Bioinformatics study of lectins : new classification and prediction in genomes

François Bonnardel

UNIVERSITÉ DE GENÈVE

UGA Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

préparée dans le cadre d'une cotutelle *entre la Communauté Université Grenoble Alpes* et *l'Université de Genève*

Spécialités: **Chimie Biologie**

Arrêté ministériel : le 6 janvier 2005 – 25 mai 2016

Présentée par

François Bonnardel

Thèse dirigée par la **Dr. Anne Imberty**
codirigée par la **Dr/Prof. Frédérique Lisacek**

préparée au sein du laboratoire CERMAV, CNRS et du Computer Science Department, UNIGE et de l'équipe PIG, SIB
Dans les Écoles Doctorales EDCSV et UNIGE

Etude bioinformatique des lectines: nouvelle classification et prédiction dans les génomes

Thèse soutenue publiquement le **8 Février 2021**,
devant le jury composé de :

Dr. Alexandre de Brevern
UMR S1134, Inserm, Université Paris Diderot, Paris, France, Rapporteur

Dr. Vincent Zoete
Molecular Modeling Group, SIB, Université de Lausanne, Switzerland, Rapporteur

Dr. Sabine Hediger
MEM, Univ. Grenoble Alpes, CEA, CNRS, Grenoble, France, Présidente

Dr. Stéphane Téletchéa
UFIP, Université de Nantes, UMR CNRS 6286, Nantes, France, Examinateur

Pr. Amos Bairoch
CALIPHO, SIB, Université de Genève, Switzerland, Examinateur

Dr. Valérie Barbié
Clinical Bioinformatics, SIB, Université de Genève, Switzerland, Examinateur

Dr. Anne Imberty
CERMAV, CNRS, Univ. Grenoble Alpes, Grenoble, France, Directeur de thèse

Dr. Frédérique Lisacek
Proteome Informatics Group, SIB, Université de Genève, Switzerland
Co-directeur de thèse

Acknowledgements

I would like to express my deepest gratitude to both co-directors Anne Imberty and Frédérique Lisacek who gave me the amazing opportunity of working together on the lectins with bioinformatics tools, their time, advice and the long day to finish the articles in time. I also owe a very important debt to Serge Perez for all the explanations on sugars, all the moments shared to make the virtual reality works, and for the development of the database. But the best times will always be the restaurant in families all together.

I would like to thank Vincent Zoete and Alexandre de Brevern for the time spent as rapporteur of my manuscript and the members of the jury: Sabine Hediger, Stéphane Téletchéa, Amos Bairoch and Valérie Barbié, who have been kind enough to evaluate this work.

I am very grateful for the committee members of my PhD for their Insight with William Helbert, Yves Vandenbrouck, Sylvie Ricard-Blum for asking me the right questions and contributing to the good development of my thesis, even if there are still possibilities for my successor to explore.

Annabelle Varrot gave me a lot of valuable insight while working together on the publications. The environment in both labs is a large family for me and I want to also thanks a lot Martin Lepsik, Simona Notova, Raphael Bermeo, François Portier, Aurore Cabanettes, Rubal Ravinder for all the time passed together at the Cermav and during conferences or at the bar. Also, the colleagues from Geneva Thibault Robin and Julien Mariethoz for all difficulties in the code tackled together.

I would also like to thank Isabelle Caldara, Magali Pourtier and Anne-Isabelle Giuntini for their efficient administrative work which allowed the smooth progress of my thesis.

Finally, I would also like to express my gratitude to my family and my friends for their support outside of work.

Table of contents

Abbreviations

AC: accession number, identifier of a molecule in a database

BLAST: Basic Local Alignment Search Tool

CATH: Protein Structure Classification database

CAZY: Carbohydrate active enzymes database

CBM: Carbohydrate-Binding Module

CFG: Consortium for Functional Glycomics

FASTA: text-based format for representing either nucleotide sequences or amino acid (protein) sequences

Gal: Galactose

GalNAc: N- acetylgalactosamine

Glc: Glucose

GlcNAc: N-acetyl glucosamine

HMM: Hidden Markov Model

Man: Mannose

NCBI: National Center for Biotechnology Information

ORF: Open Reading Frame

PDB: Protein Data Bank, file format containing atom position and type in a 3D space

PFAM: Protein family database

SCOPe: Structural Classification of Proteins — extended

List of Tables

List of Figures

List of Publications

Article I: UniLectin3D, a database of carbohydrate-binding proteins with curated information on 3D structures and interacting ligands

Article II: Structural Database for Lectins and the UniLectin Web Platform

Article III: Structure and engineering of tandem repeat lectins

Article IV: Architecture and Evolution of Blade Assembly in β-propeller Lectins

Article V: Identification and characterization of a β-trefoil lectin from lower eukaryote with an aerolysin domain

Article VI: LectomeXplore, an update of UniLectin for the discovery of carbohydrate-binding proteins based on a new lectin classification

Article VII: Proteome-wide prediction of bacterial carbohydrate-binding proteins as a tool for understanding commensal and pathogen colonisation of the vaginal microbiome

Article VIII: Exploration of the lectins in the fungal kingdom: relation with the host recognition and ecology

Prologue

The structural and molecular glycobiology team at CERMAV, localised at Grenoble, carries research on proteins function relationships using 3D structures, in interaction with complex glycans such as glycoconjugates and polysaccharides. At Geneva the Proteome Informatics Group (PIG) team, associated with the SIB, is specialized on the development of software and database dedicated to in-silico analysis, exploration and integration of the relations between glycan and protein respectively in the fields of glycomics and proteomics.

Lectins have been widely explored in model organisms due to their interesting and important capacity to recognize blood groups and their importance in pathogen to cell interactions. But in the large number of newly available genomes every month, only the curated lectins available in the reference genomes are identified. Sadly, many lectins, even when their 3D structure is solved and their function is known, are not properly annotated in protein web databases.

The research conducted focuses on the identification of new lectins, through crystallisation together with glycans and the use of lectin or glycan arrays to define the specificity of the lectin glycan interactions. Lectins are used for further development of drugs to mimic the glycan with a better affinity to the lectin to limit host-pathogen interactions through lectins.

Biochemists need from bioinformatic -omics databases to provide a high coverage of lectins in order to choose the most adapted lectin tool for their experiments. A first step is the identification of lectins, based on the 3D structure available, in genomes of isolated species of interest.

Six articles have been published and are available online and two other publications are currently in press.

Résumé

Les domaines de la bioinformatique utilisent des concepts mathématiques et des outils informatiques pour démêler les connaissances dans les données biologiques. Lorsque la bioinformatique est appliquée aux glycanes et à la glycobiologie, elle est appelée glyco-informatique. Les nouvelles technologies permettent le séquençage massif des génomes de nouvelles espèces et des métagénomes d'échantillons environnementaux. Mais tous les génomes nouvellement découverts et les protéines encodées ne sont que partiellement annotés d'une fonction biologique, récupérée par similarité à partir des organismes de référence.

La glycobiologie est le domaine de recherche consacré à l'étude des glycanes/glucides, composés d'un ou de plusieurs monosaccharides. Les lectines sont des protéines capables de se lier de manière réversible aux glycanes, et sans fonctions enzymatiques. Les lectines sont des outils puissants pour la reconnaissance des glycanes dans les échantillons, et elles sont également des cibles pour les composés thérapeutiques en raison de leur implication dans le cancer, l'immunologie et les infections.

Cette thèse vise à utiliser la bioinformatique pour développer de nouveaux outils in-silico pour l'étude des lectines. Elle a pour objectif de fournir, dans une nouvelle base de données en ligne, des informations sur les lectines pour les organismes de référence et les nouveaux génomes appartenant à d'autres organismes.

Pour fournir une classification des structures 3D des lectines et leur annotation dans les génomes, un portail web dédié a été développé, appelé UniLectin. Le module UniLectin3D fournit des structures 3D classées et stockées manuellement, ainsi que leurs glycanes en interaction. En raison de la difficulté d'identifier les lectines répétées en tandem dans les génomes, une méthode spécifique a été mise au point pour permettre la prédiction de ces lectines particulières, maintenant disponibles dans les modules PropLec et TrefLec. Enfin, le module LectomeXplore fournit des lectines prédites basées sur les 107 classes de UniLectin3D, dans les génomes disponibles du NCBI et d'UniProt. Cela a permis l'étude des lectomes de différents environnements par le biais de la collaboration décrite dans la dernière partie de la thèse.

Summary

Bioinformatics uses mathematical concepts and informatics tools to unravel the knowledge hidden in biological data. When bioinformatics is applied to glycans and glycobiology, it is called glyco-informatics. New technologies allow mass sequencing of new species genomes and of environmental samples metagenomes. But all newly discovered genomes and encoded proteins are only partially annotated with biological function assessed by similarities to reference organisms.

Glycobiology is the research field dedicated to the study of glycan/carbohydrate compounds, composed of one or multiple monosaccharides. Lectins are proteins able to bind reversibly to glycans, and without enzymatic functions. Lectins are powerful tools for the recognition of glycans in samples, and they are also targets for therapeutic compounds due to their involvement in cancer, immunology and infections.

This thesis aims to use bioinformatics for developing new *in silico* tools for the study of lectins. More specifically, it addresses the need for a new online database covering curated information on lectins for both reference organisms and newly sequenced genomes belonging to other organisms.

To provide a curated classification of lectin 3D structures and their annotation in genomes, a dedicated web portal called UniLectin, was developed and includes several modules. The UniLectin3D module provides manually curated and classified 3D structures together with their interacting glycans. Due to the difficulty of identifying tandem repeated lectins in genomes, a specific method has been developed for the prediction of those particular lectins, now available in the PropLec and TrefLec modules. Finally, the LectomeXplore module includes lectin predictions based on 107 classes defined on the basis of UniLectin3D content, and resulting from screening available sequences stored in the reference protein databases NCBI-nr and UniProt. This made the study of lectomes in different environments possible as collaborative work described in the last part of the thesis.

1 General introduction

Modern biology uses genomic, transcriptomic and proteomic approaches to elucidate the function of each part of the genomes. The aim of omics is to characterise the structure and functioning of biological molecules in organisms. The corresponding suffix -ome defines the objects of study. Model organisms are used in biology and for -omics studies, such as *E. coli* characterised by a rapid growth and a small sequenced genome of 4000 genes. A model organism can be studied at several levels in omics, including genomics, transcriptomics, proteomics, metabolomics, glycomics, epigenomics and phenomics (sets of phenotypes).

With high-speed sequencing, it is now possible to sequence DNA and then assemble the sequences to reconstitute a complete genome. This leads to an exponential increase in the number of available genomes that must then be analysed. In addition, technological advances make it easy to obtain transcriptomes, proteomes and network analyses on which protein-protein or protein-molecule interactions can be tested. It is also possible to access personalised data on the health of patients in hospitals, which also generates important data that must then be analysed.

Recent major breakthroughs in modern biology are represented in Figure 1.

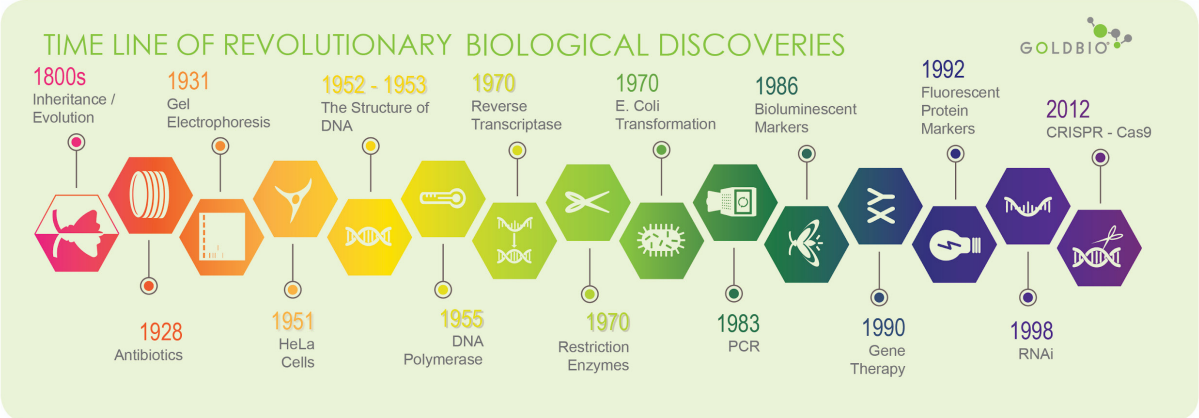


***Figure 1: Recent major breakthroughs in modern biology* (*GoldBio*).**

Bioinformatics is a discipline that combines biology, mathematics, statistics and computer science in the development of methods and tools for the study and understanding of biological data, in particular the study of large datasets. The fully sequenced Human genome represented lots of information: 3 billion base-pairs and 20 thousand protein-coding genes (Pennisi 2001; Abdellah et al. 2004). Informatics is required to determine and analyse the sequence, for the application, among others, to the biomedical fields. Bioinformatics includes both tools developed for biology but also pipelines

which are a series of computational steps necessary to filter, process and interpret data, particularly in genomics and -omics. The main technological advances in microbial community studies, represented in Figure 2, gave rise to modern biology and made the current breakthrough in genome editing and gene therapy possible.

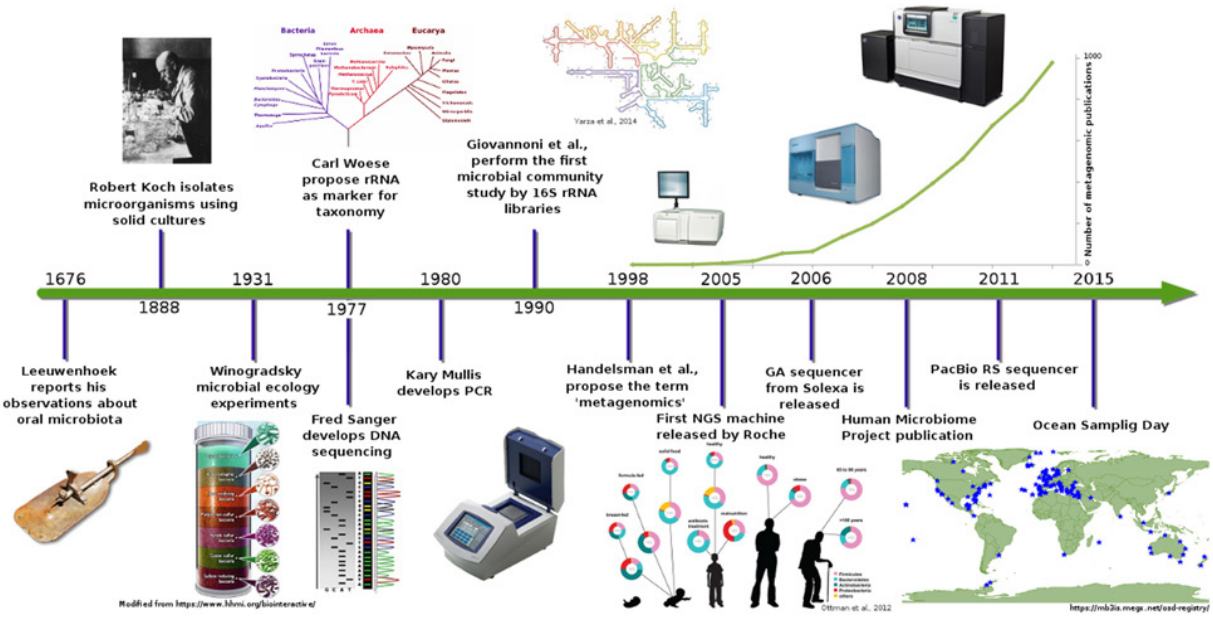


Figure 2: Metagenomics timeline and milestones. Timeline showing advances in microbial communities studies (Escobar-Zepeda et al. 2015).

New sequencing methods for DNA were commercialised as DNA sequencers around the 2000s. They were called the "next-generation sequencing" (NGS) methods, in order to distinguish them from the earlier methods, including Sanger sequencing. An entire genome can be sequenced when broken into fragments called ‘reads’ that are automatically sequenced, and then reassembled into ‘contigs’ and if possible whole chromosomes by software.

The diversity of proteins in nature leads to a massive yearly increase of the size of protein datasets with unknown functions which must be elucidated.

Sugars are essential to life, they are most abundant, most diversified and but less studied biomolecules compared to genomics and proteomics because of their complexity. Sugars are important in all aspects of life: in cell architecture (cell walls, ...), energy storage (starch, glycogen..), energy metabolism and stability of glycoproteins. The surfaces of cells are covered by glycans which are important for intermolecular and intercellular recognition. In industry and healthcare, they are used for biomaterials and medicines. Databases and tools dedicated to the study of sugars are just emerging compared to the tools and databases available for the study of genomes and proteins.

Glyco-Bioinformatic is a transversal discipline gathering the methods and tools for processing data associated with the study of sugars and related molecules (Figure 3). These data can be presented in different forms, such as glycan arrays, glycosylation of proteins, enzymatic reactions, etc.

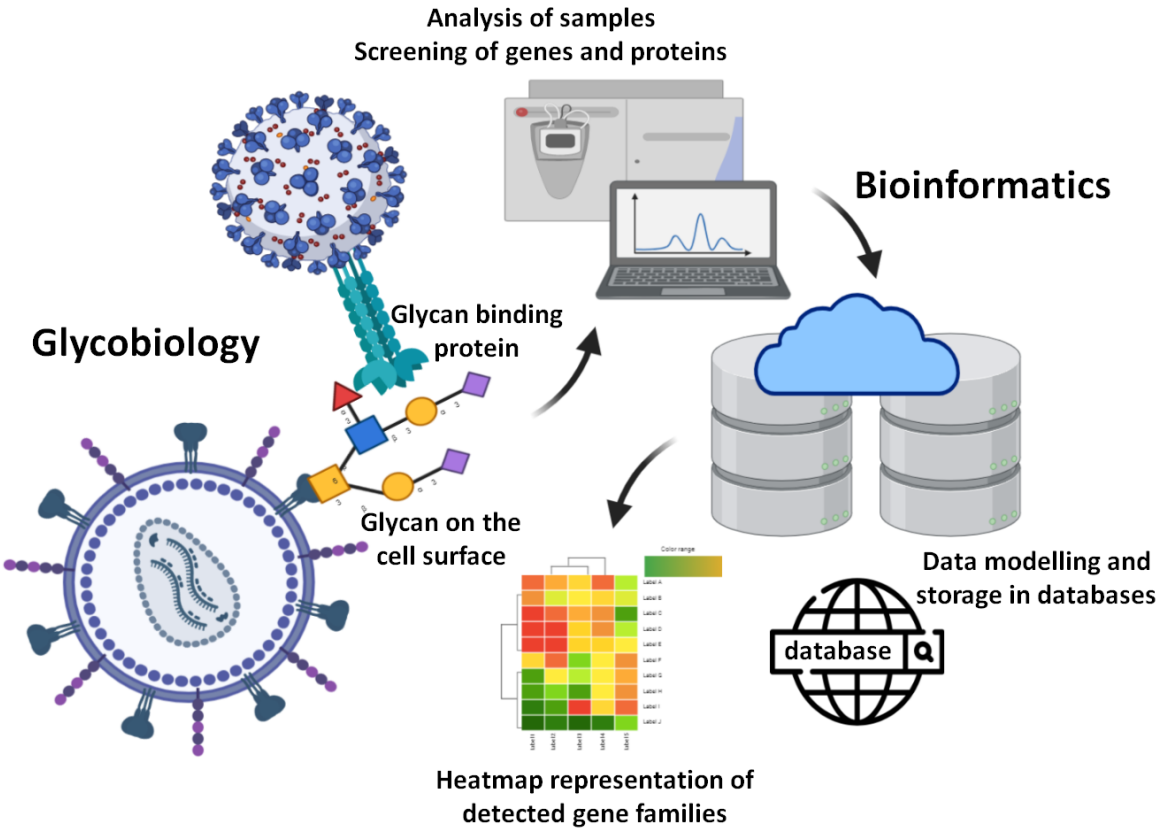


Figure 3: Glyco-Bioinformatics: at the crossroad between glycobiology and bioinformatics.

2 Proteins and their 3D structures: databases, classification and methods

Protein classification and investigation in genomes are important to improve the understanding of organism physiology at the molecular level. To help in the investigation of proteins, bioinformatics databases and tools are developed: protein and protein family databases were developed to provide and share curated information on reference organisms and allow manual curation of newly discovered organisms. Protein annotation uses sequence alignment methods to identify similarities between proteins, which might share a similar biological function. Databases also provide proteins related information such as encoding genes and genomes, or at the metabolic level with enzymes and metabolic pathways. Protein functions are investigated at different levels including their 3D structures, obtained with different methods and stored in databases. Associated tools were developed for the visualization and manipulation of 3D structures.

2.1 Sequence datasets: from genomes to proteins

The number of genomes, genes and proteins available in databases increased during the last decade with exponential dynamics thanks to technological progress, which is represented in Figure 4.

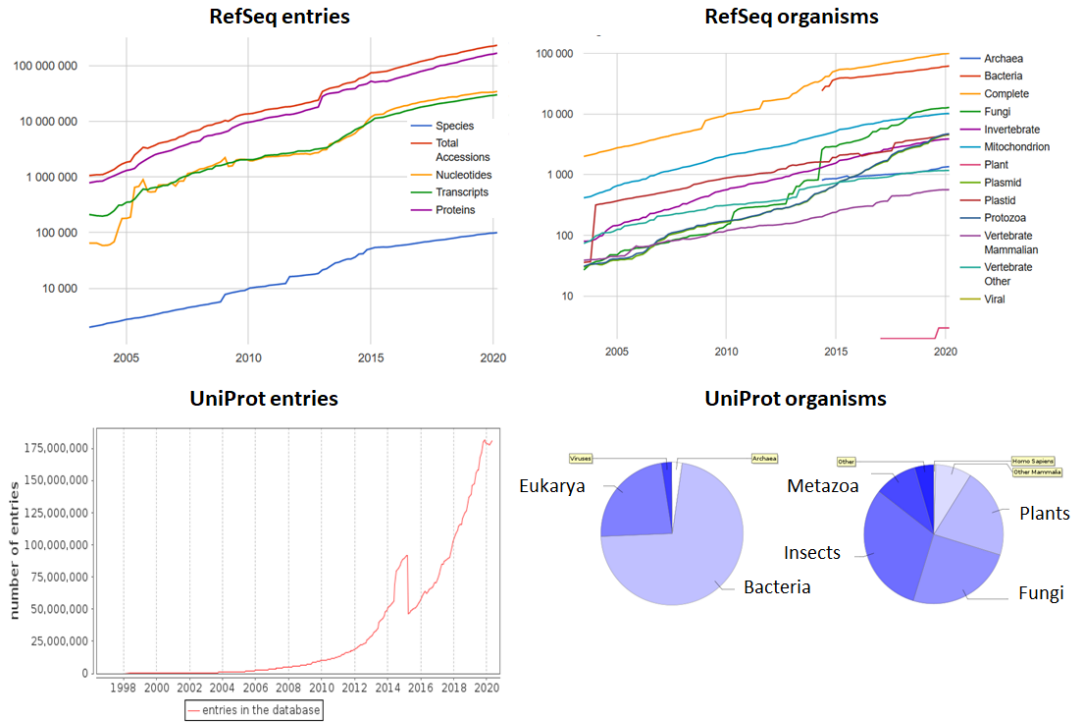


Figure 4: Evolution of genes and proteins available in RefSeq and UniProt databases, the representation is logarithmic as the number of entries grows exponentially (O'Leary et al. 2016; Bateman 2019).

Name	**Description**	**Content** (december 2020)	**Ref**
NCBI GenBank	genetic sequence database of unreviewed or computer-annotated sequences submitted from individual laboratories and large-scale sequencing projects	contains over 1.6 billion nucleic entries for 450 000 species	(Sayers et al. 2020)
EMBL-EBI	portal of omics sequences from multiple sources	Patent protein sequences from EPO, JPO, KIPO, UPSPTO. Nucleotide sequences from ENA and other sources	(Madeira et al. 2019)
NCBI - RefSeq	GenBank sequences that are manually curated by the NCBI staff.	36,167,417 nucleic entries 189,864,508 proteins 55 000 organisms including >4800 viruses, >40 000 prokaryotes, >10 000 eukaryotes	(O'Leary et al. 2016)
UniProt - trEMBL	computer-annotated supplement of SwissProt that contains all the translations of EMBL nucleotide sequence entries	209 157 139 proteins 84 387 species	(Bateman 2019)
UniProt - SwissProt	Manually curated annotation and reviewed protein sequences	563 972 curated proteins	(Bateman 2019)
PDB	Manually curated, experimentally determined structures of proteins, nucleic acids, and complex assemblies	172175 3D structures including: 168546 of Proteins, 7923 DNA, 5237 RNA. Main species: 50603 for Human, 7192 for the Mouse, and 6227 for E.coli	(Burley et al. 2017)
JGI	access point to all JGI -omic databases	732 genomes 918 assembled genomes	(Nordberg et al. 2014)
Mycocosm	JGI fungal -omic database	1,781 genomes, 1200 published	(Grigoriev et al. 2014)
Ensembl	Ensembl is a genome browser for vertebrate genomes	153 annotated assemblies 29 049 transcripts	(Cunningham et al. 2019)
Ensembl genomes	Providing genome data for non-vertebrate species	44 048 Bacterial genomes, 237 Protists, 1014 Fungi, 67 Plants, 78 Metazoa (non vertebrate)	(Howe et al. 2020)
UCSC Genomes	Reference organism genomes	195 assemblies and 105 species	(Haeussler et al. 2019)
ProGenomes	Prokaryotic genomes http://progenomes.embl.de/	87 920 high-quality genomes 12 221 species	(Mende et al. 2020)

Table 1: Repositories and manually curated protein and proteome databases

Multiple -omics platforms using distinct structural and functional annotation pipelines are available online, with different level of curation that either cover all available genomic knowledge; cluster the species and keep the ones with highly assembled genomes; or focus on reference model organism,

They also either cover all kingdoms or specialize on specific kingdoms described in Table 1. Based on genomic information and the contained protein-coding genes, protein datasets are built. They give access to the genetic sequences and also to the protein ones. They provide the option to download the species proteomes in fasta files or for some of them to download a unique non-redundant protein database covering a large number of species/kingdoms.

The content of the trEMBL section of UniProt is highly redundant, with proteins from coding sequences (CDS) submitted to the EMBL-Bank/GenBank/DDBJ nucleotide sequence resources. UniProt also provides the UniRef non-redundant database of proteins, which contains a representative sequence for protein clusters sharing 90, 80 or 50 percent of sequence identity. It avoids sequence redundancy at the cost of species-specific information.

2.2 Protein domain definition and databases

Proteins are made of one or multiple functional domains that have a large variety of functions such as synthesis, hydrolysis, modification, binding, transport, recognition. A protein domain is able to fold independently from the rest of the protein. It can be visualized at the 3D level and correspond to a 2D section in the protein sequence as represented by Figure 5. Protein families gather sequences from different species that display high protein sequence similarity and a similar function. These protein families and domains are available in databases with each database having a specificity either in the method to obtain the domains or in the living kingdoms they are focusing on. As such, protein families can be used to generate motif profiles.

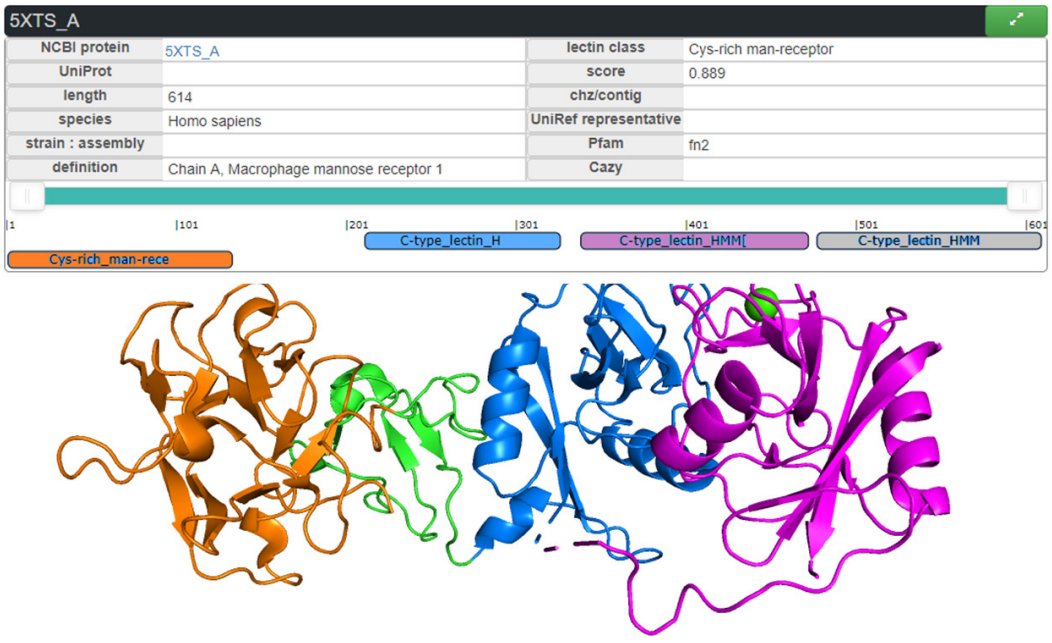


Figure 5: Representation of a protein and its multiple domains

Different types of functional domains represented by conserved peptidic motifs are available in protein family databases. InterPro federates these various sources (https://www.ebi.ac.uk/interpro/; (Mitchell et al. 2019)) and the corresponding list of its components is provided in Table 2, describing the geographical origin, the source of the protein dataset used to construct the family.

