**) and, in all cases, the two highly similar blades are predicted as expected, but in each model the C-terminal region appears to provide additional blade-like supersecondary structures. Indeed, HHsearches with just the C-terminal region showed significant matches to segments of three consecutive WD40-like blades. This suggests that this region may correspond to a disrupted β-propeller that is quickly diverging, similarly to those described in a previous work for WRAP homologs (Dunin-Horkawicz et al., 2014).

4 The Evidence for Disrupted β-Propellers

While it is possible that globally repetitive β-propeller fragments may occur within a non-propeller context, most of the sequences we found are located less than 50 residues away from one of the termini of their full-length protein, more often the C- than the N-terminus (**Figure 8**), with 781 located less than 30 residues away from both. Given earlier evidence of highly repetitive WRAP-like β-propellers disrupted by either frameshifts or in-frame stop codons, and the overall continuous distribution of blade numbers in our dataset, we therefore searched for putative β-propeller blades encoded in the 5′ or 3’ regions as evidence for disrupted genes.

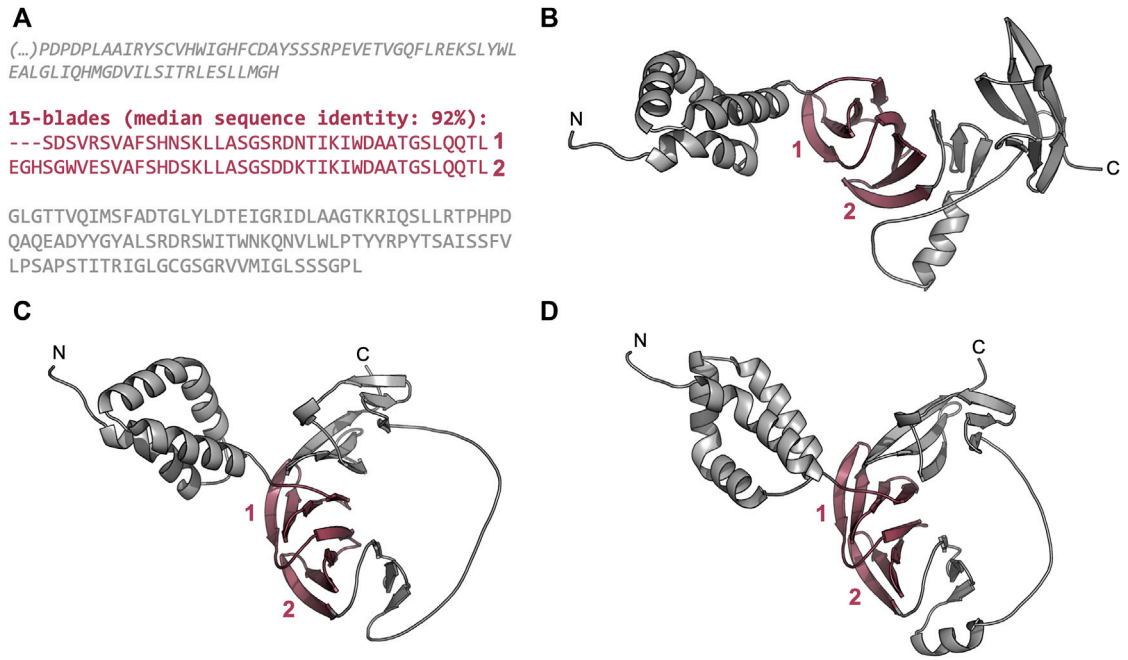

FIGURE 7 | The globally highly repetitive β-propeller of Menosporascus sp. CRB-9-2 hypothetical protein DL7770_005,219 (RYP84315.1). **(A)** Annotated β-propeller sequence, highlighting the two highly identical, consecutive blades and the flanking non-propeller sequences. The individual strands in each β-propeller blade are highlighted in grey. **(B–D)** Three-dimensional models of the β-propeller and immediate flanking regions predicted with AlphaFold v2.1.2 (Jumper et al., 2021). Modelling was carried out with the full-length protein sequence but only the region whose sequence is shown in **(A)** are depicted. **(B)** The top ranking model (QMEANDisCo Global (Studer et al., 2020) score of 0.57 ± 0.05, average pLDDT (Jumper et al., 2021) of 80.1), **(C)** the second best model (QMEANDisCo Global score of 0.58 ± 0.05, average pLDDT of 75.9), and **(D)** the third best model (QMEANDisCo Global score of 0.58 ± 0.05, average pLDDT of 71.0).

