Architecture and Evolution of Blade Assembly in β-propeller Lectins

François Bonnardel, Atul Kumar, Michaela Wimmerova, Martina Lahmann, Serge Pérez, Annabelle Varrot, Frederique Lisacek, Anne Imberty

Architecture and Evolution of Blade Assembly in β-propeller Lectins

François Bonnardel [1,2,3], Atul Kumar [1,5], Michaela Wimmerova [5,6], Martina Lahmann [7], Serge Perez [8], Annabelle Varrot [1], Frédérique Lisacek [2,3,4]*, and Anne Imberty [1]*

1. Univ. Grenoble Alpes, CNRS, CERMAV, 38000 Grenoble, France.
2. Swiss Institute of Bioinformatics, CH-1227 Geneva, Switzerland.
3. Computer Science Department, UniGe, CH-1227 Geneva, Switzerland.
4. Section of Biology, UniGe, CH-1205 Geneva, Switzerland.
5. CEITEC, Masaryk University, 625 00 Brno, Czech Republic
6. NCBR, Fac.Sci, Masaryk University, 625 00 Brno, Czech Republic
7. School of Chemistry, University of Bangor, LL57 2UW Bangor, United Kingdom,
8. Univ. Grenoble Alpes, CNRS, DPM, 38000 Grenoble, France.

* To whom correspondence should be addressed. Anne Imberty (anne.imberty@cermav.cnrs.fr, Tel: +33 476 03 76 40, Twitter: @AnneImberty) Frédérique Lisacek (frederique.lisacek@isb-sib.ch, Tel: +4122 379 58 98)

Abstract

Lectins with a β-propeller fold bind glycans on the cell surface through multivalent binding sites and appropriate directionality. These proteins are formed by repeats of short domains, raising questions about evolutionary duplication. However, these repeats are difficult to detect in translated genomes and seldom correctly annotated in sequence databases. To address these issues, we defined the blade signature of the five types of β-propellers using 3D-structural data. With these templates, we predicted 3887 β-propeller lectins in 1889 species and organised this new information in a searchable online database. The data reveals a widespread distribution of β-propeller lectins across species. Prediction also emphasises multiple architectures and led to uncover a novel β-propeller assembly scenario. This was confirmed by producing and characterizing a predicted protein coded in the genome of Kordia zhangzhouensis. The crystal structure shows a new intermediate in the evolution of β-propeller assembly and demonstrates the power of our tools

Keywords

β-propeller, lectins, oligomerisation, carbohydrate binding protein

Introduction

Among the players in glycobiology, lectins are protein receptors that can bind at least one carbohydrate, and with no enzymatic function (Lis and Sharon, 1998). Lectins are generally multivalent and such multiplicity of carbohydrate binding sites favours the strong avidity to glycoconjugates available in multiple copies on all cell surfaces. Lectins are involved in a range of biological processes taking place between cells. For example, they participate in the interaction between microorganisms and hosts cells (pathogenicity, symbiosis…). Despite such a prevalent role, lectins are rather poorly characterised in protein databases. To overcome this shortcoming, we launched the Unilectin3D database (Bonnardel et al., in press) that includes a large number of classified and manually curated lectin 3D-structures, with information on their fold, oligomeric structure and carbohydrate binding site(s). The Unilectin3D collection highlights the diversity of folds that lectins adopt, and the high frequency of the occurrence multimeric structures. However, for some lectins, multivalency is not created by oligomerization, but by tandem repeat of conserved carbohydrate binding domains. Such tandem repeats are observed in the so-called β-propeller lectins.

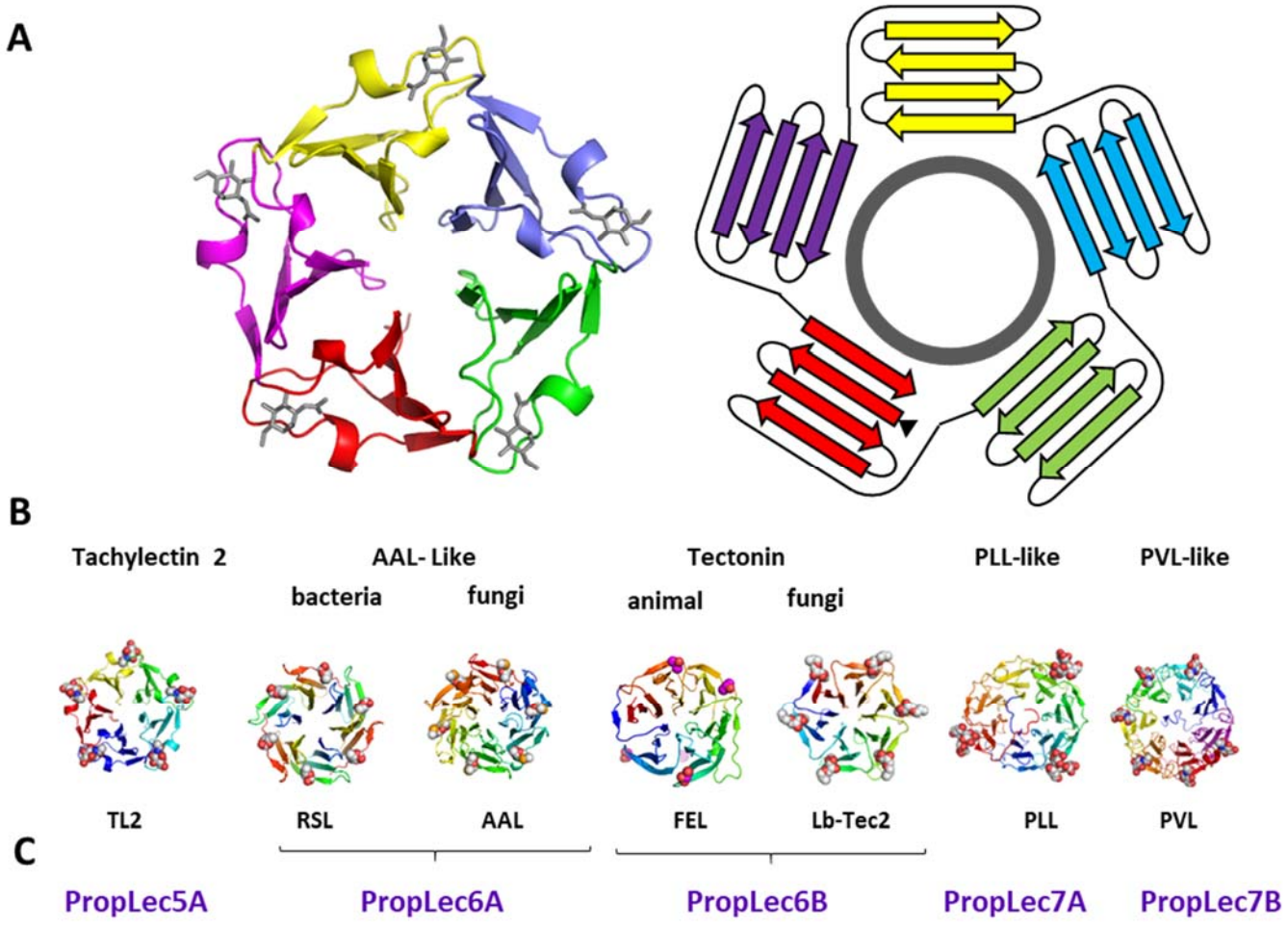


Figure 1. *A. Example of lectin β-propeller structure: the 5-bladed tachylectiy-2 (1TL2) complexed with 5 GlcNAc residues and its schematic representation. B. Structures of the seven classes of PropLecs in Unilectin3D (see Table S1 for details on each structure). C. Simplified nomenclature for the five families in the PropLec database.*

The β-propeller is a fold widely distributed in Nature (Chen et al., 2011; (Fulop and Jones, 1999). β-propeller proteins adopt a donut shape made of four to ten repeats (or blades) of four-stranded β-sheets (Chen et al., 2011; (Fulop and Jones, 1999; (Jawad and Paoli, 2002) (Figure 1A). Their functions are broad, generally related to an enzymatic active site located in the centre of the structure. Although very variable in amino acid sequences, β-propellers have been proposed to derive from a single peptide through multiple episodes or duplication and diversification (Chaudhuri et al., 2008; (Kopec and Lupas, 2013). The β-propeller fold is a very stable arrangement of repeats and allows for optimum presentation of multiple binding sites. Such topology is perfectly suited to bind carbohydrate epitopes on glycoconjugates presented on cell surfaces. It is therefore not surprising that this fold has successfully been adopted by nature for lectin functions. At the present time, Unilectin3D contains 52 X-ray structures from 13 different β-propellers proteins (PropLec) with five to seven blades that have been classified in seven different groups (Figure 1B).

Tachylectin-2, isolated from horseshoe crab, is the only 5-blade PropLec structurally characterized (Beisel et al., 1999). It binds to *N*-acetylglucosamine (GlcNAc), a glycan epitope present in the cell wall of pathogens, and is thought to be involved in the innate immunity of invertebrates (Kawabata and Iwanaga, 1999). *Aleuria aurantia* lectin (AAL) from orange peel mushroom is a 6-blade β-propeller (Wimmerova et al., 2003) that binds to fucose (Fuc). AAL-like β-propellers have also been structurally crystallized from pathogenic fungi, such as *Aspergillus fumigatus* (Houser et al., 2013), where they play a role in eliciting host immune response (Kerr et al., 2016; (Richard et al., 2018). Bacteria such as the plant pathogen *Ralstonia solanacearum* and the human pathogen *Burkholderia ambifaria* produce lectins with high similarity to AAL but containing only 2-blades (Audfray et al., 2015; (Kostlanová et al., 2005). These are the only known examples of natural β-propellers formed by oligomerisation, representing probably some ancestral form of the fold. Tectonin, a 6-blade β-propeller that binds to methylated monosaccharides generally associated with pathogens, has been structurally characterized (Capaldi et al., 2015). It is present in fish (FEL), with a proposed role in the antibacterial protection of the eggs, as well as in the mushroom *Laccaria bicolor* (Lb-Tec2) (Sommer et al., 2018). In the latter case, four tectonins oligomerize in a virus-like shape that is involved in defence against worms feeding on mushrooms (Sommer et al., 2018; (Wohlschlager et al., 2014) . The same anti-feeder role of the 7-blade PropLec in *Psathyrela velutina* (PVL) or *Agrocybe aegerita* (AAL2) mushrooms that bind GlcNAc (Cioci et al., 2006; (Ren et al., 2015) is likely. A different 7-blade β-propeller has been characterized in two species of *Photorhabdu*s bacteria (PHL and PLL) with evidence for dual specificity for Fuc and galactose (Gal) in different binding sites (Jancarikova et al., 2017; (Kumar et al., 2016).

PropLecs are of high interest for their role in defence and self-immunity. Since some of them are involved in host-pathogen recognition, they are also promising targets for glycomimetic compounds that could present anti-infectious properties. Designing multivalent molecules that fit the specific binding sites arrangement of β-propellers in pathogenic micro-organisms has been key to obtain high-affinity inhibitors (Goyard et al., 2018; (Jancarikova et al., 2018; (Machida et al., 2017). Because of their ability to bind strongly to glycoconjugates on cell surfaces, PropLecs are also useful biomarkers, for probing the glycosylation of proteins (Liu et al., 2018; (Machon et al., 2017), for labelling cancer cells (Audfray et al., 2015), or as tools to study the dynamics of glycolipids in membranes (Arnaud et al., 2013). Finally, PropLecs have been engineered, dissected in smaller pieces and reassembled to build artificial proteins for understanding stability and folding processes (Arnaud et al., 2014; (Yadid and Tawfik, 2007, 2011).

β-propeller structures are easily identified by their characteristic shape. As a result, β-propellers are in general well described in structure databases. For example, the CATH-GENE3D database (Dawson et al., 2017) has categories for propellers from 3 to 8 blades, yet not all PropLecs are included. In fact, β-propeller lectins are difficult to identify based on their amino acid sequence. The presence of short repeated peptide motifs (30 to 50 amino acids) challenges classical search programs that are based on sequence alignment. This setback, in turn, impacts the definition of family as well as the reliability of sequence-based genome mining tools. For example, the Pfam protein family database (Finn et al., 2016) defines family profiles based on domain similarity, but Pfam profiles matching PropLecs cover either part(s) of one blade (each blade is 46 to 58 amino acids long) or the whole propeller. As a result, no current tool can, as is, efficiently mine β-propellers, and they usually miss the conserved carbohydrate binding sites of PropLecs.

We developed here a precise method to detect automatically PropLecs in sequence databases. Robust peptide motifs corresponding to the repeating unit of each PropLec family were derived from the alignment of the blade sequences whose boundaries are delineated in the 3D structures. Conserved regions set the definition of family profiles and the HMMER profile search tool (Finn et al., 2015) was used to search for similar proteins in the non-redundant protein dataset from Uniprot /Uniref100 (Suzek et al., 2015). The likelihood of predicted PropLecs is scored. This prediction tool can be used to identify new targets for antibacterial drugs, association between the carbohydrate-binding and enzymatic domains, and new protein oligomerization forms. The examination of predicted results led us to unveil an alternate scenario of blade assembly hitherto not observed. We validated this potential novel way to assemble β-propeller in a predicted PropLec of *Kordia zhangzhouensis* by solving the crystal structure of this new lectin.

Results

Definition of conserved motif in each β-propeller lectin family

The presence of repeated domains in PropLecs challenges their automatic detection in genomes. Our strategy was to turn this into an advantage by defining conserved motifs corresponding to the blade signature in each family, and then to search multiple and successive occurrences of these motifs in genomes. The seven sub-groups of PropLecs that are described in Unilectin3D were defined based on structural similarity and taxonomy. By focusing only on structural and sequence similarity, we reduced this number to five PropLec families (Figure 1C). To simplify the nomenclature, each family has been named according to the number of constituting blades, e.g. PropLec5A, PropLec6A, PropLec6B, PropLec7A and PropLec7B

The structural information in the 13 different PropLecs that have been crystallized so far was used to identify the blade signature of each PropLec family (Table S1 in supplemental information). The peptide sequences were first processed with the RADAR software (Heger and Holm, 2000) in order to align the repeated regions. This alignment was refined on the basis of 3D-structural information, which entailed the adjustment of repeat boundaries to the definition of blades. When necessary, alignments were shifted along the sequence so as to centre each blade on the 3D structure. The resulting blade sequence alignments are displayed in Supplementary information (Fig S1 to S5 in supplemental information). They served as the basis for determining conserved motifs and defining characteristic profiles in the form of Hidden Markov Models (HMM). These models were

generated with the HMMbuild tool of the HMMER software suite (Finn et al., 2015). HMM profiles identify similar domains depending on the amino acid frequencies at each position of the blade and on the amino acids in previous positions.