Name / Ref	Information	Content	Link	Ref
CDD	Search Conserved Domains and Protein Classification, developped by the NCBI	4,700 models curated by the CDD group, and a total of 52,910 protein domains	https://www.ncbi.nlm.nih.gov/cdd/	(Lu et al. 2020)
CATH-Gene3D	domain families from PDB structures, developed by UCL, London, UK	151 million protein domains classified into 5,481 superfamilies	http://gene3d.biochem.ucl.ac.uk/	(Dawson et al. 2017)
SCOPe	classification of large macromolecular structures in the structural classification of proteins-extended database	1232 classified 3D fold, 2026 protein superfamilies, 4919 families	https://scop.berkeley.edu/	(Chandonia et al. 2019)
Panther	Gene ontology classification system, developed by University of Southern California, CA, US.	PANTHER™ 15.0, released 2020-02-14) 15702 PANTHER™ families 123989 subfamilies	http://www.pantherdb.org/	(Mi et al. 2019)
Pfam	Search protein families from Pfam database, developed by EMBL	Pfam identifies 17,929 families in release 32.0. Only DUF domains are undefined	http://pfam.xfam.org/	(El-Gebali et al. 2019)
Pirsf	Search against fully curated PIRSF families with HMM models, developed by Georgetown University	4,500 preliminarily curated and 3,900 fully curated families	https://proteininformationresource.org/pirwww/dbinfo/pirsf.shtml	(Nikolskaya et al. 2006)
Prints	Search protein fingerprints, developed by University of Manchester, UK	2156 fingerprints, encoding 12 444 individual motifs	http://130.88.97.239/PRINTS/index.php	(Attwood et al. 2012)
Smart	Simple Modular Architecture Research Tool, developed by EMBL	1300 curated protein domains, 200 million domains	http://smart.embl-heidelberg.de/	(Letunic and Bork 2018)
Superfamily	structural and functional annotation for proteins and genomes, developed by University of Bristol, UK	27 623 HMMs of protein families	http://supfam.org/SUPERFAMILY/	(Pandurangan et al. 2019)
Hamap	classification and annotation system of protein sequences, developed by SIB	Number of family profiles: 2'350 Number of annotation rules: 2'382	https://hamap.expasy.org/	(Pedruzzi et al. 2015)
Prosite	protein families and domains, developed by SIB	1311 patterns, 1296 profiles and 1328 ProRule	https://prosite.expasy.org/	(Sigrist et al. 2009)

Table 2: Protein family databases

Among those protein family databases, enzymes cover many characterized families. Enzymes are proteins with a biological catalyst activity, transforming a substrate into a product. In 1833, Anselme Payen was the first to report a Diastase that breaks starch into maltose. Series of enzymatic reactions form metabolic pathways that are necessary for example for energy storage, energy liberation or the creation of essential bricks of life. Some most popular metabolic pathway databases are described in **Table 3**, with different information available depending on the database, providing known enzymes with an active role on metabolites.

Name	Description	Link	Ref
BRENDA	enzyme and metabolic information: 83 000 enzymes from 9800 organisms	http://www.brenda.uni-koeln.de	(Jeske et al. 2019)
BioCyc	Dedicated to reference organism with 160 genome databases (PGDBs) predicted operons and a synteny tool	https://biocyc.org/	(Karp et al. 2018)
MetaCyc	Developed together with BioCyc, curated database of experimentally elucidated metabolic pathways from all domains of life	https://metacyc.org/	(Caspi et al. 2020)
Reactome	human pathways and reactions including signal transduction, transport, DNA replication, metabolism and other cellular processes	http://www.reactome.org	(Jassal et al. 2020)
KEGG	encyclopedia of genes and genomes, with associated information including gene ontology and manually drawn pathway maps	https://www.genome.jp/kegg/pathway.html	(Kanehisa et al. 2017)
Pathway Commons	contains data from 22 databases with 4794 detailed human biochemical processes (i.e. pathways) and ~2.3 million interactions	https://www.pathwaycommons.org/	(Rodchenkov et al. 2020)
Intenz / ExplorEnz	enzyme nomenclature, and enzymes database providing 4102 current enzyme entries	https://www.enzyme-database.org/forms.php	(McDonald et al. 2009)
WikiPathways	scientific community driven curated biological knowledge in pathway models	https://www.wikipathways.org/	(Martens et al. 2020)

Table 3: Main metabolic pathway databases

2.3 Sequence alignments methods for proteins

2.3.1 Protein alignment tool and multiple proteins alignments

Bioinformatic tools are required to align protein sequences. They can be used to search for similar homologous sequences or to study, for example, their phylogeny and evolution. A global alignment consists of taking two nucleic or RNA or peptidic sequences and aligning the two of them using their full length. Local alignment methods search the best alignment between regions of the two sequences. Heuristics are used to find an optimal solution in a reasonable computation time. Needleman–Wunsch algorithm was proposed in 1970 for global alignment. The algorithm assigns a score to every possible

alignment and keeps the best ones. The Smith-Waterman method has been used since 1981; the algorithm performs local sequence alignment, compares segments of all possible lengths and optimizes the similarity measure.

Pairwise sequence alignment (PSA) aligns pairs of sequences also with global or local approaches. PSA aims to find the alignment that maximizes the similarity between the two input (protein) sequences; the similarity is a number indicating how much two sequences are similar as calculated based on the number of identical residues and the number of most probable substitutions. The identity score shows the number of identical residues between two sequences. PSA methods are for example EMBOSS (Olson 2002), CD-HIT (Fu et al. 2012), UCLUST (Edgar 2010), etc.

Multiple sequence alignments (MSA) aim to find the best alignments for more than two sequences with also global and/or local approaches, either by aligning step by step using a successive pairwise alignment approach or using direct multiple sequence alignment. MSA methods are for example MUSCLE (Edgar 2004), MAFFT (Katoh and Standley 2013), CLUSTAL (Sievers and Higgins 2014), PROMALS (Pei and Grishin 2007), (Lassmann et al. 2009), DIALIGN-TX (Subramanian et al. 2008), HAlign (Wan and Zou 2017), etc. Tools also use both sequence and secondary structure: SPEM (Zhou and Zhou 2005) and 3D structure based: T-Coffee (Di Tommaso et al. 2011), Expresso (Armougom et al. 2006), PROMALS3D (Pei et al. 2008). MSA allows to identify protein families and construct conserved peptide motifs. Both PSA and MSA alignments are represented in Figure 6. More than two sequences can be aligned with multiple sequence alignment methods. Generally, these tools align the closer sequences pairwise, then the representative sequences of each group are aligned together. Muscle also has a refinement step that avoids gaps formed by the alignment between groups of sequences.

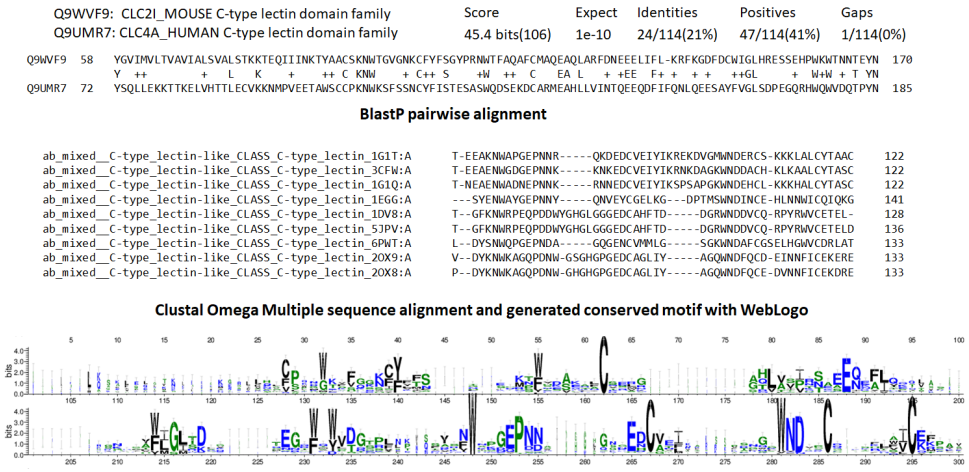


Figure 6: Pairwise sequence alignment, MSA and WebLogo (Crooks et al. 2004)

The Basic Local Alignment Search Tool (BLAST (Johnson et al. 2008)) is a sequence database mining tool used to identify sequence similarities. It is maintained by the NCBI and available as a web server and as a standalone software.

2.3.2 Profile generation methods from conserved peptide motifs

Profiles are informatic models of protein families / conserved domains. To generate a model, a domain MSA is converted into a regular expression (regex), a matrix, an HMM profile or other representations. A regular expression (regex) is a linear sequence of characters that define a search pattern. Regex patterns are used by regex methods for "find and replace" operations on strings, or for input validation. They have the advantage of searching for the best match, suitable for conserved protein motifs, and were used by Prosite first release. Matrix representation of a conserved domain MSA is called a position-specific scoring system (PSSMs). At each position in the alignment, the amino acids are scored according to their frequency. Substitution matrices (such as BLOSUM matrices) can be used to represent the higher and lower probability of substitutions by adding weighting scores.

Detecting protein homologs with high sequence identity is much easier than detecting those with low sequence identity, and is fundamental for the prediction of protein function. Profile HMMs also convert MSA into a position-specific scoring system, but has the advantage of capturing position-specific substitutions, insertions and deletions scores. Profile Hidden Markov models (HMM) method inner working is illustrated in Figure 7. HMMs are widespread methods for detecting sequence similarity to find related and particularly distantly related sequences. A few related aligned sequences can be used to construct an HMM profile to screen large sequence databases (Eddy 1998).

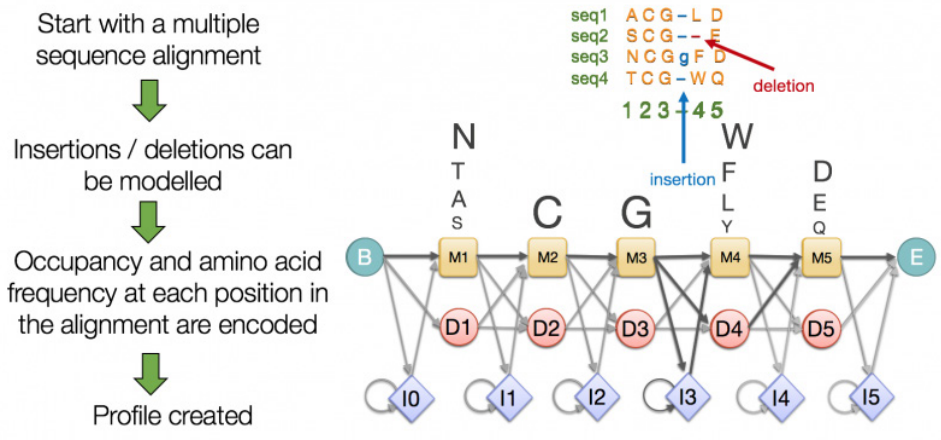


Figure 7: A profile HMM modelling a multiple sequence alignment. For each position of the HMM model, the M state corresponds to a match, the D state for deletion or gap; and the I state to insertion(s). Arrows represent statistical transition between states (El-Gebali et al. 2018)

HMMER analysis tool is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called “profile hidden Markov models” (profile HMMs) to detect remote homologs as sensitively as possible. It is often used together with a profile database, such as Pfam, and can also work with query sequences. HMMER can be downloaded or is available on a web server (Potter et al. 2018).

2.3.3 Tandem repeat detection tools

Gene duplication is a central mechanism that allows new genetic material to be generated during molecular evolution. Gene duplications can happen due to several types of errors in DNA replication and repair machinery or due to biological actions of retrotransposons. Biological errors are mainly due to recombination during the meiosis with an error during chromosome duplications or separation, or slippage during DNA replication.

The mechanisms behind protein domain repeats are not yet fully understood, but they could have been generated by an internal duplication process within a gene, where a region or protein domain is duplicated and allocated next to its origin. This protein internal duplication process is also known as tandem duplication and distinct from gene duplication process. It might explain the frequent observation of several domain repeats from the same family in eukaryotic genomes (Moore et al. 2008).

Many large proteins have evolved by internal duplication, and the resulting internal sequence repeats correspond to functional and structural units (Nacher et al. 2010). Such repeated sequences are not easy to detect by classical alignment tools, and some specific ones have been developed. Among other tools, RADAR allows the automatic detection and alignment of repeats in protein sequences (Heger and Holm 2000) and is illustrated on a β-propeller lectin in Figure 8 where all 7 blades of the propeller are properly identified. Other tandem repeat detection tools are described by the review (Pellegrini 2015). RepeatsDB also provides classified protein tandem repeat structures (Paladin et al. 2017).

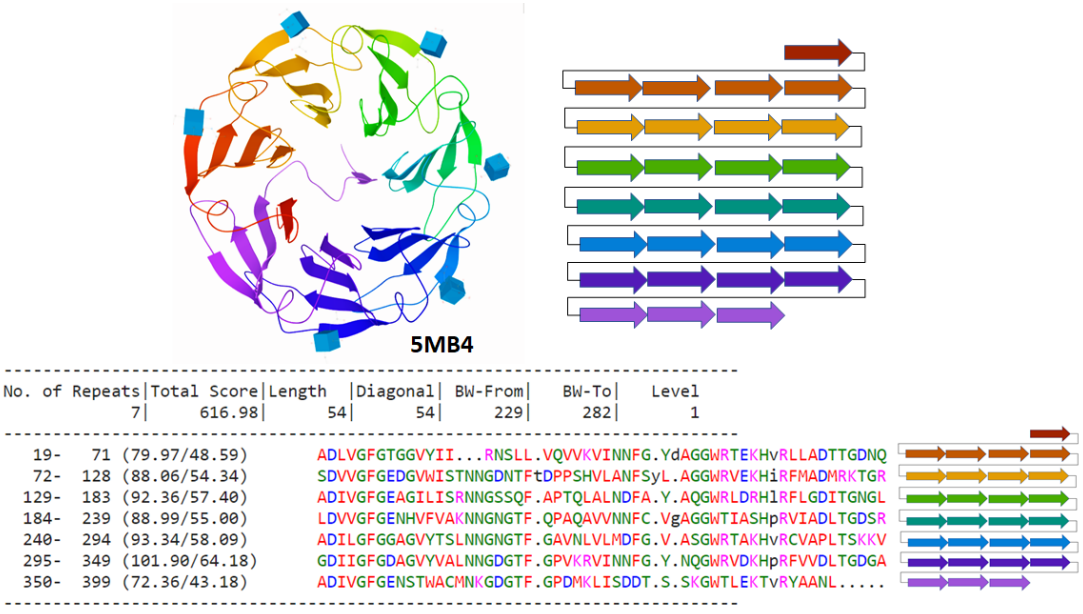


Figure 8: Propeller blade repeats as identified in RADAR, on the PDB structure 5MB4 from the species Psathyrella asperospora

2.4 3D structure of proteins: definition, databases and visualization

2.4.1 3D structure definition and obtention methods

A protein 3D structure describes the relative 3D position of all atoms of the protein and the interacting molecules if present. 3D structures of proteins can be obtained using different experimental methods: X-ray or neutron crystallography, nuclear magnetic resonance spectroscopy (NMR) or electron microscopy (EM). X-ray crystallography uses diffraction principles to determine the positions of the atoms in the crystalline state, whereas NMR spectroscopy is the only method that allows the determination of three-dimensional structures of proteins molecules in the solution phase. Electron microscopy (EM) allows high-resolution images of proteins and the new developments in cryo-EM allow for atomic resolution. 3D structures of proteins can also be modelled using de-novo approaches either from only a protein sequence and conserved motif or based on a reference 3D structure. CASP14, which evaluates protein structure prediction methods, demonstrated the efficiency of AlphaFold2 with up to 80% of a protein 3D structure correctly predicted at the cost of high computation power (CASP14 2021).

2.4.2 Evolution of the Protein Data Bank

In 1958-1960, the first three-dimensional (3D) crystal structures of proteins (hemoglobin and myoglobin) were determined (Kendrew et al. 1958; Perutz et al. 1960). Protein structures brought insights into protein function and evolution. The Protein Data Bank (PDB) was hosted at Brookhaven

National Laboratory (BNL) by Walter Hamilton (Protein Data Bank, 1971). In 1972, David Phillips announced structural biology's "coming of age" (Cold Spring Laboratory Press, 1972). The timeline of Key PDB Events from the release until 2011 is illustrated in Figure 9.

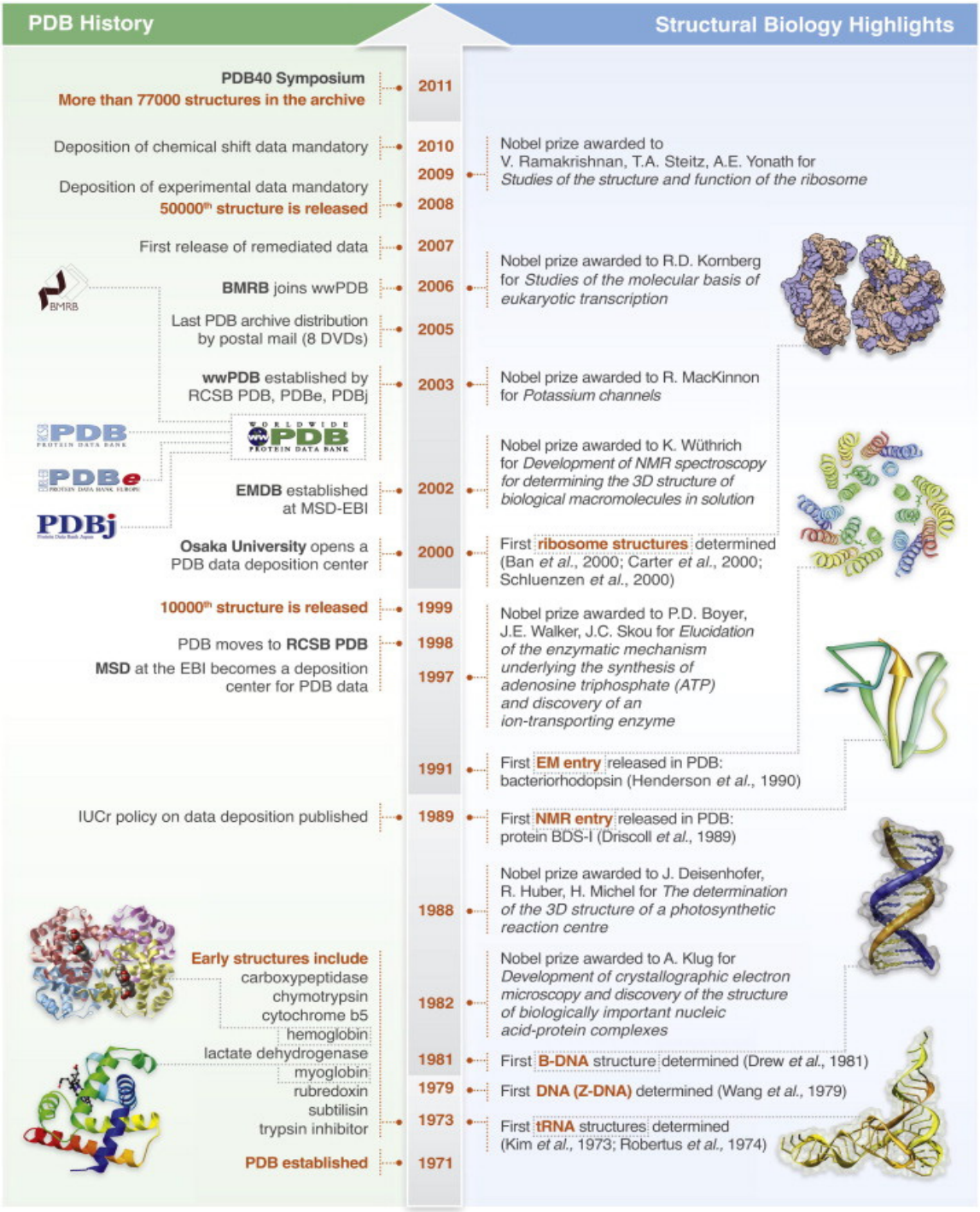


Figure 9: Timeline of Key PDB Events and Structural Biology Highlights, 1971–2011 (Left) Key events in the evolution of the PDB. (Right) selected key structures in the field of structural biology (Berman et al. 2012)

To complete bioinformatic approaches, the function of proteins can be studied in an integrative manner, bringing together 3D structure with other experimental methods from molecular biology, biological engineering and biochemistry domains. Thanks to the 3D structures of a protein, it is possible to accurately identify its single or multiple functional domains. 3D structures allow us to efficiently visualize and study the interactions between the protein and the ligands. These protein-ligand interactions are very important for the transport, synthesis and degradation of molecules. Besides, the manipulation of the protein or ligand allows the development of new drugs and antibodies by pharmaceutical companies.

2.4.3 3D structure databases

Since 1971 the protein 3D structures are managed by the Worldwide PDB (wwPDB) and stored in a world-wide repository PDB archive. This ensures that the PDB is freely and publicly available to the global community. 3D structures can be proteins, nucleic acids, viruses, polysaccharides, and other complex biomolecular assemblies. Thanks to technological progress lots of 3D structures of molecules are released every year as shown in Figure 10.

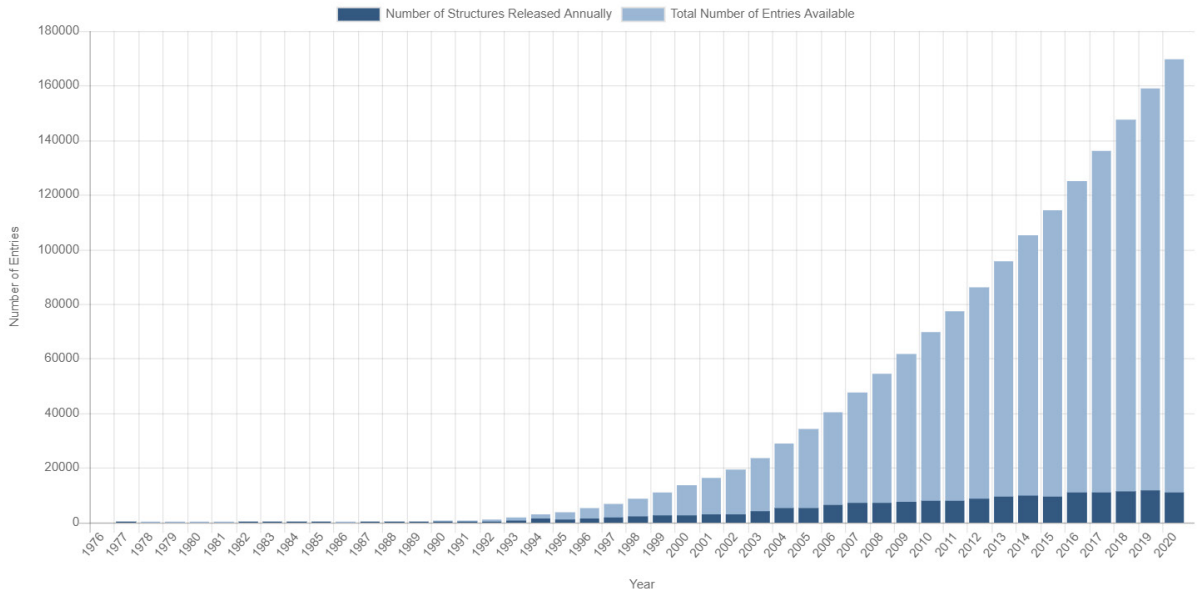


Figure 10: Increase in 3D structures of molecules available in the PDB. Their facility of obtention and resolution have improved significantly with 11178 new structures in 2018 and 11513 new structures in 2019; which is 32% more compared to the 7747 structures released in 2010 (Burley et al. 2018).

The 3D structures of proteins are stored in .pdb or .mmcif format files and require databases and interfaces to make them accessible to the public worldwide with their annotations and publications.

In addition, these structures must be linked to information about the corresponding protein and associated molecules. Databases are needed for this purpose, such as the RCSB, PDBe, PDBj sites.

The web databases RCSB PDB (USA, (Burley et al. 2018)), PDBe (Europe, (Armstrong et al. 2020)), PDBj (Japan, (Kinjo et al. 2017)) and BMRB for NMR structures (USA, (Ulrich et al. 2008)) act as data deposition, processing and distribution centers for PDB data (Burley et al. 2017). All three portal RCSB, PDBj and PDBe provide direct information on the 3D structures: 3D image and an interactive 3D viewer, files with the nature and spatial coordinates of the atoms, the nature of chemical groups/component of the 3D structure (amino acids, ligands), covalent and non-covalent linkages between the protein and its ligand or a simplified representation of the residues surrounding each ligand, the symmetry of the 3D structure. They also provide related information with external links protein UniProt AC, sequences, bibliography, taxonomy. Finally during the last years 2017-2020 all the three portals have improved the glycan identification, when present, and provide both the whole glycan and the composing monosaccharides (Refer to glycan section).

EM databases are now being developed thanks to the emergence of EM and cryoEM technologies: EMDataResource (EMDR; https://emdataresource.org, (Lawson et al. 2020)). The Electron Microscopy Data Bank (EMDB) at the European Bioinformatics Institute (EBI) was launched in 2002 (Abbott et al. 2018).

2.4.4 Tools for 3D visualization

Files of atom coordinates require tools for their visualization, to allow the 3D analysis of molecules, proteins and ligands. For proteins, different types of visualization are available: balls and sticks for atoms and bonds, ribbon representation for secondary structures with α-helix and β-sheets, or surface representations for whole domains. The ribbon representation was implemented in 1977 by (Richardson 1977). During the 1980's, Evans & Sutherland manufactured a computer system for crystallographers, which displayed the electron density map and enabled an amino acid sequence to be fitted manually. In 1982, Arthur M. Lesk and co-workers developed the first program for automatic generation of ribbon diagrams using PDB files (Lesk and Hardman 1982). In 1983, the molecular surface representation was developed by Michael Connolly (Connolly 1983). In 1992, the Richardsons described the kinetic image and their supporting programs MAGE and PREKIN (Richardson and Richardson 1992), followed by RasMol freeware in 1994 (Saqi and Sayle 1994) and VMD in 1995 (Humphrey et al. 1996).

In 2000, Warren Lyford DeLano launched the PyMOL open-source molecular viewer, distributed by Schrödinger since 2009, developed to drive the discovery of new medicines. The reviews from

Johnson and Hertig (Johnson and Hertig 2014) discuss available tools and help identify the best visualization methods for addressing specific biological questions (Martinez et al. 2019). 3D structures visualization can be done using pre-installed molecular graphics software or web interactive tools available on PDB databases.

The recent development of 3D visualization methods and the development of both graphics hardware, computational powers of computers and smartphones and rapid growth of the web allowed new interactive web visualization in 3D of molecules and proteins with tools such as JSmol (Hanson and Lu 2017), NGL (Rose and Hildebrand 2015), LiteMol (Sehnal et al. 2017) and PV (https://github.com/biasmv/pv).

Used by the PDBe, LiteMol is a streamlined structure viewer which enables a PDB structure to be conveniently displayed. For structures determined by X-ray crystallography, there is also the option to display electron density. LiteMol uses CoordinateServer to load the 3D structure of a protein in an efficient way, allowing it to load only the data required to display the part of the molecule the user wishes to visualize. The viewer is illustrated in Figure 11.

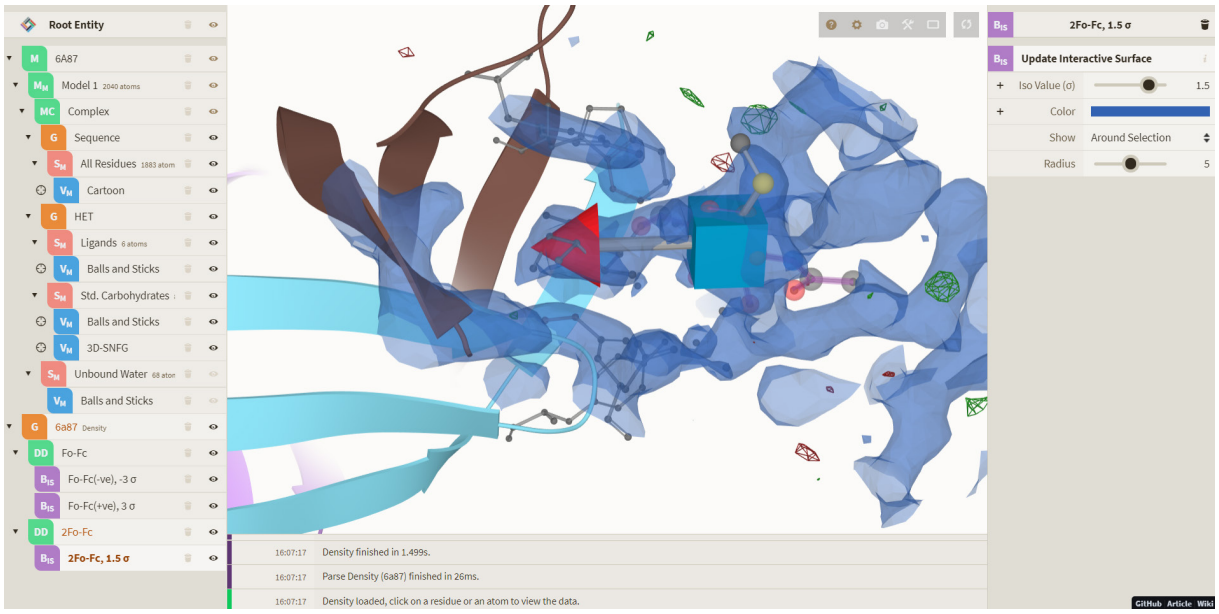


Figure 11: LiteMol viewer with the represented 6A87 structure, focused on the bound glycan with the electron density displayed (Sehnal et al. 2017).

RCSB uses JSmol to display the 3D structures in an interactive viewer and NGL is a WebGL based 3D viewer powered by MMTF. JSmol is less reactive than LiteMol. The NGL Viewer is a web application for the visualization of macromolecular structures. PV is a JavaScript viewer to display protein structures directly in the browsers. PV uses WebGL for rendering. It has been implemented with high performance and is able to render very large molecules at interactive frame rates.

The combination of such web 3D viewer with servers for predictions of protein-protein interactions, protein ligands binding pockets and docking allowed the emergence of new types of 3D web services such as PrankWeb: a web server for ligand binding site prediction in 3D structures and their visualization (Jendele et al. 2019).

The Protein-ligand interaction profiler (PLIP) is a web service for fully automated detection and visualization of relevant non-covalent protein-ligand contacts in 3D structures, freely available at www.projects.biotec.tu-dresden.de/plip-web (Salentin et al. 2015). As it focuses on non-covalent linkage, covalent ligands are filtered. This is very important to detect and analyse lectins that interact with glycans in only a non-covalent way. It returns a list of detected interactions on a single atom level, covering seven types of interactions (hydrogen bonds, hydrophobic contacts, pi-stacking, pi-cation interactions, salt bridges, water bridges and halogen bonds).

Thanks to the new hardware for virtual reality games, 3D representations of complex carbohydrates and polysaccharides in virtual reality is possible and has been implemented by UnityMol (Lv et al. 2013). It uses virtual reality helmets to allow a complete 3D visualization of 3D objects. The SweetUnityMol version includes tools for the visualisation of glycans (Pérez et al. 2015). VRmol is an integrative web-based virtual reality system to explore macromolecular structure (Xu et al. 2020).