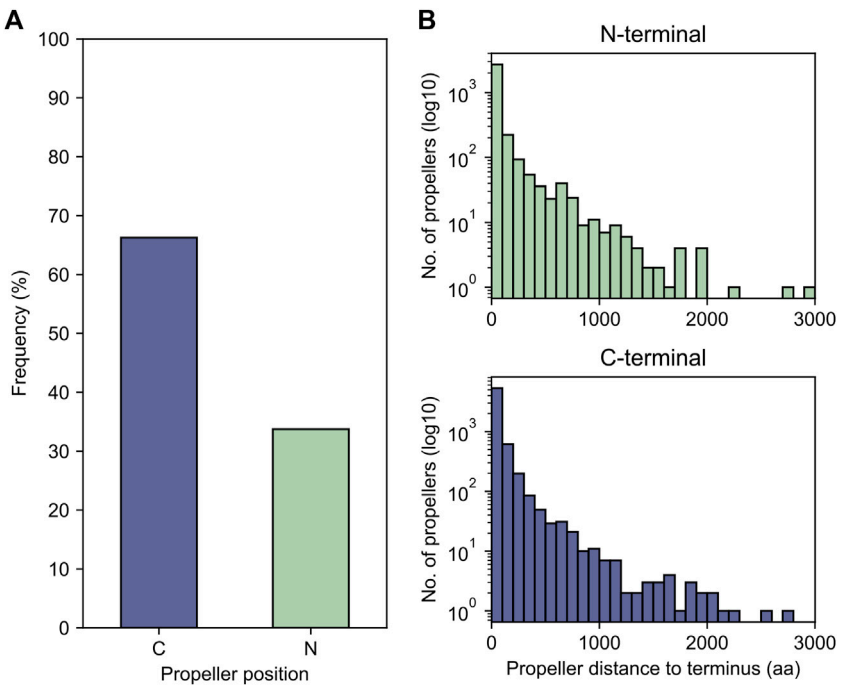

FIGURE 8 | The proximity of globally highly repetitive β-propellers to the termini. **(A)** The proportion of β-propellers in the dataset localised close to one of the termini of the corresponding host protein. For each globally repetitive β-propeller in the dataset, the distance to both the N- and C-termini was computed and the closest chosen as the closest to the β-propeller position. **(B)** Histogram of the number of amino acids separating the β-propeller from its closest terminus.