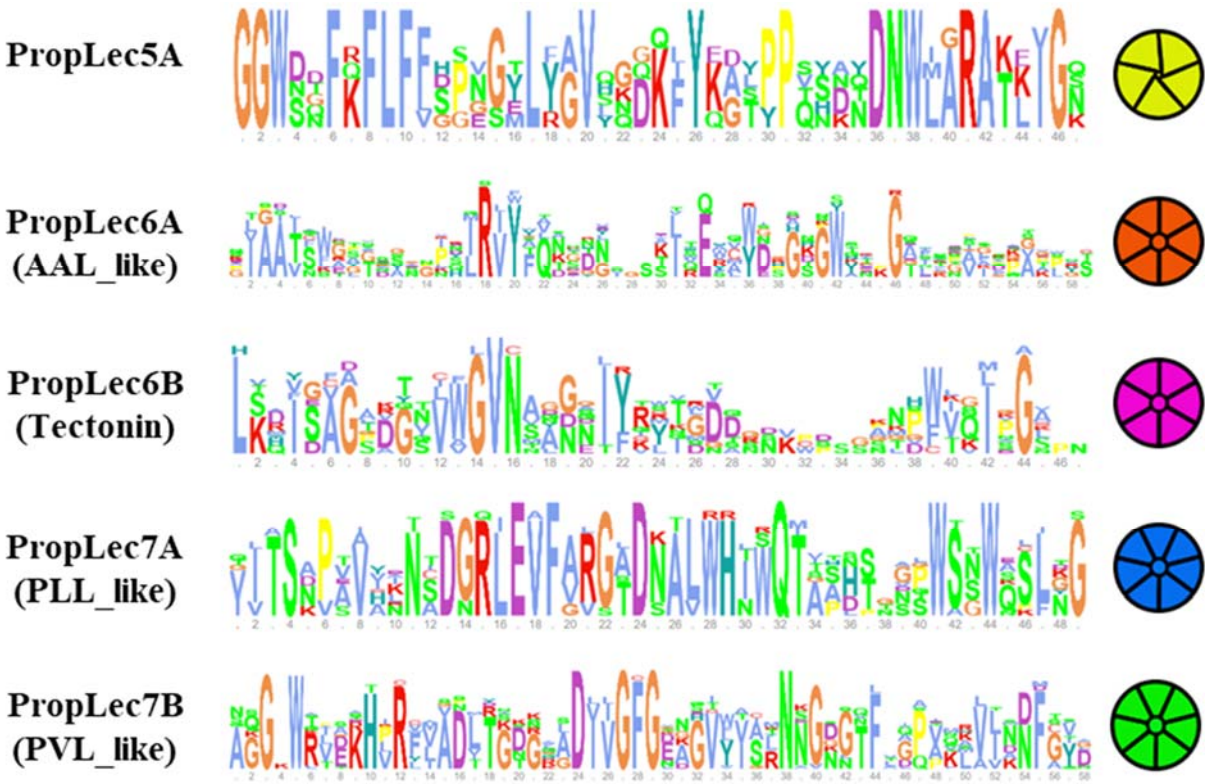


Figure 2. *Signature motif extracted from blade alignment for the 5-blade family of PropLec. Amino acids in one-letter code are coloured by class of properties, and the size of the letter corresponds to the frequency of the amino acid in the alignment. Complete sequence alignments are provided in Supp. Info (Figure S1 to Figure S5)*

As seen in Figure 2, each of the five PropLecs families have very different HMM motifs. Interestingly, the most conserved amino acids often correspond to the ones involved in the binding of the carbohydrate ligand, which indicates the conservation of function, in addition to structure.

The PropLec database

In order to identify PropLecs in other organisms, the designed motifs were fed into HMMSEARCH to process the UniRef100 non-redundant protein database (12/09/18 version containing 124 million distinct protein sequences). The predicted protein sequences were filtered with an e-value set to 0.01 while other parameters were left to default values. This search returned 3877 putative PropLec sequences containing a total of 20090 conserved blades domains (Figure S6).

A dedicated interface for mining the PropLec database is available at https://www.unilectin.eu/propeller/. For each predicted protein, information is displayed using an in-house sequence viewer and an amino acid conservation plot or sequence logo redeveloped with D3JS (Maguire et al., 2014). Both the reference and the predicted blades were aligned with the MUSCLE software (Edgar, 2004). The resulting multiple sequence alignment (MSA) is used to define one score that evaluates the similarity of each blade with the defined reference motif (see Methods section and Figure S7). A key parameter for analysing results is the cut-off value for this score on the third quartile. The information stored in the database can then be searched using different filters, such as the similarity score mentioned above, but also information related to lectin families, sequence (number of blades..), biological origin (species..) and others. The search interface is displayed in Figure S6. By default, sequences are filtered using a minimum similarity score of 0.25 and synthetic genes or partial sequences are excluded. This filtering resulted in 3605 proteins of interest. Most of them are predicted to belong to families with 7-blades (54% for PropLec7A and 22% for PropLec7B). The two 6-blade families are evenly populated (13% for PropLec6B and 8% for PropLec6A) and less than 3% belongs to the 5-bladed tachylectin family (PropLec5A).

Searches in the database generate lists of sequences with information covering family, number of blades, score, sequence length and taxonomy. Each entry can then be expanded to a full page showing further information on gene and protein sequences with cross-links to external resources as well as details of the alignment and amino acid conservation in the form of histograms. The latter highlight the comparison of the family reference and predicted motifs and allow for a visual check equivalent to the similarity score. Furthermore, the amino acids involved in the carbohydrate-binding site of the reference lectin are singled out below the alignment, as an instant evaluation of the likelihood of a lectin function. Figure 3 exemplifies a sequence from the freshwater bacterium *Kordia zhangzhouensis*. Zooming in and out is made possible both for the in-house simplified protein viewer and the NCBI gene viewer (Brown et al., 2015). Further information, such as the details of binding site contact with different carbohydrates (if available), the full alignments of all blades (predicted protein and reference) and details on neighboring genes are also visualized on the page.

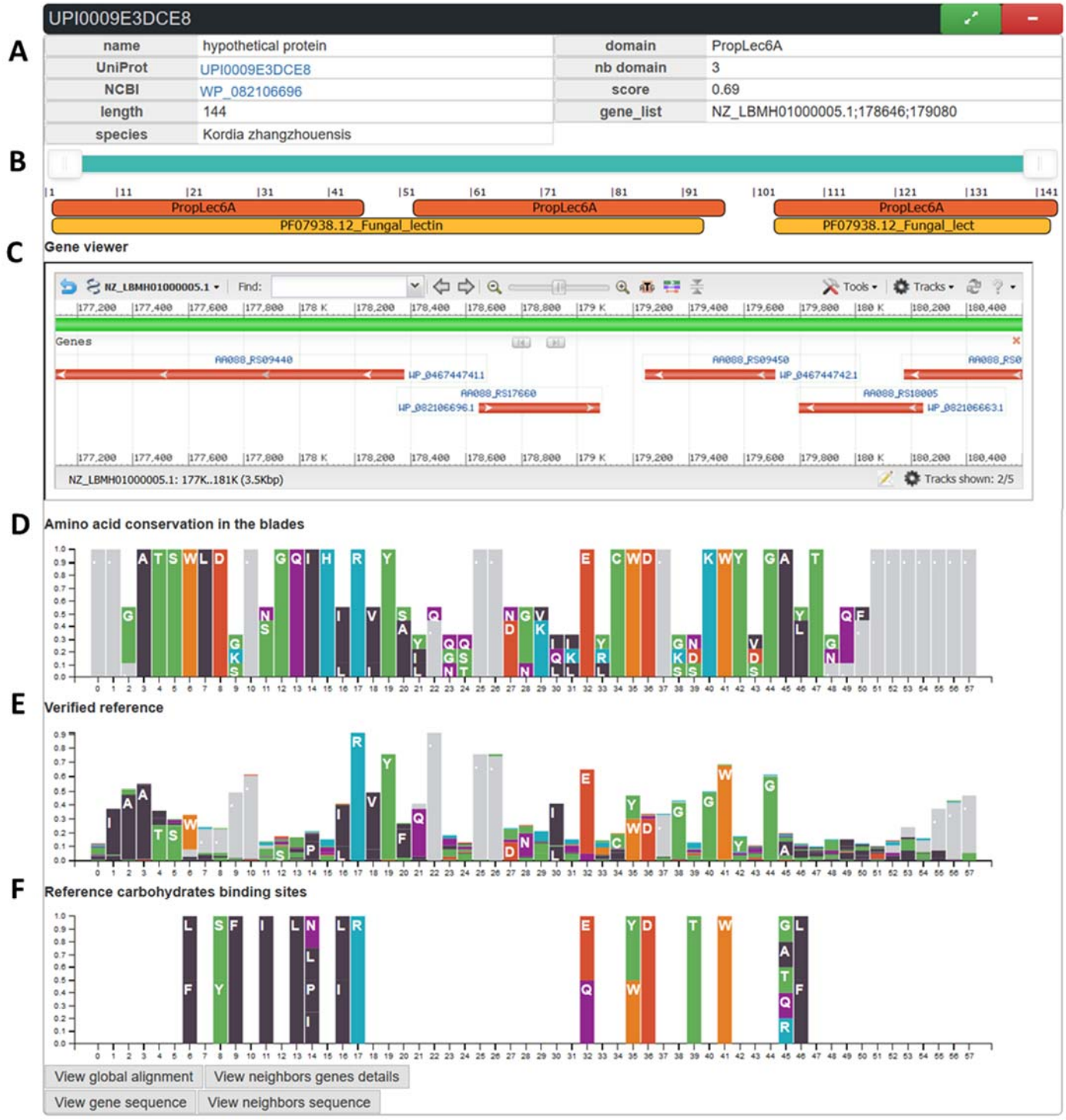


Figure 3. *Example of an entry for a predicted lectin sequence in the PropLec database. A. Information about the sequence and species. B. 2D sequence feature viewer with localisation of the predicted blades and potential Pfam domains, with a drag and drop button to zoom in on the sequence. C. NCBI viewer for the corresponding gene and chromosome D. Barchart of the amino acid conservation between blades of the predicted protein. E. Amino acid conservation of the reference blade. F Amino acids involved in carbohydrate recognition in the reference blade.*

Occurrence of PropLecs in the living kingdom

The distribution of PropLecs in the tree of life can be analysed through the interface, with both sunburst and tree representation available (Figure 4). No significant bias is observed for the two main branches with 75% PropLecs sequences of bacterial origin and 24% of eukaryotes (the non-redundant NCBI database reports 76% and 21% sequences from bacteria and eukaryotes, respectively (O'Leary et al., 2016)). Only 28 proteins have been predicted in Archae (0.7%) that appear to be under-represented. Interestingly, no PropLec sequence is identified in virus genomes with the exception of a synthetic one used in phage display (Yadid and Tawfik, 2007) that has been therefore filtered out of the database. Bias occurs in eukaryote subgroups, with an over-representation of PropLecs in fungi genomes. As much as 11% of PropLecs are in fungi (404 proteins) while fungal sequences represent less than 3% of the RefSeq database. Plant genomes do not contain any PropLec sequence and they are rare in algae. Similarly, we could not identify any sequence in birds and mammals, which comforts the hypothesis that PropLecs play mainly a role in innate immunity that has been partially replaced by acquired immunity in more evolved organisms. It should be noted that putative PropLecs were proposed in human as members of the PropLec6B family and referred as leukolectin or hTectonins (Low et al., 2009). However, the sequence of human leukolectin (GenBank Accession: ACM77812) is 100% identical with the salmon tectonin. Searching the human genome (BLAT search on UCSC browser: https://genome.ucsc.edu/cgi-bin/hgBlat) with the leukolectin gene sequence did not return any hit. Altogether, these observations point to a probable contamination problem during RNA sequencing. The other putative human tectonins (Low et al., 2009) have not been demonstrated to fold as β -propellers and they do not present any of the conserved motif that we identified.

The different families of PropLecs do not occur equally in Nature (Figure 4). All five families are present in bacteria and eukaryotes, albeit with very different populations. PropLec5A (tachylectin) is an exclusive animal lectin, identified in invertebrates (Cnidaria and crabs), xenops and fishes. Archaea and bacteria genomes contain mostly PropLec7A, the fucose/galactose lectin recently identified from several *Photorhabdus* species. Eukaryotes genomes contain all five families of PropLecs but the distribution is different in fungi, where a majority of PropLec6A (the AAL lectin) is observed, in contrast with animals, that contain mostly PropLec6B (tectonin).

Since many pathogens use lectins for recognition and adhesion to host tissues, such lectins are considered as targets for the development of anti-adhesive compounds (Imberty, 2011; (Sharon, 2006) and their identification may have a therapeutic relevance. A list of microorganisms that cause diseases in human is available from NIH NIAID Emerging Infectious Pathogens. Afilter in the main page of the database allows for selecting only PropLecs in such organisms. We identified PropLecs in more than 20 pathogenic microorganisms, and the ones that are more threatening for human health or characterized as emergent threats are listed in Table 1. Among them, only two lectins, AFL in *Aspergillus fumigatus* and BambL in *Burkholderia ambifaria,* have been fully characterized (Audfray et al., 2012; (Houser et al., 2013). AFL was demonstrated to be located on the fungal conidia and to play a role in host defence by interacting with the inflammation response (Houser et al., 2013; (Kerr et al., 2016). The lectins listed in Table 1 would therefore be of high interest for the understanding of pathogen-host interactions.

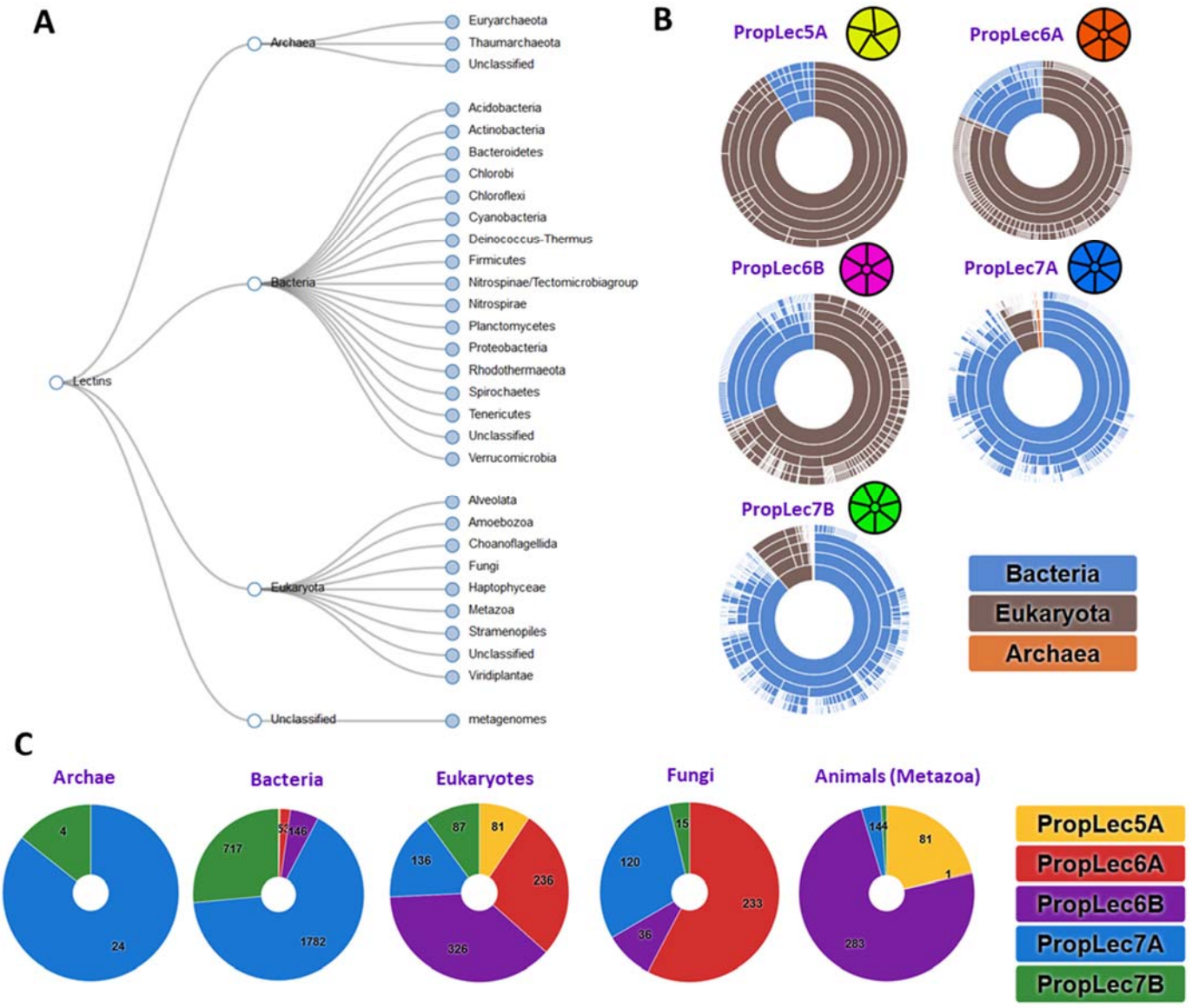


Figure 4. *Occurrence of PropLec sequence in genomes. A. Searchable tree in the PropLec database. B. Sunburst statistics for the origin in each PropLec family. C. Sunburst statistic for PropLec families in selected domains of life.*