2.4.5 Tools for structural alignment and superimposition of proteins

Protein structural alignment method is used to compare two structures and identify similarities in their 3D conformations, instead of comparing their peptidic sequences. Protein structural alignment is used to identify ligand binding pockets, classify proteins and study their evolution, or identify convergent evolution. Compared to structural superposition based on shared atoms between two structures, structural alignment does not require the selection of common atoms as a reference.

Structural alignment refers to the alignment, in three dimensions, between two or more molecular models. In the case of proteins, this is usually performed without reference to the sequences of the proteins. The root-mean square-deviation (RMSD) score reflects the divergence between two aligned structures. When the models align well, it suggests evolutionary and functional relationships that may not be discernible from sequence comparisons, as the structure is more conserved than the sequence (Russell et al. 1997).

CATH and SCOPe protein databases, RCSB and PDBe 3D structure portals, iSARST (Lo et al. 2009) and the older DALI (Holm and Laakso 2016) each provide tools that use a PDB structure to search for similar PDB structures or identify structural domains. For example, RSCB provides the structure

alignment using the algorithms CE and FATCAT. Reviews of algorithm and method for protein alignment (Poleksic 2009). Aligned fragment pairs (AFPs) are based on similarities in local geometry and used by the algorithm to search for 3D similarities. CE uses a 'rigid-body' based alignment. In contrast to this, FATCAT allows the introduction of 'twists' into the alignment for more flexibility. Multiple tools are available for the alignment of 3D structure with structure only approaches (Kawabata 2003), or both structure and sequence approaches (Gelly et al. 2011), other tools for multiple structure comparison such as mulPBA (Léonard et al. 2014).

2.5 3D based classification of folds and proteins

The 3D structure has the advantage of being more conserved and allows identifying similarities between foreign proteins. Moreover, recent years research with more efficient methods and popularity for drug design produced a lot of novel structures. It is a relevant first level for protein classification.

Regularities are present in secondary structures assembly (Levitt and Chothia 1976) and the topologies of the polypeptide chains (Sternberg and Thornton 1976; Richardson 1977). They come from the physical and chemical properties of proteins (Chothia 1984) and can be used to classify protein folds.

The CATH database provides a classification of protein domain structures based on sequence and structure similarity (Sillitoe et al. 2019). There are four main levels: class, architecture, topology and homologous family. Homologous families have either significant sequence similarity (≥35% identity) or high structure similarity (SSAP > 80) and 20% identity. Structural similarity is evaluated by an automatic method (SSAP), which assigns a score of 100 for identical proteins and generally returns scores above 80 for homologous proteins. Scores for more distant folds are higher than 70 (topology or fold level). To name fold classes, CATH uses either SCOPe fold names or only SCOPe numbers.

The SCOPe database provides a classification of the structure of proteins, based on the automatic and curated classification of PDB structures. The SCOPe (Sillitoe et al. 2019) database hierarchically classified domains from the most studied protein folds based on their structural and evolutionary relationships. Protein structures are classified using a combination of manual curation and automated methods. The release SCOPe 2018, classifies 90 992 PDB entries (about two-thirds of PDB entries). Both CATH and SCOPe classification are compared in Figure 12.

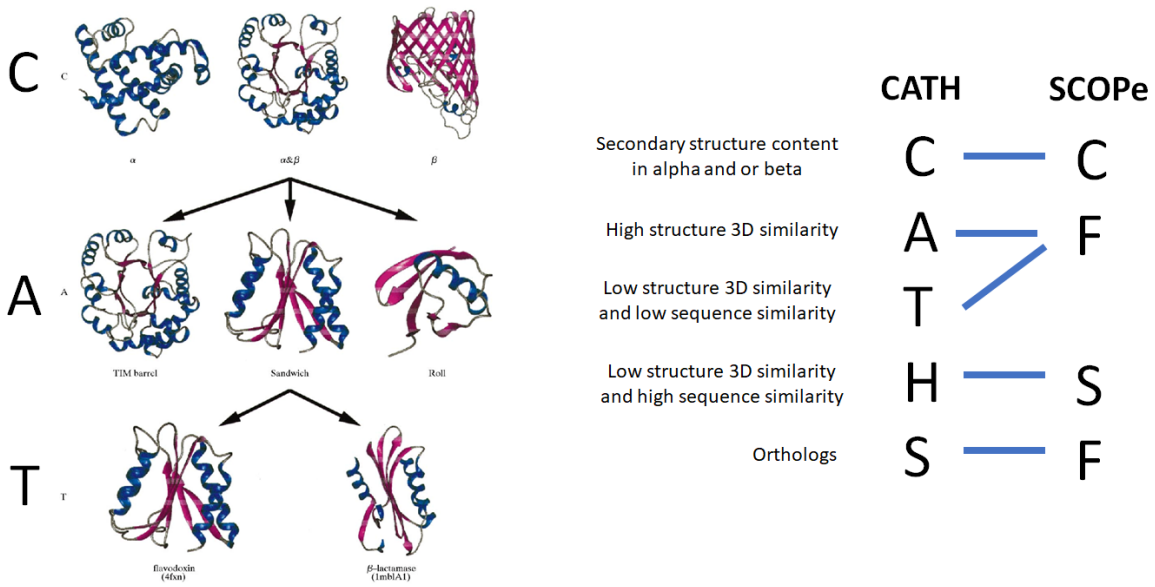


Figure 12: CATH classification compared to SCOPe classification

The mapping of the protein folds between the SCOP and CATH resources have differences in the way the domains have been grouped at the fold level. At the superfamily level, 70% can be mapped between the two resources. The SUPERFAMILY resource is analogous to Gene3D, and it was created to separate structural and evolutionary relationships, to allow connection between proteins that are evolutionarily related to having different folding arrangements.

Another recent structural classification is the Evolutionary Classification of Protein Domains, or ECOD, by the Grishin group database of 3D structures for remote homologs identification. As of March 2016 it contained 116,441 PDB structures ((Cheng et al. 2014); http://prodata.swmed.edu/ecod).

3 Glycosciences and Glyco-Bioinformatics

This section will present glycosciences and glycans followed by the development of glyco-informatics resources and tools. Finally, the lectins that are glycan-binding proteins will be presented together with the bioinformatic databases and tools for lectins.

3.1 Glycoscience, glycobiology and glycans

Glycoscience defines all research fields related to glycans. Glycobiology addresses the biological mechanism of synthesis, transport, recognition, and catalysis of glycans. Glycans are composed of one or multiple monosaccharides assembled in chains or complexed branched structures. Cells from all organisms are covered by glycoconjugates, which play an important role in their interactions with other cells and circulating molecules. Many proteins, including enzymes, transporters, and receptors interact with the glycan. Biochemists need databases online providing information on glycans and their receptors, and therefore glyco-bioinformatics is a rapidly growing domain.

Glycoscience research drives innovation potentials in existing drugs. It has future possibilities in Glyco-based precision medicine diagnostics. Glycans are also used as vaccine components (Corolleur et al. 2020). Glycoengineering is an essential contribution to drug development (Beck et al. 2017).

3.1.1 Glycan definition and functions

Glycan is a generic term to describe linear or branched biomolecules where each building block (monosaccharide) is attached to the next through a glycosidic linkage. As such, the term glycan refers to any form of mono, oligo and polysaccharides, a linear or branched polymer consisting of monosaccharide residues, such as cellulose. The relationships between monosaccharides, oligosaccharides and polysaccharides are similar to amino acids for proteins, or nucleotides for nucleic acids (Varki 2017). All oligosaccharides or glycan structures are created by linking different monosaccharides through glycosidic linkages to form linear or branched structures. All these features generate fantastic molecular diversity.

To improve the visual communication of glycan structures, the SNFG standardized representation of glycans was developed to associate and represent each residue and monosaccharides by simplified symbols ((Neelamegham et al. 2019); www.ncbi.nlm.nih.gov/glycans/snfg.html) as represented in Figure 13. It allowed tremendous progress for glycan communications.

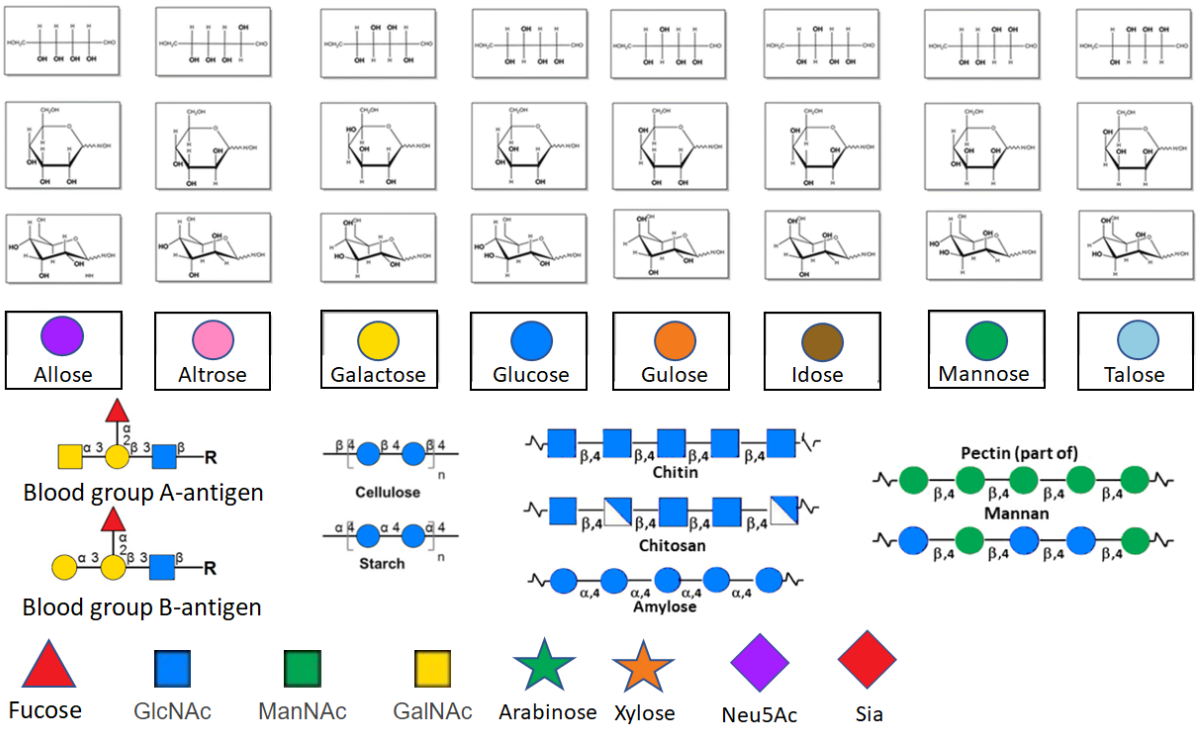


Figure 13: SNFG representation of Hexose compared to classical chemical representation. Example with glycans widely present in nature.

Carbohydrates have many structural roles with principally the glycocalyx a substantial physical barrier that covers all eukaryotic cells and the polysaccharide coats of various prokaryotes represent. Cellulose is a natural polysaccharide consisting of a linear chain of D-glucose molecules. This biopolymer is the main constituent of plant cell walls, including wood. Chitin is a nitrogenous polysaccharide resulting from the polymerization of N-acetylglucosamine. It plays an essential structural role in several eukaryotic taxa, for example, the external cuticle of insects and shells of crustaceans, or the cell wall of fungi.

Carbohydrate have energy storage roles. For example the Starch composed of chains of D-glucose molecules is found in the reserve organs of many plants: seeds, roots, potatoes, sweet potatoes, fruits, bananas. Glycogen is a complex carbohydrate polymer of glucose, used by animals and fungi to store chemical energy.

Carbohydrates have numerous biological functions in a variety of biological pathways, including cell adhesion, signaling, and recognition (Cummings 2019), self / non self recognition. Glycans are part of the protein folding in eukaryotic cells. They play an important role in cell-cell interactions, and have a signaling role for pathogens or for signaling for example the degradation. Recognition and protein inner control biological functions of glycans are represented in Figure 14.

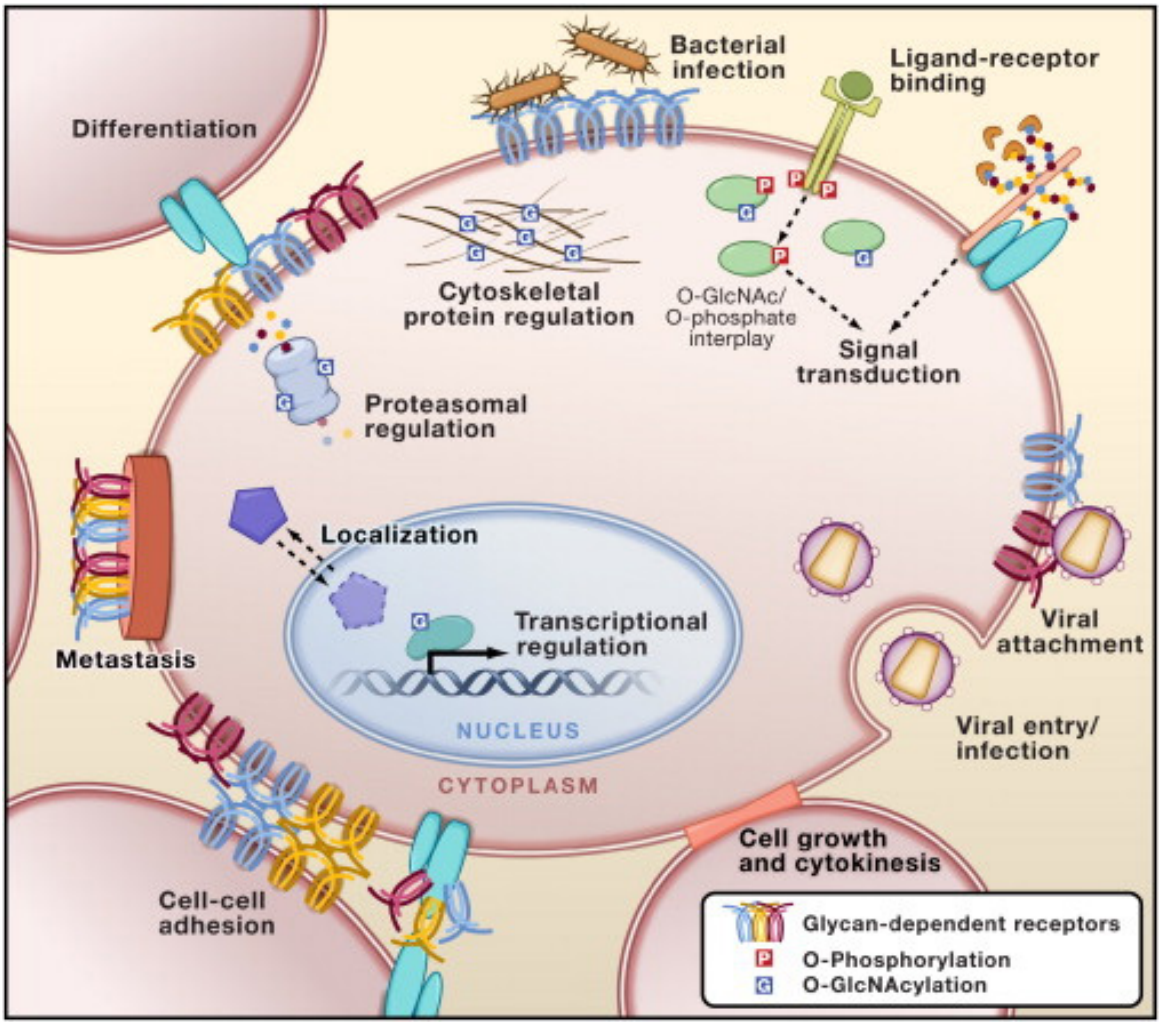


Figure 14: Simplified representation of some role of complex glycans. Complex glycans at the cell surface are targets of microbes and viruses, regulate cell adhesion and development, influence metastasis of cancer cells, and regulate a myriad of receptor-ligand interactions. Glycans within the secretory pathway regulate protein quality control, turnover, and trafficking of molecules to organelles (Hart and Copeland 2010).

Glycoconjugates are covalently linked with other chemical compounds such as peptides, proteins, lipids, respectively named glycopeptides, glycoproteins and glycolipids. A glycoprotein is a glycoconjugate in which a protein carries one or more glycans covalently attached to the polypeptide backbone, usually via N- or O-linkages in mammalian organisms. An N-glycan makes a glycosidic linkage with the side-chain nitrogen of an asparagine residue that is a part of a consensus peptide sequence NX(S/T). An O-glycan makes a glycosidic linkage with the terminal oxygen of a serine or threonine residue. Glycosaminoglycans are specific types of glycans also attached using O-glycosylation, shaped as long linear polysaccharides consisting of repeating disaccharide units.

3.1.2 Glycans of the cell membrane

The cell membrane surrounds the cytoplasm and nucleoplasm. It is composed of phospholipids, glycolipids and transmembrane proteins which can be glycosylated. Except for animals, a second layer named cell wall is present in most prokaryotes, algae, eukaryotes including fungi and plants. The cell wall provides structural support, protection, and filtering mechanism. A major function is stabilizing the pressure when water enters. The cell wall is composed of a large diversity of glycans. The different types of cell membranes and cell wall are represented for the plant in Figure 15, for the Fungi in Figure 16 and for the Bacteria in Figure 17. Cell membranes are made up of lipids and proteins. Membrane lipids have one hydrophilic and one hydrophobic pole; the most abundant are phospholipids, glycolipids and possibly cholesterol. Glycolipid is formed by attaching a lipid and a glycan. Glycans on the external surface of the cell wall (or the cell membrane especially for Animals) have a very important role in cell cell interactions, as they are recognized by glycan binding proteins at the surface of other cells (Hart and Copeland 2010).

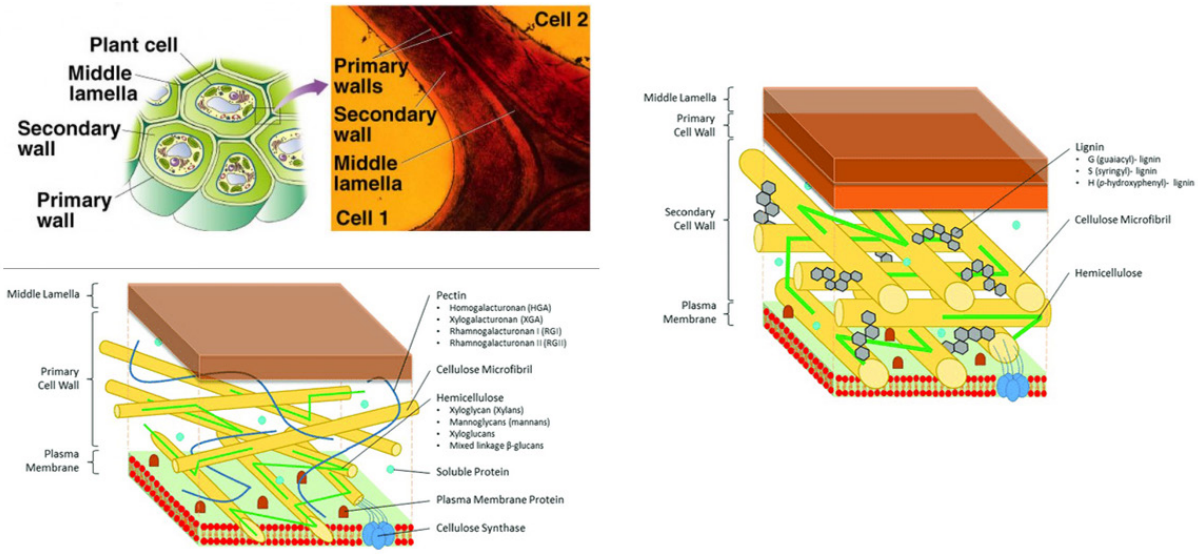


Figure 15: Composition of the plant cell wall. Cellulose is a major component of the plant cell wall, composed of linearly linked D-glucose units (Loix et al. 2018).

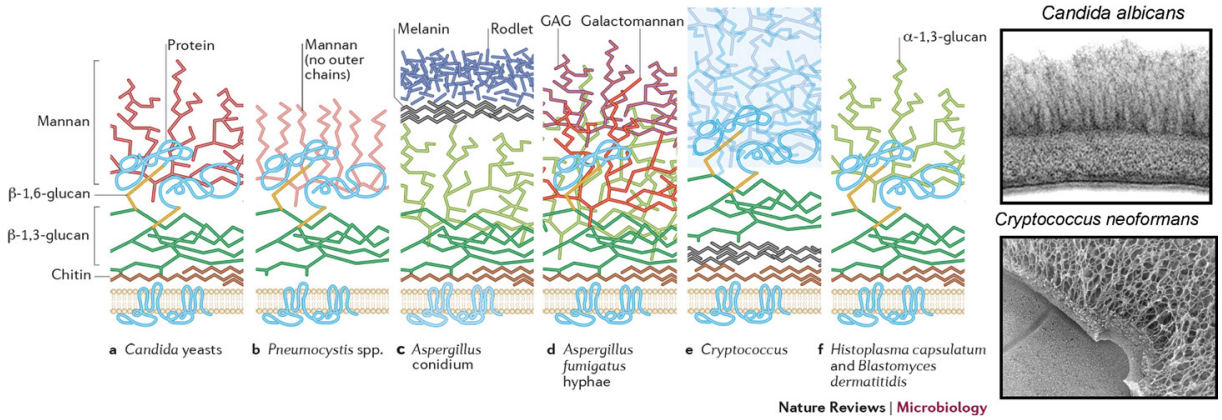


Figure 16: Structural organization of the cell walls of fungal pathogens (Gow et al. 2017). The glycan composition of the fungal cell wall varies depending on the species.

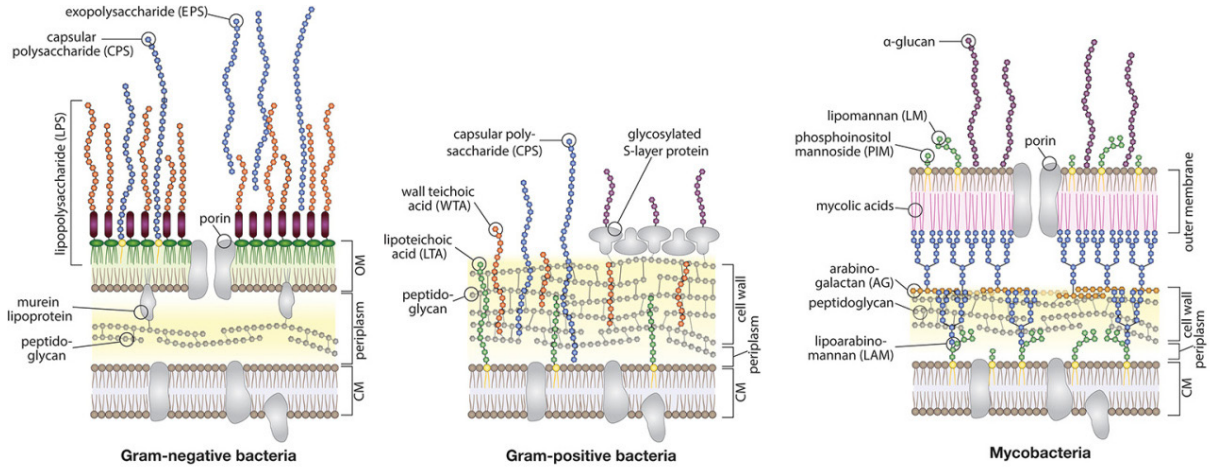


Figure 17: Different cell surfaces between Bacteria. Cell walls polysaccharides in bacterial and fungi are implicated in the interactions with the host receptors (Whitfield et al. 2017).

3.2 Glyco-Bioinformatics

Informatics plays a critical role in modern biology, using genes and proteins similarity it is possible to predict the function of proteins. Contrary to glycans, genes and proteins are typically linear molecules suitable for powerful informatics tools. Compared to genomics and transcriptomics, glycomics and the associated glyco-bioinformatics are less explored. The development of informatics tools for glycobiology is complex as Glycan structures have no template from which to be predicted; they are assembled and degraded by cellular metabolism and the active glyco-enzyme. Glycans are built block by block in a tree-like pattern, assembled by enzymes such as glycosyltransferases and the resulting structures can be extremely complex.

Glyco-Bioinformatics aims to organise glycan and glyconjugate information to generate *in silico* new knowledge through the interconnection of available resources. The determination of glycan

composition *in silico* with analytical methods (e.g. mass spectrometry) supports the characterization of glycans and glycoconjugates in a given sample. This information is then used to elucidate their functions in the inner working of an organism. Glycan diversity is generated through biosynthetic pathways. Pathway-based prediction of glycans is seriously investigated in the Atlas described in (Narimatsu et al. 2019) where human glycosyltransferases are organized in glycosylation pathway maps used to reconstruct a large part of the structural diversity of the human glycome. Bioinformatics played an essential role in the integration of genomics, transcriptomics and proteomics but glycomics and metabolomics are still isolated due to their complexity and lack of knowledge. But this also shows that Glycomics has a lot of potential.

These glycans can also be glycoconjugates with various databases. Finally, functional glycomics has also been largely developed to study the interactions between the glycan and the proteins. Different databases and tools are available to store, integrate and study glycan composition, glycosylation sites. glycoprotein, glycan–protein interactions, glycogenes. The different fields of glyco-informatics are represented by Figure 18. To connect the different categories of information, new portals are developed (Li et al. 2020; Abrahams et al. 2020). New approaches are developed towards glyco 3D structural data generation (Scherbinina and Toukach 2020).

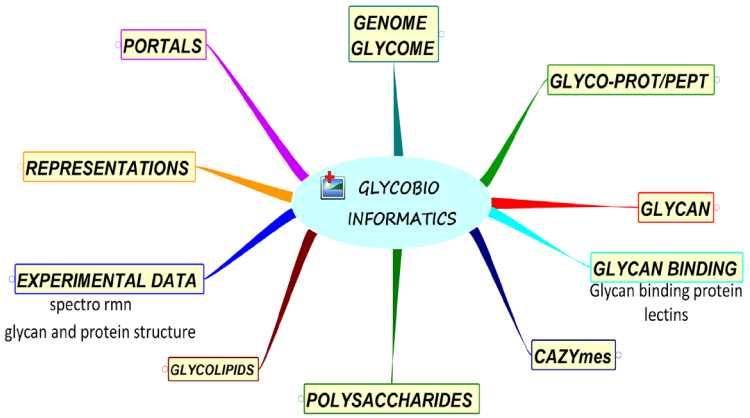


Figure 18: The different domain of Glyco informatics (Davide Alocci, Frederique Lisacek 2019)

Glycan-binding databases and tools will be discussed later in the Lectin section. Will not be discussed as it is not required for the subject:

- Databases and tools for experimental results of glycan studies
- Glycoproteomics tools, ie. predict possible glycosylations sites
- Polysaccharides databases: EPS (Birch et al. 2019), EK3D (Kunduru et al. 2016), MATRIX-DB (Clerc et al. 2019a), POLYSACDB (Aithal et al. 2012), Polysac3DB (Sarkar and Pérez 2012)
- Glycolipids tools

3.2.1 Informatic formats of glycans

Many bacterial glycoconjugates are on the cell envelope/surfaces, essential for viability. Surface glycoconjugates drive crucial interactions for immune defences. They show a wide diversity in structures and use sugars not found elsewhere in nature. Antigens of many pathogenic microorganisms have carbohydrate origin. The recognition of bacteria by the host immune system is determined by the structure of these compounds. The bacterial saccharide sequences have a greater diversity of monosaccharides, with species-specific monosaccharides. These structures are difficult to visualize and encode.

Informatics formats were developed to store the glycan structural information with different formats adopted due to the difficulty to find a suitable format to complex glycans. Multiple formats representing glycan structures have been developed successively to improve the encoding of glycan structures; they are listed in Table 4. Conversion between formats is possible, thanks to tools such as GlycanFormatConverter (Tsuchiya et al. 2019).

Format	Description	Style	Human readability	Reference
IUPAC	The International Union of Pure Applied Chemistry (IUPAC) has specified the 'Nomenclature of Carbohydrates' to describe complex oligosaccharides based on a three-letter code to represent monosaccharides	linear text	✓	(McNaught 1997)
KCF	Kyoto Encyclopedia of Genes and Genomes (KEGG) uses the format KEGG Chemical Function (KCF) to represent glycans	connection table		(Hashimoto et al. 2006)
LinearCode	LinearCode is a linear syntax for glycoconjugates used by CFG	linear text	✓	(Raman et al. 2006)
GlycoCT	GlycoCT was developed by the EuroCarbDB project.	connection table		(Al Jadda et al. 2015)
WURCS	Web3 Unique Representation of Carbohydrate Structures (WURCS) was proposed as a new linear notation for carbohydrates for the Semantic Web of the GlyTouCan project	linear text		(Tanaka et al. 2014)
Linucs	It is the first to use a connection table approach. Linucs is another linear notation from 2001	linear text	✓	(Bohne-Lang et al. 2001)

Table 4: features of glycan informatic codes. The glycan is either described in a connection table that first describes the monosaccharides and then the linkages, or a linear text representing the whole glycan structure. 'Human readability' indicates whether the text is in general human readable.

3.2.2 Glycans databases

Glycan structure refers to the composition and construction in monosaccharides but not to the glycan 3D structure. Combination of methods is needed to elucidate glycan structural information and their

conjugate. It includes liquid chromatography (LC), capillary electrophoresis (CE), nuclear magnetic resonance (NMR), mass spectrometry (MS), and (micro)arrays. A large variety of tools are developed to address the difficulty of analysing glycoproteins and glycans (more than 50% of all proteins are modified by glycans). Improvements in analytical methods the qualitative and quantitative data of glycans, glycosites, glycopeptides, and glycoproteins have increased tremendously and requires the development of databases and informatics tools to store and integrate those data (Gray et al. 2019). Generated data can be submitted to a glycan database with specific standards, such as GlycoPost (Watanabe et al. 2020).

