Sequence analysis

The VCBS superfamily forms a third supercluster of β-propellers that includes tachylectins and integrins

Joana Pereira and Andrei N. Lupas

Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Tübingen 72076, Germany

Abstract

Motivation: β-Propellers are found in great variety across all kingdoms of life. They assume many cellular roles, primarily as scaffolds for macromolecular interactions and catalysis. Despite their diversity, most β-propeller families clearly originated by amplification from the same ancient peptide—the ‘blade’. In cluster analyses, β-propellers of the WD40 superfamily always formed the largest group, to which some important families, such as the α-integrin, Asp-box and glycoside hydrolase β-propellers connected weakly. Motivated by the dramatic growth of sequence databases we revisited these connections, with a special focus on VCBS-like β-propellers, which have not been analysed for their evolutionary relationships so far.

Results: We found that VCBS-like form a supercluster with integrin-like β-propellers and tachylectins, clearly delimited from the superclusters formed by WD40 and Asp-Box β-propellers. Connections between the three superclusters are made mainly through PQQ-like β-propeller. Our results present a new, greatly expanded view of the β-propeller classification landscape.

Contact: pereira.joanam@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Proteins with a β-propeller domain are found in all kingdoms of life (Fig. 1c). They are involved in diverse biological processes, from adhesion to transcription regulation (Chen et al., 2011; Falup and Jones, 1999; Guruprasad and Dhamayanthi, 2004; Pons et al., 2003). In them, the β-propeller acts mostly as a recognition site for different biomolecules, but may also carry catalytic activity. These repetitive domains (Andrade et al., 2001; Soding and Lupas, 2003) adopt a toroid fold, where between 4 and 12 (Fig. 1d) copies of a widespread supersecondary structure, the 4-stranded β-membrane, are arranged radially around a central channel (Fig. 1a, b). These repeats, whose strands are labelled A–D (Fig. 1b), are called ‘blades’ and the toroids they form correspondingly ‘propellers’. Blades carry specific sequence motifs which allow the classification of cognate β-propellers into a hierarchy of families and superfamilies (Chaudhuri et al., 2008; Chen et al., 2011; Falup and Jones, 1999; Guruprasad and Dhamayanthi, 2004; Pons et al., 2003).

Despite their wide sequence diversity (Fig. 1c, f), most β-propeller families are related to each other and emerged by independent amplification from a set of homologous ancestral blades, in a process that is still visibly ongoing (Afanasieva et al., 2019; Alva et al., 2015; Chaudhuri et al., 2008; Dinin-Horkawicz et al., 2014; Kopec and Lupas, 2013). Classification studies (Chaudhuri et al., 2008; Kopec and Lupas, 2013) suggested that most β-propeller families form a supercluster centred on WD40 β-propellers, a large superfamily characterized by a Trp-Asp motif at the end of strand C (in position 40). Proteins assigned to this supercluster in previous studies included the 7-bladed β-subunits of G-proteins, the 6-bladed low-density lipoprotein (LDL) receptors, the 6-bladed protein kinase PknD and the 5-bladed tachyletin-2 family, which comprises eukaryotic lectins involved in the innate immunity of cnidarians and crustaceans (Beisel et al., 1999; Hayes et al., 2010; Neer et al., 1994). Some peripheral groups connected weakly to this supercluster (Chaudhuri et al., 2008; Kopec and Lupas, 2013), such as the 7-bladed β-propeller domain of α-integrins, characterized by a Ca²⁺- binding DxDxDG motif in the loop connecting strands A and B (loop AB) and an FG-GAP/Cage motif, which is contiguous in space but not sequence, covering the N-terminal end of strand A and the C-terminal end of strand B (Chouhan et al., 2011; Rigden and Galperin, 2004). This connection was proposed to be weakly mediated by Asp-Box β-propellers, most of whose members are characterized by a SxDxDGTW motif in the loop connecting strands C and D (loop CD) (Quistgaard and Thiirup, 2009).