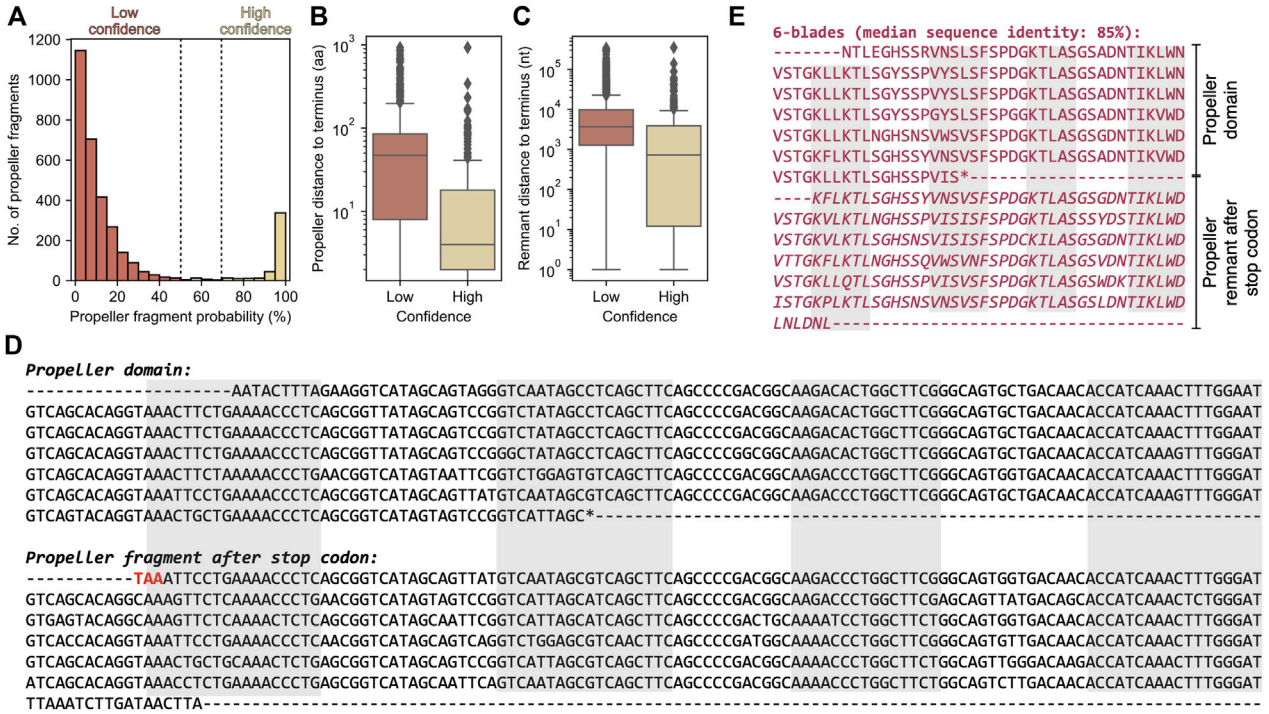

FIGURE 9 | The β-propeller fragments found in the genomic neighbourhood of globally highly repetitive β-propellers. **(A)** Histogram of the HHsearch probability of the identified putative fragments. Boxplots for the distance of the **(B)** highly repetitive β-propeller (in amino acids) and **(C)** corresponding putative β-propeller fragment (in nucleotides) to the closest host protein/gene terminus, separated by confidence level. **(D)** Nucleotide sequence and genomic context of the exemplary globally repetitive β-propeller from Nostoc sp. HK-1 hypothetical WD40 repeat-containing protein BBD61132.1. The starting codon is highlighted in blue and in frame, disruptive, stop codons in red. The regions coding for the strands making up the different blades are highlighted in grey. **(E)** Amino acid sequence of the 6 highly similar blades annotated at the C-terminus of BBD61132.1 plus the additional 6 found beyond the stop codon (and that make the β-propeller fragment).

We could map 7,329 (76%) globally repetitive β-propellers to 3,175 unique genome assemblies and found putative β-propeller fragments in the neighbourhood of 21% of these (corresponding to 1,553 globally repetitive β-propellers across 584 assemblies). While 83% of the fragments are of low confidence (**Figure 9A**), as the probability of their HHsearch match is lower than 50%, the probability distribution is bimodal and 13% (corresponding to 428 fragments close to 320 globally repetitive β-propellers) have a HHsearch probability above 70%. High confidence fragments tend to be closer to the gene containing the globally repetitive β-propellers, which per se are closer to the terminus (**Figures 9B,C**). There is no clear family or taxonomic preference for such putatively disrupted β-propellers (**Supplementary Figure S7**).

One high-confidence example is shown in **Figure 9D**. It is exactly downstream from the open reading frame (ORF) for Nostoc sp. HK-1 hypothetical WD40 repeat-containing protein BBD61132.1, which contains 6 WD40-like blades with a median sequence identity of 85% (**Figure 9E**). Aligning the fragment nucleotide sequence with that coding the β-propeller in BBD61132.1 revealed that it is the result of a deletion that placed a premature stop codon in frame. Right after this stop codon there is an out-of-frame ORF with six other blades (the fragment), which are not only highly similar within themselves, but also to those in BBD61132.1, and a C-terminal TPR repeat that is partially assigned to ORF BBD61130.1. It is thus likely that BBD61132.1 was a longer protein that contained a C-terminal TPR repeat and an at least 12-bladed globally repetitive β-propeller that was disrupted.

Another example is shown in **Supplementary Figure S7**. In this case, a hypothetical protein, DMF28_08,825 (PYL67592.1) from an unknown Verrucomicrobia bacterium, contains a globally repetitive β-propeller with four Kelch-like blades of median sequence identity of 76% followed by a segment without similarity to any known protein (**Supplementary Figure S8A**). The analysis of the nucleotide sequence shows that there was most likely a large deletion in the fifth blade, which moved the gene out of register and caused a sixth blade to be encoded in another frame, hence its amino acid sequence not matching anything in the protein database. Following an in-frame stop codon there, there is a further seventh blade and the part required to complete the first blade (**Supplementary Figures S7B,C**). We conclude that this protein used to have a 7-bladed Kelch-like β-propeller, which is now disrupted by a frameshift and a stop codon.

Our analysis provides a conservative look at disrupted β-propellers, as we did not identify additional blade-like sequences in the genomic vicinity of every β-propeller we think may be disrupted, such as in the case of the highly repetitive β-propeller in the hypothetical protein DL7770_005,219 described above.