Name	Description	Link	Ref
GlyTouCan	repository of glycan structures with unique accession identifiers. Allows the deposition of glycan structure with minimal metadata.	https://glytoucan.org/	(Fujita et al. 2020)
Glyco3D-BioOligo	contains representations, 3D structures and NMR spectra of most occurring glycans	http://glyco3d.cermav.cnrs.fr/search.php?type=bioligo	(Pérez et al. 2015)
GlycoStore	Provides a centralised resource that combines glycan structure information with chromatographic separation and electrophoretic data.	https://glycostore.org/	(Zhao et al. 2018b)
KEGG GLYCAN	collection of experimentally determined glycan structures associated with KEGG pathways and CarbBank	https://www.genome.jp/kegg/glycan/	(Aoki- Kinoshita and Kanehisa 2015)
UniCarb-DB	glycan sequences with associated MS fragmentation database for glycans, with information about structure, taxonomy, tissue, and associated protein	https://unicarb-db.expasy.org/	(Hayes et al. 2011)
Carbohydrate Structure DataBase CSDB	structures and taxonomy of carbohydrates present in nature	http://csdb.glycoscience.ru/database/	(Toukach and Egorova 2016)
Glycosciences.de	database of carbohydrate sequences from literature that provides taxonomy, linking to PDB, NMR and 3D modules	http://glycosciences.de/	(Böhm et al. 2019)

Table 5: glycan databases

Informatics tools and databases were first developed to store, process these glycans data and associate them with glycan structures. For example UniCarb-DB associates analytical MS and MS/MS Data into a glycan structure, which can then be used for further exploration. The different glycans databases, which address different analytical methods (NMR, MS, LC) and are currently functional and available online, are listed and described in Table 5. Glycan (structure) databases provide a vast amount of data on monosaccharide composition.

3.2.3 Glycoproteomics databases

Glycosylation modifies proteins or lipids to which glycans are linked. Because of their biosynthesis and structural complexity, it is not currently possible to accurately predict the structures of glycans from genomic and transcriptomic data. Instead, the identity of each glycan in a biological sample must be identified using analytical methods. These are the requirements to understand the biological roles and consequences of glycans, glycoproteins, and protein-glycan interactions. This depends on the availability of databases that allow these structures to be archived, organized, searched, and annotated. However, the complexity and diversity of glycan structures make the development of these databases challenging.

Name	Description	Link	Ref
GlyConnect	Characterization of protein glycosylation at several level: information on glycan structures and compositions, glycoproteins, glycosylation sites, taxonomy, tissue expression, and diseases in cross reference with UniProt, NextProt, GeneCard and the PDB	https://glyconnect.expasy.org/	(Alocci et al. 2019)
GlycoFish	Database of Zebrafish N-linked Glycoproteins Identified Using SPEG Method Coupled with LC/MS	http://betenbaugh.jhu.edu/GlycoFish	(Baycin-Hizal et al. 2011a)
GlycoFly	database for Drosophila N-linked glycoproteins identified using SPEG—MS techniques.	http://betenbaugh.jhu.edu/GlycoFly	(Baycin-Hizal et al. 2011b)
GlyProt	Tool. N-glycan conformations to be attached to all the spatially accessible potential N-glycosylation sites	http://www.glycosciences.de/glyprot/	(Bohne-Lang and Von der Lieth 2005)
O-GlycBase	A database of glycoproteins with O-linked glycosylation sites	http://www.cbs.dtu.dk/databases/OGLYCBASE/	(Gupta et al. 1999)
UniCarbKB	Integrating glycan structure abundance and their association with proteins, compositional glycoproteomics data, and disease associations	http://unicarbkb.org	(Campbell and Packer 2016)

Table 6: Glycoproteomics databases

The different Glycoproteomics databases describe experimental sources of glycoproteins, glycans, and glycosylation sites data. For example, GlyConnect can be used to identify a glycan associated protein through glycosylation sites, view them in a 3D structure if available, and find associated diseases in specific tissue. The databases currently functional and available online are described in Table 6.

3.2.4 Glycan drawing tools, 3D builder and 3D structure verification

Glycan drawing tools allow the use of the simplified SNFG representation to build the glycan of interest, which is much easier than directly writing manually a IUPAC code, or to generate a 3D structure. Using the graphical representation constructed by the user, all types of text can be generated to represent the glycan. The different Glycan drawing tools, 3D builder of glycans, validation tools for glycan 3D conformation are described in Table 7.

Name	Description	Link	Ref
DrawGlycan-SNFG	tool to render glycans and glycopeptides with fragmentation information	http://www.virtualglycome.org/DrawGlycan/	(Cheng et al. 2017)
SugarSketcher	Online Glycan Drawing	https://glycoproteome.expasy.org/sugarsketcher/	(Alocci et al. 2018b)
GlyTouCan graphical search	Glycan Drawing to search in Glytoucan	https://glytoucan.org/Structures/graphical	(Tiemeyer et al. 2017)
GlycoGlyph	glycan structures using a GUI and providing the linear nomenclature and SVG SNFG representation	https://glycotoolkit.com/Tools/GlycoGlyph/	(Mehta and Cummings 2020)
CT23D, MatrixDB	translate glycosaminoglycan sequences into 3D models	http://matrixdb.univ-lyon1.fr/	(Clerc et al. 2019)
POLYS-GLYCAN BUILDER	An intuitive application to build 3D structures of polysaccharides	http://glyco3d.cermav.cnrs.fr/builder.php	(Engelsen et al. 2014)
Privateer from CCP4	conformational validation of carbohydrate structures	http://legacy.ccp4.ac.uk/html/privateer.html	(Agirre et al. 2015)
PDB-CARE	conformational validation of carbohydrate structures	http://www.glycosciences.de/tools/pdb-care/	(Lütteke and von der Lieth 2004)
Rosetta	suite for macromolecular modeling, suitable for correcting 3D errors in carbohydrates	https://www.rosettacommons.org	(Das and Baker 2008)

Table 7: Glycan drawing tools, 3D builder, 3D conformation validation tools

Based 2D model of a glycan's monosaccharides and linkages, and by selecting for each glycan-glycan linkage the angles using predefined maps, it is possible to construct a glycan or a glycosaminoglycan 3D structure model, as represented in Figure 19.

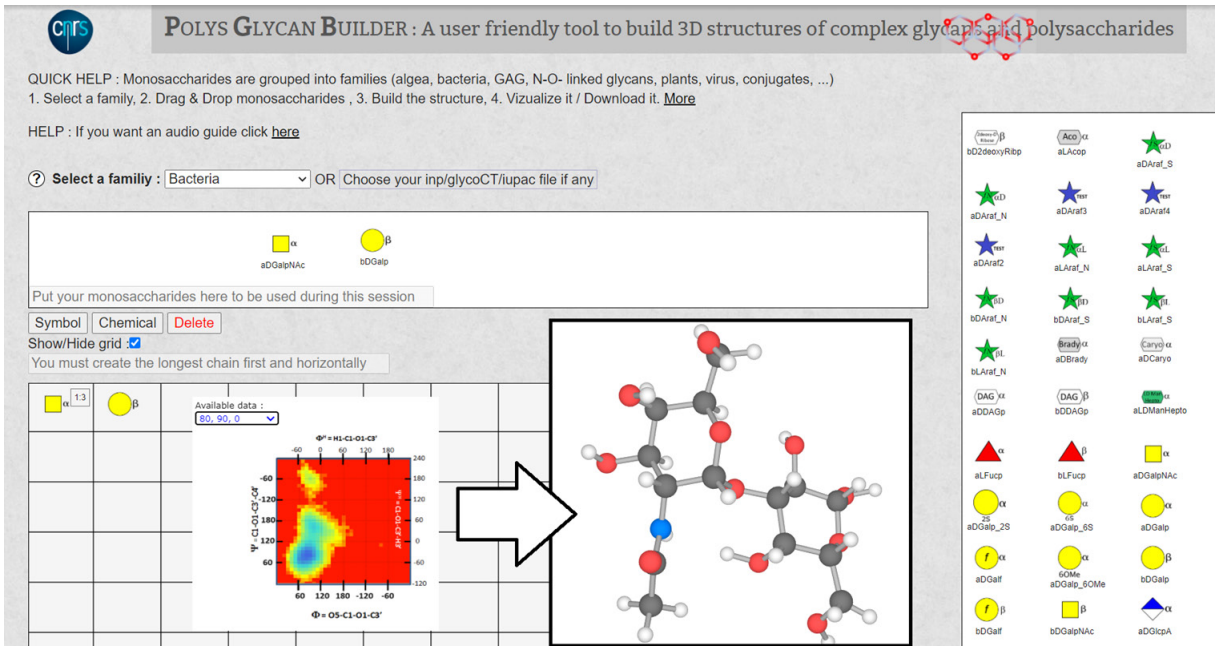


Figure 19: Polys glycan builder

3.2.5 Metabolic pathways of glycans: databases and tools

The exploration of protein glycosylation by analyzing and targeting enzymes involved in glycosylation processes thanks to major advances with new methods in quantitative transcriptomics, proteomics and gene editing allowed to produce in silico models of cellular glycosylation. Glycosylation pathways maps facilitate genetic approaches to create personalized glycosylation pathways (Schjoldager et al. 2020). Glycans enzymes and metabolic pathways databases and related tools are described in Table 8. To complement pathway databases, pathway-based prediction of glycans is investigated in the Atlas (Narimatsu et al. 2019).

Carbohydrate-Active enZymes (CAZymes) assemble, breakdown, and modify glycans and glycoconjugates using their catalytic and binding modules (Lombard et al. 2014). Each CAZymes family of the classification was created using protein domains experimentally characterized in the literature. For each family, predicted related proteins from protein sequence databases are available and are defined using both module modelling/calibration and manual curation.

Name	Description	Link	Ref
CAZy	Curated carbohydrate-active enzyme database, with protein families identified across species and associated genes	http://www.cazy.org/	(Helbert et al. 2019)
O-Glycologue	simulator of the enzymes of O-linked glycosylation	https://glycologue.org/	(McDonald et al. 2010)
GPP glycan pathway predictor	Prediction of glycan structures from gene expression data based on glycosyltransferase reactions	http://rings.t.soka.ac.jp/gpp.html	(Aoki-Kinoshita 2015)
glycomaple	visualization tool for pathways	https://glycomaple.glycosmos.org/	(Yamada et al. 2020)
glycompare	decompose glycan to a minimal set of intermediate substructures	https://github.com/LewisLabUCSD/GlyCompare	(Bao et al. 2020)
GNAT	describe glycans and glycosylation reaction networks	https://virtualglycome.org/gnat	(Liu et al. 2013)

Table 8: Metabolic pathways of glycans: databases and tools

3.2.6 Glycoscience portals

Glycobiologists and researchers have many resources technically available to them, but these are often hard to find. To provide a simpler unified access to glycoscience ressources and improve communication between databases and tools, glycoscience portals have been recently released and are listed in Table 9. They aim to connect the available resources but also incite the users to provide glycomics data using standardization rules.

Name	Description	Link	Ref
GlyCosmos	integration of omics data including glycogenes, glycoconjugates such as glycoproteins and glycolipids, molecular structures, and pathways.	https://glycosmos.org	(Yamada et al. 2020)
RINGS	algorithmic and data mining tools to aid glycobiology research	http://rings.t.soka.ac.jp/	(Akune et al. 2010)
GlyGen	Computational and Informatics Resources for Glycoscience	https://www.glygen.org/	(York et al. 2020)
Glycomics@ExPASy	connect and integrate glycomics data including glycoproteins, molecular structures, disease, tissues, glycan binding proteins, glyco-epitopes	https://www.expasy.org/	(Mariethoz et al. 2018)
Consortium for Functional Glycomics- CFG	data from the screening of the consortium's glycan array platform, glycogene microarray of tissues and cells, MALDI-MS screening of glycans from mouse and human tissue with information about histology, immunology, hematology, and metabolism	www.functionalglycomics.org	(Venkataraman et al. 2015)
Glycosciences.de	annotated data collection linking glycomics and proteomics data	http://glycosciences.de/	(Böhm et al. 2019)
Glyco3D	portal for databases covering 3D monosaccharides, disaccharides, oligosaccharides, polysaccharides, glycosyltransferases, lectins, monoclonal antibodies against carbohydrates, and glycosaminoglycan-binding proteins	http://glyco3d.cermav.cnrs.fr/home.php	(Pérez et al. 2015)

Table 9: Glycoscience portals

The Proteome Informatics Group (PIG) is part of the Swiss Institute of Bioinformatics (SIB) and the University of Geneva, develops bioinformatics databases to store, analyse glycobiology data and connect it with the other fields of bioinformatics, genomics, transcriptomics and proteomics. Glycomics@ExPASy portal provides access to Glyconnect and Sugarbind databases along with the tools: GlycoSiteAlign, Epitope Extractor and Glydin tool to generate an epitope network (Mariethoz et al. 2018). The database and tools relationships are represented by Figure 20.

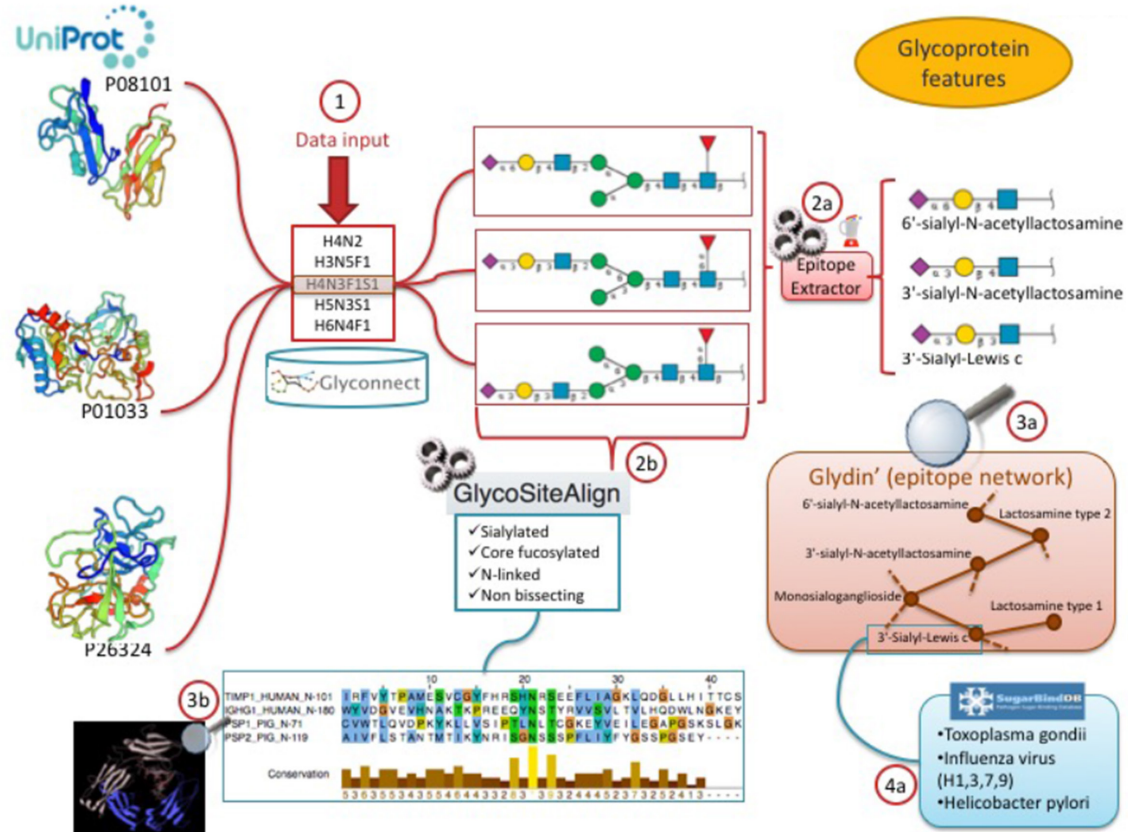


Figure 20: Glycomics@ExPASy platform. From composition to glycoprotein features. (1) A list of glycan compositions is given to GlyConnect, which retrieves all the related glycoproteins and glycan structures corresponding to this composition. Glycan structures can be further processed to extract contained glycan epitopes using EpitopeXtractor (2a). Glycoepitope results can be recognized by lectins in SugarBindDB (4a), where further information on the pathogens can be browsed (Mariethoz et al. 2018).

3.3 Lectins

3.3.1 Historical evolution of the definition

In 1888, the toxicity of castor bean (*Ricinus communis*) was demonstrated to originate from a hemagglutinating factor called “ricin.” in the PhD work of H Stillmark (for an history or ricin see (Polito et al. 2019)). In 1898, Elfstrand introduced in his PhD manuscript, the term “hemagglutinin” as a common name for all plant proteins that clump blood cells. Raubitschek reported in 1907 for the first time the presence of nontoxic lectins in the beans of several leguminous plants *Phaseolus vulgaris*, *Pisum sativum, Lens culinaris*, and *Vicia sativa*. The discovery that certain haemagglutinins selectively agglutinate the erythrocytes of a particular human blood group in the ABO system introduced the term "lectin". However, haemagglutinins are also called agglutinins. In 1952, Watkins and Morgan (Bhende et al. 1951) demonstrated that the agglutination properties of lectins are based on a specific sugar binding activity. Lectins were finally considered to be sugar-binding proteins that

could be distinguished from other proteins. The analysis of the interaction between lectins and carbohydrates is based on the agglutination of erythrocytes or other cell types. The carbohydrate specificity of lectins was then explored by indirect methods, such as inhibition of cell precipitation or agglutination by sugars or glycoconjugates. More recently, the introduction of high-performance techniques, such as front end affinity chromatography and glycan chips, have enabled the high throughput screening of carbohydrate collections with only small amounts of purified lectin.

In 1980, Lectins were defined as carbohydrate-binding proteins of non-immune origin that agglutinate cells and/or precipitate glycoconjugates (Hampton et al. 1980). In 1995, plant lectins were defined as "all plant proteins that possess at least one non-catalytic domain that binds reversibly to a specific mono- or oligosaccharide" (Peumans and Van Damme 1995). A distinction is made between hololectins and chimerolectins. Hololectins are proteins with only carbohydrate-binding domains (merolectin for one and superlectins for multiple domains). Chimerolectins are proteins with a lectin domain and other functional domains (Lis and Sharon 1986).

In this thesis is used the current definition of Lectins proposed by Sharon in 1993: “Lectin” are proteins without enzymatic functions that bind glycans specifically non-covalently and that are not antibodies.

In comparison to lectins, antibodies are manufactured by the immune system in response to an infection or inoculation. Individually, a lectin interacts weakly with sugars, but lectins can assemble in multimers or use series of repetition of the binding domain to multiply the number of weak interactions and create a stronger multivalent interaction, such as scratch/velcro. Using this definition, the lectins do not include carbohydrate-binding modules CBM (binding pocket of the enzyme) of glyco-enzymes such as oxydoreductases, transferases, esterases, glycosidases, proteases, other hydrolases, lyases, isomerases, and ligases. At present, at least 60 families of CBM are known (Helbert et al. 2019).

Lectins are ubiquitous in Nature, being found in all kinds of organisms, from viruses to humans (Sharon 2008). They have a biotechnological interest and are used as reagents for the study of glycoconjugates in solution and on cells, for cell characterization and separation and for the definition of glycan structure (Ambrosi et al. 2005). Lectins have biological functions in the infection process, immune response, and inflammation. Recent developments include the advent of glycomimetic or allosteric small molecule inhibitors for this important protein class and their use in chemical biology and drug research (Meiers et al. 2019).

3.3.2 Lectin structures and binding sites

The lectin domain that contains the binding pocket is called CRD for carbohydrate recognition domain, with different types of fold, such as β-helix, β-trefoil, α-β barrels. Present on the cell surface and in the cytoplasm, the proteins with lectins domains exhibit an important diversity of domain combinations, that are named architectures. Lectin structural classes are defined based on the different CRD fold and their architectures. Such diversity is illustrated using C type lectins.

The C-type lectin class is Ca++ dependent, used as a bridge between the binding pocket amino acids and the hydroxyl groups of sugars. The C-type lectin class is composed of a large variety of proteins with diverse oligomerization and composition in peptidic domains forming long tails in the extracellular matrix, as represented in Figure 21. For example, the collectin mannan-binding proteins are involved in the recognition of pathogens. The selectins consist of a CRD combined with an epidermal growth factor-like domain and a variable number of complement regulatory protein-like repeat domains and are involved in recruiting leukocytes at inflammation sites (Cummings and McEver 2017).

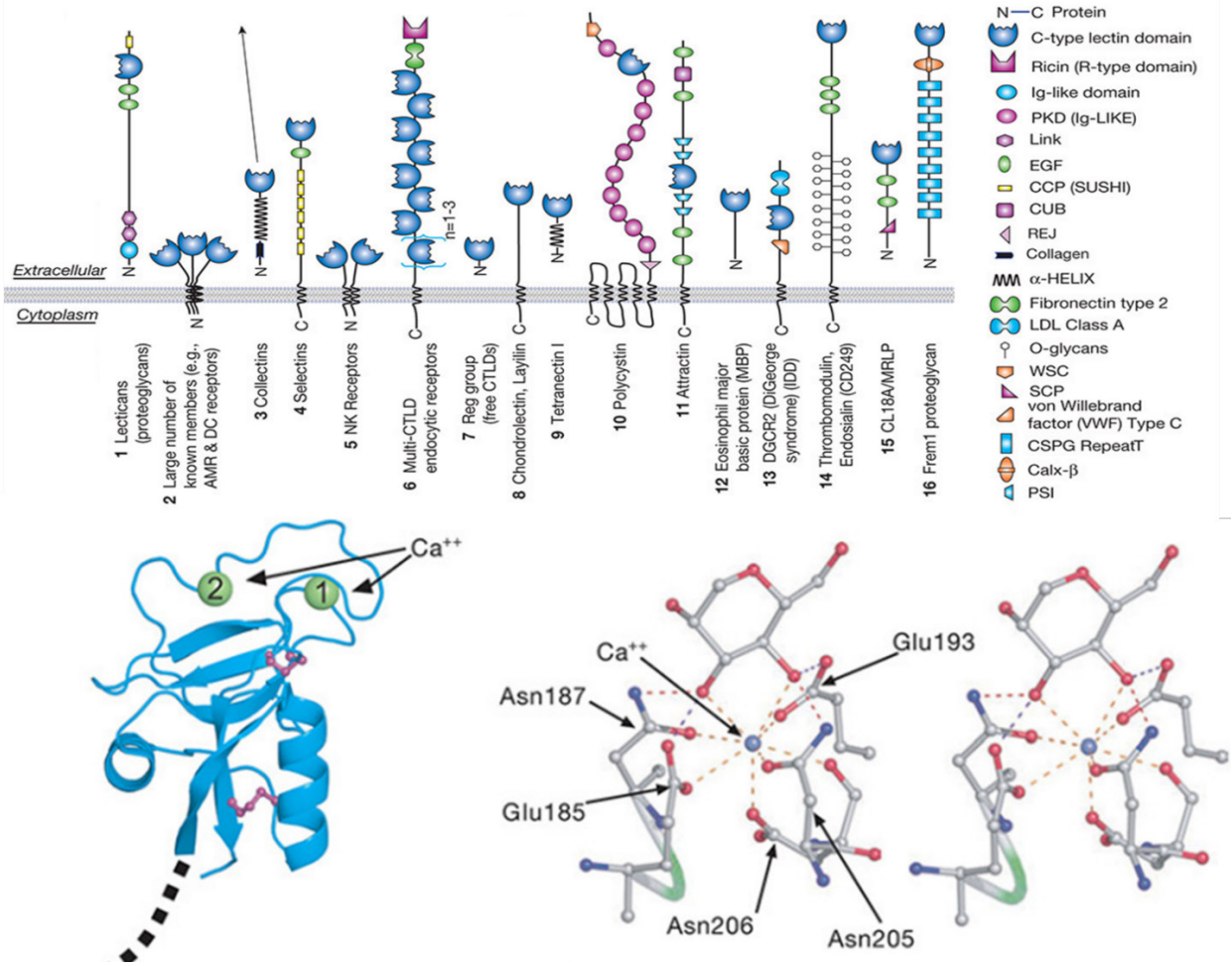


Figure 21: C type lectins variety of architectures and presentation mode of the CRD and represented interactions in the binding pocket, showing the importance of the calcium atom (Cummings and McEver 2017).

Due to the low affinity of the lectin domain for glycans, lectins assemble into oligomers or use the tandem repetition of CRD to increase the binding affinity to glycans, as shown in Figure 22. Oligomeric lectins either have their binding sites distributed in all directions; or a circular shape with all binding sites on one side which is adapted to the membrane surface (Brinda et al. 2004). For example, β-propeller lectins are proposed to have a common ancestor as a blade that evolved by duplication (Yadid and Tawfik 2007).

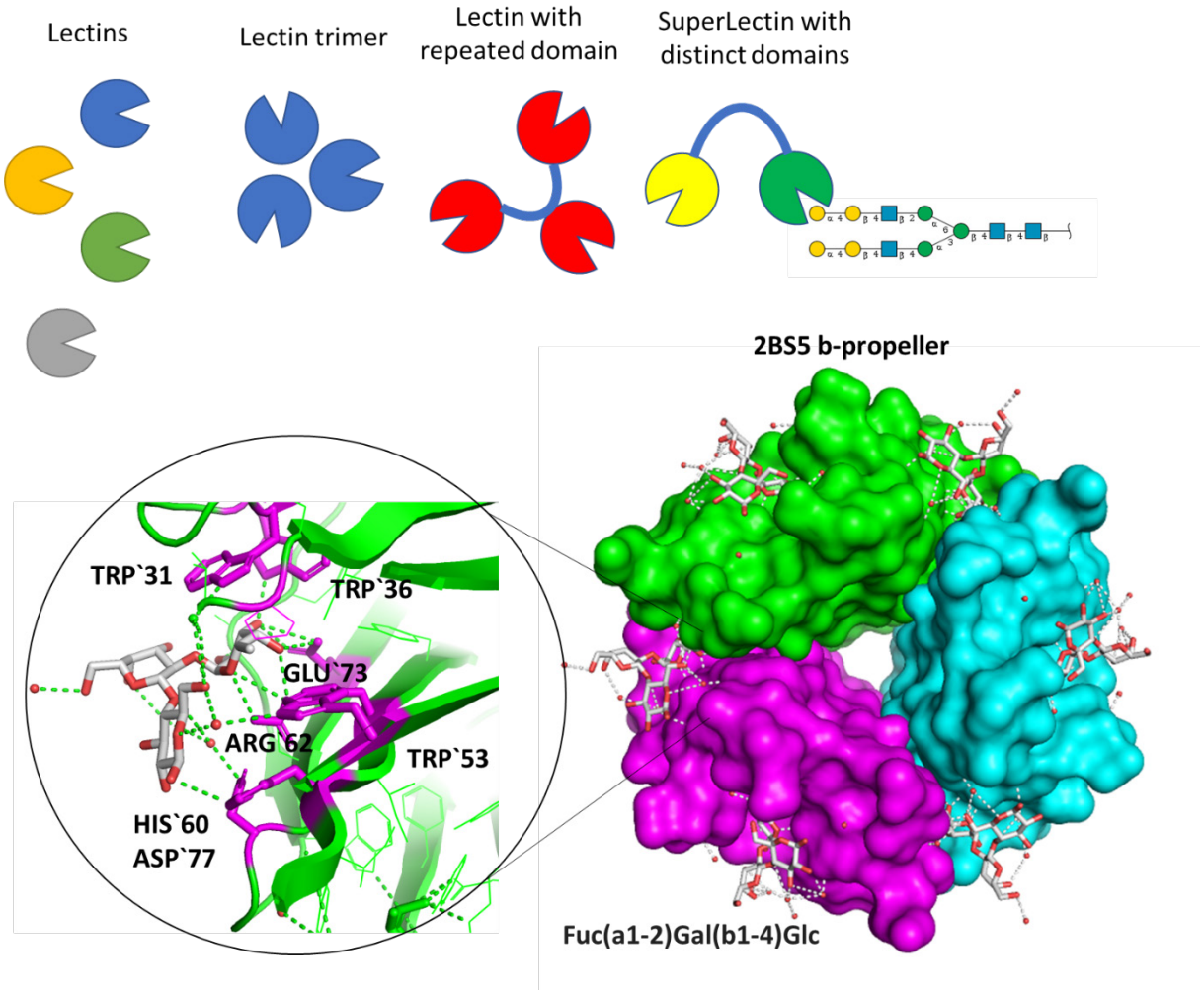


Figure 22: Lectin multivalency using oligomerization, repetition of domains and combination of distinct lectin domains. Example: 3D structure of the two blade repeat from Ralstonia solanacearum that assembles in trimer to form a six bladed β-propeller with zoom on the interaction in the glycan binding site).

3.3.3 Lectin functions across kingdoms

The lectins have common functions that are found in all kingdoms, represented in Figure 23. They are involved in cellular interactions and play a role in self/non-self recognition. Lectins are often part of defence systems: they are involved in the innate immunity of many organisms, especially invertebrates, and are toxic to predators of other organisms (fungi, plants). Pathogens, such as bacteria or viruses, use lectins for recognition and adhesion to glycans on the cell surface of a target organism.

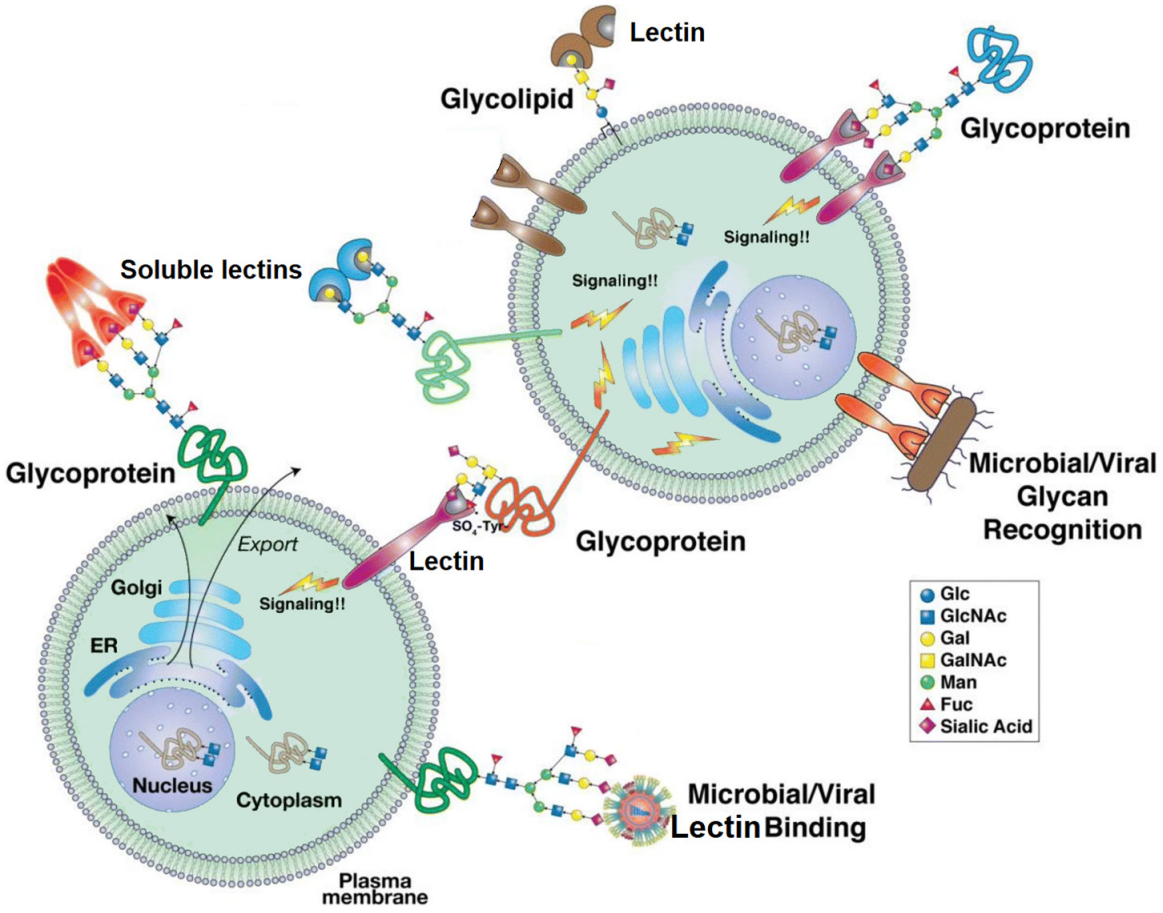


Figure 23: Schematic representation of possible function of lectins (Cummings 2019).