Missing from these studies were β-propellers of the Vibrio, Coluella, Bradyrhizobium and Shewanella (VCBS) family (Pfam: PF13517), a poorly described group that has hitherto not been analysed systematically for its evolutionary relationships. VCBS encompasses the 7-bladed β-propellers in aldol-2-ulose dehydratases (AUDH) (Claesson et al., 2012), ABC toxin component B (TcB)

Received on October 14, 2020; revised on December 14, 2020; editorial decision on December 18, 2020; accepted on December 21, 2020

Associate Editor: Pier Luigi Martelli

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
sequences, were re-clustered at P-values of 10^{-18} (Fig. 2b) and 10^{-20}, respectively (Supplementary Fig. S1a).

In order to evaluate the domain environments of the β-propellers in each subcluster, their parent full-length proteins were collected and binned by size, with a step of 100 residues. A representative for each bin was collected and domains annotated iteratively with HHsearch as above. A maximum of four iterations was carried out, where sequence regions not yet mapped to a domain were searched individually. Only the best match, at a probability better than 70% and larger than 40 residues were considered. Signal peptide prediction was carried out with Phobius (Käll et al., 2004).

For HMM comparisons, the full-length sequences of the β-propellers composing the clusters and subclusters depicted in Figure 2 were used. For each group, the sequences were aligned with MUSCLE (Edgar, 2004) and the alignment trimmed with trimAl (Capella-Gutierrez et al., 2009), removing columns where >25% of the positions were a gap (gap score of 0.75) and sequences that only overlapped with less than half of the columns populated by 80% or more of the other sequences. HMM profiles were built with HHmake and aligned with HHalign (Söding, 2005), using default parameters without secondary structure scoring. The alignments were then inspected and segments corresponding to the best conserved individual blades were used to build Figure 3b. Structural alignments were carried out with TM-align (Zhang and Skolnick, 2005).

3 Results

PSI-BLAST searches with 13 β-propellers of known structure, chosen to represent the families described above (Supplementary Table S1), yielded a total of 5996 sequences from bacteria, archaea and eukaryotes (see Methods). When clustered by pairwise similarity (Fig. 2), these sequences form three superclusters organized around cores of WD40, Asp-Box and VCBS-like β-propellers, respectively. The WD40 and Asp-Box superclusters were expected, based on previous analyses (Chaudhuri et al., 2008; Kopec and Lupas, 2013), but we were struck by the clear grouping of the other β-propeller families into a third supercluster, centred on VCBS and clearly delimited from the other two.

The core of the VCBS supercluster comprises prokaryotic β-propellers from diverse hypothetical protein families (Supplementary Fig. S1), which carry a signal sequence and may contain several β-propeller domains, accompanied by domains associated with biomolecular interactions (mostly immunoglobulin-like domains, but also armadillo repeats and jelly-roll-like lectins, Supplementary Fig. S1). The VCBS core group is connected to a large periphery of VCBS-like families, including PVL, TcB and AUDH, as well as to diverse hypothetical β-propellers, which have hitherto remained unstudied (Fig. 2b and Supplementary Fig. S1). β-Propeller families in this periphery are found in a variety of hypothetical proteins, whose domain composition suggests an involvement in biomolecular interactions and catalysis (Supplementary Fig. S1a). The most peripheral families that still connect directly to the VCBS core are the integrin-like β-propellers and the bacterial RGL11 family (rhagmologacturonan lyase YesX, ECOD: 001396995). Two other important β-propeller families complete the VCBS supercluster, comprising tachylectins and PQQ β-propellers, respectively. These connect to each other, and also to the VCBS core via RGL11, in the case of PQQ and a β-propeller family we have named VCBS actinolectins, in the case of tachylectins. We chose the name ‘VCBS actinolectins’ given their exclusive occurrence in actinobacteria and evolutionary connection to tachylectins (Fig.1b and Supplementary Fig. S1), but no member of this family has as yet been characterized functionally or structurally. These β-propellers are found in proteins that carry a signal sequence and either consist of the single β-propeller domain or of the β-propeller preceded by a TIM barrel (Supplementary Fig. S1a). Their connection to the tachylectin cluster is mediated by a core of bacterial tachylectin-like sequences, which are found in secreted proteins often containing additional domains involved in catalysis. Two groups radiate from this core, the eukaryotic tachylectins-2 and a...
second family of actinobacterial β-propellers, both of which are comprised of secreted proteins consisting of the β-propeller domain alone. The identification of these multiple tachylectin-like families was a striking result as tachylectin β-propellers have been considered for long time as near-orphans and have so far only been reported in eukaryotes (Beisel et al., 1999; Hayes et al., 2010; Smock et al., 2016).