DISCUSSION

The range of internal sequence symmetries in β-propellers has led us to propose that their different lineages arose by independent cycles of amplification and differentiation (Chaudhuri et al., 2008). Here we analysed whether these cycles are still ongoing within individual lineages by tracking the occurrence of highly symmetrical β-propellers. We find that β-propellers with a median internal sequence identity of more than 60% are widespread across all major β-propeller lineages, including the three superclusters formed by WD40, VCBS and Asp-Box proteins, respectively (Pereira and Lupas, 2021). They are also widespread with respect to their phylogenetic distribution, occurring in all kingdoms of life, with a higher incidence however in individual branches, such as Actinobacteria, Cyanobacteria and Proteobacteria among the bacteria, Ascomycota and Basidiomycota among the eukaryotes, and Euryarchaeota among the archaea.

The cutoff for our analysis at 60% internal sequence symmetry was chosen because there was only one natural β-propeller exceeding this value and this protein was the one that had prompted this study in the first place. In fact, though, the range of internal sequence symmetries in β-propellers goes in a continuum from less than 20% up to 100%. The focus on those with high internal sequence symmetry was therefore motivated by our effort to substantiate the independent amplification of a large number of diverse β-propellers. In the process, we also encountered a fair number of β-propellers that appear to be disrupted, indicating that the cycle may include an element of decay as well.

This prompts the question of the evolutionary mechanisms that govern this cycle.

β-Propellers are a closed fold, in which a variable number of repeats bring about an overall globular shape. The expectation from other repetitive globular folds is that an ancestral amplification at the origin of the fold was followed by an extended period of differentiation through descent with modification. This does not, however, appear to apply to all β-propellers (Chaudhuri et al., 2008). In ancient β-propellers, like the one forming the G protein β subunit, blade similarity is indeed higher between equivalent blades in different homologs than within the same β-propeller, indicating the expected ancestral amplification, followed by gradual differentiation. In highly repetitive β-propellers however blade similarity is higher between blades of the same β-propeller than to any other blade, indicating a recent origin, independent of other β-propellers in the database.

We are thus faced with the question of temporal balance between the amplification and differentiation events. As we show for the WRAP proteins, even extremely similar β-propellers may have been amplified independently, being only homologous at the level of the blade, but analogous in the fully amplified form. Judging from the range of internal sequence symmetry in β-propellers, we are thus led to conclude that amplification events happen continuously in many lineages and are still ongoing today.

In our dataset, we encountered mostly forms that have been amplified globally from one single blade (92% of the dataset), although there was one instance of a form that had arisen from the amplification of a 2-bladed fragment, the first instance ever observed to our knowledge. The remaining 8% of the dataset, however, were instances of local amplifications within otherwise fully differentiated β-propellers. The most surprising of these resulted from the recent duplication of an exon in an ancient protein of bilateria, which appears to be species-specific even within the silk moths, where it occurred. We conclude that while global amplification is the dominant mechanism for new β-propellers, local amplification of individual blades is substantial and opens new evolutionary possibilities for otherwise fully specialised forms.

In mature β-propellers, particular lineages tend to prefer a well-defined number of blades (Kopec and Lupas, 2013). Since there is no counting system associated with the amplification process, we would anticipate that our set of recently amplified β-propellers would contain forms with more or fewer blades than preferred by that particular lineage. Indeed, we observe that repetitive β-propellers of a given lineage mostly have the blade number preferred by that lineage, but additionally manifest many larger and smaller ones. As seen from protein engineering studies, β-propellers with smaller numbers may reach a folded structure by oligomerisation (Smock et al., 2016; Afanasieva et al., 2019; Vrancken et al., 2020). These studies also show that β-propellers with larger numbers can fold to the preferred number by leaving one or more blades unstructured. While the presence of unstructured material is likely to be disfavoured and rapidly removed by in-frame deletion, it is less clear why oligomeric forms should be disfavoured, but the observation is that only a minute fraction of natural β-propellers are oligomeric. In either case, the newly arisen β-propeller will be likely to converge rapidly onto the preferred number of blades, either through further amplification or deletion, but one cannot exclude the hypothesis that in the meantime such β-propellers may also form higher order structures.