Animal lectins cover at least 12 structural classes, including the most studied galectin and the Ca2+ dependent C-type lectin described above. They are involved in cell adhesion, growth regulation, differentiation and survival, with the main function of innate immune system recognition molecules (Kilpatrick 2002). Animal lectins have complex architectures with repetition of the lectin domain and combination with other functional domains. The calnexins, calreticulins and malectins are chaperones involved in protein folding in the endoplasmic reticulum (Gabius 1997; Kaltner and Gabius 2001).

Fungal lectins serve as storage proteins as in plants and are involved in the growth, development and morphogenesis of fungi. They mediate the host recognition necessary for ectomycorrhizal symbiosis and association with algae or cyanobacteria in the lichen, or for yeast flocculation. Fungal lectins are also involved in the defence of fungi (against parasites and eaters) since some have toxic activities such as insecticides, vermicides or antivirals, and their interaction with host glycoconjugates may be involved in the infection process of pathogenic fungi (Varrot et al. 2013). Their roles and potential applications are illustrated in Figure 24.

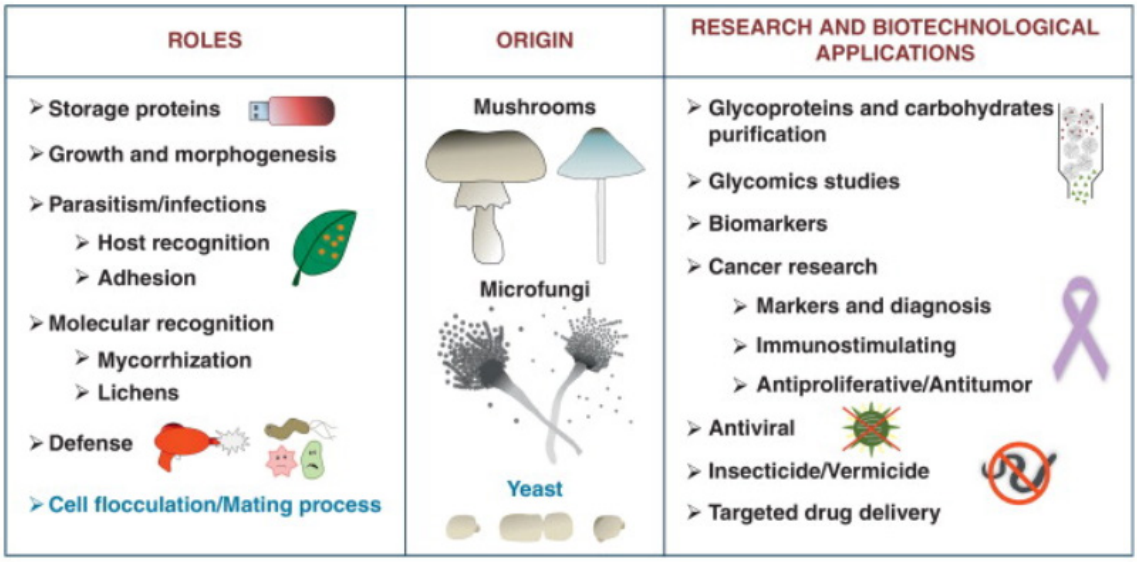


Figure 24: roles and potential applications of fungal lectins (Varrot et al. 2013).

Algae lectins bind specific carbohydrates to the aggregation of erythrocytes, yeasts, bacteria and various unicellular algae. In freshwater and marine algae, lectins are essential for the recognition and adhesion of gametes during sexual reproduction. They are also involved in symbiosis and defence (Liao et al. 2003). A large number of algae possess agglutinins, lectins that cause blood agglutination (Boyd et al. 1966). Compared to plant lectins, algal lectins have a strong oligosaccharide binding specificity, which allows them to be used as probes against carbohydrates on the cell surface and in drug targeting. Algae/marine lectins are described in more detail in the review (Singh et al. 2015).

Plant lectins have insecticidal, antifungal, antibacterial, and antiviral functions, and sometimes participate in biotic stress responses. (Peumans and Van Damme 1995; Van Damme 2014). They are commonly used for crop improvement, biomedical research and glycobiology tools (Tsaneva and Van Damme 2020).

Bacterial lectins are often used for bacterial adherence to host glycoconjugate at the surface of target cells, to initiate the infection (Sharon 1987). Heterotrophic bacteria depend on saprophytism,

symbiosis or pathogenicity for their source of energy. These lifestyles require the specific recognition for adhesion and subsequent invasion (Imberty et al. 2005).

Virus lectins such as flu hemagglutinin are involved in the fusion of the viral envelope with the plasma membrane and the uptake of the virus into cells. Influenza virus hemagglutinin binds to sialic acid (Neu5Ac) containing glycans, Human noroviruses capsid lectin bind to histo-blood group antigens (HBGAs) (Koromyslova et al. 2015). Rotavirus also can bind to sialic acid residues of newborn infants.

Protist cell surface lectins bind glycoconjugates on marine planktonic protists control the chemical interactions between predatory planktonic protists and their prey. Phagocytosis is initiated by lectin receptors on the surface (Roberts et al. 2011), as represented in Figure 25.

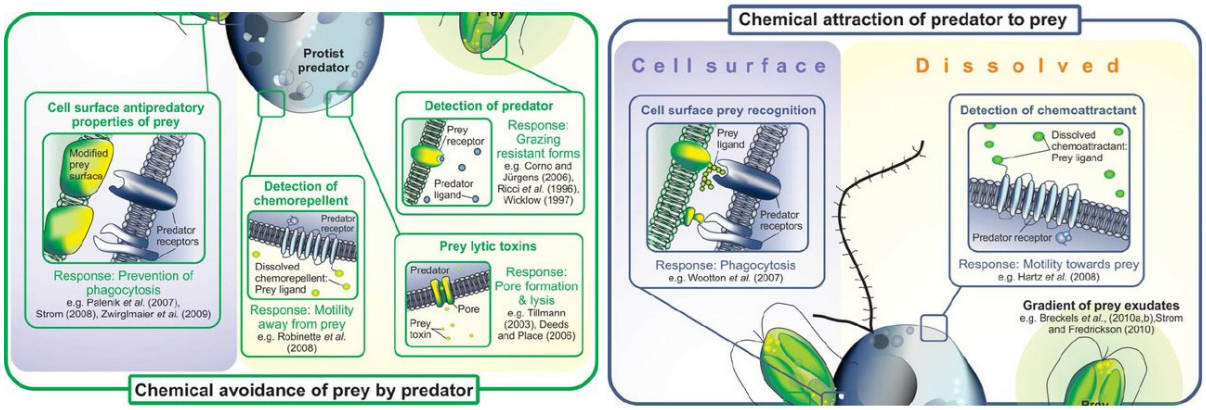


Figure 25: Mechanisms underlying chemical interactions between predatory planktonic protists and their prey (Roberts et al. 2011).

3.3.4 Applications to biotechnology, therapeutics and agriculture

Lectins can be used for biotechnological and therapeutic applications: separation of glycoconjugates, tissue labelling, chemical techniques for the visualisation of biological structures (histocytochemistry), targeting the compound on cancer cells (future development), as an antiviral (griffithsin, banana lectin...), as shown in Figure 26.

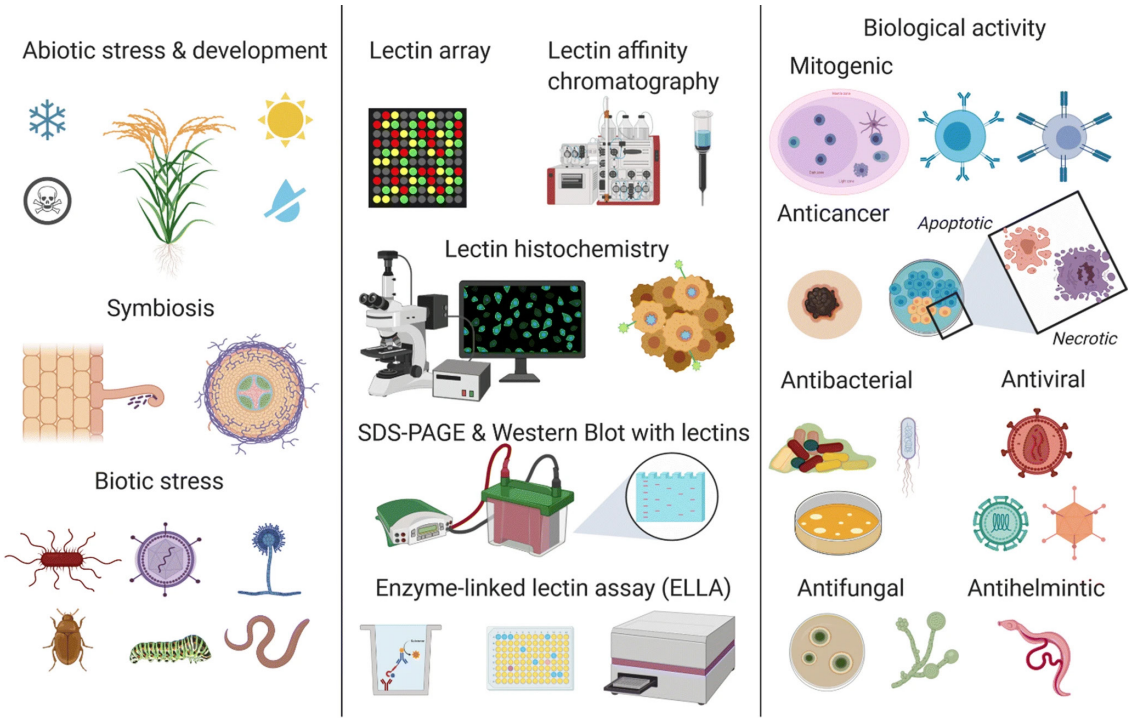


Figure 26: Schematic representation of lectin applications in agriculture, glycobiology and medicine (Tsaneva and Van Damme 2020).

Glycomimetic drugs can be designed to occupy the lectin binding sites for release therapy, or to trigger the associated signalling pathway in the lectin-associated cell. For example, bacteria use pili adhesins, such as FimH lectin, to bind to the cell surface for colonization but trigger inflammation on the host. By understanding their affinity for glycan, specific glycomimetics can be created that have a better affinity for targeting the adhesin (Sattin and Bernardi 2016).

3.3.5 Classification of lectins

Different classification methods can be used for lectins and have been proposed by (Lis and Sharon 1998). Lectins can be classified according to their glycan specificity, sequence, fold and binding pockets. With the increasing number of structures, classification according to structure seems to be the most relevant. A general classification system has been proposed for animal lectins by (Gabius 1997) defining five distinct families of animal lectins, based on the CRD fold/structure and the corresponding composition in peptide domains. In 1989, Drickamer proposed another classification for animal lectins based on amino acid sequence similarities and other properties (Drickamer 1989). Plant lectins can be classified into 12 families, based on the lectin CRD (Peumans et al. 2001; Fujimoto et al. 2014). Lectins are also classified on the basis of their glycan specificity (Wu et al. 2009).

Recent advances in molecular biology allow an easier classification of lectins by amino acid sequences, as described by E. Van Damme (Van Damme 2014). Some lectins distributed in several kingdoms share the same protein patterns. Structural biology allows to define a finer classification of lectins, necessary to create a new inter-species classification. In 2014, Hirabayashi updated the classification of lectins into 48 families (Fujimoto et al. 2014) based on their three-dimensional structures and using entries in Pfam (El-Gebali et al. 2019) and InterPro (Mitchell et al. 2019). The latest classification of lectins is presented in Essentials of glycobiology 3 (Figure 27).

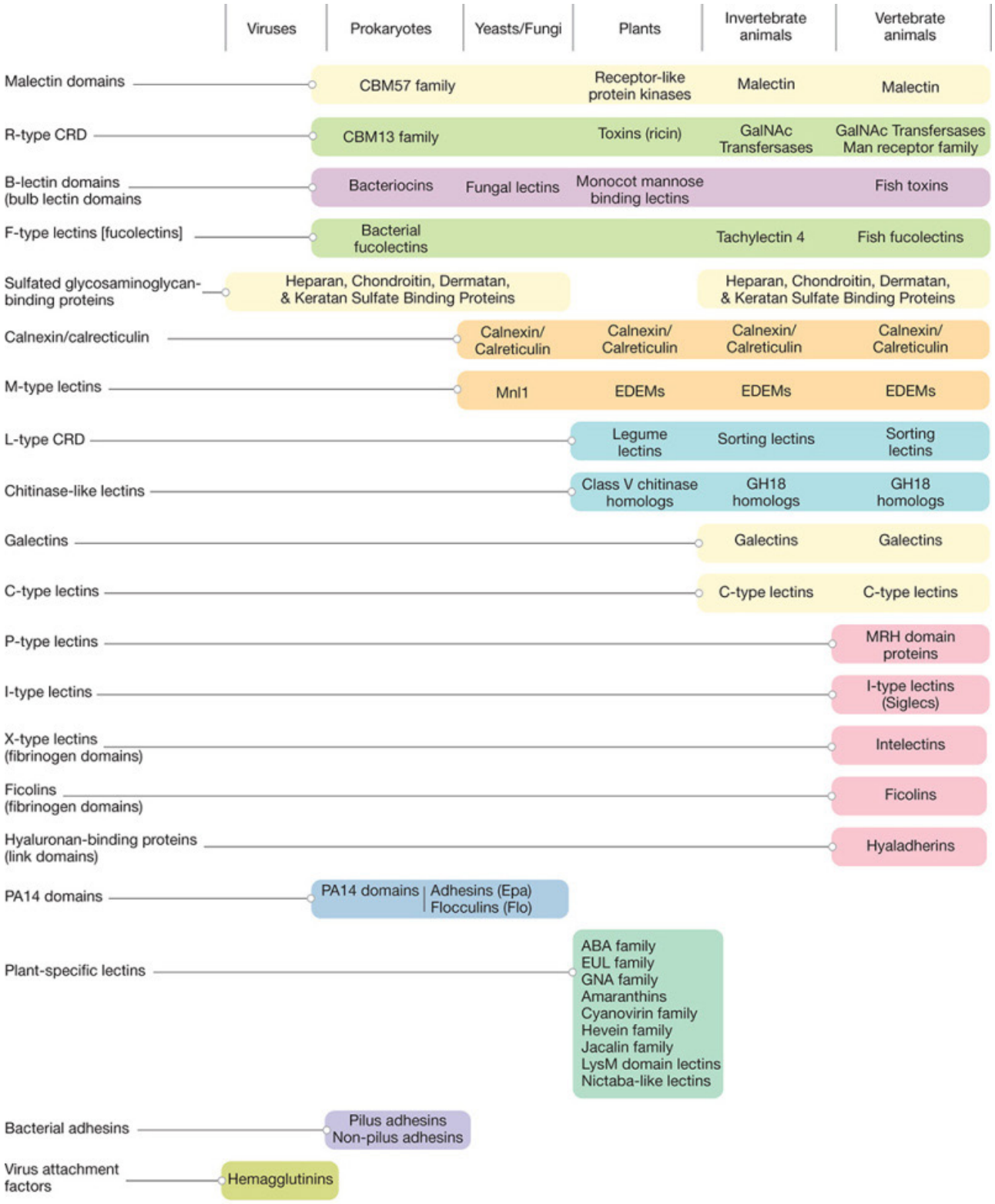


Figure 27: Several prominent structural families of glycan-binding proteins (GBPs) and their biological distributions, classified based on the structures of the CRDs (Taylor et al. 2017).

3.4 State of lectin and glycan-binding bioinformatics

Functional glycomics studies the functions of glycans and their recognition by glycan-binding proteins, whether free or present in cells, bacteria and viruses. Glycan-binding proteins include glycoenzymes, antibodies and lectins. Databases and tools have been developed for these different families of proteins, sometimes with overlaps. Information on lectins can be found in lectin databases, but also in databases on glycan networks and glyco-epitopes, which are the extremity of complex glycan interacting with lectins.

3.4.1 Issues with lectin annotation in online databases

Lectins are well-known proteins, characterised in reference organisms and for which we have associated 3D structures with interacting glycans. Despite this, lectins lack curative annotations in reference genomes and are therefore not easily identifiable in newly available genomes. Indeed, the new genomes are annotated using sequence similarity to reference sequences with annotations of variable quality. The lack of lectin curated annotation is partly due to the lack of correct classification of lectins in the available protein family databases. Protein family databases such as Pfam provide a classification of proteins based on their sequence similarity, while CATH provides a classification of proteins based on their structure. In both cases, there is a lack of manual processing of protein classification and families. UniProt, which contains both polymerised proteins and automatically annotated proteins, uses the protein family for protein annotation, thus providing possible uncharacterised lectin information to all new proteins obtained on the basis of new genomes. The CAZy database provides knowledge on the glycan-active enzymes involved in the synthesis and modification of glycan hydrolysis. There is no comparable database for lectins where genomics, proteomics and X-ray structures are unified.

3.4.2 Lectin databases

Several databases describing lectin data have been proposed in recent years. They aim to provide different levels of information, such as specificity, sequences or 3D structures. Some lectin information can also be found in more general databases on ProCarbDB (Copoiu et al. 2020) and the carbohydrate binding proteins of the Carbohydrate Binding Domain (Cazy). Information on CAZy CBM can cover lectins that have evolved from loss of enzyme function such as chi-lectins.

Several databases have been discontinued: AnimalLectinDb, BacterialLectinDb, CancerLectinDB. (Damodaran et al. 2008; Kumar and Mittal 2011, 2012). Although several lectin databases have been developed and put online, few are still accessible, including the Glyco3D lectin database, the Lectin Frontier database, the GlyCosmos and Procaff lectin list as presented in Table 10.

Name	Description	Link	Ref
LectinDB	provides for plant lectin the taxonomy, domain architecture, molecular sequence, and structural details as well as carbohydrate and blood group specificities	http://proline.physics.iisc.ernet.in/lectindb/search.html	(Chandra et al. 2006)
SugarBind	interconnection between structural information on lectin and other bioinformatics resources related to lectins, ligands and function	https://sugarbind.expasy.org/	(Mariethoz et al. 2016)
Glyco3D Lectin database	1500 curated 3D structures for 350 lectins with a curated kingdom based classification, associated glycans and publication	http://glyco3d.cermav.cnrs.fr/search.php?type=lectin	(Pérez et al. 2015a)
Lectin Frontier Database	provides 398 lectins including 180 lectins with associated glycan information, and interactions with glycans of glycoproteins	https://acgg.asia/lfdb2/	(Hirabayashi 2004)
GlyCosmos lectin	Protein entries annotated as lectins in UniProt.	https://glycosmos.org/lectins/index	(Yamada et al. 2020)
Procaff	binding affinity of protein-carbohydrate complexes	https://web.iitm.ac.in/bioinfo2/procaff/	(Siva Shanmugam et al. 2020)
Cazy CBM	Carbohydrate Binding Module within glyco-enzymes	http://www.cazy.org/Carbohydrate-Binding-Modules.html	(Terrapon et al. 2017)

Table 10: Lectin and CBM databases

The development of high throughput technologies, such as lectins and glycans, has generated large amounts of data on glycan-protein interactions, available in glycan databases. Information on the specificity of lectins can be retrieved from databases that bring together the results of experimental approaches. Lectin Frontier DataBase (Hirabayashi et al. 2015) provides a quantitative interaction obtained by frontal affinity chromatography. The Consortium for Functional Genomics (CFG) provides a service for glycan networks with available specificity data (Venkataraman et al. 2015). The different databases providing information on lectin affinity and glycan arrays are listed in Table 11.

Name	Description	Link	Ref
CFG glycan array	glycan profiling database	http://www.functionalglycomics.org/	(Venkataraman et al. 2015)
MCAW-DB	glycan profiling database containing 1081 glycan microarray samples collected from the CFG	https://mcawdb.glycoinfo.org/	(Hosoda et al. 2018)
GlyMDB	comprehensive glycan microarray database, 5203 glycan microarray from the CFG	http://www.glycanstructure.org/glymdb/	(Cao et al. 2020)
LM-GlycomeAtlas	web tool visualizing the data from Lectin Microarray	https://glycosmos.org/lm_glycomeatlas/index	(Nagai-Okatani et al. 2019)

Table 11: lectin affinity and lectin-glycan array databases

3.4.3 Related glycan-binding epitope databases and tools

Lectins recognize and interact with simple glycans or the extremity of complex glycans, called glyco-epitopes, that may be found on the surface of cells. Such epitopes are available in databases and can also be generated by breaking the complex glycan into small fragments of possible epitopes. The different databases providing epitopes recognised by lectins, and the tools that extract potential epitopes are listed in Table 12. Sugarbind database also provides the epitope recognized by lectins.

Name	Description	Link	Ref
Glyco-CD	Provides a collection of lectins and carbohydrates, information on 63 clusters of differentiation (CD) antigens	www.glycosciences.de/glyco-cd	(Kumar et al. 2012)
GlycoEpitope	Contains useful information on carbohydrate antigens, i.e. glyco-epitopes, and antibodies has been assembled as a compact encyclopedia	www.glycoepitope.jp	(Okuda et al. 2015)
PACDB	Pathogen Adherence to Carbohydrate Database	jcggdb.jp/search/PACDB.cgi	Not published
Glydin	compiles maps information relative to glycoepitopes (glycan determinants) as published in the literature or reported in databases	https://glycoproteome.expasy.org/epitopes/	(Alocci et al. 2018)
GlyS3	matches any substructure such as glycan determinants to a large collection of structures recorded in GlyConnect and SugarBindDB.	https://glycoproteome.expasy.org/substructuresearch/	(Alocci et al. 2018)
Epitope Extractor	Decompose a complex glycan into epitopes	https://glycoproteome.expasy.org/epextractor/	(Alocci et al. 2018)

Table 12: glyco epitopes databases and tools

3.4.4 Prediction of lectins

The identification of new lectins can be carried out either to search for the occurrence of a specific domain among species or, on the contrary, to search for all possible lectins in an organism or group.

The first approach uses classic bioinformatics tools such as BLAST and HMMER. For example, it has been shown that F-type lectins, first identified in the eel, are not only present in fish, but also participate in the adhesion of pathogenic bacteria to fucosylated glycoconjugates (Vasta et al. 2012), (Vasta et al. 2017). Similarly, galectins have been screened using Blast in humans and Chordata species to reconstruct their phylogeny (Houzelstein et al. 2004). Another analysis focused on the different domains of plant lectins identified in the genomes of five representative central angiosperms (*Arabidopsis thaliana, Glycine max, Cucumis sativus, Oryza sativa ssp. japonica and Oryza sativa ssp. indica*) (Van Holle et al. 2017). Thanks to advances in genome analysis tools, when a new lectin is discovered the sequence is compared with databases to identify similar proteins.

The other approach, which consists of predicting the entire lectome, i.e. all the lectins present in a species, is more complex. The quality of the annotation of the genome is not sufficient, and it is therefore necessary to rely on a library of lectin sequences to launch a search. Some lectin families are available in protein family databases such as Pfam and CATH, but they do not cover all carbohydrate recognition domains. However, a few attempts have been made in recent years. A comprehensive bioinformatics study of Archean lectins, using a research database constructed from the lectin sequences listed in Glyco3D-Lectin3D with PSI-BLAST against 165 Archean genomes, showed the presence of lectins with the same fold indicating their ancient origin long before the divergence of the three branches (Abhinav et al. 2016). A study of the whole lectin genome across nematode species, using the Pfam profiles of lectins available for HMMER analysis against nematode genome databases, also led to an interesting study of nematode lectins (Bauters et al. 2017). The cucumber genome was analysed for lectin domains (Dang and Van Damme 2016). Sixty-four sequences containing lectin domains with homologues of known 3D structures were identified thanks to a research on the genomes of mycobacteria (Abhinav et al. 2013).

To help the identification of lectin domains in genomes, the tool PLecDom: was developed but covered only the plant lectin domains (Shridhar et al. 2009). Another tool focuses on the prediction of protein−carbohydrate binding sites which include lectin domains (Taherzadeh et al. 2016).

4 Objectives

Lectins are used to target specific sugars and decipher the glyco code on the cell surface, and can also be used as glyco-epitope/antigen recognition tools. Lectins are present in all living kingdoms, including the human body, where we are seeking to improve our understanding of cellular interactions. Because of their role in cellular interactions and innate immunity, lectins are of great interest for the development of anti-infectious compounds. They are also used as biomarkers to label glycoconjugates and as vectors to transport drugs to specific sites. The future of research using lectins would allow practical applications in the fields of food, agriculture, health and pharmaceutical research, including the design of lectin to target cancer cells, or their use as antimicrobial and antifungal drugs.

Issue #1: no universal lectin classification

Until recently, the classification of lectins did not involve a hierarchy with well-defined rules compatible across the different kingdoms of life. Such a classification would facilitate exploration and research on lectins that are often enough, studied separately if their origin is different. Observed similarities between lectins across kingdoms is however incompatible with a kingdom-based classification. This question was addressed with the new classification proposed in 2014, based on the 3D structures of lectins and peptidic sequences using non curated Pfam domains. A classification based on glycan specificity is also limited because a lectin recognises the epitope of a wide variety of complex glycans.

Issue #2: no universal lectin database

Sparse information is currently available online in databases that cover lectin knowledge, such as SugarBind (pathogen lectins only) and GlyCosmos-Lectins (federates and lists a selection of ~2000 lectins), but do not provide comparative criteria. Furthermore, the lack of curated and automatic annotations of lectins in databases, including UniProt, RefSeq, Pfam, CATH, and SCOPe further illustrates the need to collect information in a single resource and cross-reference it with major bioinformatics sites. Finally, PDB databases, including RCSB, PDBe and PDBj, provide structural information on protein-ligand interactions and despite an improved glycan description, these details are rarely considered in other bioinformatics resources.

Issue #3: no lectin prediction method from sequence data

Genome annotation usually entails running several prediction tools that detect genes and their products, based on profiles. However, up to now, no profile has been defined to achieve the comprehensive identification of lectins in a genome, or the screening of all lectins in a sequence database

To solve the three cited issues, multiple objectives have been met in this thesis.

- The definition of a hierarchical classification with defined rules to organize the lectins and facilitate their exploration. Such a classification requires the manual curation of available protein and 3D structures of lectins.
- The development of a database based on this classification, to provide 3D structures of lectins, curated lectin annotation, and complete information on interacting ligands to scientists worldwide.
- The development of a lectome prediction method applicable to genome screening. The outcome may lead to candidate proteins for drug design and more generally to more complete annotation of lectins in screened genomes.
- The development of a database to store and support the exploration of candidate lectins. Dedicated and efficient tools are required to explore the large number of genomes available in the living kingdoms. Their validation can be achieved by screening large datasets selected by experts who can evaluate and interpret the results.

The issues, objectives are represented in a graphical abstract (Figure 28) together with an overview of the results.

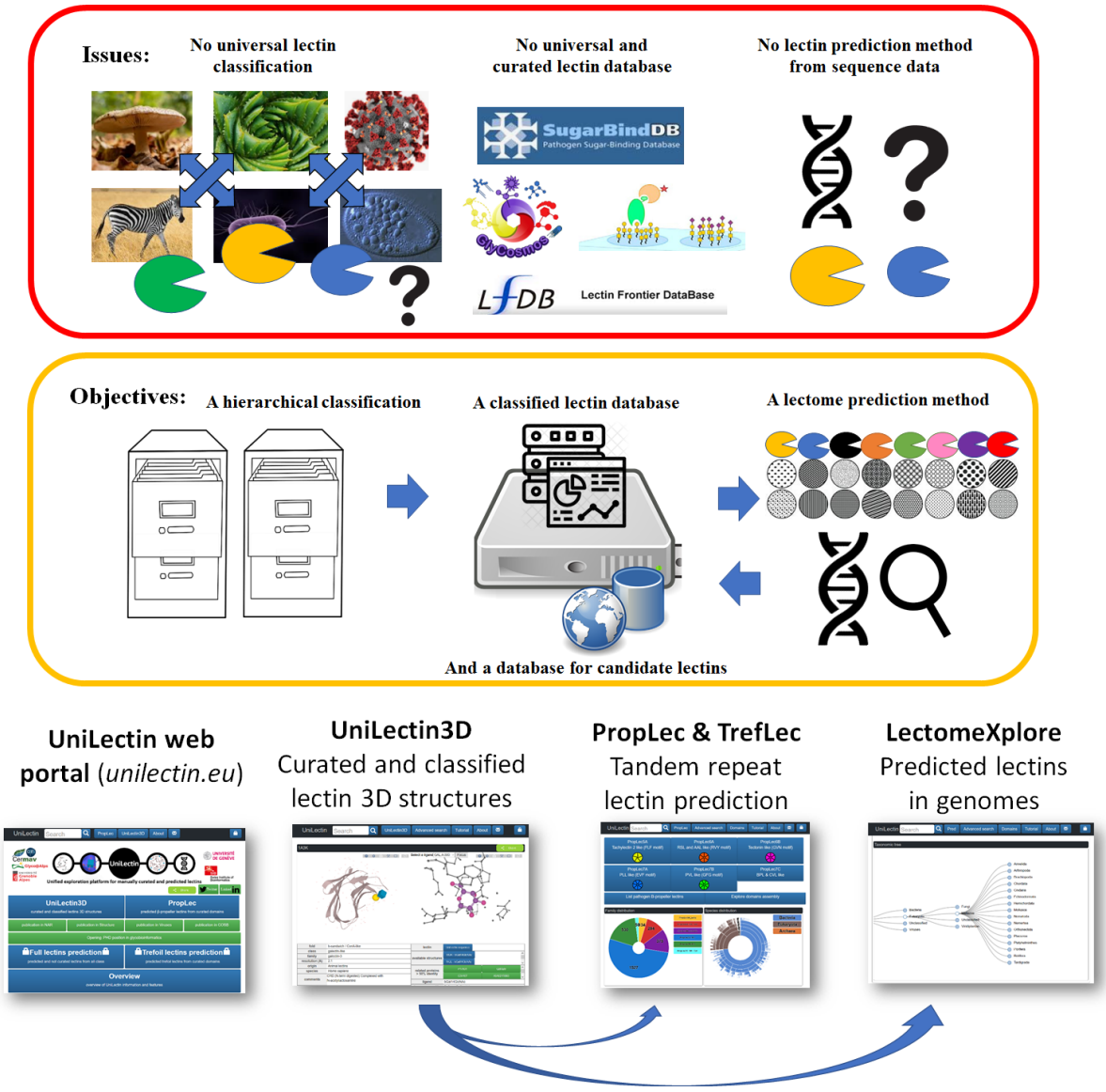


Figure 28: Graphical abstract of the thesis

5 Results I: UniLectin3D database and classification

A large number of 3D structures of lectins are described each year and provide new information on glycan recognition and binding methods. In order to provide the scientific community with annotations and classification of 3D lectin structures, the Glyco3D lectin database was published in 2015 but is no longer maintained. The UniLectin portal and the UniLectin3D database have been developed to meet the need for an actualised curated data set and an interactive web database, using the 1500 3D structures in the Glyco3D lectin database. To facilitate the maintenance, a simple and efficient lectin administration tool was required, which led to the development of this new web database.