Table 1: Identification of PropLecs in the genomes of pathogenic micro-organisms.

	species	propfamily	disease	PMID
Gram+ bacteria	*Bacillus cereus*	PropLec7A	Food poisoning	23488744
	Clostridium botulinum	PropLec7A	Botulism, food poisoning	28800585
	C. tetani	PropLec7A	Tetanus	25638019
	Nocardia mikamii	PropLec7B	Opportunistic lung infection	19915112
Gram- bacteria	*Burkholderia ambifaria, B. cepacia*	PropLec6A	Opportunistic lung infection	22170069
	B.ubonensis	PropLec6A, PropLec6B	Opportunistic lung infection	27303639
	Coccidioides immitis	PropLec7A	"Valley fever", meningitis	28597822
	Ralstonia pickettii	PropLec6A	Emerging nosocomial infection	16337309
Fungi	*Aspergillus fumigatus*	PropLec6A	Aspergillosis, lung infection	10194462
	Fonsecaea erecta	PropLec6A	Chromomycosis, skin infection	11204152
	Phialophora attae	PropLec6A, PropLec7A	Chromomycosis, skin infection	26586868
	Trichophyton tonsurans	PropLec6A, PropLec7A	Dermatophytosis, scalp infection	23053563
Oomycetes	*Pythium insidiosum*	PropLec6B	Pythiosis, multisystemic infection	20800978

Prediction of topology and modular associations

Since the family motifs have been defined to correspond to one blade length, the search procedure can predict the number of blades that are conserved in the sequences. In the database, the number of blades varies from 1 to 26, but most sequences are predicted to include 6 or 7 blades, in agreement with the known 3D-structures (Figure 5). A significant number of sequences show a lower number of blades than expected, such as 6 blades for PropLec7A and PropLec7A, which is explained by variation in amino acids in one blade of the protein (degeneration of sequences). Larger number of blades generally corresponds to the tandem repeat of several propellers in the sequence, explaining the highest occurrence for 12 and 18 blades, corresponding to 2 or 3 propellers in the same sequence.

Carbohydrate-binding domains or modules (CBM) are related to lectins, since they bind to carbohydrate, but they are usually monovalent. CBMs act as substrate binding and can be combined with carbohydrate-active enzymes (Boraston et al., 2004). It is therefore of interest to analyse the modular architecture of the predicted PropLecs to check if they could also associate with enzyme-active domains. The database interface has been designed to search for the occurrence of such modules. Twenty-nine distinct domains not overlapping with PropLecs domains were identified, some of them are listed in Table 2. Glycosyl hydrolases, or other enzymes acting on carbohydrates are often attached to PropLecs, which are then supposed to act as substrate recognition

modules. Other enzymes are also identified such as peptidases or peroxidases. Interestingly, PropLec can also tandem with other carbohydrate-binding proteins, such as C-type lectins.

Table 2: Selection of functional domains identified with PropLecs with a modular design on the same peptide.

Family	Architecture	*Species*	Associated protein	Pfam
PropLec6A		*Aspergillus lentulus*	Aldo-keto reductase yakc	PF00248
PropLec6A		*Actinopolymorpha singaporensis*	Cysteine peptidase	PF00112
PropLec6B		*Branchiostoma belcheri* (lancelet)	C-type lectin	PF00059
PropLec6B		*Branchiostoma belcheri*	Animal haem peroxidase	PF03098
PropLec7A		*Streptomyces sp*	Melibiase 2 (galactosidase)	PF16499
PropLec7A		*Frigoribacterium sp*	Arabinosidase, galactosidase	PF04616
PropLec7B		*Streptomyces davaonensi*	Peptidase S8	PF00082
PropLec7B		*Scytonema hofmannii*	Chitinase	PF00704

Occurrence of novel assembly fold for β-propeller

As described above, β-propellers are generally consisting of one peptide presenting a tandem-repeat. The only exception occurred in the PropLec6A family: these lectins have been characterized in three fungi (see Table S1) with six blade repeats for a domain approximately 300 amino acid-long, but also in bacteria with two blade repeats in a 90 amino acid domain, that trimerizes to form the same 6-blade propeller (Audfray et al., 2012; (Kostlanová et al., 2005). This is the only case of natural β-propeller assembled by oligomerization. The bimodal distribution of blade numbers in PropLec6A family, with maxima at 6-blade and 2-blade is shown in Figure 5 and in supplemental information (Figure S8). However, from the graph distribution, we predicted that 3-blade domains could also exist, which would correspond to a β-propeller formation by dimerization that was never observed before.

The predicted 3-blade sequences of PropLec6A were therefore analysed to select those with a high similarity score, an approximate size of 150 amino acids (three repeats) and correct gene start and ending. Four sequences were selected, and annotated as 3-blades lectins : UPI0009E3DCE8 in *Kordia zhangzhouensis* (Du et al., 2015) and A0A2T6C3M6 in *K. periserrulae* (Choi et al., 2011), bacteria from freshwater and marine environment, respectively, as well as A0A1V6N7V4 in *Penicillium polonicum* and A0A124GTL0 in *P. freii*, two filamentous fungi responsible for the production of mycotoxins (Mills et al., 1995). The alignment of blade sequences of the *K. zhangzhouensis* lectin (KozL) and *P. polonicum* one (PepL) are displayed in Figure

6. Both proteins present conservation of all the amino acids involved in fucose binding and can therefore be annotated as putative lectins. Analysis of the identity matrix at the blade level (Figure S9) demonstrates a strong conservation of blades within the KozL sequence (55 to 62% identity), higher than in the other sequences of PropLec6A group. Internal conservation is low within PepL blade sequences (9 to 25%). Blade sequences of KozL present stronger similarity to bacterial lectin (BambL) than to fungal ones, as expected.

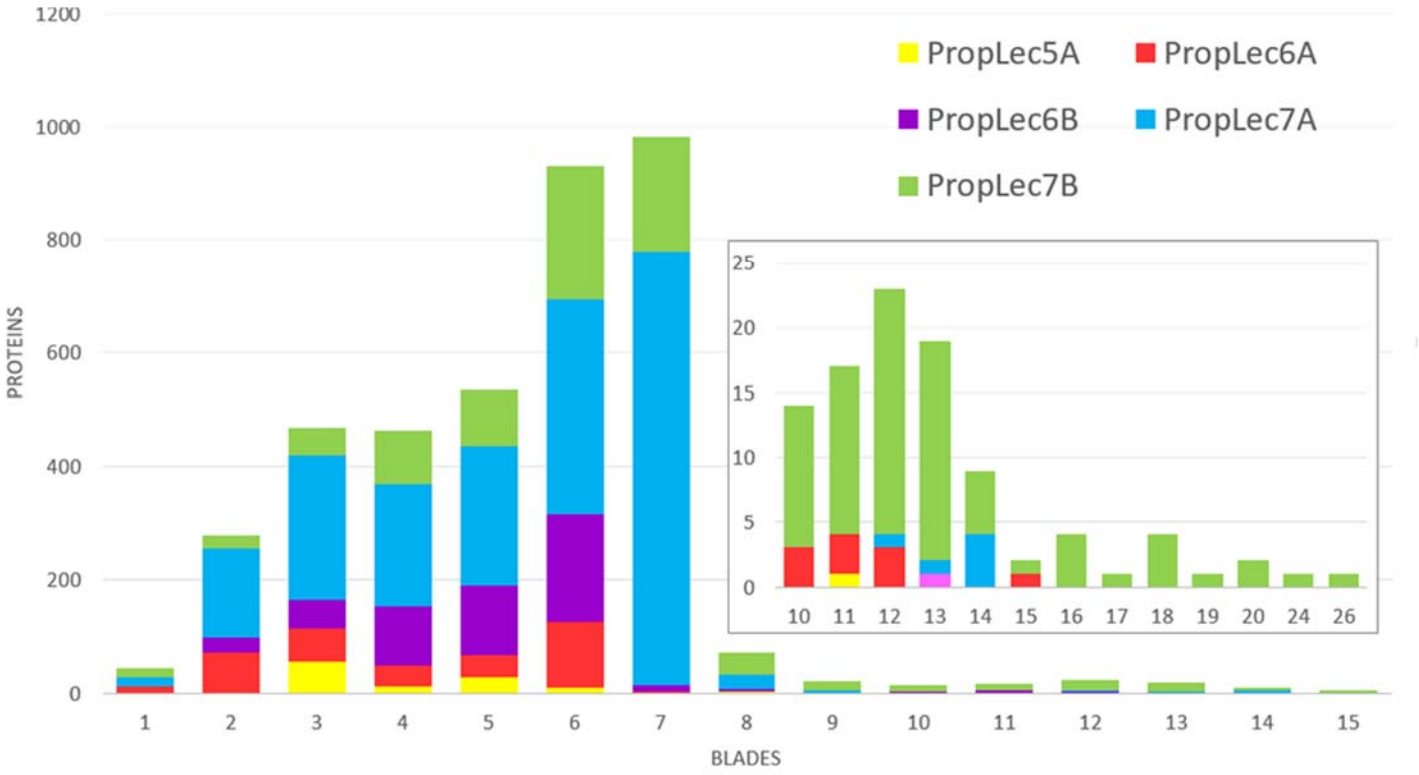


Figure 5. *Analysis of number of adjacent blades in predicted PropLecs*

The genes coding for KozL and PepL were synthetized after appropriate codon optimization and expressed in *Escherichia coli.* Although PepL formed inclusion bodies, Kozl was obtained in a soluble form with expected size of 16 kDa. It was produced and purified on a carbohydrate-affinity column as previously described for RSL (Kostlanová et al., 2005). The protein is fully functional with very strong affinity for fucose as determined by titration microcalorimetry (figure 7A). A dissociation constant (Kd) of 0.86 µM is obtained for methyl-α-L-fucoside (MeFuc), in agreement with affinity previously measured with RSL and BambL. Titration microcalorimetry is also suited for measuring the molar ration of ligand to protein, and a value of 3.2 was obtained, confirming the presence of three active binding sites on each KozL protomer.

Crystals of KozL complexed with MeFuc were obtained by co-crystallisation. Diffraction data were collected on beam line PX1 at Soleil synchrotron to 1.55 Å resolution in $P22_12_1$ space group. Attempts to solve the structure by molecular replacement method were not successful. A methyl-α-L-selenofucoside derivative (SeFuc), synthetized as previously described (Kostlanová et al., 2005), was cocrystallised for SAD phasing

and data were collected at 2.65 Å resolution. Statistics for both complexes are described in Table S2 (sup. Info), and only the structure of KozL with αMeFuc was fully refined and described herein.

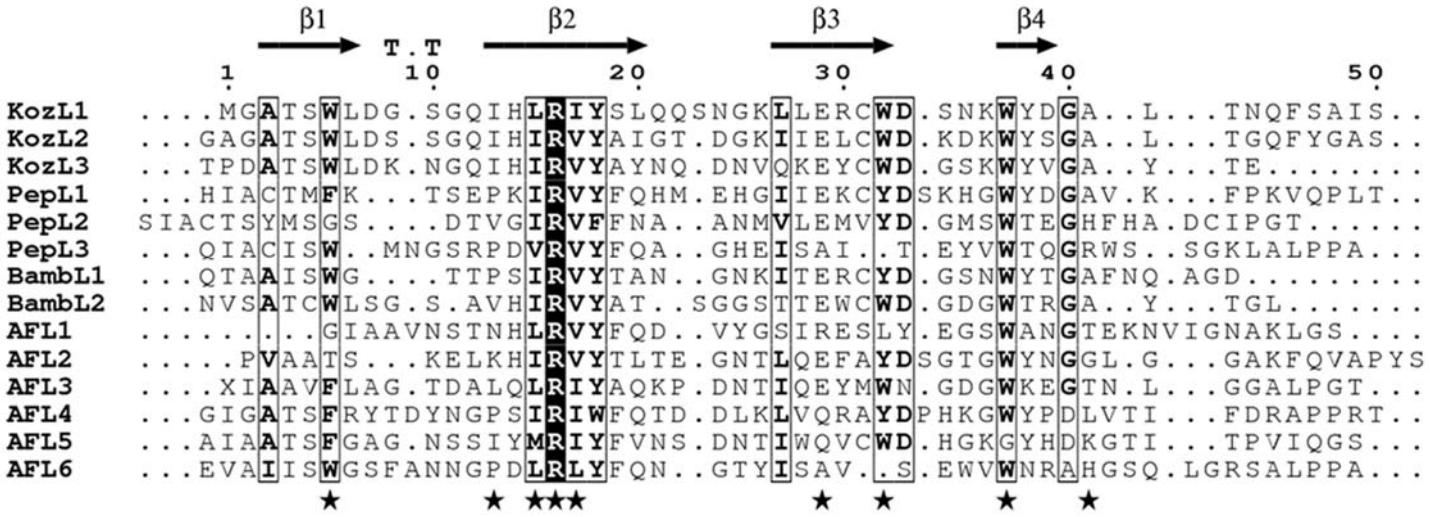


Figure 6. *Alignment of sequences for the 3-blade proteins KoZL and PepL with selected 2-blade and 6-blade members of PropLec6A family*

The asymmetric unit contains four monomers of KozL assembled in two β-propellers, and two additional monomers that form another dimer of β-propeller when applying the 2-fold symmetry of the space group. The tetramer formed by chains ABCD (Figure 7C) presents an interface of 13 000 $Å^2$ as calculated by PISA (PDBe.org). The oligomeric state in solution was confirmed by analytical ultracentrifugation (Figure S10). The dominant peak of KozL (90% of the total signal) has a sedimentation coefficient of 4.4 S (4.6 S at standard conditions) and corresponds to the tetrameric species with a moderately elongated shape (f/f0 = 1.3).

Electron density clearly indicates the presence of 18 MeFuc residues (three per monomers), with few additional molecules of crystallizing agents (2-methyl-2,4-pentanediol, ethanediol and nonaethylene glycol) and water molecules. The binding sites are located between the blades, with two intramolecular and one intermolecular sites. The amino acids involved in fucose binding are fully conserved in the three blades and are very similar to what has been observed in BambL and RSL. Fucose is stabilized by hydrogen bonds to side chains of Arg (16/67/117 for the three sites), Glu (29/79/129) and Trp (87/137/37*) and to main chain of Ala (41/91/141) and by hydrophobic interactions with another Trp indol ring (82/132/32*) and Ile (64/114/13*) (Figure 7D).