HMM comparisons highlight the sequence motifs behind the connections described here (Fig. 3). The most prominent motif is the aspartate-rich DxDxDG sequence of loop AB (Figs 3b and 4) (Chouhan et al., 2011; Cioci et al., 2006; Rigden and Galperin, 2004). While in PVL and α-integrin, this loop binds Ca2+ (Fig. 4b), in other members, it may recognize also other metal cations (Chouhan et al., 2011; Claesson et al., 2012; Meusch et al., 2014; Rigden and Galperin, 2004). Also conspicuous are two non-contiguous, highly conserved residues of loop CD, G and W (Fig. 3b). Their functional role is uncertain, but in integrin-like β-propellers, the G coordinates a water molecule involved in Ca2+.

Fig. 2. Classification landscape of representative β-propeller families. (a) Cluster map of all 5996 sequences collected. Clustering was carried out with CLANS in 2D until equilibrium at a BLASTp P-value of 10−5, with connections represent similarities at this P-value (the darker, the more similar). Different regions of the map are annotated with the name of the sequences within the corresponding cluster or, when a cluster encompasses multiple families, by the β-propeller family as in ECOD and Pfam. (b) Cluster map of the 2662 sequences in the VCBS supercluster. Clustering was carried out as in (a) but a BLASTp P-value of 10−18, in order to expand it and uncover its internal structure. Connections are shown at a BLASTp P-value of 10−10. Dots are coloured based on the family name (f-name) of the best match in HMM searches against ECOD. Multiple colours within the same cluster correspond to sequences that match multiple close β-propeller families. HP stands for ‘hypothetical propeller’

Fig. 3. HMM comparison of β-propeller groups. (a) Sequence homology matrix of β-propeller groups selected from the cluster maps, as measured by the probability of the alignment of full-length HMM profiles with HHAlign. (b) Multiple alignment of the HMM consensus sequences, focused on representative single-bladed regions. Sequence motifs common to the VCBS supercluster are highlighted in grey and summarized on top. Their function in members of known structure is depicted: a grey circle with Me+ represents ‘metal binding’ and a grey hexagon ‘sugar binding’. The Asp-Box motif is highlighted in light red. Arrows depict the four strands of blade and are named accordingly. This annotation was carried out based on the known structures of families shown, but represent only a consensus as, due to structural deviations or especial structural features, the specific start and end of these strands may be shifted.
binding (Chouhan et al., 2011), and in tachylectin-2 the W anchors a short β-helix involved in forming the sugar-binding pocket (Fig. 4). A fourth prominent motif is a GW in loop DA’ (the loop that connects strand D from one blade to strand A of the next) (Figs 3b and 4a, c), which in tachylectin-2 and PVL is involved in forming the sugar-binding pocket (Supplementary Fig. S2) (Cioci et al., 2006; Kawabata and Tsuda, 2002).

While widely represented in the families of the VCBS supercluster, none of these motifs is universal. Thus, for example, the aspartate-rich motif of loop AB is not found in tachylectin-like and PQQ β-propellers. These are connected to other families in the supercluster by the sequence of loop CD and, in the case of tachylectin-like β-propellers, by the GW motif of loop DA’.

4 Conclusions

Our results confirm the relationship conjectured between fungal PVL lectins, tachylectin-2 and integrin-like β-propellers (Cioci et al., 2006). We find that all three of these eukaryotic protein families are satellites of larger prokaryotic clusters, from which they are presumably descended. Jointly with these, they are part of a supercluster of β-propeller families, centred on the large group of prokaryotic VCBS β-propellers. This supercluster had not been recognized in previous studies (Chaudhuri et al., 2008; Kopec and Lupas, 2013) because most relevant proteins could not be included, primarily due to the lack of relevant sequences of known structure. Many of these sequences are of bacterial origin and resulted from metagenomic studies, highlighting the importance of such efforts for the better understanding of protein evolution paths and the structure of the β-propeller sequence space.

Acknowledgements

The authors thank Felipe Merino for stimulating discussions.