5.1 Unilectin 3D construction using old Glyco3D lectin data

UniLectin3D has been built from the glyco3D-lectin3D database. The new database has been extended in terms of the number of entries (2200 in December 2020 against 1500 in Lectin3D) but also in terms of the information contained. New features have been included, with each lectin entry associated with information on its origin, fold, class, origin and a UniProt AC that has similar proteins. Graphical information is provided on the glycan linkage site thanks to a collaboration with the developers of the PLIP tool and SwissModel. The database also provides external links to the RCSB, PDBe, PDB-CARE, GlyTouCan, GlyTouCan, GlyConnect, SugarBind, PubMed and CFG glycan networks, selected from the available resources presented in the introduction.

The entries in the Glyco3D database have been checked and their origin, species and fold have been corrected if necessary. Associated monosaccharide glycans are now described using the IUPAC nomenclature which can be converted to GlycoCT using the GlyConnect tools. The presentation of lectin folds has been modified to be consistent with the definition used in SCOPe. The fold names are assigned manually when no corresponding definition has been found in SCOPe.

For retrocompatibility with Glyco3D portal, UniLectin3D data is stored in Glyco3D's SQL database. PHP scripts generate the pages of the HTML database according to the data contained in the SQL database. By default, all the lectins are displayed and a search interface allows the user to restrict the displayed results. The homepage provides statistics on the lectins illustrated by sunbursts generated using the D3JS JavaScript interactive graphics libraries. This allows the user to navigate and interact with the sunburst graphics and classification tree. The database uses bootstrap3 for the CSS rendering of the interface.

5.2 Improving the lectin classification

Historically, the classification of lectins is based on species and lectin domain (CRD) structure. The latest classification of Glyco3D is organised into three levels, with a first level for the domains of life: bacteria, fungi, animals, algae, plants and viruses. The second level separates the different types of lectins known historically and according to their structure. The third level separates close (neighbouring) species for which the same lectin has been found. This classification is historical but arbitrary, with no concrete criteria to be respected.

Current classifications of lectins lack rules: they all have drawbacks and do not reflect the variety of structures and functions observed. It was stressed that a well-defined classification is necessary. The use of a classification based on the sequence of structures would help to build more reproducible patterns of lectin families and to better define the quality of the predicted lectins. A first proposition was made in 2014 by Hirabayashi based on the Pfam domains (Fujimoto et al. 2014).

Proteins have explored a wide variety of structures by natural selection, which has been demonstrated to be more conserved than the protein sequence. In the case of lectins, not all known folds are used, but lectins have already been identified for about 30 distinct folds. The lectins can be classified from these 30 folds, represented in Figure 29. The first obvious separation is between α-helix, β-sheet and the α/β mixed domain. For each of these types, different patterns can be formed by the combination of several helices, sheets and hairpins. The most distinctive are the β-helix, the β-barrel and the β-prism. In many lectins, the assembly of four β-leaves can form a Greek key, while the assembly of eight such leaves forms a jelly roll. A flat superposition of two β-leaves in a β-sandwich is also often observed.

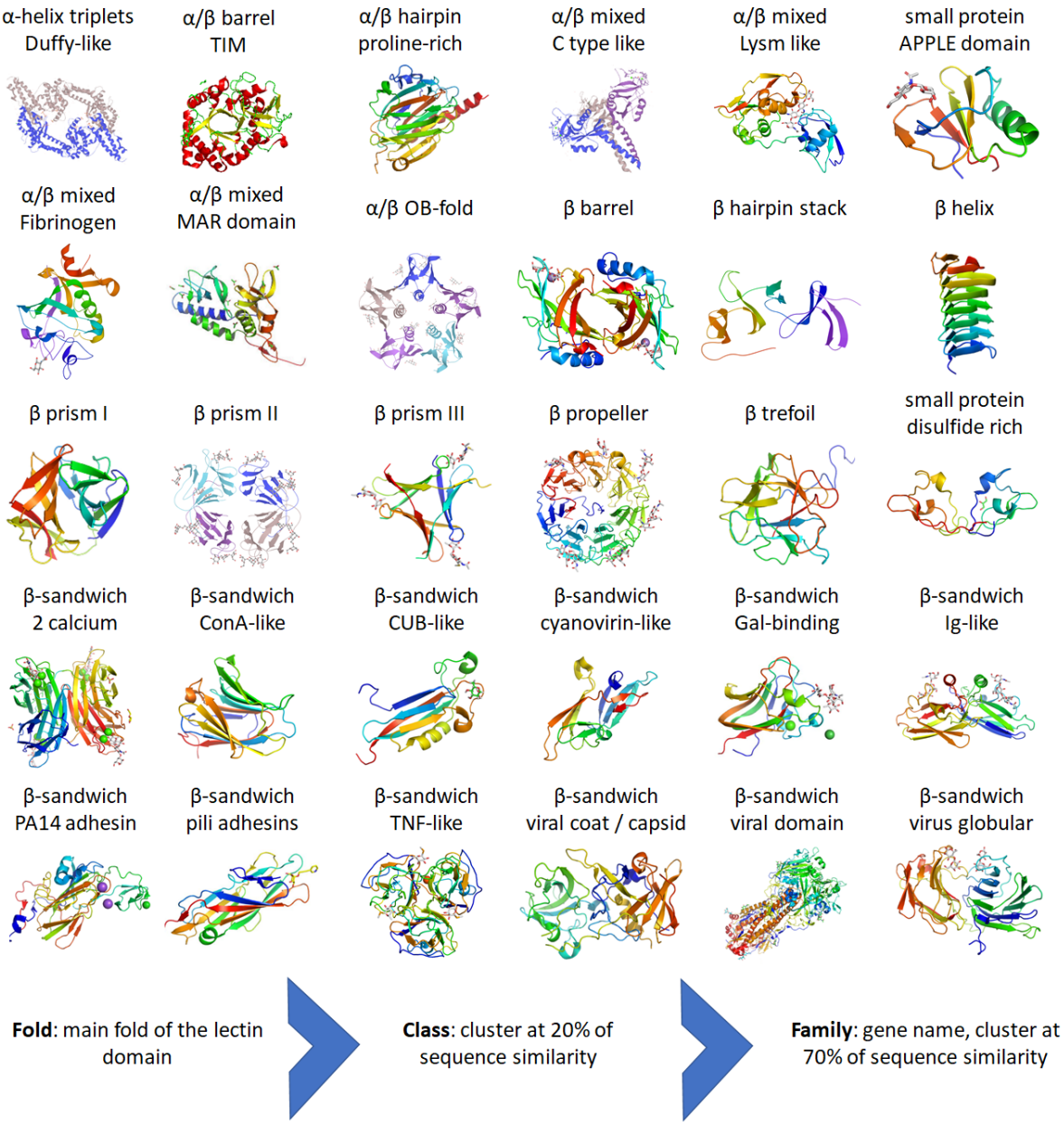


Figure 29: Representation of lectin folds along with the new classification

Two years after the first version of Unilectin3D and the publication of the article describing it (Article I), it has become clear that the "historical" classification, based on a mixture of origin and specificity, will not be sufficient for future development. The updating of the classification was decided as follows (described in Article VI page 2). The classification of lectines has been revised to be based on both structures and 3D sequences. It has 3 levels: (i) the folds of the lectin domain, (ii) the classes based on 20% of sequence similarity (iii) the families based on 70% of sequence similarity. Thanks to the new classification, it is possible to observe the fold distribution of lectins by kingdoms, as

shown in Figure 30, with the β-sandwich conA-like fold being the most characterized in Plant and Animal lectins and β-sandwich pili adhesins for the Bacterial lectins.

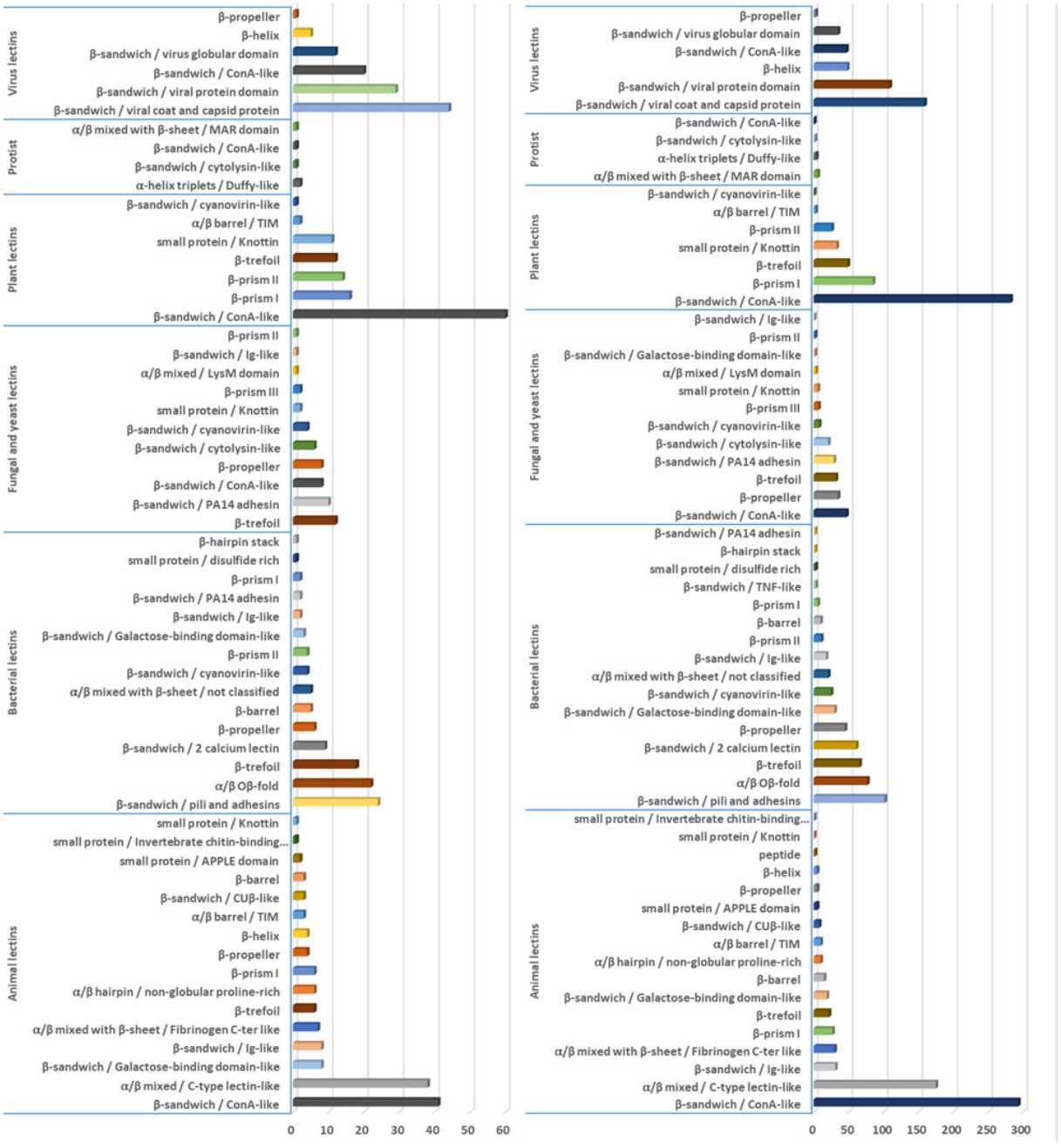


Figure 30: Lectin distribution by kingdoms and folds (Left); 3D lectin structures distribution by kingdoms and folds (Right).

5.3 Administration of the database and tutorial

Administration tools are needed to maintain the databases. This represents a significant workload that is not clearly reflected in the publications (Article I). A set of administrative tools has been created to easily add, modify or delete database entries and associated information. The public and

administrative sections of the database have been used by several users who have provided appropriate feedback to each of them; with the incorporation of the feedbacks a fully functional and user friendly database has been released.

The UniLectin 3D database is updated by direct inspection of the weekly output of the PDB (300 to 500 structures with 2 to 20 readings each week) for the entry of new 3D lectin structures. The presence of a carbohydrate residue with no covalent link to the protein is an effective selection criterion to identify lectins in the PDB, after filtering out all carbohydrate-active enzymes. For lectin without interacting glycan, the associated publication can provide information on the lectin glycan-binding activity.

Manual curation is necessary to assign each structure in the classification and to detect the glycan in the binding site, i.e. to save it in IUPAC format. A data entry interface has been created, which automatically loads most of the required information from the PDB code, including the species, the publication if available and associated UniProt AC. A new lectin structure can therefore be added by manually entering the PDB ID, classification, ligand and any specific comment. For modifications, a dedicated page displays all available entries in a table and allows the user to directly modify the values. The details of each lectin can also be consulted to modify the values and to delete or add images.

The use of Unilectin3D is not limited to a few lectin specialists and should be accessible to a wider audience of glycobiologists and glyco-chemists. A tutorial page has been integrated into the site and published as a methodological article (Article II).

5.5 Article I: The UniLectin3D database

UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands

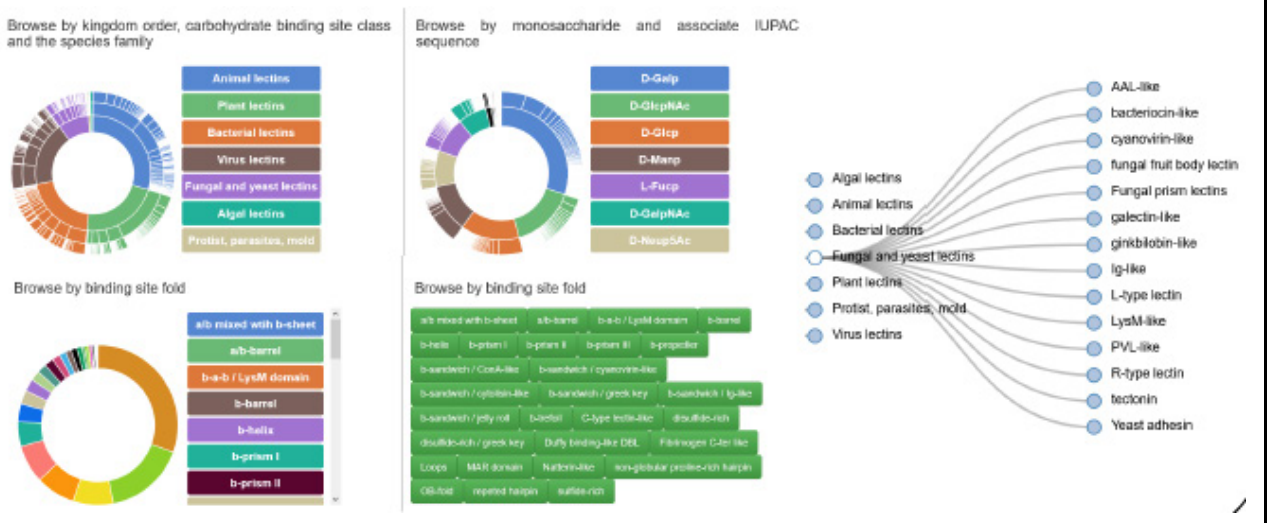
Bonnardel, F., Mariethoz, J., Salentin, S., Robin, X., Schroeder, M., Perez, S., Lisacek, F & Imberty, A. (2019). UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. **Nucleic Ac. Res.**, 47(D1), D1236-D1244. https://doi.org/10.1093/nar/gky832

Abstract

Lectins, and related receptors such as adhesins and toxins, are glycan-binding proteins from all origins that decipher the glycocode, i.e. the structural information encoded in the conformation of complex carbohydrates present on the surface of all cells. Lectins are still poorly classified and annotated, but since their functions are based on ligand recognition, their 3D-structures provide a solid foundation for characterization. UniLectin3D is a curated database that classifies lectins on origin and fold, with cross-links to literature, other databases in glycosciences and functional data such as known specificity. The database provides detailed information on lectins, their bound glycan ligands, and features their interactions using the Protein–Ligand Interaction Profiler (PLIP) server. Special care was devoted to the description of the bound glycan ligands with the use of simple graphical representation and numerical format for cross-linking to other databases in glycoscience. We conceived the design of the database architecture and the navigation tools to account for all organisms, as well as to search for oligosaccharide epitopes complexed within specified binding sites. UniLectin3D is accessible at https://www.unilectin.eu/unilectin3D.

Personal contributions

Development of UniLectin portal, UniLectin3D databases and visualisation features, UniLectin3D management tools including automatic loading of lectin features. Draft of the manuscript and figures.



UniLectin3D, une base de données de protéines de liaison aux sucres avec des informations sur les structures 3D et les ligands en interaction

Bonnardel, F., Mariethoz, J., Salentin, S., Robin, X., Schroeder, M., Perez, S., Lisacek, F & Imberty, A. (2019). UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. **Nucleic Ac. Res.**, 47(D1), D1236-D1244. https://doi.org/10.1093/nar/gky832

Abstract

Les lectines, et les récepteurs apparentés tels que les adhésines et les toxines, sont des protéines d'origine variées qui déchiffrent le glycocode, c'est-à-dire les informations structurelles codées dans la conformation des glucides complexes présents à la surface de toutes les cellules. La classification et l'annotation des lectines laisse à désirer, mais comme leurs fonctions sont basées sur la reconnaissance de ligands, leurs structures 3D fournissent une base solide pour la caractérisation. UniLectin3D est une base de données régulièrement maintenue qui classe les lectines selon leur origine et leur repliement, avec des liens croisés avec la littérature, d'autres bases de données en glycosciences et des données fonctionnelles telles que la spécificité connue. La base de données fournit des informations détaillées sur les lectines, leurs ligands glycanniques, et décrit leurs interactions à l'aide du serveur PLIP (Protein-Ligand Interaction Profiler). Un soin particulier a été apporté à la description des ligands glycannique, avec l'utilisation d'une représentation graphique simple et d'un format numérique permettant de les relier à d'autres bases de données en glycosciences. Nous avons conçu l'architecture de la base de données et les outils de navigation pour prendre en compte tous les organismes, ainsi que pour rechercher les épitopes d'oligosaccharides dans des sites de liaison spécifiques. UniLectin3D est accessible à l'adresse https://www.unilectin.eu/unilectin3D.

Contributions

Développement du portail UniLectin, des bases de données et des fonctionnalités de visualisation UniLectin3D, des outils de gestion UniLectin3D incluant le chargement automatique des fonctionnalités de lectines. Ecriture de la première version et réalisation des figures.

5.6 Article II: Tutorial chapter for Unilectin3D database

Structural Database for Lectins and the UniLectin Web Platform

Bonnardel, F., Perez, S., Lisacek, F., & Imberty, A. (2020). Structural database for lectins and the UniLectin web platform. **Methods Mol. Biol**. *2132*, 1-14.doi:10.1007/978-1-0716-0430-4_1

Abstract

The search for new biomolecules requires a clear understanding of biosynthesis and degradation pathways. This view applies to most metabolites as well as other molecule types such as glycans whose repertoire is still poorly characterized. Lectins are proteins that recognize specifically and interact noncovalently with glycans. This particular class of proteins is considered as playing a major role in biology. Glycan-binding is based on multivalence, which gives lectins a unique capacity to interact with surface glycans and significantly contribute to cell–cell recognition and interactions. Lectins have been studied for many years using multiple technologies and part of the resulting information is available online in databases. Unfortunately, the connectivity of these databases with the most popular omics databases (genomics, proteomics, and glycomics) remains limited. Moreover, lectin diversity is extended and requires setting out a flexible classification that remains compatible with new sequences and 3D structures that are continuously released. We have designed UniLectin as a new insight into the knowledge of lectins, their classification, and their biological role. This platform encompasses UniLectin3D, a curated database of lectin 3D structures that follows a periodically updated classification, a set of comparative and visualizing tools and gradually released modules dedicated to specific lectins predicted in sequence databases. The second module is PropLec, focused on β-propeller lectin prediction in all species based on five distinct family profiles. This chapter describes how UniLectin can be used to explore the diversity of lectins, their 3D structures, and associated functional information as well as to perform reliable predictions of β-propeller lectins.

Personal contribution

Development of UniLectin3d, formatting of the tutorial as a web page in UniLectin3D.
Draft of the manuscript and figures.

Base de données des structure 3D de lectines et la plate-forme web UniLectin

Bonnardel, F., Perez, S., Lisacek, F., & Imberty, A. (2020). Structural database for lectins and the UniLectin web platform. **Methods Mol. Biol**. *2132*, 1-14.doi:10.1007/978-1-0716-0430-4_1

Abstract

La recherche de nouvelles biomolécules nécessite une bonne compréhension de la biosynthèse et des voies de dégradation. Cette approche s'applique à la plupart des métabolites ainsi qu'à d'autres types de molécules telles que les glycanes dont le répertoire est encore mal caractérisé. Les lectines sont des protéines qui reconnaissent spécifiquement les glycanes et interagissent de manière non covalente avec eux. Cette classe particulière de protéines est considérée comme jouant un rôle majeur en biologie. La liaison aux glycanes est basée sur la multivalence, ce qui confère aux lectines une capacité unique d'interaction avec les glycanes de surface et contribue de manière significative à la reconnaissance et aux interactions entre les cellules. Les lectines ont été étudiées pendant de nombreuses années en utilisant de multiples technologies et une partie des informations qui en résultent est disponible en ligne dans des bases de données. Malheureusement, la connectivité de ces bases de données avec les bases de données en omique les plus populaires (génomique, protéomique et glycomique) reste limitée. De plus, la diversité des lectines est étendue et nécessite la mise en place d'une classification flexible qui reste compatible avec les nouvelles séquences et structures 3D qui sont continuellement publiées. Nous avons conçu UniLectin pour apporter une nouvelle vision de la connaissance des lectines, de leur classification et de leur rôle biologique. Cette plate-forme comprend UniLectin3D, une base de données de structures 3D de lectines qui suit une classification régulièrement mise à jour, un ensemble d'outils de comparaison et de visualisation et des modules progressivement mis à disposition dédiés à des lectines spécifiques prévues dans des bases de données de séquences. Le deuxième module, PropLec, est axé sur la prédiction des lectines de toutes les espèces à l'aide de cinq profils de famille distincts appartenant aux β-propeller. Ce chapitre décrit comment UniLectin peut être utilisé pour explorer la diversité des lectines, leurs structures 3D et les informations fonctionnelles associées, ainsi que pour effectuer des prédictions fiables des lectines β-propeller.

Contributions

Developpement de UniLectin3d, mise en forme du tutoriel sous la forme d’une page web dans UniLectin3D. Ecriture de la première version et réalisation des figures.

6 Results II: Prediction of tandem repeats lectins in genomes

6.1 Tandem repeat lectins: β-propeller, β-trefoil, β-prism and others

The multivalence of lectins is required to achieve a high avidity for glycoconjugates, and can be obtained by oligomerisation, by multiple presentations on the cell surface, or by peptide repetitions in the protein sequence. The latter case is frequent and introduces complexity for the definition of the conserved peptidic motifs. Tandem repeating lectins are presented as double repeats: cyanovirin, β-prism; triple repeats: β-trefoil; and multiple repeats: β-propellers, β-prism I. β-trefoil and β-propeller as shown in Figure 31, present a wide range of specificities and multivalencies (see Article III). These protein folds are not limited to lectins and have been successfully adopted for a wide range of functions. They are thought to evolve from the duplication of a single domain and have the advantage of high stability (Yadid and Tawfik 2011). Such folds are of interest in synthetic biology as scaffolds with variable internal symmetry (Hu et al. 2015; Hirabayashi and Arai 2019).

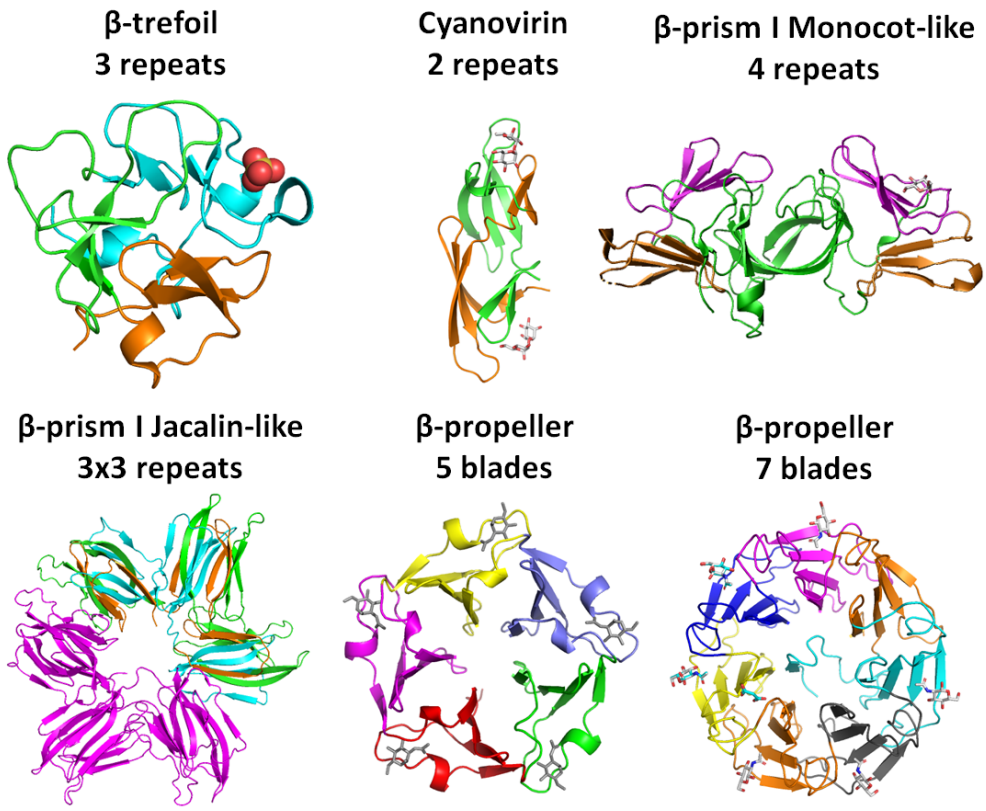


Figure 31: 3D structures of different types of tandem repeats lectins, repeats are represented by different colors

6.1.1 β-Propeller lectins

β-propellers are composed of a repeated domain that can be described as "blades". They are found in all areas of life, with a strong preponderance of eukaryotes. β-propellers have a wide variety of sequences and are classified by blade numbers in the SCOP and CATH databases. β-propeller proteins are common in nature with 4 to 10 symmetrical blades, a diversity explained by the evolutionary process of gene duplication and fusion (Mylemans et al. 2020). The shape of the β-helix is well adapted to lectins as it allows 5 to 7 binding sites (based on the lectin currently known) oriented to bind to glycosylated surfaces. The different classes and families of β-propeller lectins are represented in Figure 32. The β-propellers have one binding site per blade, sometimes located at the interface between two blades. In some cases, one of the binding pockets is not active because the symmetry is altered by the closure of the β-propeller.

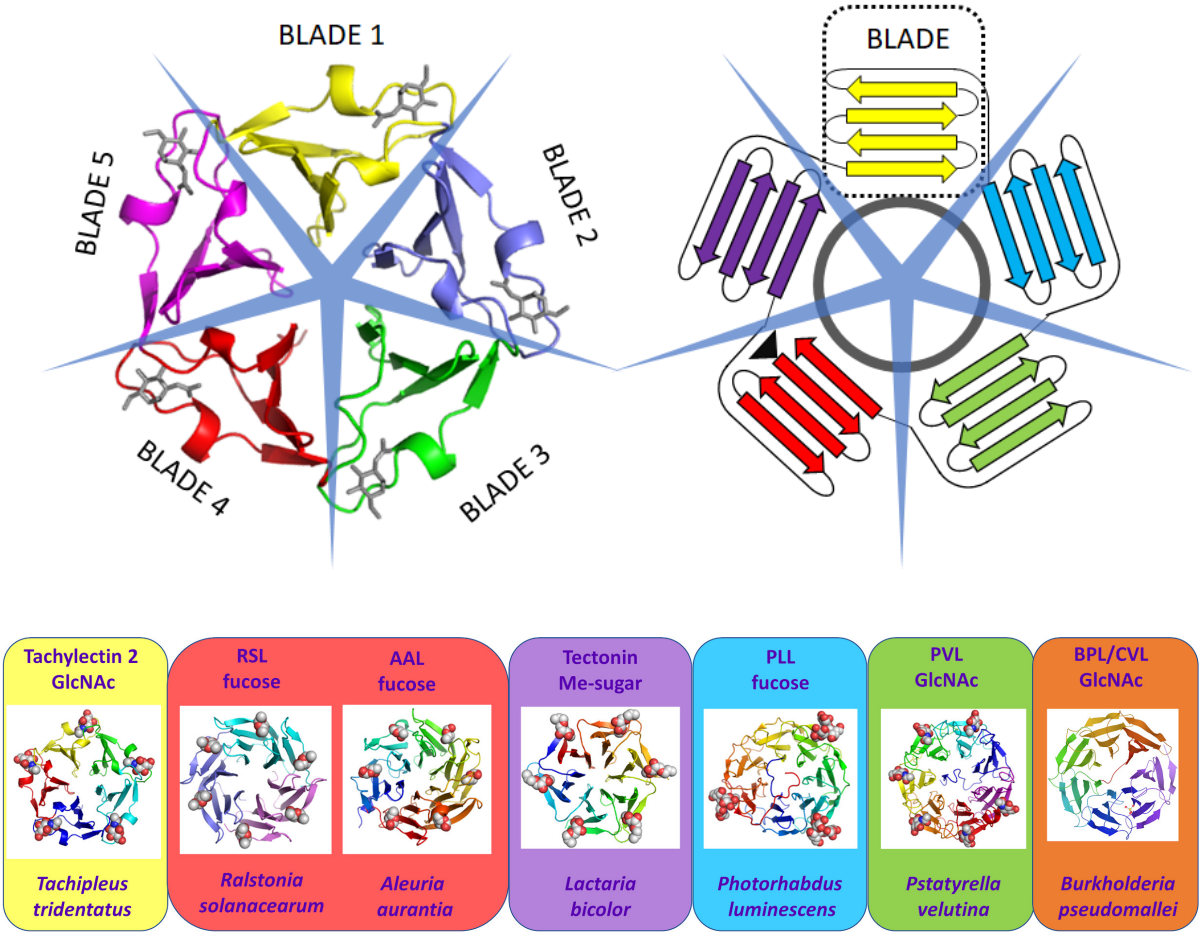


Figure 32: β-propeller lectin 3D structures and the different β-propeller lectin classes

An exception to the β-propeller lectin is the viral haemagglutinin of the measles virus with a sialic acid binding site in the middle. This lectin has evolved from a neuraminidase which has lost its

enzymatic activity. This type of lectin, or lectenzyme, is difficult to identify in the genome since sequence similarity search makes it possible to find lectins, but also all the similar active enzymes. This particular class is therefore not included in the β-propeller database.