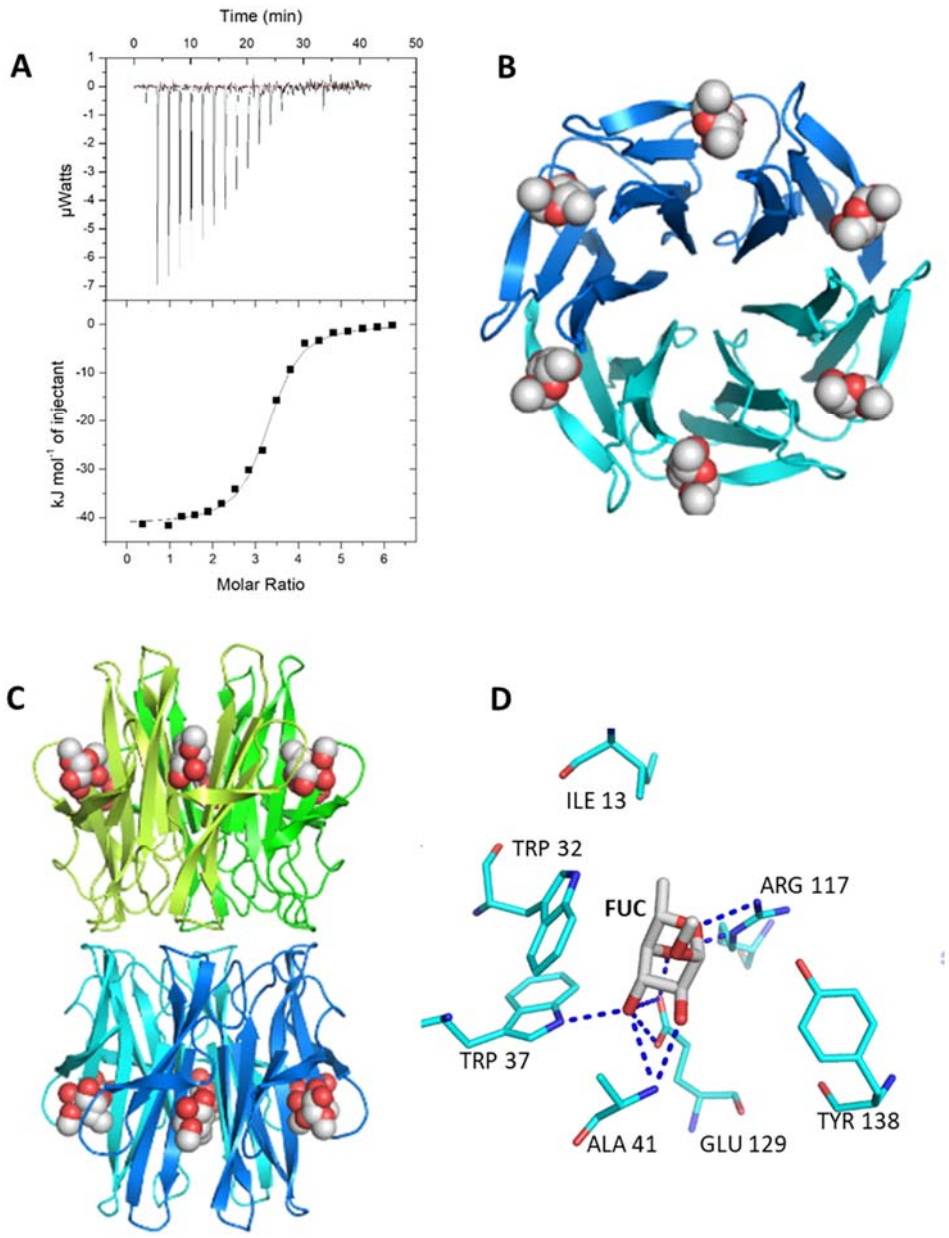


Figure 7. *Function and structure of KozL. A. ITC data with thermogram (top) and integrated peaks (bottom), B. Dimer of KozL assembles in a β-propeller structure with each chain represented by a shade of blue, and MeFuc ligand represented by spheres. C. Tetramer of KozL assembled in dimeric association of β-propellers. D. The fucose binding site in one of the three binding sites with hydrogen bonds is represented by blued dashed lines.*

Discussion and Conclusion

The blade signatures that have been designed in the present study allowed for the identification of new PropLec sequences in a wide collection of genomes. The KozL protein, that was not annotated previously as a lectin, provided an experimental validation of our approach. The protein function was confirmed by measuring its strong affinity for fucose. Apart from the conservation of binding sites and the 4-strand β-sheet repeats, KozL is rather different from other PropLec6A structures, especially the loops on both side of the donuts (Figure 8A). The β-propeller of KozL is formed by dimerization of two 3-blade domains, which has

never been reported before for a natural protein, but could only be obtained as artificial high-symmetry engineered pizza protein (Voet et al., 2014). The validation of such evolution intermediate in bacteria is of high interest. As illustrated in Figure 8B, in the PropLec6A family, the donut shape of the β-propeller can be formed by dimerization (KoZL), trimerization (for bacterial lectins BambL/RSL) or can be monomeric (for fungal lectins AFL/AAL/AOL). Evolution used symmetry in a very efficient way to build the same objects from different numbers of domain repeats.

.

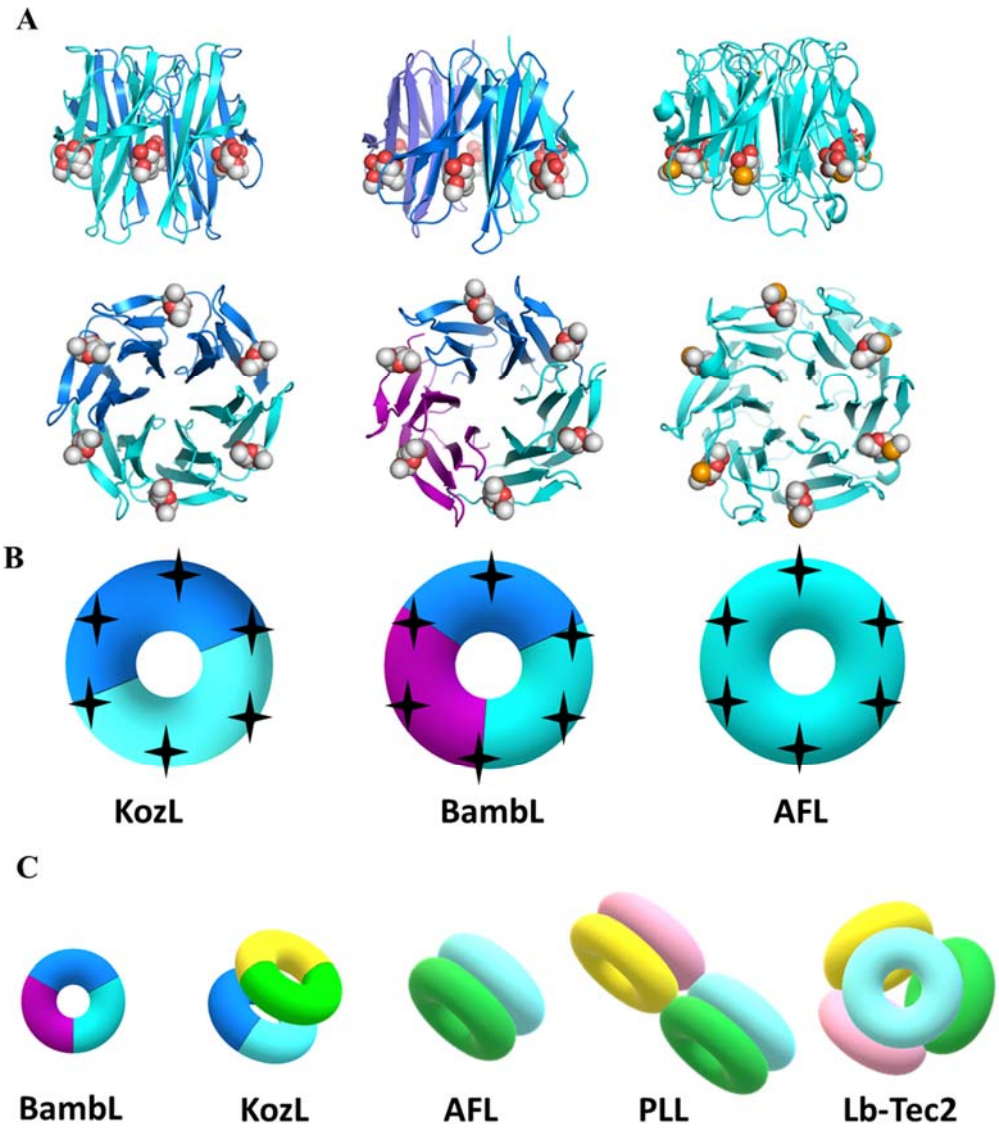


Figure 8. *A. Different oligomerisation modes for the creation of6-propeller structure in PropLec6A family. In the donut schematic representation, the stars denote glycan binding sites. B. Different assemblies of β-propellers observed in PropLec 3D-structures.*

Furthermore, β-propeller lectins have the ability to form supra-molecular assemblies by oligomerisation of the donut shapes, resulting in the different organisation of carbohydrate binding sites in space. Figure 8C schematizes the different oligomerization modes that have been observed so far. Some PropLecs such as BambL in the PropLec6A family, but also PVL in the PropLec7B family, occur as single β-propeller in

solution, while others, such as KoZL and AFL (PropLec6A) and PHL (PropLec7A) are in the form of back-to-back propellers, that present binding sites in opposite directions. The tetrameric association of β-propellers is observed in PLL (PropLec7A), with stabilization by disulphide bridges, and in fungal tectonin Lb-Tec2 (PropLec6B) where four β-propellers form a round-shaped virus-like assembly with 24 carbohydrate binding sites evenly partitioned on the surface.

In this study, we identified almost 4000 sequences of putative PropLecs and we validated our approach with the experimental characterization of a novel structure with strong interest for evolution. Clearly, the wealth of new sequences identified opens the way to research on the evolution of β -propeller folds. Furthermore, the donut shape of PropLecs is a very robust protein structure that can be used as scaffold for building multivalent protein structures and PropLecs from pathogenic organisms are likely to be involved in host-glycan recognition and can be used as target of anti-infectious compounds.

Acknowledgements. The authors acknowledge support by the ANR PIA Glyco@Alps (ANR-15-IDEX-02) and the Alliance Campus Rhodanien Co-Funds (http://campusrhodanien.unige-cofunds.ch). A.K., M.W. and A.I. are grateful for the support of EEC Bison project (H2020-TWINN-2015-692068) for financing the stay of A.K. in Grenoble. A.K. and M.W. acknowledge the CEITEC 2020 project (LQ1601) from MEYS CR. We acknowledge the CF Biomolecular Interactions and Crystallization supported by the CIISB research infrastructure (LM2015043 funded by MEYS CR) for their support with obtaining AUC data. We are grateful to synchrotron SOLEIL (Saint Aubin, France) for access and technical support at beamline PROXIMA 1 and for the help of Pierre Legrand.

Author contributions. F.B. developed the database and build the interface under the guidance of F.L., A.I. and S.P. A. K. produced the lectin and characterized it under guidance from A.V. M.L. synthetized the selenoligand. A.V. solved the crystal structure and refined it. M.W. performed and analysed ultracentrifugation experiments. F.B., A.I and F.L. wrote the manuscript and prepared figures with the critical input of S.P., A.V. and M.W.

Declaration of interest. The authors declare no competing interests

References

Agirre, J., Iglesias-Fernandez, J., Rovira, C., Davies, G.J., Wilson, K.S., and Cowtan, K.D. (2015). Privateer: software for the conformational validation of carbohydrate structures. Nat. Struct. Mol. Biol. *22*, 833-834.

Arnaud, J., Claudinon, J., Tröndle, K., Trovaslet, M., Larson, G., Thomas, A., Varrot, A., Römer, W., Imberty, A., and Audfray, A. (2013). Reduction of lectin valency drastically changes glycolipid dynamics in membranes, but not surface avidity. ACS Chem. Biol. *8*, 1918-1924.

Arnaud, J., Tröndle, K., Claudinon, J., Audfray, A., Varrot, A., Römer, W., and Imberty, A. (2014). Membrane deformation by neolectins with engineered glycolipid binding sites. Angew. Chem. Int. Ed. *53*, 9267–9270.

Audfray, A., Beldjoudi, M., Breiman, A., Hurbin, A., Boos, I., Unverzagt, C., Bouras, M., Lantuejoul, S., Coll, J.L., Varrot, A.*, et al.* (2015). A recombinant fungal lectin for labeling truncated glycans on human cancer cells. PLoS One *10*, e0128190.

Audfray, A., Claudinon, J., Abounit, S., Ruvoën-Clouet, N., Larson, G., Smith, D.F., Wimmerová, M., Le Pendu, J., Römer, W., Varrot, A.*, et al.* (2012). The fucose-binding lectin from opportunistic pathogen *Burkholderia ambifaria* binds to both plant and human oligosaccharidic epitopes. J. Biol. Chem. *287*, 4335-4347.

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

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

Bonnardel, F., Mariethoz, J., Salentin, S., Robin, X., Schroeder, M., Perez, S., Lisacek, F., and Imberty, A. (in press). UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. Nucleic Acids Res.

Boraston, A.B., Bolam, D.N., Gilbert, H.J., and Davies, G.J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem. J. *382*, 769-781.

Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R*., et al.* (2015). Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. *43*, D36-42.

Capaldi, S., Faggion, B., Carrizo, M.E., Destefanis, L., Gonzalez, M.C., Perduca, M., Bovi, M., Galliano, M., and Monaco, H.L. (2015). Three-dimensional structure and ligand-binding site of carp fishelectin (FEL). Acta Crystallogr. D. Biol. Crystallogr. *71*, 1123-1135.

Chaikuad, A., Knapp, S., and von Delft, F. (2015). Defined PEG smears as an alternative approach to enhance the search for crystallization conditions and crystal-quality improvement in reduced screens. Acta Crystallogr. D. Biol. Crystallogr. *71*, 1627-1639.

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

Chen, C.K., Chan, N.L., and Wang, A.H. (2011). The many blades of the beta-propeller proteins: conserved but versatile. Trends Biochem. Sci. *36*, 553-561.

Choi, A., Oh, H.M., Yang, S.J., and Cho, J.C. (2011). Kordia periserrulae sp. nov., isolated from a marine polychaete Periserrula leucophryna, and emended description of the genus Kordia. Int J Syst Evol Microbiol *61*, 864-869.

Cioci, G., Mitchell, E.P., Chazalet, V., Debray, H., Oscarson, S., Lahmann, M., Gautier, C., Breton, C., Pérez, S., and Imberty, A. (2006). b-Propeller crystal structure of *Psathyrella velutina* lectin: An integrin-like fungal protein interacting with monosaccharides and calcium. J. Mol. Biol. *357*, 1575-1591.

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr. D. Biol. Crystallogr. *62*, 1002-1011.

Cowtan, K. (2010). Recent developments in classical density modification. Acta Crystallogr. D. Biol. Crystallogr. *66*, 470-478.

Dawson, N.L., Sillitoe, I., Lees, J.G., Lam, S.D., and Orengo, C.A. (2017). CATH-Gene3D: Generation of the resource and its use in obtaining structural and functional annotations for protein sequences. Methods Mol. Biol. *1558*, 79-110.

Du, J., Liu, Y., Lai, Q., Dong, C., Xie, Y., and Shao, Z. (2015). Kordia zhangzhouensis sp. nov., isolated from surface freshwater. Int J Syst Evol Microbiol *65*, 3379-3383.

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

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D. Biol. Crystallogr. *66*, 486-501.

Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., and Eddy, S.R. (2015). HMMER web server: 2015 update. Nucleic Acids Res. *43*, W30-38.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A*., et al.* (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*, D279-285.

Fulop, V., and Jones, D.T. (1999). Beta propellers: structural rigidity and functional diversity. Curr. Opin. Struct. Biol. *9*, 715-721.

Gorrec, F. (2009). The MORPHEUS protein crystallization screen. J. Appl. Crystallogr. *42*, 1035-1042.

Goyard, D., Baldoneschi, V., Varrot, A., Fiore, M., Imberty, A., Richichi, B., Renaudet, O., and Nativi, C. (2018). Multivalent glycomimetics with affinity and selecivity towards fucose-binding receptors from emerging pathogens. Bioconjug. Chem. *29*, 83−88

Heger, A., and Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. Proteins *41*, 224-237.

Houben, K., Marion, D., Tarbouriech, N., Ruigrok, R.W., and Blanchard, L. (2007). Interaction of the C-terminal domains of sendai virus N and P proteins: comparison of polymerase-nucleocapsid interactions within the paramyxovirus family. J. Virol. *81*, 6807-6816.

Houser, J., Komarek, J., Kostlanova, N., Cioci, G., Varrot, A., Kerr, S.C., Lahmann, M., Balloy, V., Fahy, J.V., Chignard, M.*, et al.* (2013). A soluble fucose-specific lectin from *Aspergillus fumigatus* conidia - Structure, specificity and possible role in fungal pathogenicity. PloS ONE *8*, e83077.

Imberty, A. (2011). Bacterial lectins and adhesins: structures, ligands and functions. In E-book : Synthesis and Biological Applications of Glycoconjugates, O. Renaudet, and N. Spinelli, eds. (Bentham Science Publishers Ltd), pp. 3-11.

Jancarikova, G., Herczeg, M., Fujdiarova, E., Houser, J., Kover, K.E., Borbas, A., Wimmerova, M., and Csavas, M. (2018). Synthesis of alpha-l-Fucopyranoside-Presenting Glycoclusters and Investigation of Their Interaction with Photorhabdus asymbiotica Lectin (PHL). Chemistry.

Jancarikova, G., Houser, J., Dobes, P., Demo, G., Hyrsl, P., and Wimmerova, M. (2017). Characterization of novel bangle lectin from *Photorhabdus asymbiotica* with dual sugar-binding specificity and its effect on host immunity. PLoS Pathog *13*, e1006564.

Jawad, Z., and Paoli, M. (2002). Novel sequences propel familiar folds. Structure *10*, 447-454.

Kabsch, W. (2010). Xds. Acta Crystallogr. D. Biol. Crystallogr. *66*, 125-132.

Kawabata, S., and Iwanaga, S. (1999). Role of lectins in the innate immunity of horseshoe crab. Dev. Comp. Immunol. *23*, 391-400.

Kerr, S.C., Fischer, G.J., Sinha, M., McCabe, O., Palmer, J.M., Choera, T., Lim, F.Y., Wimmerova, M., Carrington, S.D., Yuan, S.*, et al.* (2016). FleA Expression in Aspergillus fumigatus Is Recognized by Fucosylated Structures on Mucins and Macrophages to Prevent Lung Infection. PLoS Pathog *12*, e1005555.

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

Kostlanová, N., Mitchell, E.P., Lortat-Jacob, H., Oscarson, S., Lahmann, M., Gilboa-Garber, N., Chambat, G., Wimmerová, M., and Imberty, A. (2005). The fucose-binding lectin from *Ralstonia solanacearum*: a new type of -propeller architecture formed by oligomerisation and interacting with fucoside, fucosyllactose and plant xyloglucan. J. Biol. Chem. *280*, 27839-27849.

Kumar, A., Sykorova, P., Demo, G., Dobes, P., Hyrsl, P., and Wimmerova, M. (2016). A novel fucose-binding lectin from P*hotorhabdus luminescens* (PLL) with an unusual heptabladed beta-propeller tetrameric qtructure. J. Biol. Chem. *291*, 25032-25049.

Lis, H., and Sharon, N. (1998). Lectins: Carbohydrate-specific proteins that mediate cellular recognition. Chem. Rev. *98*, 637-674.

Liu, W., Han, G., Yin, Y., Jiang, S., Yu, G., Yang, Q., Yu, W., Ye, X., Su, Y., Yang, Y.*, et al.* (2018). AANL (Agrocybe aegerita lectin 2) is a new facile tool to probe for O-GlcNAcylation. Glycobiology *28*, 363-373.

Low, D.H., Ang, Z., Yuan, Q., Frecer, V., Ho, B., Chen, J., and Ding, J.L. (2009). A novel human tectonin protein with multivalent beta-propeller folds interacts with ficolin and binds bacterial LPS. PLoS One *4*, e6260.

Machida, T., Novoa, A., Gillon, É., Zheng, S., Claudinon, J., Eierhoff, T., Imberty, A., Römer, W., and Winssinger, N. (2017). Dynamic cooperative glycan assembly blocks binding of bacterial lectins to epithelial cells Angew. Chem. Int. Ed. *56*, 6762-6766.

Machon, O., Baldini, S.F., Ribeiro, J.P., Steenackers, A., Varrot, A., Lefebvre, T., and Imberty, A. (2017). Recombinant fungal lectin as a new tool to investigate O-GlcNAcylation processes. Glycobiology *27*, 123-128.

Maguire, E., Rocca-Serra, P., Sansone, S.-A., and Chen, M. (2014). Redesigning the sequence logo with glyph-based approaches to aid interpretation. Paper presented at: Eurographics Conference on Visualization (EuroVis).

McCoy, A.J. (2007). Solving structures of protein complexes by molecular replacement with Phaser. Acta Crystallogr. D. Biol. Crystallogr. *63*, 32-41.

Mills, J.T., Seifert, K.A., Frisvad, J.C., and Abramson, D. (1995). Nephrotoxigenic Penicillium species occurring on farm-stored cereal grains in western Canada. Mycopathologia *130*, 23-28.

Murshudov, G.N., Skubak, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F., and Vagin, A.A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr. D. Biol. Crystallogr. *67*, 355-367.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D.*, et al.* (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733-745.

Ren, X.M., Li, D.F., Jiang, S., Lan, X.Q., Hu, Y., Sun, H., and Wang, D.C. (2015). Structural basis of specific recognition of non-reducing terminal N-acetylglucosamine by an *Agrocybe aegerita* Lectin. PLoS One *10*, e0129608.

Richard, N., Marti, L., Varrot, A., Guillot, L., Guitard, J., Hennequin, C., Imberty, A., Corvol, H., Chignard, M., and Balloy, V. (2018). Human bronchial epithelial cells inhibit *Aspergillus fumigatus* germination of extracellular conidia via FleA recognition. Sci. Rep. *8*, 15699

Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. *42*, W320-324.

Schneider, T.R., and Sheldrick, G.M. (2002). Substructure solution with SHELXD. Acta Crystallogr. D. Biol. Crystallogr. *58*, 1772-1779.

Schuck, P. (2000). Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. Biophys. J. *78*, 1606-1619.

Sharon, N. (2006). Carbohydrates as future anti-adhesion drugs for infectious diseases. Biochim. Biophys. Acta *1760*, 527-537.

Sheldrick, G.M. (2010). Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. Acta Crystallogr. D Biol. Crystallogr. *66*, 479-485.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Sommer, R., Makshakova, O.N., Wohlschlager, T., Hutin, S., Marsh, M., Titz, A., Kunzler, M., and Varrot, A. (2018). Crystal structures of fungal tectonin in complex with O-methylated glycans suggest key role in innate immune defense. Structure *26*, 391-402.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., and UniProt, C. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926-932.

Voet, A.R., Noguchi, H., Addy, C., Simoncini, D., Terada, D., Unzai, S., Park, S.Y., Zhang, K.Y., and Tame, J.R. (2014). Computational design of a self-assembling symmetrical beta-propeller protein. Proc. Natl. Acad. Sci. U. S. A. *111*, 15102-15107.

Wimmerova, M., Mitchell, E., Sanchez, J.F., Gautier, C., and Imberty, A. (2003). Crystal structure of fungal lectin: Six-bladed b-propeller fold and novel recognition mode for *Aleuria aurantia* lectin. J. Biol. Chem. *278*, 27059-27067.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G., McCoy, A.*, et al.* (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D. Biol. Crystallogr. *67*, 235-242.

Wohlschlager, T., Butschi, A., Grassi, P., Sutov, G., Gauss, R., Hauck, D., Schmieder, S.S., Knobel, M., Titz, A., Dell, A.*, et al.* (2014). Methylated glycans as conserved targets of animal and fungal innate defense. Proc. Natl. Acad. Sci. U. S. A. *111*, E2787-2796.

Yadid, I., and Tawfik, D.S. (2007). Reconstruction of functional beta-propeller lectins via homo-oligomeric assembly of shorter fragments. J. Mol. Biol. *365*, 10-17.

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

STAR Methods

Key resources table

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Bacterial Strains		
BL21(DE3)	Merk-Novagen	69450
Chemicals and Recombinant Protein		
Cryoloops	Molecular Dimensions Ltd	MD7-133 / MD7-134
a-methyl-fucoside	TCI	M1051
Mannose agarose column	Sigma-Aldrich	M6400-10ML
Morpheus I	Molecular Dimension Ltd	MD1-46
Deposited Data		
PDB accession code	This paper	6HTN
Recombinant DNA		
KozL and PepL genes	Eurofins Genomics	N/A
pET-TEV	(Houben et al., 2007)	N/A
Software and Algorithms		
XDS package	(Kabsch, 2010)	http://xds.mpimf-heidelberg.mpg.de/
CCP4i	(Winn et al., 2011)	http://www.ccp4.ac.uk/
ShelXC/D/E	(Sheldrick, 2010)	http://shelx.uni-goettingen.de/
PHASER	(McCoy, 2007)	https://github.com/secastel/phaser
Parrot	(Cowtan, 2010)	http://parrot.cgu.edu.tw/
Buccaneer	(Cowtan, 2006)	http://www.ysbl.york.ac.uk/~cowtan/buccaneer/buccaneer.html
Refmac 5.8	(Murshudov et al., 2011)	https://www.ccp4.ac.uk/
COOT	(Emsley et al., 2010)	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/
Privateer	(Agirre et al., 2015)	http://www.ccp4.ac.uk/
Clustal omega	(Sievers et al., 2011)	https://www.ebi.ac.uk/Tools/msa/clustalo/
ESPript 3.0	(Robert and Gouet, 2014)	http://espript.ibcp.fr/ESPript/ESPript/
PyMOL	Schrödinger	https://pymol.org
UniRef100	(Suzek et al., 2015)	https://www.uniprot.org/uniref/
HMMSEARCH	(Finn et al., 2015)	http://hmmer.org/
RADAR	(Heger and Holm, 2000)	https://www.ebi.ac.uk/Tools/pfa/radar/
SequenceLogoVis	(Maguire et al., 2014)	https://github.com/ISA-tools/SequenceLogoVis
MUSCLE	(Edgar, 2004)	https://www.ebi.ac.uk/Tools/msa/muscle/
NCBI gene viewer	(Brown et al., 2015)	https://www.ncbi.nlm.nih.gov/projects/sviewer/
PFAM-A	(Finn et al., 2016)	https://pfam.xfam.org/
CATH-Gene3D	(Dawson et al., 2017)	http://www.cathdb.info/
RefSeq	(O'Leary et al., 2016)	https://www.ncbi.nlm.nih.gov/refseq/

Database construction

The features, taxonomy and identified domains of the predicted lectins (from UniRef100 database release of 12/09/18 with HMMSEARCH version 3.2) are stored in distinct tables to preserve the reactivity of the web platform and avoid computing information on the run. Predicted protein information was collected from UniProt and corresponding RefSeq entry including data on the related 1889 species. 18545 Pfam domains from 194 Pfam families have also been identified on the predicted proteins (from PFAM-A release of 15/10/18). The information is fetched using the predicted protein UniProt AC and PYTHON 3 scripts and the information files are loaded in the database with PHP scripts to facilitate maintenance and update.

Web module construction

The UniLectin web platform (https://www.unilectin.eu) is dedicated to the classification and curation of lectin structures (UniLectin3D module) and prediction of lectin sequences in genomes. The module dedicated to β-popeller lectins (PropLec) is available on the UniLectin platform. The interface has been developed with PHP version 7, Bootstrap version 3, and MySQL database version 5.6. Interactive graphics are developed in JavaScript based on D3JS libraries version 3 and dynamically generated to match the research criteria selected by the user.

Defining a similarity score

HMMER default scores is not comparable between predicted proteins with a different number of blades. To avoid the bias we defined a new similarity score. We use the alignment of the reference seed and the predicted blades performed with MUSCLE to define a quality score for each predicted protein. The similarity score is shown below. To control the bias due to variable numbers of predicted blades and different lengths of conserved blade domain in distinct families, the calculations are centered on those two criteria.

$$score = \sum_{i}^{1:len(MSA)} REF\_FREQ[MAX(AAi)] * PRED\_FREQ[MAX(AAi)]$$

$$Sim_{score} = \frac{score - MEAN(scores\_family\_nbblades)}{SD(scores\_family\_nbblades)}$$

Where REF_FREQ is the frequency of the most frequent amino acid (MAX (AAi)) at position i in the reference/seed domain and PRED_FREQ is the frequency of the most frequent amino acid at position i in the predicted protein. Score distributions by family are shown in Figure S7

Cloning of Kum and Kordia genes

The peptide sequences of the putative lectins KozL from *Kordia zhangzhouensis* (UPI0009E3DCE8) and PepL from *Penicillium polonicum* (A0A1V6N7V4) were translated into nucleotide sequence and were

synthesized after codon optimization for expression in Escherichia coli (Eurofins Genomics,Germany). The genes were introduced in the pET-TEV expression vector using NcoI and XhoI restriction sites and enzymes (New England Biolabs) (Houben et al., 2007). The pET-TEV-KozL and pET-TEV-PepL vectors were transformed into *E. coli* BL21(DE3)

Production and purification of KoZL

E. coli BL21 (DE3) cells harboring the plasmid pET-TEV-KozL were cultured in LB Broth medium with 30 $\mu g.mL^{-1}$ kanamycin at 37°C. When the culture reached an A600nm of 0.6-0.8, protein expression was induced with 0.1 mM isopropyl β-D-thiogalactoside. After 3 hours at 37 °C, cells were harvested by centrifugation at 6000 x g for 10 min and frozen at -20 °C. The pellet from 1 liter culture was resuspended in 30 mL of buffer A composed of 30 mM Tris-HCl pH 8.5 and 500mM NaCl prior addition of 1µl of Benzonase endonucluease (Sigma-Aldrich). After 15 min incubation at room temperature, cells were disrupted at a pressure of 1.9 kBar (Constant Cell Disruption System). The cell lysate was centrifuged at 24000 x g for 30 min at 4°C and the supernatant was filtered on 0.45 µm prior loading on a 10 ml mannose agarose column (Sigma-Aldrich) pre-equilibrated with buffer A. The column was washed with buffer A to remove unbound proteins and elution was performed with buffer A supplemented with 20 mM mannose. Purity was checked on 15% SDS-PAGE gel (15 %) before pooling of the appropriate fractions for dialysis with buffer B composed of 20 mM Tris pH 8.5, 250 mM NaCl. KozL was concentrated by centrifugation using a Vivaspin (3KDa, Sartorius) and stored in fridge for further use. For the long-term storage at -20°C, KozL was dialysed against ultrapure water, and lyophilised and kept in deep fridge. The total yield of purified KozL was 35 mg from 4.5 g of cells. All expression conditions tested for pET-TEV-KozL led to the formation of inclusion bodies to date.