While the possible effects of sequencing errors in our results cannot be discarded, in our dataset, there is a non-trivial number of β-propellers disrupted by frameshift and in-frame stop codons. Since these are highly repetitive forms, it would seem that they are decaying soon after their origin and might therefore represent proteins that were unable to converge on a form useful to the cell in a relevant timeframe. There is however also the possibility that their usefulness was present, but of limited duration, for example if they were involved in innate immunity or self-recognition. In this context, we note that there is a large number of disrupted β-propellers among the WRAP branch of WD40, which has been implicated in bacterial innate immunity, and among VCBS proteins, which may be involved in bacterial self-recognition (Dunin-Horkawicz et al., 2014).

Overall, this study has revealed a number of features that are not widely observed in protein evolution. Most conspicuously, the continuous, open-ended amplification of new forms from smaller, subdomain-sized fragments is observed in fibrous folds, such as coiled coils (Hernandez Alvarez et al., 2019), and solenoids, such as TPR (Zhu et al., 2016), but so far not

in globular folds. β-propellers appear to be unique in forming closed structures while yet undergoing continuous amplification from single blades. As a flip side of the continuous amplification, β-propellers also offer many examples for the decay of young protein coding genes, which either have never reached usefulness or have lost it within brief evolutionary time. For those that have not started decaying soon after their genesis, we are also afforded the opportunity to observe the accumulation of synonymous and non-synonymous mutations. In conclusion, β-propellers are an ideal model system in which to study the evolutionary cycle of proteins at warp speed.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data are in public databases at NCBI and UniProt. The code used and the results generated during this project are available through: https://github.com/JoanaMPereira/RecentlyAmplifiedProps.

AUTHOR CONTRIBUTIONS

AL initiated the project, JP performed the analysis with contributions from AL, both authors wrote the manuscript.

FUNDING

This work was supported by the Volkswagenstiftung (grant number 94810) and institutional funds of the Max Planck Society.

ACKNOWLEDGMENTS

We would like to thank the Bioinformatics group of the Department of Protein Evolution at the Max Planck Institute for Developmental Biology, especially Laura Weidmann-Krebs and Hadeer Elhabashy, for stimulating discussions. JP would also like to thank the Schwede group and sciCORE at the University of Basel for providing computational resources and system administration support with protein structure modelling with AlphaFold2 using af2@scicore.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.895496/full#supplementary-material

REFERENCES

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

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi:10.1093/nar/25.17.3389

Alva, V., Söding, J., and Lupas, A. N. (2015). A Vocabulary of Ancient Peptides at the Origin of Folded Proteins. *eLife* 4, e09410. doi:10.7554/eLife.09410.001

Amara, N., Krom, B. P., Kaufmann, G. F., and Meijler, M. M. (2011). Macromolecular Inhibition of Quorum Sensing: Enzymes, Antibodies, and beyond. *Chem. Rev.* 111 (1), 195–208. doi:10.1021/cr100101c

Andrade, M. A., Perez-Iratxeta, C., and Ponting, C. P. (2001). Protein Repeats: Structures, Functions, and Evolution. *J. Struct. Biol.* 134 (2–3), 117–131. Academic Press. doi:10.1006/jsbi.2001.4392

Beisel, H.-G., Kawabata, S.-I., 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 (9), 2313–2322. doi:10.1093/emboj/18.9.2313

Bezerra, G. A., Dobrovetsky, E., Seitova, A., Fedosyuk, S., Dhe-Paganon, S., and Gruber, K. (2015). Structure of Human Dipeptidyl Peptidase 10 (DPPY): A Modulator of Neuronal Kv4 Channels. *Sci. Rep.* 5. doi:10.1038/srep08769

Biegert, A., and Söding, J. (2008). De Novo identification of Highly Diverged Protein Repeats by Probabilistic Consistency. *Oxf. Univ. Press* 24 (6), 807–814. doi:10.1093/bioinformatics/btn039

Blatch, G. L., and Lässle, M. (1999). The Tetratricopeptide Repeat: A Structural Motif Mediating Protein-Protein Interactions. *BioEssays* 21 (11), 932–939. doi:10.1002/(sici)1521-1878(199911)21:11<932::aid-bies5>3.0.co;2-n

Bliven, S. E., Lafita, A., Rose, P. W., Capitani, G., Prlić, A., and Bourne, P. E. (2019). Analyzing the Symmetrical Arrangement of Structural Repeats in Proteins with CE-Symm. *PLoS Comput. Biol.* 15 (4), e1006842. doi:10.1371/journal.pcbi.1006842

Brown, N. G., Chow, D.-C., Ruprecht, K. E., and Palzkill, T. (2013). Identification of the β-Lactamase Inhibitor Protein-II (BLIP-II) Interface Residues Essential for Binding Affinity and Specificity for Class A β-Lactamases. *J. Biol. Chem.* 288 (24), 17156–17166. Â© 2013 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology. doi:10.1074/jbc.M113.463521