β-propeller repeats are difficult to detect in translated genomes and are rarely properly annotated in sequence databases. The PropLec database, and associated prediction method, was therefore developed to identify β-propellers in genomes while taking into account sequence repetition in tandem repeats (Article IV).

After the publication, a new 7-bladed β-propeller structure was described (Sýkorová et al. 2020). This new class is included in the online version of the PropLec module. In addition, the prediction has been updated with the latest version of the NCBI and UniProt protein databases.

6.1.2 β-Trefoil lectins

The β-trefoil proteins are formed in a trilobal organisation consisting of peptide repeats named leaves. Each leaf has at one side a barrel structure and on the other side a triangular arrangement of hairpins. Many β-trefoil repeats contain a QxW motif whose aromatic residue is involved in the hydrophobic core. The triple symmetry can be degenerated, and many β-trefoil lectins have only one or two active binding sites for carbohydrates, the highly stable and degenerated β-trefoil are shown in Figure 33. Based on 20% of sequence similarity between the lectin domains, and excluding associated domains such as toxin and aerolysin domains, the β-trefoil lectins can be separated into 11 different classes as described in Table 13, the largest being the Ricin-like.

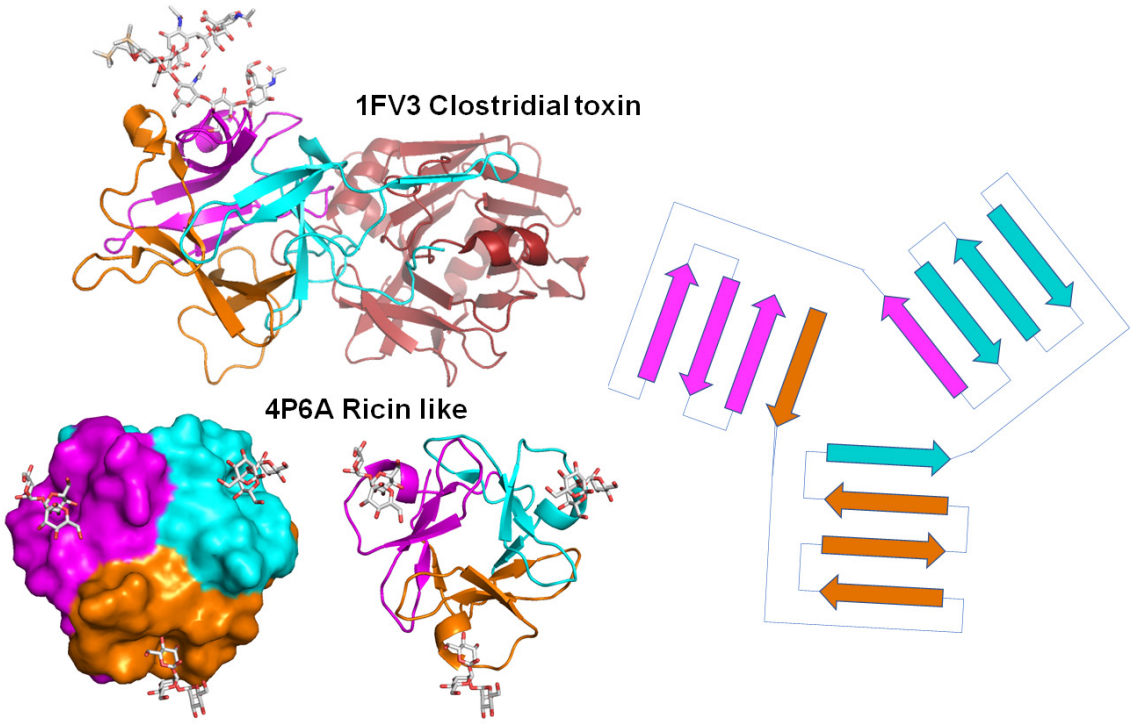


Figure 33: β-trefoil 3D structure and topological representation. The clostridial toxin lectins have very low sequence identity between leaves, since one is involved in connecting to the toxin domain. The Ricin like Actinohivin lectin from actinomycete has three high symmetrical leaves with conserved binding sites to the glycan.

Class	Family	Origin	UniProt AC	IUPAC
Amaranthin-like	Amaranthin	Plant lectins	Q71QF2	Gal(b1-3)GalNAc,
Boletus and Laetiporus β-trefoil lectin	LSLa, BEL	Fungal and yeast lectins	Q7Z8V1, R4GRU5, R4GRU4, R4GRU6, R4GRU9	Gal(b1-4)Glc, Gal(b1-4)GlcNAc, , Gal, GalNAc, Gal(b1-3)...
Clitocybe lectin-like	CNL	Fungal and yeast lectins	B2ZRS9	Gal(b1-4)Glc, GalNAc(b1-4)GlcNAc
Clostridial toxin	TeNT, BoNT/B, BoNT/A, BoNT/C, BoNT/F, BoNT/G, BoNT/D	Bacterial lectins	P04958, P10844, P0DPI0, Q9LBR1, P0DPI1, A7GBG3, Q60393, P...	GalNAc, NeuAc, Gal, Gal(b1-4)Glc, NeuAc(a2-3)Gal(b1-4...
Coprinus β-trefoil lectin	CCL2	Fungal and yeast lectins	B3GA02	GlcNAc(b1-4)[Fuc(a1-3)]GlcNAc, NeuAc(a2-3)GlcNAc(b...
Cys-rich man-receptor	Cys-rich domain man-receptor	Animal lectins	Q61830	GalNAc, Gal(b1-4)[Fuc(a1-3)]GlcNAc
Earthworm lectin	earthworm EW29	Animal lectins	O96048	Gal(b1-4)Glc, NeuAc(a2-6)Gal(b1-4)Glc, GalNAc
Fungi and Clostridium β-trefoil lectin	(HA1) HA-33/D and C, (HA1) HA-33/A, MOA, (HA3) HA70/C and others	Bacterial lectins, Fungal and yeast lectins	P0DPR0, Q45871, Q9LBR4, Q8X123, P46085, Q75WT9, F6KMV5, L...	Gal(a1-3)Gal(b1-4)GlcNAc, NeuAc, GalNAc, Gal, Gal(a1-...
Mytilectin	mussel lectin, synthetic mussel lectin	Animal lectins	H2FH31, B3EWR1, A0A0P0E482,	GalNAc, Gal, GalN, Gal(a1-4)Gal(b1-4)Glc
Ricin-like	Abrus agglutinin, abrin-a, VAA, TKL-1, SNA-II, ricin V and others	Plant lectins, Animal lectins, Bacterial lectins	P11140, Q6ITZ3, Q7SIF1, Q9AVR2, P81446, P06750, Q868M7, P...	Gal, Gal(b1-4)Glc, GalNAc, Fuc, Man(a1-2)Man, Man(a1-2...
Sclerotinia lectin-like	Sclerotinia	Fungal and yeast lectins	A7XUK7	Gal(b1-3)GalNAc

Table 13: β-Trefoil lectin classes associated families and glycans

6.2 Development of PropLec and TrefLec modules

6.2.1 Manual curation of tandem repeats

For lectin structures of the same class, tandem repeats must be selected manually to define the retained pattern, which is explained in more detail in Articles IV and V. Pymol is used for the visualization and manipulation of the 3D structures. A protein PDB structure can contain one or more protein chains or peptides. In the case of lectins, it is important to identify the appropriate chain containing the lectin domain. Pymol allows the selection of specific chains or residues to extract or hide them. The visualization of the interacting glycan allows finding the binding pockets of the lectin domain, and interactions with the closest residues can be highlighted. In the case of tandem repeated domains, they can be superposed on the basis of the 3D structure, as shown in Figure 34.

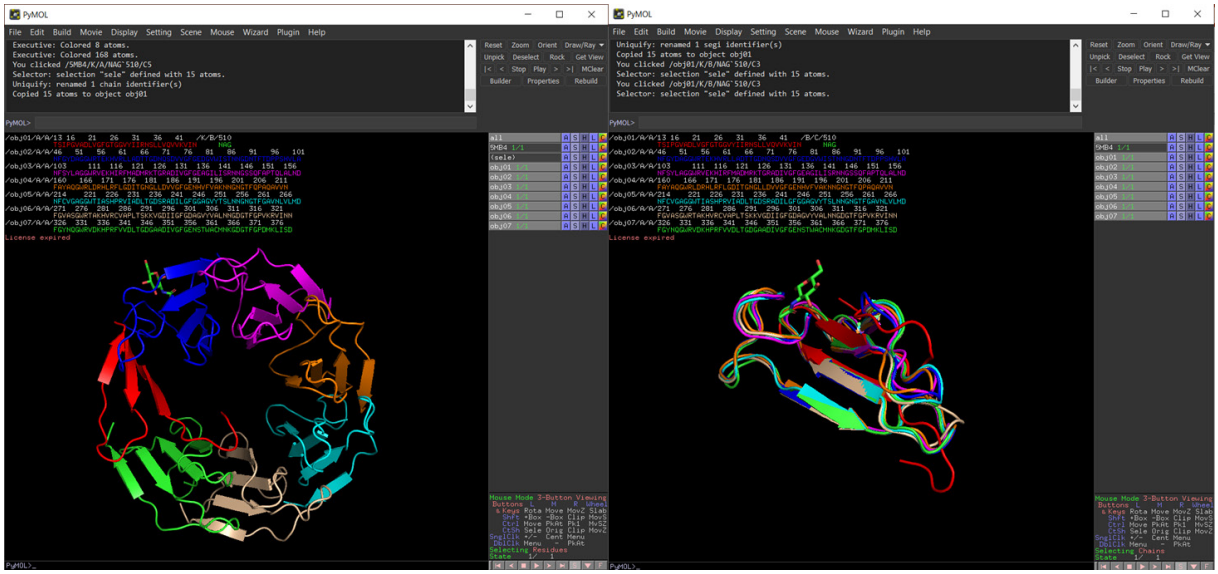


Figure 34: Propeller blades superposition in Pymol using CEalign method.

The manually defined tandem repeats in available β-propeller or β-trefoil 3D structures are aligned by MUSCLE to generate a conserved motif for each lectin class. The motif is given to HMMER to generate a HMM profile used by HMMER to screen selected protein datasets, NCBI-nr and UniProt databases, to identify candidate lectins.

6.2.2 Scoring system for tandem repeat lectins

To ensure the quality of the predictions, HMMER tool compute a p-value used as a first reporting threshold. The resulting predicted lectins are then filtered to eliminate the small domains identified (less than 10 residues). A large number of remaining predicted β-propeller or β-trefoil are ordered by quality in the web database.

The HMMER score is not suitable for the web database, as it is not normalised across domains and not easy to read. The HMMER score is the sum of the position scores, with for each position a score from 0 (mismatch) to 9 (match, conserved residue).

For a score simple to read in the 0-1 range and to interpret, a similarity score is used. It is calculated between two consensus sequences: from the reference pattern; and from the predicted propeller blades. The reference consensus sequence is generated by HMMER alignments, and for the predicted domain, the identified domains are aligned by MUSCLE to obtain a multiple sequence alignment. Due to the overall lower quality of the β-trefoil reference domains, the scores are lower than for the β-propeller.

6.2.3 Web database features

The sequences of the predicted β-propeller and β-trefoil lectins are stored in an SQL database, and separated respectively in the PropLec and TrefLec database. A homepage provides statistics on predicted lectins by lectin class and by taxonomy using sunburst graphic. The distribution of the number of repeats is also displayed. The taxonomy can be browsed through an interactive tree. An advanced search page allows the user to make more specific searches, and for each predicted lectin, a detailed page is provided with protein characteristics and graphs for quality control, used to compare the amino acid motifs of the reference and predicted domain, as well as the glycan binding residues. D3JS is used for the development of those interactive graphs.

6.2.4 Manual verification of the candidate lectin quality

Even if a lectin is predicted using curated domains, low quality sequences come from the source protein database and are filtered using different keywords, such as "partial" or "fragment". To guarantee the quality of the protein sequence which is in fact identified by tools based on genomic and transcriptomic data, it is advisable to analyse the corresponding gene and its source genomic assembly (either a chromosome / plasmid or at least the contig), using the genomic source of the protein or an explosion against reference genomes. Indeed, different types of errors may be present: a shift during sequencing that adds or deletes a domain/fragment to the protein and errors during the identification of the protein on the source assembly.

6.3 Article III : Structure and engineering of tandem repeat lectins

Structure and engineering of tandem repeat lectins

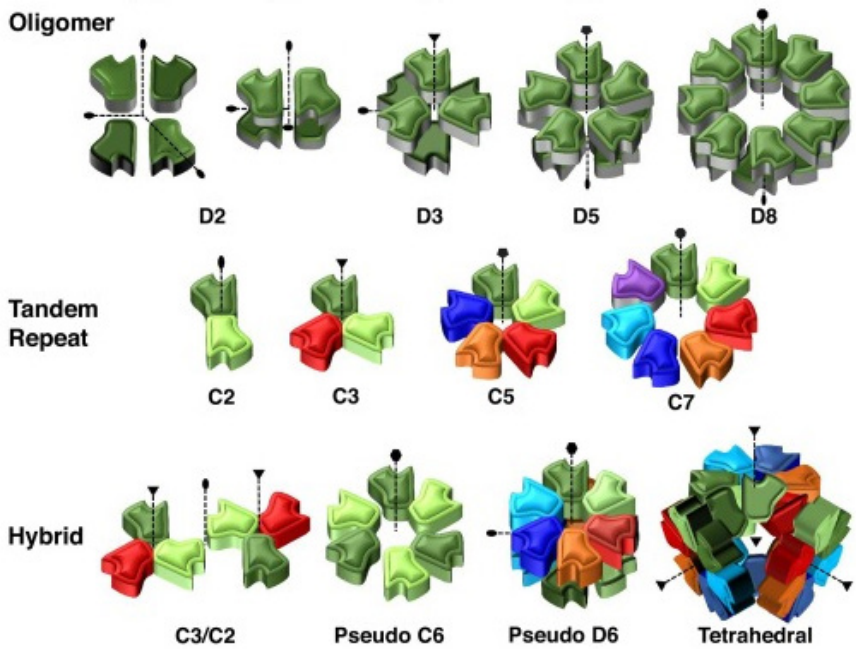
Notova, S., Bonnardel, F., Lisacek, F., Varrot, A., & Imberty, A. (2020). Structure and engineering of tandem repeat lectins. **Curr. Op. Struct. Biol.**, *62*, 39-47. https://doi.org/10.1016/j.sbi.2019.11.006

Abstract

Through their ability to bind complex glycoconjugates, lectins have unique specificity and potential for biomedical and biotechnological applications. In particular, lectins with short repeated peptides forming carbohydrate-binding domains are not only of high interest for understanding protein evolution but can also be used as scaffold for engineering novel receptors. Synthetic glycobiology now provides the tools for engineering the specificity of lectins as well as their structure, multivalency and topologies. This review focuses on the structure and diversity of two families of tandem-repeat lectins, that is, β-trefoils and β-propellers, demonstrated as the most promising scaffold for engineering novel lectins.

Personal contributions

First draft and figures of the β-propellers section, and 3D models



Structure et ingénierie des lectines à répétition de séquences

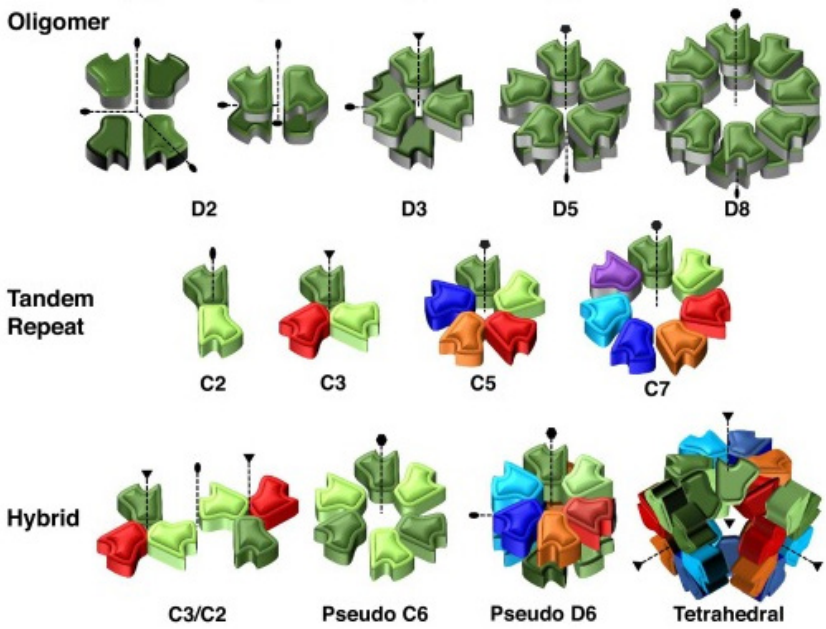
Notova, S., Bonnardel, F., Lisacek, F., Varrot, A., & Imberty, A. (2020). Structure and engineering of tandem repeat lectins. **Curr. Op. Struct. Biol.**, *62*, 39-47. https://doi.org/10.1016/j.sbi.2019.11.006

Abstract

Grâce à leur capacité à lier des glycoconjugués complexes, les lectines ont une spécificité et un potentiel uniques pour les applications biomédicales et biotechnologiques. En particulier, les lectines dont les peptides courts et répétés forment des domaines de liaison aux glucides sont d'un grand intérêt pour la compréhension de l'évolution des protéines, mais peuvent également être utilisées comme support pour l'ingénierie de nouveaux récepteurs. La glycobiologie de synthèse fournit désormais les outils nécessaires à l'ingénierie de la spécificité des lectines ainsi que de leur structure, de leur multivalence et de leur topologie. Cette revue se concentre sur la structure et la diversité de deux familles de lectines à répétition en tandem, à savoir β-trefoils et β-propellers, dont il a été démontré qu'elles constituaient l'échafaudage le plus prometteur pour l'ingénierie de nouvelles lectines.

Contributions

Rédaction de la première version et préparation figures de la section des β-propellers, et des modèles 3D.



6.4 Article IV : PropLec article: prediction of β-propeller lectins

Architecture and Evolution of Blade Assembly in β-propeller Lectins

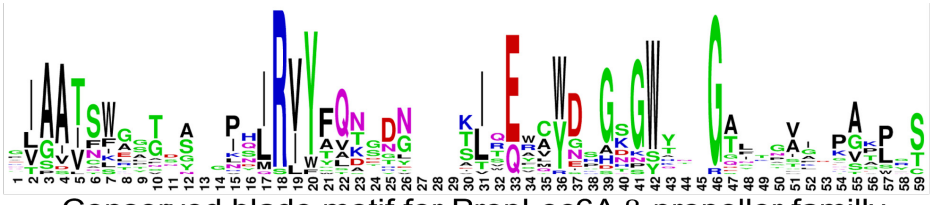
Bonnardel, F., Kumar, A., Wimmerova, M., Lahmann, M., Perez, S., Varrot, A., Lisacek, F. & Imberty, A. (2019). Architecture and evolution of blade assembly in β-propeller lectins. **Structure**, 27, 764-775. https://doi.org/10.1016/j.str.2019.02.002

Abstract

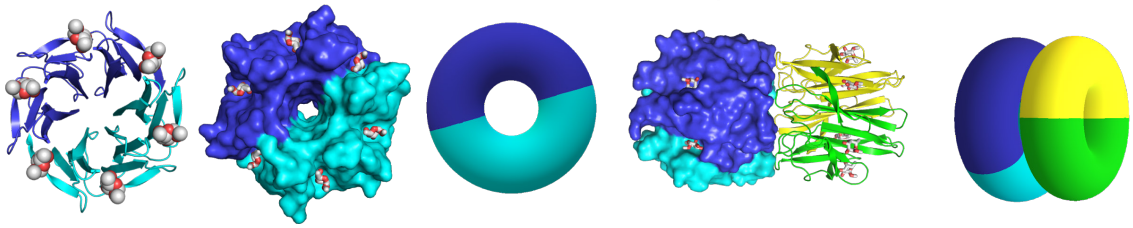
Lectins with a β-propeller fold bind glycans on the cell surface through multivalent binding sites and appropriate directionality. These proteins are formed by repeats of short domains, raising questions about evolutionary duplication. However, these repeats are difficult to detect in translated genomes 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 3,887 β-propeller lectins in 1,889 species and organized this information in a searchable online database. The data reveal a widespread distribution of β-propeller lectins across species. Prediction also emphasizes multiple architectures and led to the discovery of a β-propeller assembly scenario. This was confirmed by producing and characterizing a predicted protein coded in the genome of Kordia zhangzhouensis. The crystal structure uncovers an intermediate in the evolution of β-propeller assembly and demonstrates the power of our tools.

Personal contributions

Development of PropLec database and visualisation features. Building propeller motifs, selection of the protein databases, screening using HMMER tools. Development of the Python pipeline to process the results and integrate them into the site with filtering steps, formatting, scoring. Protein feature acquisition including the taxonomy, and loading in the SQL database. Search for new sequences of interest to be produced and characterized by the biochemists. Draft of the manuscript and figures



Conserved blade motif for PropLec6A β-propeller familly



Identification of a β-propeller lectin with a unique new assembly

Two propellers assemble together as a tetramer

Architecture et évolution de l'assemblage des pales dans les lectines β-propeller

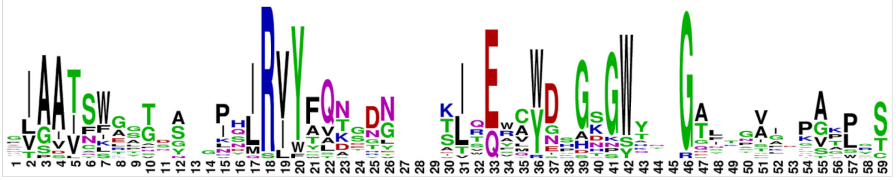
Bonnardel, F., Kumar, A., Wimmerova, M., Lahmann, M., Perez, S., Varrot, A., Lisacek, F. & Imberty, A. (2019). Architecture and evolution of blade assembly in β-propeller lectins. **Structure**, 27, 764-775. https://doi.org/10.1016/j.str.2019.02.002

Abstract

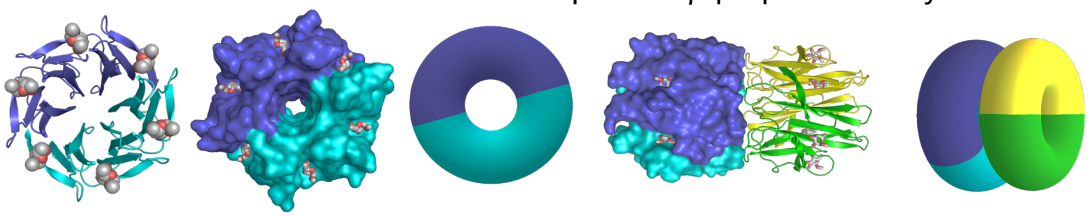
Les lectines dotées d'un repliement β-propeller se lient aux glycanes à la surface de la cellule par des liaisons multivalentes et d'une directionnalité appropriée. Ces protéines sont formées par des répétitions de domaines courts, ce qui soulève des questions sur la duplication évolutive. Cependant, ces répétitions sont difficiles à détecter dans les génomes et sont rarement annotées correctement dans les bases de données de séquences. Pour répondre à ces questions, nous avons défini la signature des répétitions des cinq types de β-propeller en utilisant des données structurelles 3D. Avec ces modèles, nous avons prédit 3 887 lectines de β-propeller dans 1 889 espèces et organisé ces informations dans une base de données consultable en ligne. Les données révèlent une large distribution des lectines β-propeller parmi les espèces. La prédiction met également l'accent sur les architectures multiples, permettant la découverte d'un scénario d'assemblage de β-propeller. Cela a été confirmé par la production et la caractérisation d'une protéine prédite codée dans le génome de *Kordia zhangzhouensis*. La structure cristalline représente un nouvel intermédiaire dans l'évolution de l'assemblage de β-propeller et démontre la puissance de nos outils.

Contributions

Développement de la base de données PropLec et de ses fonctionnalités de visualisation. Construction de motifs d'hélices, sélection des bases de données de protéines et leur criblage à l'aide des outils HMMER. Développement du pipeline Python pour traiter les résultats et les intégrer au site avec des étapes de filtrage, le formatage, la notation. Acquisition de caractéristiques des protéines, y compris la taxonomie, et le chargement dans la base de données SQL. Recherche de nouvelles séquences d'intérêt à produire et à caractériser par les biochimistes.



Conserved blade motif for PropLec6A β-propeller familly



Identification of a β-propeller lectin with a unique new assembly

Two propellers assemble together as a tetramer

6.5 Article V : TrefLec article: prediction of trefoil lectins

Identification and characterization of a β-trefoil lectin from lower eukaryote with an aerolysin domain

Simona Notova, François Bonnardel, Annabelle Varrot, Frédérique Lisacek and Anne Imberty

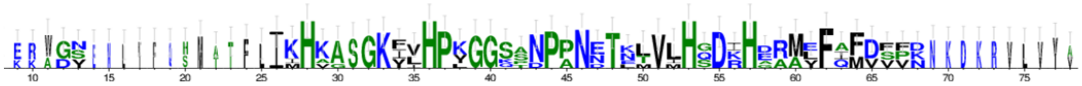
Manuscript in preparation

Abstract

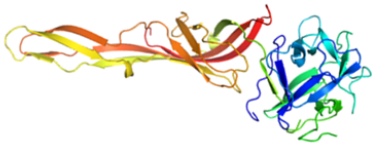
Glycan binding proteins called lectins play an important role in glycan recognition, cell-cell interactions and the immune system. The efficiency of glycan-binding lectins mainly relies on multivalency, arising either from oligomerization or tandem repeats of lectin domains. The β-trefoil fold, characterized by three leaf-shaped repeats is adopted by several classes of lectins of interest. Tandem repeats in amino acid sequences hampers the detection of coding regions in complete genomes. Based on the classification of lectin 3D structures proposed in the Unilectin3D database in particular for β-trefoil lectins, we defined the trefoil-leaf signature of 11 lectin classes. Processing the content of UniProtKB and NCBI-nr led us to identify 26994 β-trefoil lectins in 3607 species. The predicted sequences are organized in the TrefLec database available at unilectin.eu/trefoil, where they can be searched and visualized. We selected the example of SoraL, a lectin predicted in Salpingoeca rosetta, a member of the primitive colony-forming choanoflagellates that are the closest single-celled relatives of animals. SoraL belongs to the Mytilectin family, previously limited to molluscs and brachyopods. The lectin was expressed recombinantly and bind to N-acetylgalactosamine. The crystal structure was generated and it confirmed the strong structural similarity with known mollusc Mytilectins but also with fungal lectins. Finally, SaroL is associated with an aerolysin domain, predicted to play a defensive role against parasites through pore-forming.

Personal contributions

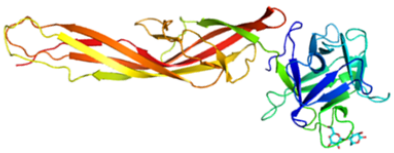
Development of TrefLec database and visualisation features. Building trefoil motifs, selection of the protein databases, screening using HMMER tools. Development of the Python pipeline to process the results and integrate them into the site with filtering steps, formatting, scoring. Protein feature acquisition including the taxonomy, and loading in the SQL database. Search for Mytilectins in genomes. Writing the part of the manuscript describing the database.