ITC experiments

ITC experiments were performed with isothermal titration calorimeters (MicroCaliTC200; Malvern). Experiments were carried out at 25 ºC ± 0.1 ºC. Methyl-α-fucoside (TCI) solution was prepared in same buffer as KozL. The ITC cell contained 0.02 mM mM of KozL and the syringue 0.6 mM of MeFuc. The ligand was added by infection of 2 µL at intervals of 2 min while stirring at 1000 rpm. Prior to sample analysis, a control experiment, where the protein sample in the calorimeter cell was substituted by buffer, was performed, resulting in insignificant heat of dilution. Integrated heat effects were analysed by nonlinear regression using a one site binding model (Microcal Origin 7). The experimental data fitted to a theoretical titration curve gave the association constant Ka and the enthalpy of binding (ΔH). The experiments were performed in duplicates

Analytical ultracentrifugation experiments

Analytical ultracentrifugation experiments were performed using ProteomeLab XL-I analytical ultracentrifuge (Beckman Coulter) equipped with An-60 Ti rotor. Before analysis, lyophilized KoZL was dissolved in the experimental buffer (20 mM Tris, 150 mM NaCl, pH 7.4) and the buffer was used as an optical reference.

Sedimentation velocity experiments were conducted in titanium double-sector centerpiece cells (Nanolytics Instruments, Germany) loaded with 380 µL of both protein sample (0.02-0.17 $mg.mL^{-1}$) and reference solution. Data were collected using absorbance optics at 20 °C at a rotor speed of 50,000 rpm. Scans were performed at 280 nm at 4 min intervals and 0.003 cm spatial resolution in continuous scan mode. The partial specific volume of protein and the solvent density and viscosity were calculated from the amino acid sequence and buffer composition, respectively, using the software Sednterp (http://bitcwiki.sr.unh.edu). The sedimentation profiles were analyzed with the program Sedfit 15.01 (Schuck, 2000). Continuous c(s) distribution model was used for the analysis.

Crystallization and structure determination of KozL

Crystallization experiments were performed using the hanging-drop vapor-diffusion method with drops made of 1 µL of protein at 10 $mg.mL^{-1}$ in buffer B and 1 µL of reservoir solution at 19 °C. Commercial screens (Morpheus I and II, Clear Strategy Screen I and II and BCS; Molecular Dimensions Ltd) led to several crystallization hits. Cocrystallisation with 20 mM MeFuc using the solution 1-40 from Morpheus 1 (Gorrec, 2009) (120 mM alcohols, 100mM buffer pH 6.5, 37.5% MPD/PEG1000/PEG3350) results in parallelepiped crystals in 3-5 days. Thick rods were obtained after cocrystallisation of KozL incubated with 1 mM methyl-α-selenofucoside (SeFuc) (Kostlanová et al., 2005) in 35% PEG smear medium, 10% isopropanol optimized from hit from solution 2-38 of the BCS screen (Chaikuad et al., 2015). Crystals were directly mounted in a Litholoop (Molecular Dimensions Ltd) and flashed freezed in liquid nitrogen. Data were collected using a Pilatus 6M detector (Dectris Ltd) on the Proxima-1 beamline at SOLEIL, Saint Aubin, France. The data were processed using XDS (Kabsch, 2010). All further computing was performed using the CCP4 suite and interfaces (Winn et al., 2011) (Table S2).

Structural determination and refinement. The structure of KozL was solved by SAD method at the selenium peak (λ = 0.97914 Å) using the signal of the selenated ligand. ShelXC/D (Schneider and Sheldrick, 2002) found 21 selenium sites with CC-All 34.09 and CFOM of 52.35. Since phases obtained by those sites were not of good enough quality to allow hand determination and initial model building using with ShelxE, 10 of the selenium sites related by non-crystallographic operator as determined using Profess were used for SAD phasing using heavy metal site in PHASER (McCoy, 2007). Density modification was then performed using Parrot (Cowtan, 2010) and initial model building with Buccaneer (Cowtan, 2006). Only protein chains with assigned sequence were then used for molecular replacement of the complex data of KozL in complex with Mefuc at 1.55 Å using Phaser. Initial autobuilding and refinement of the 6 protein chains was performed with Buccaneer followed by iterative structure refinement with Refmac5.8 (Murshudov et al., 2011) and manual model corrections in COOT (Emsley et al., 2010). 5% of the observations were set aside for cross-validation analysis and riding hydrogen atoms were added and used for geometry and structure-factor calculations. The stereochemical quality of the refined models was validated on the wwPDB Validation server: http://wwpdb-

validation.wwpdb.org and carbohydrates were checked in Privateer (Agirre et al., 2015). All figures were drawn with PyMOL Molecular Graphic System program (Version 2.0.4, Schrodinger, LLC).

Accession codes. Coordinates of the structure of KozL in complex with MeFuc and structure factors for both MeFuc and MeSeFuc complex data have been deposited in the Protein Data Bank (https://www.rcsb.org/) (Berman et al., 2000) under accession codes 6HTN.

Architecture and evolution of blade assembly in β-propeller lectins

François Bonnardel, Atul Kumar, Michaela Wimmerova, Martina Lahmann, Serge Perez, Annabelle Varrot, Frédérique Lisacek*, and Anne Imberty*

Supplemental Information

Table S1 : β-propeller lectin used in the present study (one representative PDB code is indicated for each protein)

Type	Name	Sugar	Species	Uniprot	PDB	SCOPe	Pfam	Ref	Blades
PropLec5A	TL2	GlcNac	*Tachypleus trdentatus*	Q27084	1TL2	b.67.1.1	Pf14692, Pf14517	[1]	5
PropLec6A	AAL	Fucose	*Aleuria aurantia*	P18891	1IUC	b.68.8	Pf07938	[2]	6
	AFL	Fucose	*Aspergillus fumigatus*	Q4WW81	4AGI	b.68.8	Pf07938	[3]	
	AOL	Fucose	*Aspergillus oryzae*	Q2UNX8	5EO7	b.68.8	Pf07938	[4]	
	RSL	Fucose	*Ralstonia solanacearum*	Q8XXK6	2BS5	b.68.8.1	Pf07938	[5]	3 x 2
	BambL	Fucose	*Burkholderia ambifaria*	Q8XXK6	3ZW0	b.68.8.1	PF07938	[6]	
PropLec6B	FEL	Me-sugar	*Cyprinius carpio*	P68512	4RUQ	*not classified*	PF06462	[7]	6
	Lb-Tec2	Me-sugar	*Laccaria bicolor*	B0CZL6	5FSB	*not classified*		[8]	
PropLec7A	PLL	Fucose	*Photorhabdus luminescens*	Q7N8J0	5C9L	*not classified*	Pf03984 Pf13517	[9]	7
	PAL	Fucose	*Photorhabdus asymbiotica*	C7BLE4	5MXE	*not classified*	PF03984	[10]	
PropLec7B	PVL	GlcNac	*Psathyrella velutina*	Q309D1	2BWM	*not classified*	Pf13517	[11]	7
	AAL2	GlcNac	*Agrocybe aegerita*	H6CS64	4TQJ	*not classified*	Pf13517	[12]	
	PAL	GlcNac	*Psathyrella asperospora*	A0A1U7Q1Z0	5MB4	*not classified*	Pf13517	[13]	

Table S2: Data collection and refinement statistics for x-ray structures of KozL/ligand complexes

Complex	KozL/SeFuc			KozL/MeFuc		
Data collection						
Beamline	PX1			PX1		
Wavelength (Å	0.97914			0.97857		
Space group	$P22_12_1$			$P22_12_1$		
Unit cell *a, b, c* (Å)	87.1, 90.5, 143.8			86.2, 90.7, 143.4		
Resolution (Å)	45.35-2.65 (2.78-2.65)*			39.25–1.55 (1.57-1.55)		
Nb reflections	259639			827345		
Nb uniques reflections	33043			165158		
R_{merge}	0.081 (0.757)			0.041 (0.654)		
R_{pim}	0.045 (0.411)			0.031 (0.496)		
Mean $I/\sigma I$	15.2 (2.4)			16.9 (1.7)		
Completeness (%)	99.8 (99.9)			99.8 (98.0)		
Redundancy	5.0 (5.1)			5.0 (5.0)		
$CC_{1/2}$	99.1 (99.0)			99.9 (77.3)		
Mid-slope of anom probability	1.104					
FOM	0.71					
Refinement						
Resolution (Å)				39.28–1.55		
No. reflections				156839		
No. free reflections				8260		
R_{work} / R_{free}				15.5 / 18.0		
R.m.s Bond lengths (Å)				0.0156		
Rmsd Bond angles (°)				1.8		
Rmsd Chiral ($Å^3$)				0.099		
No. atoms	Chain A	Chain B	Chain C	Chain D	Chain E	Chain F
Protein	1167	1171	1169	1160	1184	1174
Sugar	36	36	36	36	36	36
Waters	219	216	222	226	210	229
B-factors ($Å^2$)						
Protein	22.9	22.0	21.1	21.1	21.0	21.3
Sugar	20.9	22.2	18.8	20.2	21.6	19.8
Waters	36.3	35.2	34.6	34.9	33.7	34.8
Ramachandran Allowed (%)				99.1		
Favored(%)				95.6		
Outliers(%)				8		
PDB Code				6HTN		

*Values in parentheses are for highest-resolution shell.

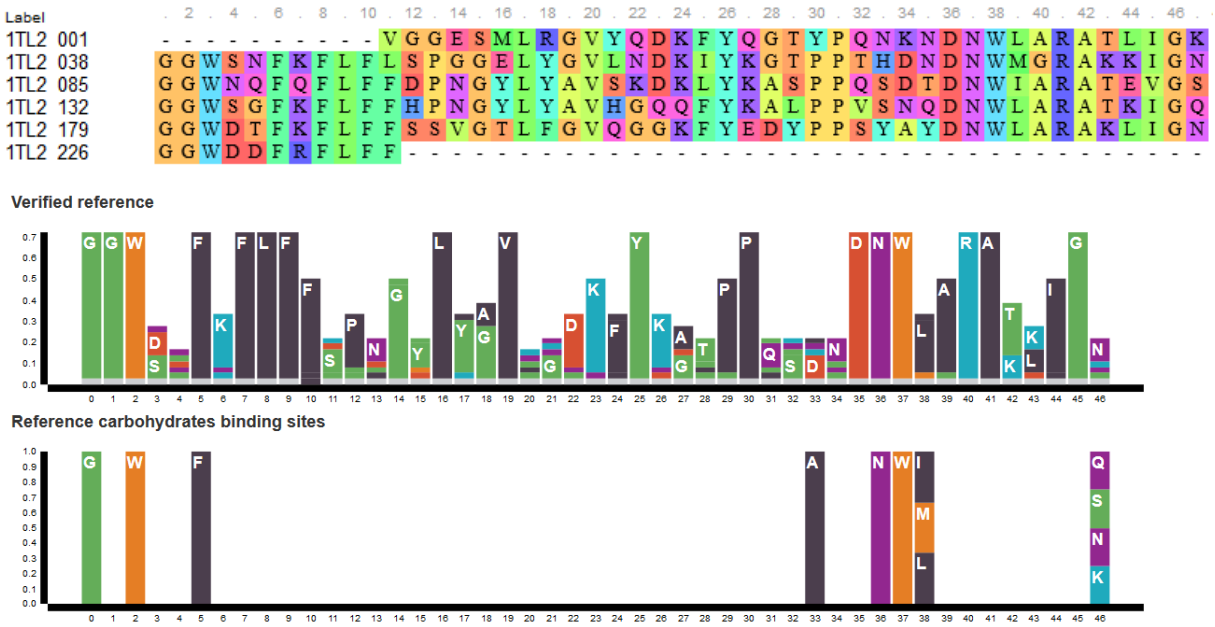


Figure S1. Signature motif of the PropLec5A family. Top panel: Multiple sequence alignment of the blades from structure 1TL2 from *Tachypleus tridentatus*. Middle panel: Resulting amino acid conservation. Bottom panel: Frequency of amino acids in contact with carbohydrate ligand.

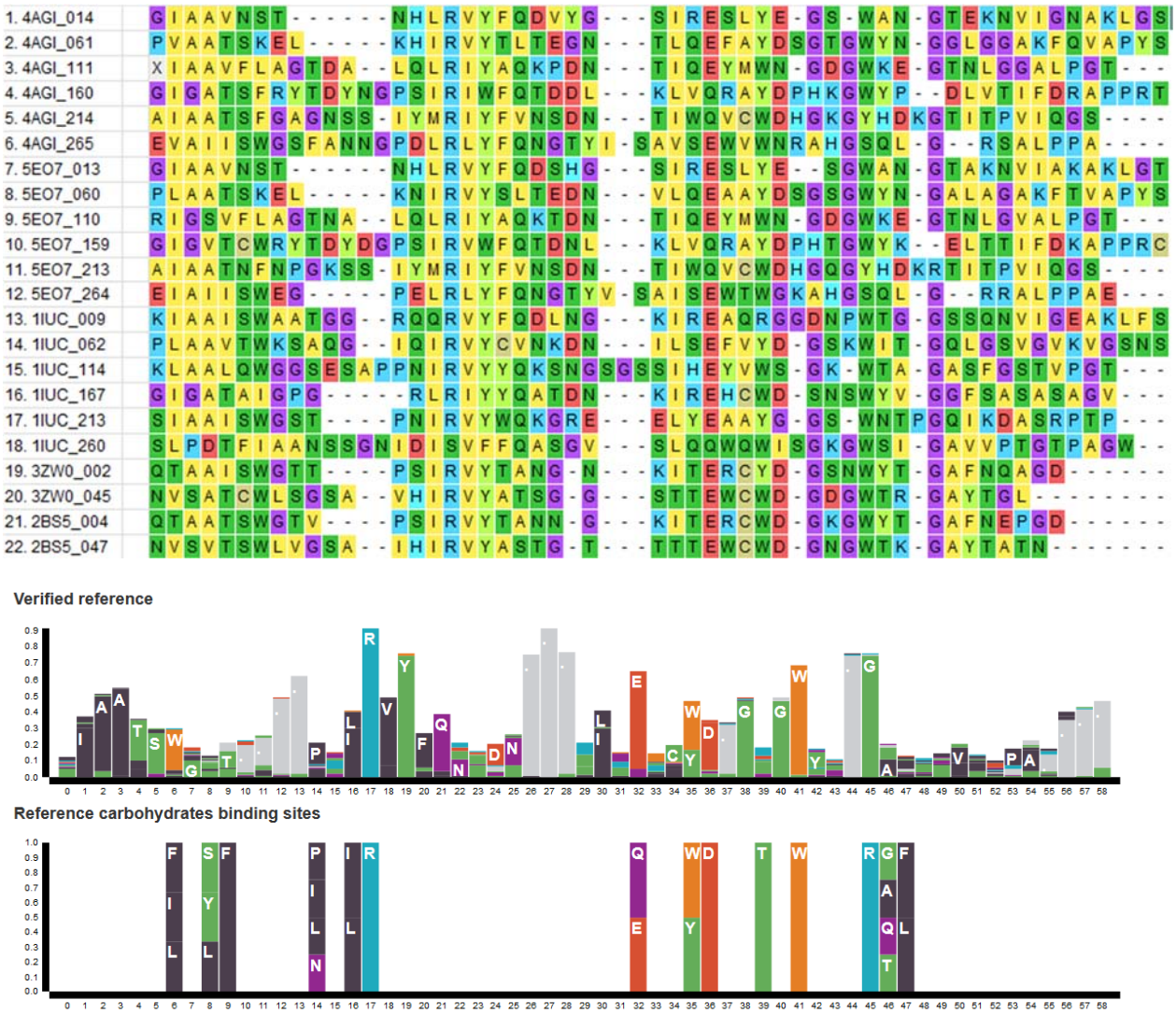


Figure S2. Signature motif of the PropLec6A family. Top panel: Multiple sequence alignment of the blades from crystal structures from *Aspergillus fumigatus* (4AGI), *Aspergillus oryzae* (5EO7), *Aleuria aurantia* (1IUC), *Burkholderia ambifaria* (3ZW0) and *Ralstonia solanacearum* (2BS5). Middle panel: Resulting amino acid conservation. Bottom panel: Frequency of amino acids in contact with carbohydrate ligand.