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. *Bioinformatics* 25 (15), 1972–1973. doi:10.1093/bioinformatics/btp348

Chaudhuri, I., Söding, J., and Lupas, A. N. (2008). Evolution of the β-propeller Fold. *Proteins* 71 (2), 795–803. doi:10.1002/prot.21764

Chen, C. K.-M., Chan, N.-L., and Wang, A. H.-J. (2011). The Many Blades of the β-propeller Proteins: Conserved but Versatile. *Trends Biochem. Sci.* 36, 553–561. doi:10.1016/j.tibs.2011.07.004

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., et al. (2014). ECOD: An Evolutionary Classification of Protein Domains. *PLoS Comput. Biol.* 10 (12), e1003926. doi:10.1371/journal.pcbi.1003926

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* 25 (11), 1422–1423. doi:10.1093/bioinformatics/btp163

Corbett, K. D., Shultzaberger, R. K., and Berger, J. M. (2004). The C-Terminal Domain of DNA Gyrase A Adopts a DNA-Bending β-pinwheel Fold. *Proc. Natl. Acad. Sci. U.S.A.* 101 (19), 7293–7298. doi:10.1073/pnas.0401595101

Das, S., Mandal, M., Chakraborti, T., Mandal, A., and Chakraborti, S. (2003). Structure and Evolutionary Aspects of Matrix Metalloproteinases: A Brief Overview. *Mol. Cell. Biochem.* 253 (1-2), 31–40. doi:10.1023/A:1026093016148

Dunin-Horkawicz, S., Kopec, K. O., and Lupas, A. N. (2014). Prokaryotic Ancestry of Eukaryotic Protein Networks Mediating Innate Immunity and Apoptosis. *J. Mol. Biol.* 426 (7), 1568–1582. doi:10.1016/j.jmb.2013.11.030

Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic acids Res.* 32 (5), 1792–1797. Oxford University Press. doi:10.1093/nar/gkh340

Frickey, T., and Lupas, A. (2004)., 20. Bioinformatics, 3702–3704. doi:10.1093/bioinformatics/bth444CLANS: A Java Application for Visualizing Protein Families Based on Pairwise Similarity*Bioinformatics*18

Fülöp, V., Böcskei, Z., and Polgár, L. (1998). Prolyl Oligopeptidase: an Unusual Beta-Propeller Domain Regulates Proteolysis. *Cell* 94 (2), 161–170. doi:10.1016/S0092-8674(00)81416-6

Fülöp, V., and Jones, D. T. (1999). β Propellers: Structural Rigidity and Functional Diversity. *Curr. Opin. Struct. Biol.* 9, 715–721. doi:10.1016/S0959-440X(99)00035-4

Ghosh, M., Anthony, C., Harlos, K., Goodwin, M. G., and Blake, C. (1995). The Refined Structure of the Quinoprotein Methanol Dehydrogenase from Methylobacterium Extorquens at 1.94 Å. *Structure* 3 (2), 177–187. doi:10.1016/S0969-2126(01)00148-4

Giuseppe, P. O., Neves, F. O., Nascimento, A. L. T. O., and Guimarães, B. G. (2008). The Leptospiral Antigen Lp49 Is a Two-Domain Protein with Putative Protein Binding Function. *J. Struct. Biol.* 163 (1), 53–60. doi:10.1016/j.jsb.2008.04.003

Gupta, V. A., and Beggs, A. H. (2014). Kelch Proteins: Emerging Roles in Skeletal Muscle Development and Diseases. *Skelet. Muscle* 4 (1), 11–12. doi:10.1186/2044-5040-4-11

Guruprasad, K., and Dhamayanthi, P. (2004). Structural Plasticity Associated with the β-propeller Architecture. *Int. J. Biol. Macromol.* 34 (1–2), 55–61. doi:10.1016/j.ijbiomac.2004.03.003

Hadjebi, O., Casas-Terradellas, E., Garcia-Gonzalo, F. R., and Rosa, J. L. (2008). The RCC1 Superfamily: From Genes, to Function, to Disease. *Biochimica Biophysica Acta (BBA) - Mol. Cell Res.* 1783, 1467–1479. doi:10.1016/j.bbamcr.2008.03.015