Funding

This work was supported by the Volkswagenstiftung [94810]; and institutional funds of the Max Planck Society.

Conflict of Interest: none declared.

Data availability

This study is based solely on data deposited in publicly available repositories, and all analyses were carried out with publicly available software.

References

Afanasieva, E. et al. (2019) Structural diversity of oligomeric β-propellers with different numbers of identical blades. Elife, 8, e49853.

Altschul, S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.

Alva, V. et al. (2015) A vocabulary of ancient peptides at the origin of folded proteins. Elife, 4, e09410.

Andrade, M.A. et al. (2001) Protein repeats: structures, functions, and evolution. J. Struct. Biol., 134, 117–131.
Beisel,H.G. et al. (1999) Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab Tachypleus tridentatus. *EMBO J.*, 18, 2313–2322.

Capella-Gutierrez,S. et al. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25, 1972–1973.

Chaudhuri,J. et al. (2008) Evolution of the β-propeller fold. *Proteins Struct. Funct. Bioinf.*, 71, 795–803.

Chen,C.-K.M. et al. (1999) Tachylectin-2: crystal structure of a specific non-self-recognizing lectin of the crab Tachypleus tridentatus. *Nature*, 397, 508–512.

Cheng,H. et al. (2014) ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.*, 10, e1003926.

Chouhan,B. et al. (2011) Conservation of the human integrin-type beta-propeller domain in bacteria. *PLoS One*, 6, e25069.

Cioci,G. et al. (2006) β-Propeller crystal structure of Phanerochaete chrysosporium with the reaction intermediate ascopyrone. *J. Mol. Biol.*, 357, 1573–1591.

Claesson,M. et al. (2012) Crystal structure of bifunctional aldose-2-ulose dehydratase/isomerase from Phanerochaete chrysosporium with the reaction intermediate ascopyrone. *M. J. Mol. Biol.*, 417, 279–293.

Dunin-Horkawicz,S. et al. (2014) Prokaryotic ancestry of eukaryotic protein networks mediating innate immunity and apoptosis. *J. Mol. Biol.*, 426, 1568–1582.

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

Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, 20, 3702–3704.

Fulop,V. and Jones,D.T. (1999) β Propellers: structural rigidity and functional diversity. *Curr. Opin. Struct. Biol.*, 9, 713–721.

Guruprasad,K. and Dhamayanthi,P. (2004) Structural plasticity associated with the β-propeller architecture. *Int. J. Biol. Macromol.*, 34, 55–61.

Hayes,M.L. et al. (2010) High amino acid diversity and positive selection at a putative coral immunity gene (tachylectin-2). *BMC Evol. Biol.*, 10, 150.

Käll,L. et al. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, 338, 1027–1036.

Kawabata,S.I. and Tsuda,R. (2002) Molecular basis of non-self recognition by the horseshoe crab tachylectins. *Biochim. Biophys. Acta Gen. Subjects*, 1572, 414–421.

Kopeć,K.O. and Lupas,A.N. (2013) β-Propeller blades as ancestral peptides in protein evolution. *PLoS One*, 8, e77074.

Li,W. and Godzik,A. (2006) Cdd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658–1659.

Makarova,K.S. et al. (2019) Antimicrobial peptides, polymorphic toxins, and self-nonself recognition systems in archaea: an untapped armory for intermicrobial conflicts. *MBio*, 10, e00715–e00719.

Meusch,D. et al. (2014) Mechanism of Tc toxin action revealed in molecular detail. *Nature*, 508, 61–65.

Neer,E.J. et al. (1994) The ancient regulatory-protein family of WD-repeat proteins. *Nature*, 371, 297–300.

Pons,T. et al.(2003) Beta-propellers: associated functions and their role in human diseases. *Curr. Med. Chem.*, 10, 505–524.

Quistgaard,E.M. and Thrup,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, 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, 971–984.

Smock,R.G. et al. (2016) De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints. *Cell*, 164, 476–486.

Soding,J. (2005) Protein homology detection by HMM – HMM comparison. *Bioinformatics*, 21, 951–960.

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

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, 33, 2302–2309.

Zimmermann,L. et al. (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.*, 430, 2237–2243.