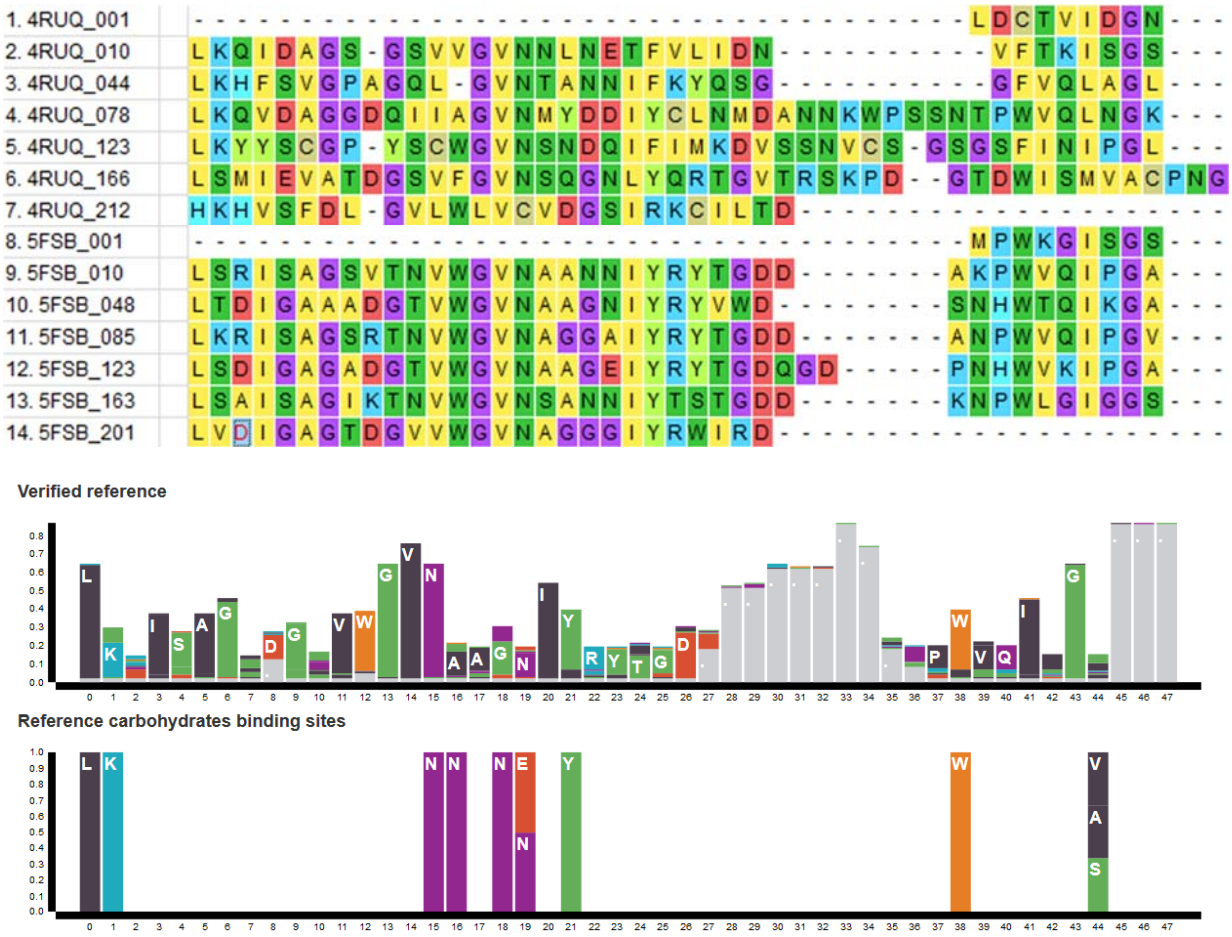


Figure S3. Signature motif of the PropLec6B family. Top panel: Multiple sequence alignment of the blades from crystal structures from *Laccaria bicolor* (5FSB) and *Cyprinius carpio* (4RUQ). Middle panel: Resulting amino acid conservation. Bottom panel Frequency of amino acids in contact with carbohydrate ligand.

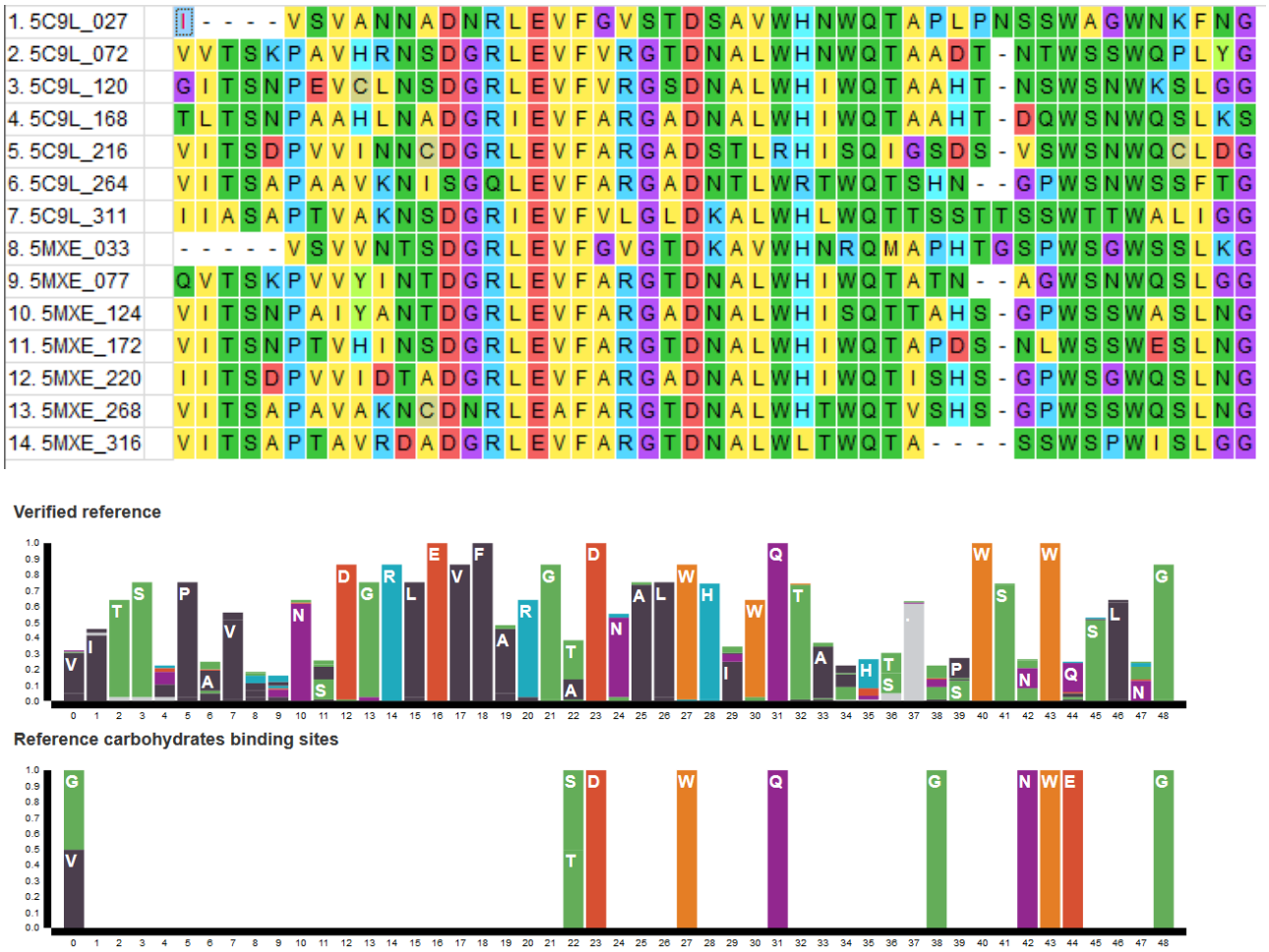


Figure S4. Signature motif of the PropLec7A family. Top panel: Multiple sequence alignment of the blades from crystal structures from *Photorhabdus luminescens* (5C9L) *and Photorhabdus asymbiotica* (5MXE). Middle panel: Frequency of amino acids in contact with carbohydrate ligand.

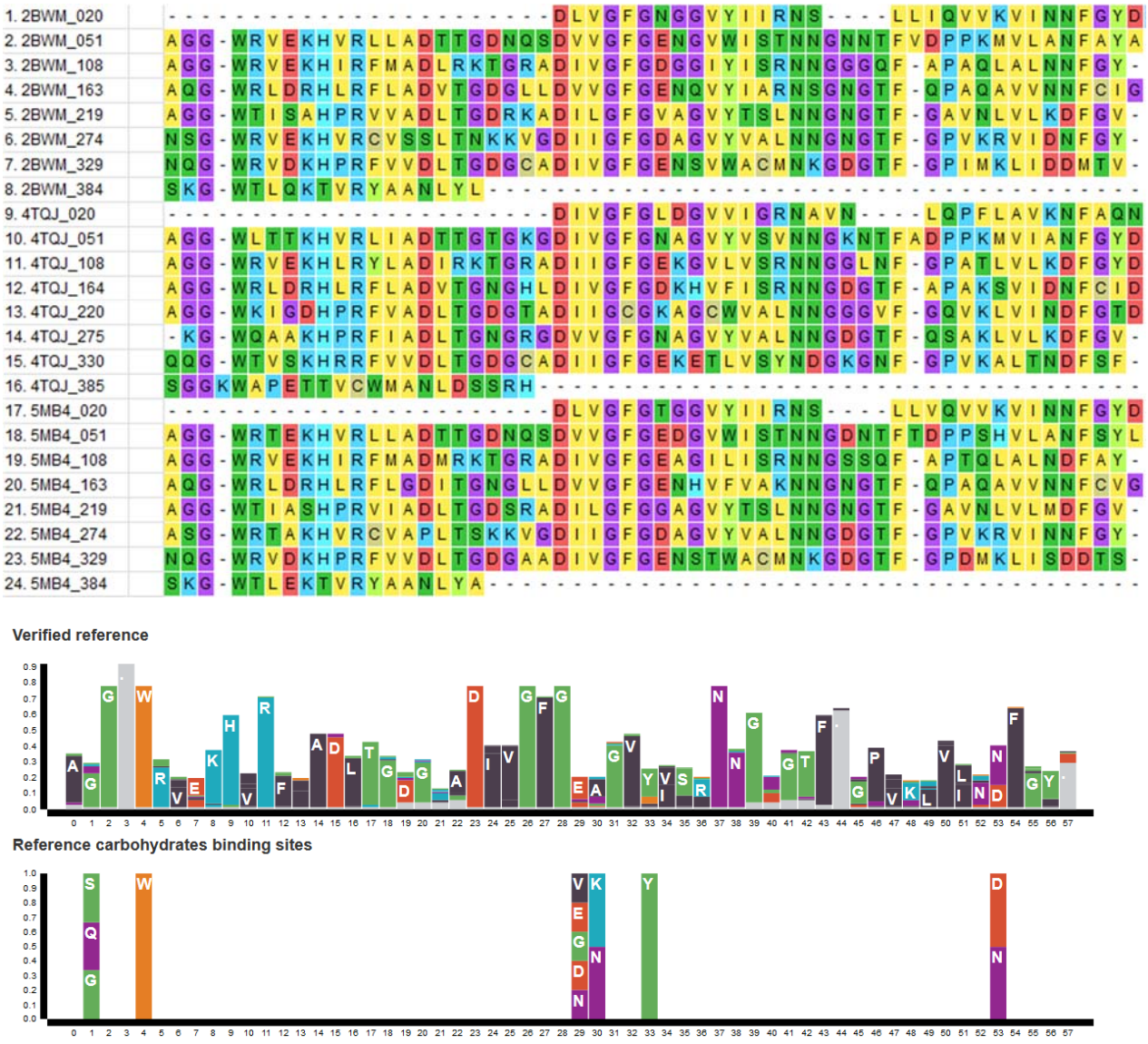
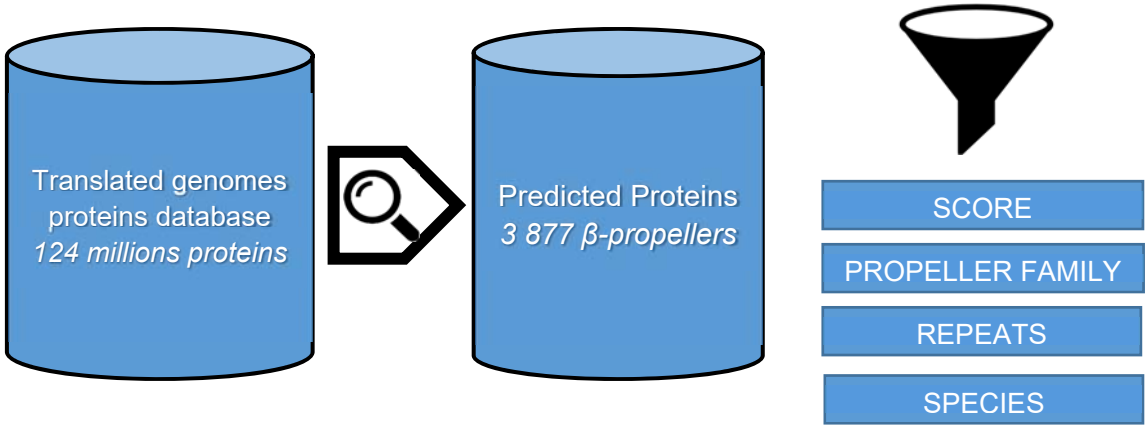


Figure S5. Signature motif of the PropLec7B family. Top panel: Multiple sequence alignment of the blades from crystal structures from *Psathyrella velutina* (2BWM), *Agrocybe aegerita* (4TQJ) and *Psathyrella asperospora* (5MB4). Middle panel: Resulting amino acid conservation. Bottom panel: Frequency of amino acids in contact with carbohydrate ligand.