Hernandez Alvarez, B., Bassler, J., and Lupas, A. N. (2019). Structural Diversity of Coiled Coils in Protein Fibers of the Bacterial Cell Envelope. *Int. J. Med. Microbiol.* 309 (5), 351–358. doi:10.1016/J.IJMM.2019.05.011

Hester, G., and Wright, C. S. (1996). The Mannose-specific Bulb Lectin fromGalanthus nivalis(Snowdrop) Binds Mono- and Dimannosides at Distinct Sites. Structure Analysis of Refined Complexes at 2.3 Å and 3.0 Å Resolution. *J. Mol. Biol.* 262 (4), 516–531. doi:10.1006/jmbi.1996.0532

Hiramatsu, H., Kyono, K., Higashiyama, Y., Fukushima, C., Shima, H., Sugiyama, S., et al. (2003). The Structure and Function of Human Dipeptidyl Peptidase IV, Possessing a Unique Eight-Bladed β-propeller Fold. *Biochem. Biophysical Res. Commun.* 302 (4), 849–854. doi:10.1016/S0006-291X(03)00258-4

Jawad, Z., and Paoli, M. (2002). Novel Sequences Propel Familiar Folds. *Structure* 10 (4), 447–454. doi:10.1016/S0969-2126(02)00750-5

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Kaus, K., Biester, A., Chupp, E., Lu, J., Visudharomn, C., and Olson, R. (2019). The 1.9 Å Crystal Structure of the Extracellular Matrix Protein Bap1 from Vibrio cholerae Provides Insights into Bacterial Biofilm Adhesion. *J. Biol. Chem.* 294 (40), 14499–14511. doi:10.1074/jbc.RA119.008335

Kim, K. H., and Paetzel, M. (2011). Crystal Structure of Escherichia coli BamB, a Lipoprotein Component of the β-Barrel Assembly Machinery Complex. *J. Mol. Biol.* 406 (5), 667–678. Elsevier Ltd. doi:10.1016/j.jmb.2010.12.020

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

Levasseur, A., and Pontarotti, P. (2011). The Role of Duplications in the Evolution of Genomes Highlights the Need for Evolutionary-Based Approaches in Comparative Genomics. *Biol. Direct* 6 (1), 11. doi:10.1186/1745-6150-6-11

Makarova, K. S., Wolf, Y. I., Karamycheva, S., Zhang, D., Aravind, L., and Koonin, E. V. (2019). Antimicrobial Peptides, Polymorphic Toxins, and Self-Nonself Recognition Systems in Archaea: an Untapped Armory for Intermicrobial Conflicts. *mBio* 10 (3), e00715–19. doi:10.1128/mBio.00715-19

Matsushima, N., Enkhbayar, P., Kamiya, M., Osaki, M., and Kretsinger, R. (2005). Leucine-Rich Repeats (LRRs): Structure, Function, Evolution and Interaction with Ligands. *Ddro* 2, 305–322. doi:10.2174/1567269054087613

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2021). ColabFold - Making Protein Folding Accessible to All. *bioRxiv*. doi:10.1101/2021.08.15.456425

Ochiai, A., Itoh, T., Maruyama, Y., Kawamata, A., Mikami, B., Hashimoto, W., et al. (2007). A Novel Structural Fold in Polysaccharide Lyases. *J. Biol. Chem.* 282 (51), 37134–37145. Â© 2007 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology. doi:10.1074/jbc.M704663200

Ochiai, A., Itoh, T., Mikami, B., Hashimoto, W., and Murata, K. (2009). Structural Determinants Responsible for Substrate Recognition and Mode of Action in Family 11 Polysaccharide Lyases. *J. Biol. Chem.* 284 (15), 10181–10189. Â© 2009 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology. doi:10.1074/jbc.M807799200

Oubrie, A., Rozeboom, H. J., Kalk, K. H., Olsthoorn, A. J., Duine, J. A., and Dijkstra, B. W. (1999). Structure and Mechanism of Soluble Quinoprotein Glucose Dehydrogenase. *EMBO J.* 18 (19), 5187–5194. doi:10.1093/emboj/18.19.5187

Park, H. U., and Lee, K. J. (1998). Cloning and Heterologous Expression of the Gene for BLIP-II, a -Lactamase-Inhibitory Protein from Streptomyces Exfoliatus SMF19. *Microbiology* 144 (8), 2161–2167. doi:10.1099/00221287-144-8-2161

Pei, J., and Grishin, N. V. (2014). PROMALS3D: Multiple Protein Sequence Alignment Enhanced with Evolutionary and Three-Dimensional Structural Information. *Methods Mol. Biol.* 1079, 263–271. doi:10.1007/978-1-62703-646-7_17