A



B

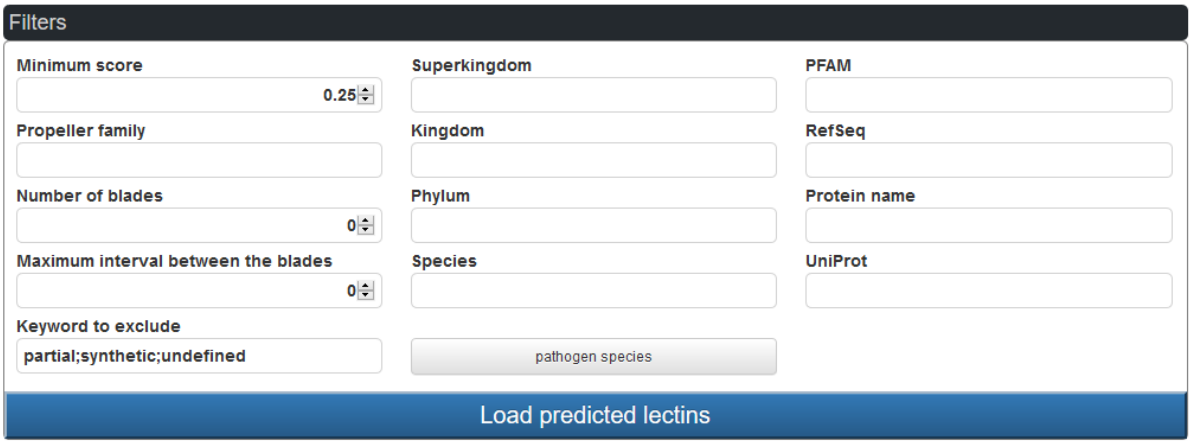


Figure S6: A. strategy for building the PropLec database. Propeller lectins are screened based on the 5 families motifs in UniRef100 database. The identified lectins are filtered based on similarities score and their features including protein name, taxonomy, number of blade. B. Advanced search interface on PropLec web module with a large numbers of criteria available to filter the predicted lectins

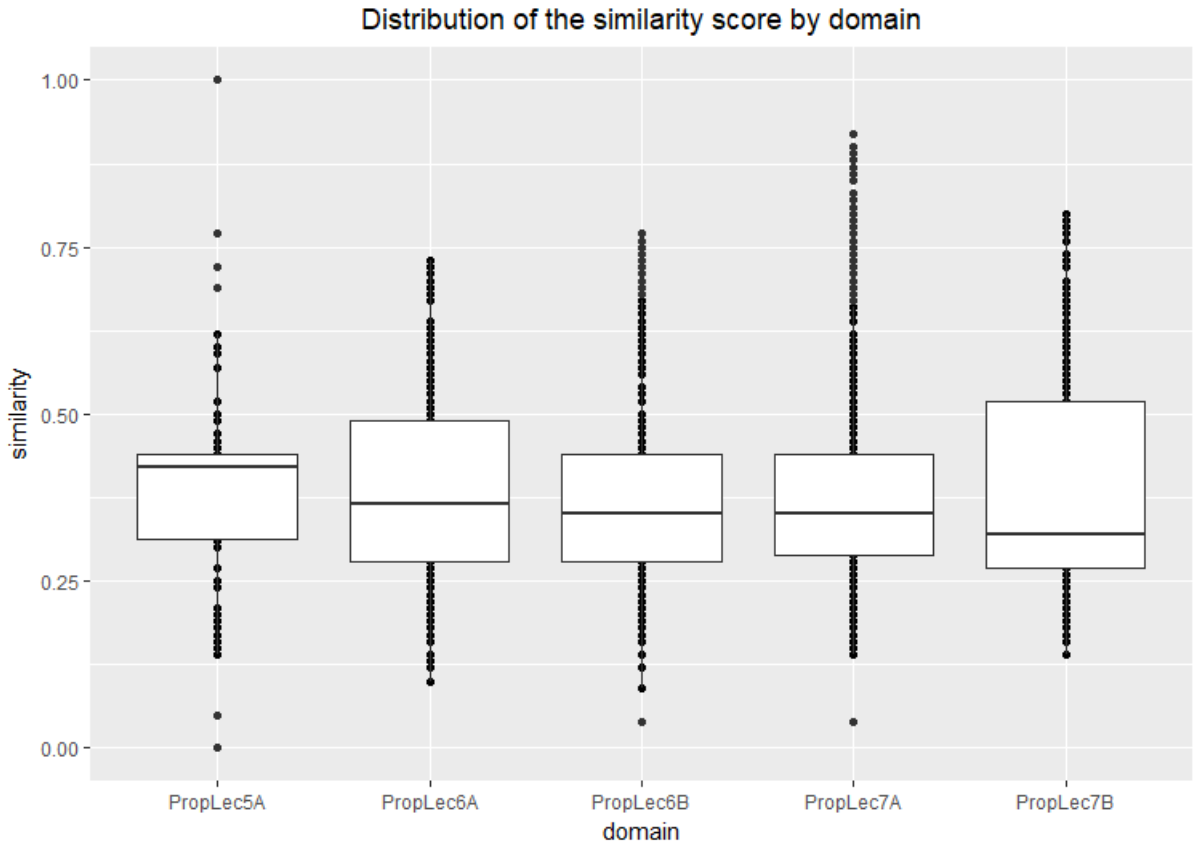


Figure S7. Boxplot distribution of the similarity score for all predicted sequences, grouped by PropLec family. Boxes represent the quartiles and the middle line represent the median.

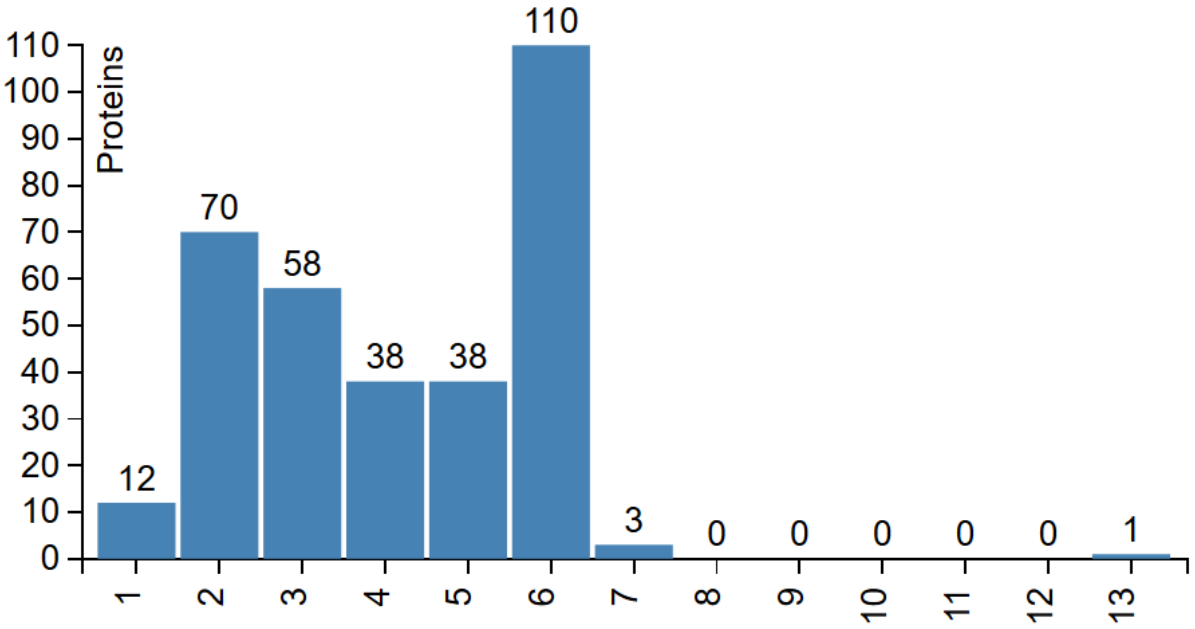


Figure S8 . Distribution of the number of blades in the predicted protein in the PropLec6A family

	KozL1	KozL2	KozL3	PepL1	PepL2	PepL3	BambL1	BambL2	AFL1	AFL2	AFL3	AFL4	AFL5	AFL6
KozL1	100.00	62.50	54.76	23.91	17.50	14.63	37.84	46.34	14.63	23.91	25.53	25.00	29.79	13.64
KozL2	62.50	100.00	59.09	31.25	19.05	20.93	41.03	52.38	17.07	26.09	31.25	24.00	30.61	15.22
KozL3	54.76	59.09	100.00	28.57	25.00	18.42	46.15	48.78	18.92	30.00	30.23	20.45	36.36	17.07
PepL1	23.91	31.25	28.57	100.00	23.26	34.88	37.50	30.00	27.50	29.17	21.28	34.00	29.17	20.00
PepL2	17.50	19.05	25.00	23.26	100.00	19.05	32.56	28.21	20.00	26.83	19.05	15.91	13.64	11.36
PepL3	14.63	20.93	18.42	34.88	19.05	100.00	35.90	21.62	26.83	17.07	20.93	19.57	17.78	50.00
BambL1	37.84	41.03	46.15	37.50	32.56	35.90	100.00	40.54	18.92	36.84	23.08	29.27	29.27	34.15
BambL2	46.34	52.38	48.78	30.00	28.21	21.62	40.54	100.00	16.67	31.58	36.59	16.67	21.43	20.00
AFL1	14.63	17.07	18.92	27.50	20.00	26.83	18.92	16.67	100.00	22.50	21.43	9.09	16.28	22.73
AFL2	23.91	26.09	30.00	29.17	26.83	17.07	36.84	31.58	22.50	100.00	26.09	20.83	15.22	11.63
AFL3	25.53	31.25	30.23	21.28	19.05	20.93	23.08	36.59	21.43	26.09	100.00	18.37	26.53	19.57
AFL4	25.00	24.00	20.45	34.00	15.91	19.57	29.27	16.67	9.09	20.83	18.37	100.00	31.37	20.41
AFL5	29.79	30.61	36.36	29.17	13.64	17.78	29.27	21.43	16.28	15.22	26.53	31.37	100.00	22.92
AFL6	13.64	15.22	17.07	20.00	11.36	50.00	34.15	20.00	22.73	11.63	19.57	20.41	22.92	100.00

Figure S9 : Percentage identity matrix resulting of a multiple sequence alignment of the blades of selected bacterial and fungal ProLec6 members computed by clustal omega.

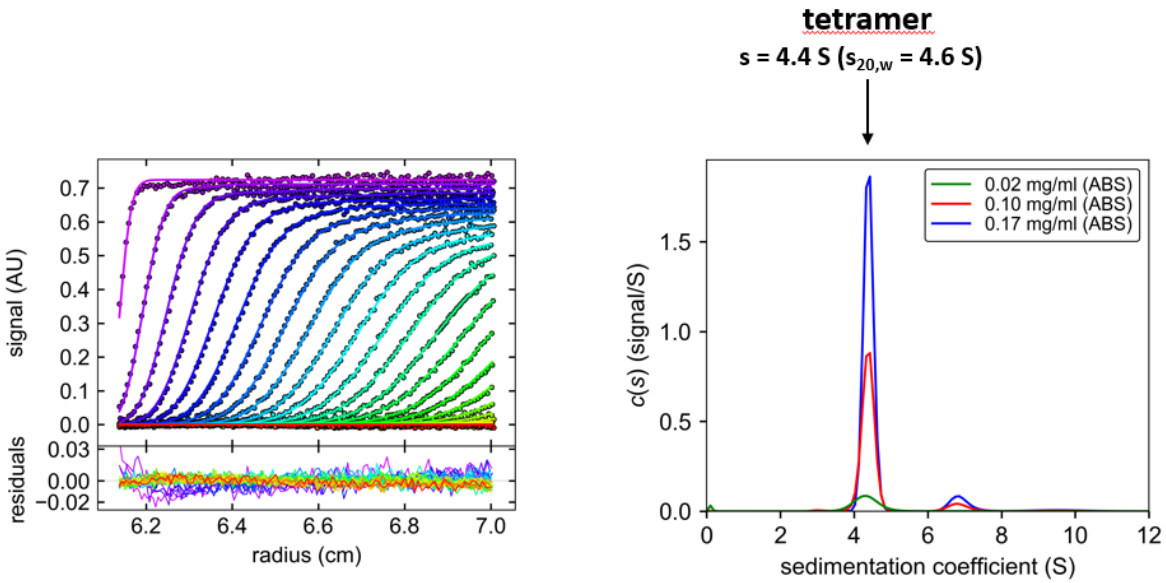


Figure S10. Analytical ultracentrifugation studies for the sedimentation velocity measurement of KozL

References

1. Beisel, H.G., Kawabata, S., Iwanaga, S., Huber, R. & Bode, W. 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 (1999).
2. Wimmerova, M., Mitchell, E., Sanchez, J.F., Gautier, C. & Imberty, A. Crystal structure of fungal lectin: Six-bladed β-propeller fold and novel recognition mode for *Aleuria aurantia* lectin. *J. Biol. Chem.* **278**, 27059-27067 (2003).
3. Houser, J. et al. A soluble fucose-specific lectin from *Aspergillus fumigatus* conidia - Structure, specificity and possible role in fungal pathogenicity. *PloS ONE* **8**, e83077 (2013).
4. Makyio, H., Shimabukuro, J., Suzuki, T., Imamura, A., Ishida, H., Kiso, M., Ando, H. & Kato, R. Six independent fucose-binding sites in the crystal structure of *Aspergillus oryzae* lectin. *Biochem Biophys Res Commun* **477**, 477-482 (2016).
5. Kostlanová, N., Mitchell, E.P., Lortat-Jacob, H., Oscarson, S., Lahmann, M., Gilboa-Garber, N., Chambat, G., Wimmerová, M. & Imberty, A. The fucose-binding lectin from *Ralstonia solanacearum*: a new type of -propeller architecture formed by oligomerisation and interacting with fucoside, fucosyllactose and plant xyloglucan. *J. Biol. Chem.* **280**, 27839-27849 (2005).
6. Audfray, A. et al. The fucose-binding lectin from opportunistic pathogen *Burkholderia ambifaria* binds to both plant and human oligosaccharidic epitopes. *Journal of Biological Chemistry* **287**, 4335-4347 (2012).
7. Capaldi, S., Faggion, B., Carrizo, M.E., Destefanis, L., Gonzalez, M.C., Perduca, M., Bovi, M., Galliano, M. & Monaco, H.L. Three-dimensional structure and ligand-binding site of carp fishelectin (FEL). *Acta Crystallogr D Biol Crystallogr* **71**, 1123-1135 (2015).
8. Sommer, R., Makshakova, O.N., Wohlschlager, T., Hutin, S., Marsh, M., Titz, A., Kunzler, M. & Varrot, A. Crystal structures of fungal tectonin in complex with O-methylated glycans suggest key role in innate immune defense. *Structure* **26**, 391-402 (2018).
9. Kumar, A., Sykorova, P., Demo, G., Dobes, P., Hyrsl, P. & Wimmerova, M. A novel fucose-binding lectin from P*hotorhabdus luminescens* (PLL) with an unusual heptabladed beta-propeller tetrameric qtructure. *J Biol Chem* **291**, 25032-25049 (2016).
10. Jancarikova, G., Houser, J., Dobes, P., Demo, G., Hyrsl, P. & Wimmerova, M. Characterization of novel bangle lectin from *Photorhabdus asymbiotica* with dual sugar-binding specificity and its effect on host immunity. *PLoS Pathog* **13**, e1006564 (2017).
11. Cioci, G. et al. β-Propeller crystal structure of *Psathyrella velutina* lectin: An integrin-like fungal protein interacting with monosaccharides and calcium. *J. Mol. Biol.* **357**, 1575-1591 (2006).
12. Ren, X.M., Li, D.F., Jiang, S., Lan, X.Q., Hu, Y., Sun, H. & Wang, D.C. Structural basis of specific recognition of non-reducing terminal N-acetylglucosamine by an *Agrocybe aegerita* Lectin. *PLoS One* **10**, e0129608 (2015).
13. Ribeiro, J.P., Ali Abol Hassan, M., Rouf, R., Tiralongo, E., May, T.W., Day, C.J., Imberty, A., Tiralongo, J. & Varrot, A. Biophysical characterization and structural determination of the potent cytotoxic *Psathyrella asperospora* lectin. *Proteins* **85**, 969-975 (2017).