Pereira, J., and Lupas, A. N. (2021). The VCBS Superfamily Forms a Third Supercluster of β-propellers that Includes Tachylectin and Integrins. *Bioinformatics* 36, 5618–5622. (January). doi:10.1093/bioinformatics/btaa1085

Polgár, L. (2002). The Prolyl Oligopeptidase Family. *Cell. Mol. Life Sci.* 59, 349–362. doi:10.1007/s00018-002-8427-5

Pons, T., Gómez, R., Chinea, G., and Valencia, A. (2003). Beta-propellers: Associated Functions and Their Role in Human Diseases. *Cmc* 10 (6), 505–524. doi:10.2174/0929867033368204

Quistgaard, E. M., and Thirup, S. S. (2009). Sequence and Structural Analysis of the Asp-Box Motif and Asp-Box Beta-Propellers; a Widespread Propeller-type Characteristic of the Vps10 Domain Family and Several Glycoside Hydrolase Families. *BMC Struct. Biol.* 9 (1), 46. doi:10.1186/1472-6807-9-46

Rigden, D. J., and Galperin, M. Y. (2004). The DxDxDG Motif for Calcium Binding: Multiple Structural Contexts and Implications for Evolution. *J. Mol. Biol.* 343 (4), 971–984. Academic Press. doi:10.1016/j.jmb.2004.08.077

Rigden, D. J., Woodhead, D. D., Wong, P. W. H., and Galperin, M. Y. (2011). New Structural and Functional Contexts of the Dx[DN]xDG Linear Motif: Insights into Evolution of Calcium-Binding Proteins. *PLoS ONE* 6 (6), e21507. doi:10.1371/journal.pone.0021507

Smock, R. G., Yadid, I., Dym, O., Clarke, J., and Tawfik, D. S. (2016). De Novo Evolutionary Emergence of a Symmetrical Protein Is Shaped by Folding Constraints. *Cell* 164 (3), 476–486. doi:10.1016/j.cell.2015.12.024

Söding, J., and Lupas, A. N. (2003). More Than the Sum of Their Parts: on the Evolution of Proteins from Peptides. *BioEssays* 25 (9), 837–846. doi:10.1002/bies.10321

Söding, J. (2005). Protein Homology Detection by HMM-HMM Comparison. *Bioinformatics* 21 (7), 951–960. doi:10.1093/bioinformatics/bti125

Stevens, T. J., and Paoli, M. (2008). RCC1-like Repeat Proteins: A Pangenomic, Structurally Diverse New Superfamily of β-propeller Domains. *Proteins* 70 (2), 378–387. doi:10.1002/prot.21521

Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., and Schwede, T. (2020). QMEANDisCo-Distance Constraints Applied on Model Quality Estimation. *Bioinformatics* 36 (6), 1765–1771. doi:10.1093/bioinformatics/btz828

Varnay, I. (2009). *Investigations on the 100 kDa Homohexameric Protein Ph1500 by NMR Spectroscopy*. Technische Universität München.

Vrancken, J. P. M., Aupič, J., Addy, C., Jerala, R., Tame, J. R. H., and Voet, A. R. D. (2020). Molecular Assemblies Built with the Artificial Protein Pizza. *J. Struct. Biol. X* 4, 100027. doi:10.1016/j.yjsbx.2020.100027

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* 46 (W1), W296–W303. Oxford Academic. doi:10.1093/nar/gky427

Yadid, I., and Tawfik, D. S. (2011). Functional β-propeller Lectins by Tandem Duplications of Repetitive Units. *Narnia* 24 (1–2), 185–195. doi:10.1093/protein/gzq053

Yadid, I., and Tawfik, D. S. (2007). Reconstruction of Functional β-Propeller Lectins via Homo-Oligomeric Assembly of Shorter Fragments. *J. Mol. Biol.* 365 (1), 10–17. Academic Press. doi:10.1016/j.jmb.2006.09.055

Zhu, H., Sepulveda, E., Hartmann, M. D., Kogenaru, M., Ursinus, A., Sulz, E., et al. (2016). Origin of a Folded Repeat Protein from an Intrinsically Disordered Ancestor. *eLife* 5, 551–560. eLife Sciences Publications Limited. doi:10.7554/eLife.16761

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., et al. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* 430 (15), 2237–2243. doi:10.1016/j.jmb.2017.12.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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