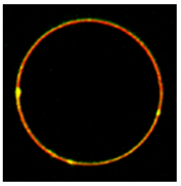
Mytilectin conserved domain used for the identification



3D structure of the identified mytilec in *Salpingoeca rosetta*



Laetiporus sulphureus lectin complexed with lactose



SaroL bind to the vesicle membrane

Identification et caractérisation d'une lectine β-trefoil des eucaryotes inférieurs avec un domaine aérolysine

Simona Notova, François Bonnardel, Annabelle Varrot, Frédérique Lisacek and Anne Imberty

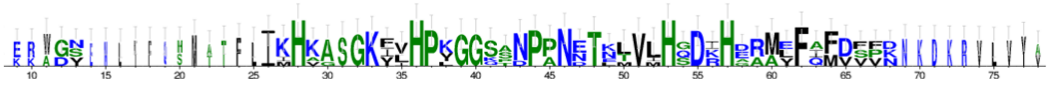
Manuscript in preparation

Abstract

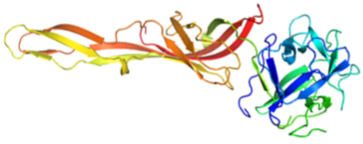
Les protéines de liaison aux glycanes, appelées lectines, jouent un rôle important dans la reconnaissance des glycanes, les interactions entre les cellules et le système immunitaire. L'efficacité des lectines liant le glycane repose principalement sur la multivalence, résultant soit de l'oligomérisation, soit de la répétition en tandem des domaines des lectines. Le fold β-trefoil, caractérisé par des répétitions en forme de trois feuilles, est adopté par plusieurs classes de lectines. Les répétitions en tandem dans les séquences d'acides aminés entravent la détection des régions codantes dans les génomes complets. En se basant sur la classification des structures 3D des lectines proposée dans Unilectin3D, en particulier pour les lectines β-trefoil, nous avons défini la signature trèfle-feuille de 11 classes de lectines. L'analyse de UniProtKB et NCBI-nr nous a permis d'identifier 26994 lectines β-trefoil dans 3607 espèces. Les séquences prédites sont organisées dans la base de données TrefLec disponible sur unilectin.eu/trefoil, où elles peuvent être recherchées et visualisées. Nous avons sélectionné l'exemple de SoraL, une lectine prédite dans Salpingoeca rosetta, un membre de la colonie primitive formant des choanoflagellés qui sont les parents unicellulaires les plus proches des animaux. SoraL appartient à la famille des Mytilectines venant des mollusques et brachyopodes. La lectine est exprimée de manière recombinante et se lie à la N-acétylgalactosamine. La structure cristalline confirme la similitude structurelle avec les Mytilectines des mollusques connus mais aussi avec les lectines fongiques. Enfin, le SaroL est associé à un domaine d'aérolysine, dont on prédit qu'il jouera un rôle défensif contre les parasites par la formation de pores.

Contributions

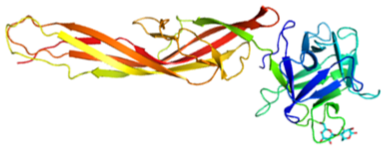
Développement de la base de données TrefLec et des fonctionnalités de visualisation. Construction des motifs de Trefoil, sélection des bases de données de protéines, criblage à l'aide des outils HMMER. Développement du pipeline Python pour traiter les résultats et les intégrer dans le site avec des étapes de filtrage, de formatage, de notation. Acquisition des caractéristiques des protéines, y compris la taxonomie, et chargement dans la base de données SQL. Recherche de Mytilectins dans les génomes. Ecriture de la partie portant sur les bases de données et prédiction.



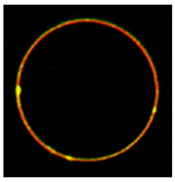
Mytilectin conserved domain used for the identification



3D structure of the identified mytilec in *Salpingoeca rosetta*



Laetiporus sulphureus lectin complexed with lactose



SaroL bind to the vesicle membrane

7 Result III: LectomeXplore database and prediction

The reference genomes lack lectin annotation based on the information available in the literature and on the 3D structure of lectins. By using sequence homology tools, it is possible to find lectins of similar structure and function between kingdoms. Homologous proteins can derive from the evolution of the same common ancestor by plasmid transfer. The similarity of fold and function can also result from convergent evolution, but then a similarity of sequence is missing.

Because of the poor annotation of lectins in genomes, it appeared necessary to develop a tool to search for lectins in sequence databases. Indeed, lectins are described in UniProt and Pfam using a variety of nomenclatures ranging from general nomenclatures (glycan-binding protein, sugar-binding, carbohydrate-binding protein, recognition domain) to more specific nomenclatures (glycan name, or lectin historical name...). A large number of keywords must be used, such as mannose, adhesin, agglutinin... ricin, which is not practical. The general terms, i.e. "Glycan binding proteins", sometimes refer to other classes of proteins such as sugar carriers or enzymes and must be filtered. The results associated with lectin-related keywords in the databases are listed in Table 14.

	lectin	glycan binding	carbohydrate binding	sugar binding	agglutinin
PFAM domain	531	197	775	930	128
CATH superfamily	110	0	40	11	17
SCOPe family	14	0	9	2	10
InterPro domain	804	47	613	454	43

Table 14: Statistics on protein family annotated with lectin related keywords.

The new lectin classification provides lectin classes that share 20% of sequence similarities, allowing robust preserved patterns to be defined which can be used to search for candidate lectins in databases. Similarly to the previous prediction of tandem repeat lectins, HMM profiles are established using the HMMER tool. To avoid solitary peripheral domains, at least 80% of lectin must be present in the conserved motif MSA. HMMER uses the HMM profiles to identify candidate lectins in the selected protein dataset. The NCBI-nr and UniProtKB protein databases are used by HMMER-search, which then uses the HMM profiles to identify candidate lectins. Only results exceeding a p-value threshold are declared by HMMER as predicted lectins.

The NCBI-nr and UniProtKB databases have the advantage of covering all kingdoms (without having to manually obtain the proteome of each species of interest) but the disadvantage of being highly redundant.

Compared to the tandem repeat score, LectomeXplore uses a similarity score calculated between the consensus sequence of the reference (generated by HMMER) and the best predicted lectin domain (in case several domains are predicted). The pipeline developed in Python, shown in Figure 35, launches BLAST on the 3D structures to generate the lectin classification, prepares HMMER input MSA files, launches and parses HMMER output and formats the results to load them in the SQL database.

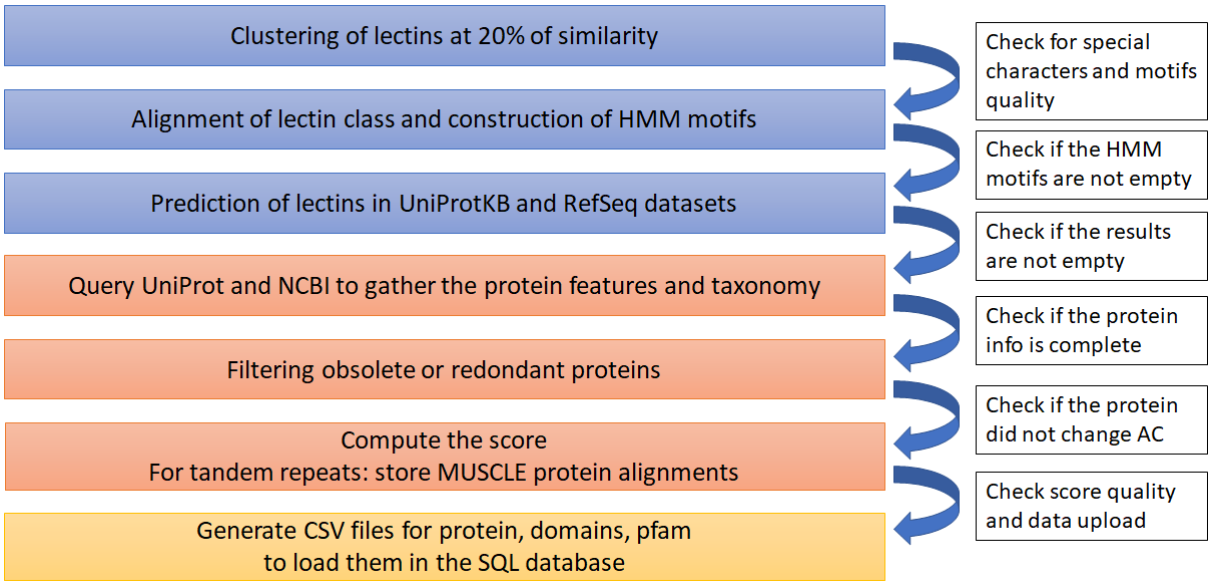


Figure 35: LectomeXplore Python pipeline

7.1 LectomeXplore database data storage

The LectomeXplore web database uses an SQL database containing the predicted lectin functions and communicates with the PHP server to generate the website pages according to the user's requests. The SQL database is composed as shown in Figure 36.

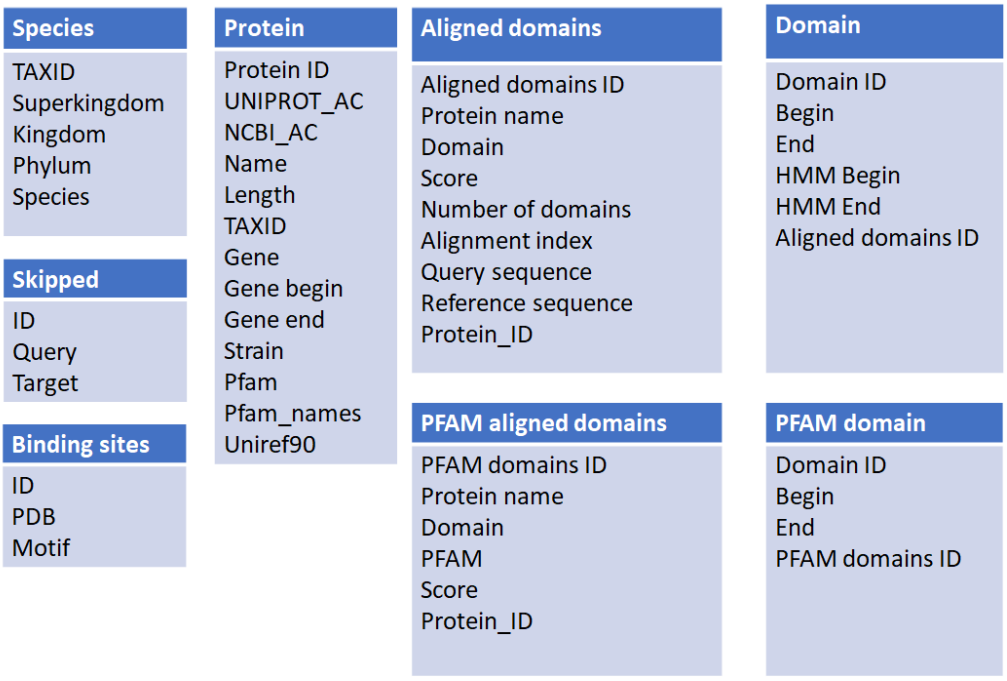


Figure 36: UML representation of LectomeXplore database

The LectomeXplore database contains several tables for all the information displayed on the web database. The Protein table contains the characteristics associated with the NCBI and UniProt AC, the protein name and length, the taxonomy of the associated species, the genomic source, and the identified Pfam name. The species table contains the taxonomy. The two tables Domains and Aligned Domain contain the identified lectin regions with the name, score, the reference consensus sequence already aligned with the best identified domain and the associated proteinID, the number of identified regions, and the positions of each region on the protein. The "skipped" table contains the filtered protein sharing a region of 100 identical residues with the already identified protein for the same species. The binding site table contains short patterns corresponding to the glycan binding pockets, used to identify the residue required for glycan binding of the reference pattern. The Pfam Aligned Domain and Pfam Domain tables contain the information of the identified Pfam domain on the protein.

7.2 Development of LectomeXplore module

Predicted lectins are mostly designated as "undefined", "uncharacterised" in UniProt, which makes it easy to see the extent to which predicted lectins are not annotated as lectins and require automatic annotation and manual processing. Predicted proteins may contain other functional domains present outside the lectin domain, which can be identified with domain protein families such as Pfam

domains. The Pfam domains of lectins are removed. Pfam also does not contain all functional domains and it would be preferable to have other domains from protein families.

For a large number of folds, lectins are predicted in other kingdoms and could be explored. The heatmap permits visualization of the presence and absence of individuals in the function of two factors. They are used to represent the predicted lectins by classes and taxonomic branches. It allows to quickly visualize the distribution of lectin between different classes or between species.

7.4 Article VI : The LectomeXplore database

LectomeXplore, an update of UniLectin for the discovery of carbohydrate-binding proteins based on a new lectin classification

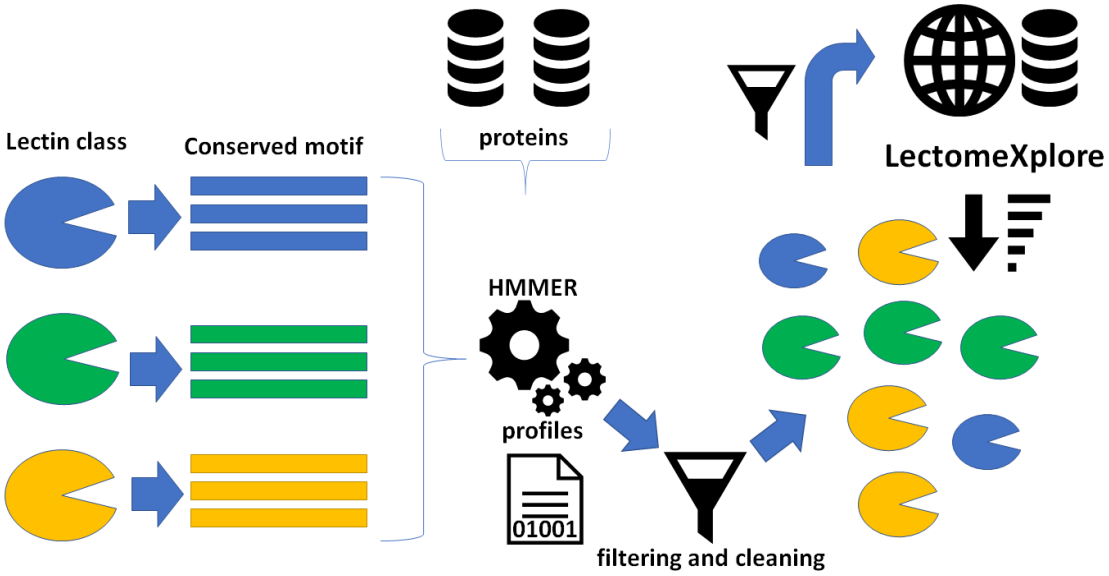
Bonnardel, F., Mariethoz, J., Pérez, S., Imberty, A., & Lisacek, F. (2021). LectomeXplore, an update of UniLectin for the discovery of carbohydrate-binding proteins based on a new lectin classification. **Nucleic Ac. Res.** (in press) https://doi.org/10.1093/nar/gkaa1019

Abstract

Lectins are non-covalent glycan-binding proteins mediating cellular interactions but their annotation in newly sequenced organisms is lacking. The limited size of functional domains and the low level of sequence similarity challenge usual bioinformatics tools. The identification of lectin domains in proteomes requires the manual curation of sequence alignments based on structural folds. A new lectin classification is proposed. It is built on three levels: 1) 35 lectin domain folds, 2) 109 classes of lectins sharing at least 20% sequence similarity and 3) 350 families of lectins sharing at least 70% sequence similarity. This information is compiled in the UniLectin platform that includes the previously described UniLectin3D database of curated lectin 3D structures. Since its first release, UniLectin3D has been updated with 485 additional 3D structures. The database is now complemented by two additional modules: PropLec containing predicted β-propeller lectins and LectomeXplore including predicted lectins from sequences of the NBCI-nr and UniProt for every curated lectin class. UniLectin is accessible at https://www.unilectin.eu/

Personal contributions

First draft of a multi-level classification based on structure and sequence. Development of the LectomeXplore database and visualization features. Grouping of lectins, construction of lectin class patterns. Selection of protein databases and their analysis using HMMER to identify lectins. Development of the Python pipeline for filtering, formatting and scoring results. Acquisition of the characteristics of the predicted proteins, including taxonomy, and their loading into the SQL database. Wrote the article and made the figure.

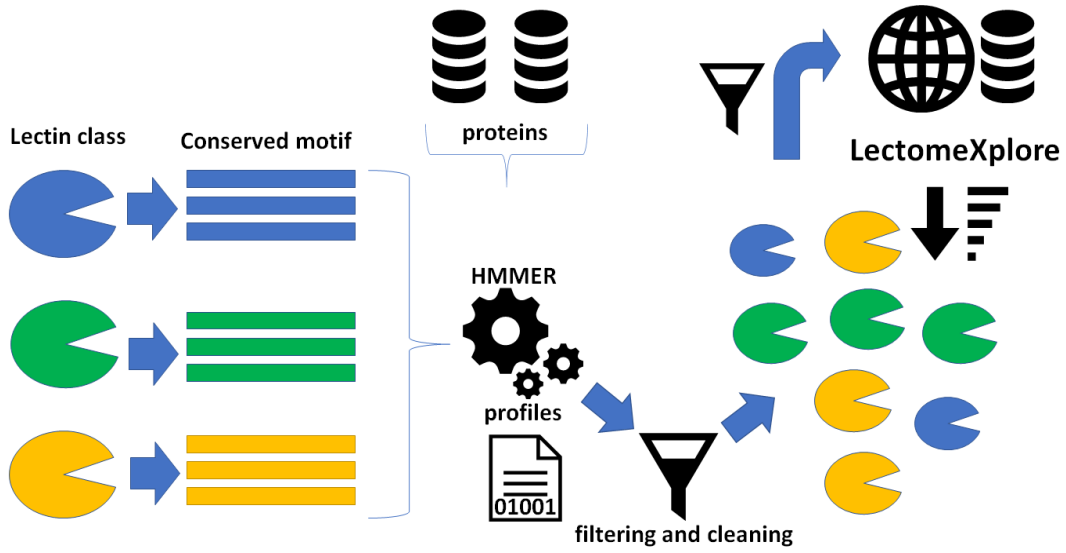


LectomeXplore, une mise à jour d'UniLectin pour la découverte de protéines liant les glucides basée sur une nouvelle classification des lectines

Bonnardel, F., Mariethoz, J., Pérez, S., Imberty, A., & Lisacek, F. (2021). LectomeXplore, an update of UniLectin for the discovery of carbohydrate-binding proteins based on a new lectin classification. **Nucleic Ac. Res.** (in press) https://doi.org/10.1093/nar/gkaa1019

Abstract
Les lectines sont des protéines non covalentes reconnaissant les glycannes et servant de médiateur dans les interactions cellulaires, mais leur annotation dans les organismes nouvellement séquencés fait défaut. La taille limitée des domaines fonctionnels et le faible niveau de similarité des séquences mettent au défi les outils bioinformatiques habituels. L'identification des domaines des lectines dans les protéomes nécessite un nettoyage manuel des alignements de séquences basés sur les repliements structurels. Une nouvelle classification des lectines est proposée. Elle est construite sur trois niveaux : 1) 35 repliements/fold de domaines de lectines, 2) 109 classes de lectines partageant au moins 20% de similarité de séquence et 3) 350 familles de lectines partageant au moins 70% de similarité de séquence. Ces informations sont compilées dans la plate-forme UniLectin qui comprend la base de données UniLectin3D décrite précédemment qui contient les structures 3D des lectines. Depuis sa première version, UniLectin3D a été mise à jour avec 485 structures 3D supplémentaires. La base de données est maintenant complétée par deux modules supplémentaires : PropLec contenant les lectines prédites de fold β-propeller et LectomeXplore comprenant les lectines prédites à partir des séquences du NBCI-nr et UniProt pour chaque classe de lectine conservée. UniLectin est accessible à l'adresse suivante : https://www.unilectin.eu/

Contributions
Première ébauche d'une classification à plusieurs niveaux basée sur la structure et la séquence. Développement de la base de données LectomeXplore et des fonctionnalités de visualisation. Regroupement des lectines, construction de motifs de classes de lectines. Sélection des bases de données de protéines et leur analyse à l'aide de HMMER, pour identifier les lectines. Développement du pipeline Python pour le filtrage, le formatage, la notation des résultats. Acquisition des caractéristiques des protéines prédites, y compris la taxonomie, et leur chargement dans la base de données SQL. Rédaction de l'article et préparation des figures.



8 Result IV: Application to the prediction of lectomes

More than 500,000 lectin candidates are available across the different kingdoms, a dataset too large to be explored by hand. In order to refine these predictions, different exploration possibilities and applications are possible. The lectome of species of particular interest such as emerging pathogens, can bring new insight into therapeutic targets. In less explored kingdoms (Algi, Invertebrates) new biomolecules for biotechnology can be discovered. The protein architectures that combine one or more lectin domain with others functional domains are very interesting. A lectin class can be used to build a phylogenetic tree to study lectin evolution. In fundamental research, the lectin can be studied in interaction with a "biome" and a particular environment.

8.1 Human microbiome lectins exploration

Healthy relationships between populations of microbes and human anatomical sites including the skin or the digestive system are called the microbiome and demonstrated to be essential for health and well-being. The multiple anatomical environments of the human microbiome are represented by Figure 37. The effect of the microbiome on human health is underestimated and recent advances in metagenomics have led to a better understanding of the relationship between disease and the human microbiome. The intestinal microbiome has received a lot of attention in recent years (MacIntyre et al. 2015; Mitra et al. 2020). The 100 trillion symbiotic microorganisms that reside in the human gut perform several vital biological functions. Imbalances in microbial populations are associated not only to intestinal disorders, but with inflammations, immunology disorders, and production of metabolites that alter other organs.

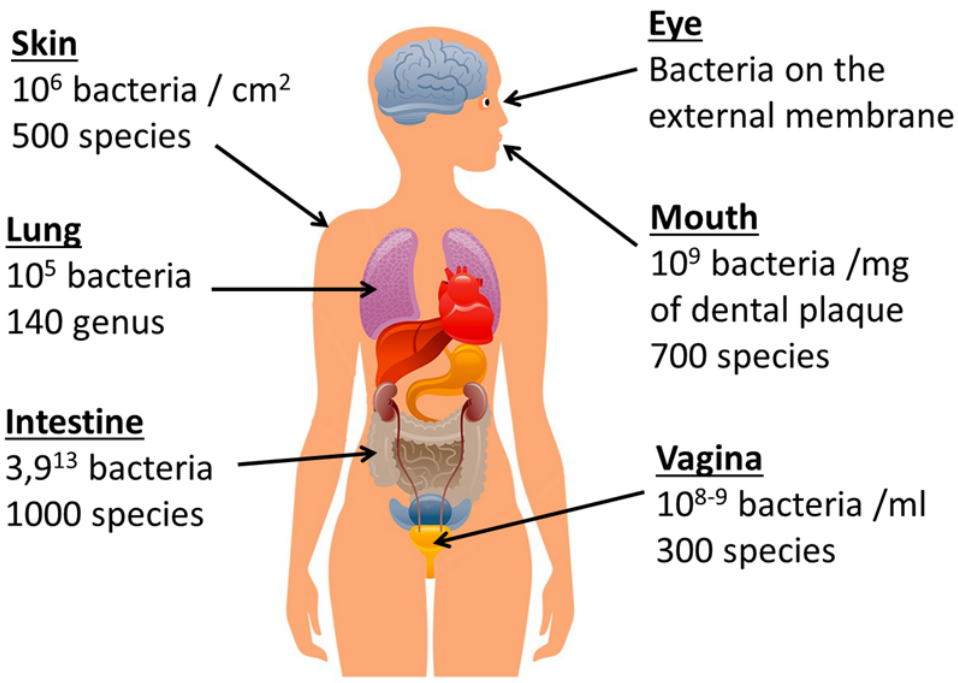


Figure 37: Atlas of the human microbiome

Less work has been dedicated to the microbiome of the reproductive system. The vaginal microbiome is particularly important for healthy pregnancies (Brown et al. 2018). It has been shown that an imbalance in the microbiome species is associated with problems of premature birth, as shown in Figure 38. Preterm birth (PTB) remains the second most common cause of neonatal death across the globe. In this section, vaginal microbiome species are discussed because of recent interest in the relationship between the human microbiome and disease. By recovering the species present in the vaginal microbiome and their strains, it is possible to analyse their proteomes in order to identify their lectome. Correlations are analysed between the dysfunctional microbiome and the presence of certain lectins (Article VII).

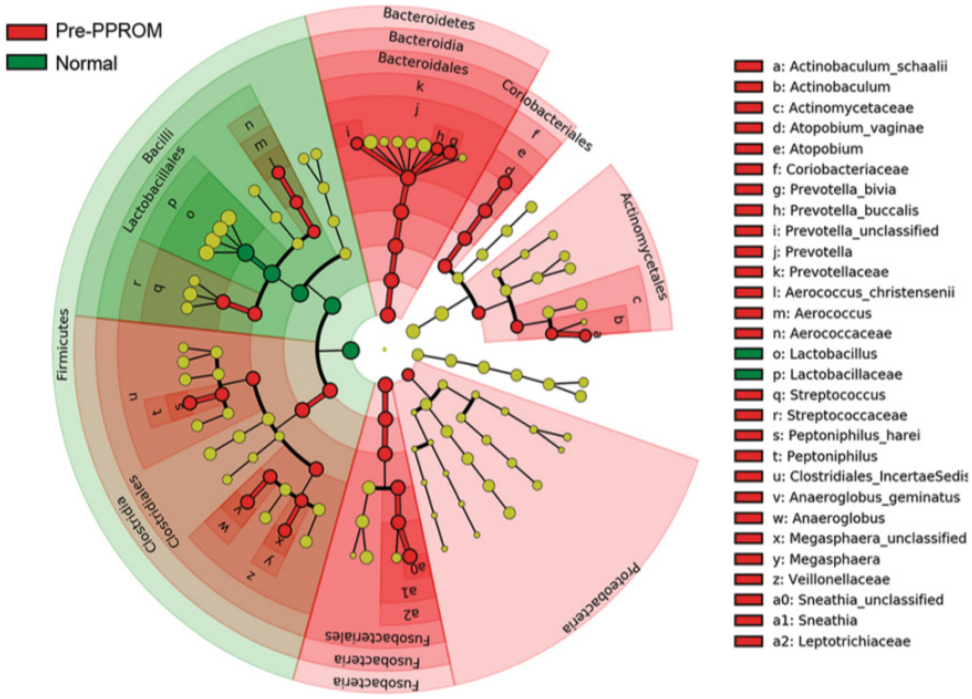


Figure 38: Bacterial taxonomic groups discriminate between normal-term delivery and women destined to experience preterm prelabour rupture of the fetal membranes (PPROM). Cladogram describing differentially abundant vaginal microbial clades and nodes observed between women subsequently experiencing normal-term delivery or PPROM (Brown et al. 2018).

8.2 Mycobiome lectins exploration in relation to fungi ecology

Mushrooms are a very diverse branch of eukaryotes, not capable of photosynthesis and therefore heterotrophic like animals. They are the main decomposers in ecological systems and spread by transmission of fungal spores in air or water. Abundant throughout the world, most fungi go unnoticed because of their small structures and way of life in the soil or on dead matter. However, they have been domesticated for food processing, for the production of antibiotics, or the industrial production of enzymes. Some fungi are used as biological pesticides to control bacteria, fungi and insects.

MycoCosm is a fungal genomics portal (http://jgi.doe.gov/fungi) dedicated to the annotation of fungal genomes through its own annotation pipeline, to fill the gaps in the fungal tree of life. It is managed by the Joint Genome Institute (JGI) of the US Department of Energy (DOE) and the international fungal community. Currently, the Mycocosm database provides 1750 fungal strains, 1200 of which have been published in the different branches of the tree of fungal life shown in Figure 39.

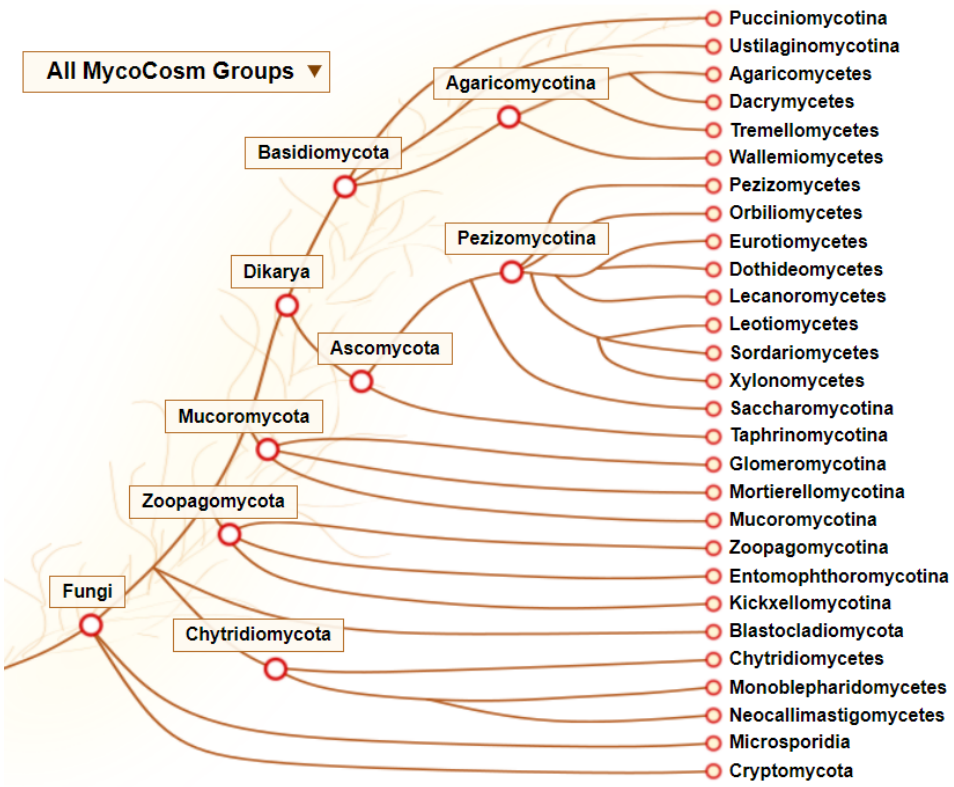


Figure 39: Mycocosm Fungi taxonomic tree

Mushrooms occupy very diverse ecological niches. Mycorrhizal fungi are plant symbionts, i.e. they live in an association of mutual interest. Saprobes are involved in the decomposition of organic matter (dead or waste). Lichens are composite organisms formed by algae or cyanobacteria living among the filaments of multiple species of fungi. Yeasts are unicellular microorganisms. Some fungi develop in symbiosis with insects in mutual relationships. Ectomycorrhizal fungi grow on the root of the plant while endophytic fungi live in a plant for at least part of their life cycle without causing any apparent disease. Ecological databases are available thanks to curated ecology and genome based ecology predictions by Funguild tool (Nguyen et al. 2016).

Fungal lectins are of great interest for the development of new drugs against opportunistic fungi and the discovery of new antifungal and antibacterial lectins. Mycocosm database provides genomes and proteomes of better quality than those provided by the NCBI. It is therefore possible to analyse fungi proteomes in order to identify their lectome. The lectomes are used to search for correlations between the ecology of fungi and the presence of certain lectins (Article VIII).

8.3 Article VII : Exploration of the vaginal microbiome lectome

Proteome-wide prediction of bacterial carbohydrate-binding proteins as a tool for understanding commensal and pathogen colonisation of the vaginal microbiome

François Bonnardel, Stuart M. Haslam, Anne Dell, Ten Feizi, Yan Liu, Virginia Tajadura Ortega, Yukie Akune, Lynne Sykes, Phillip R. Bennett, David A. MacIntyre, Frédérique Lisacek and Anne Imberty

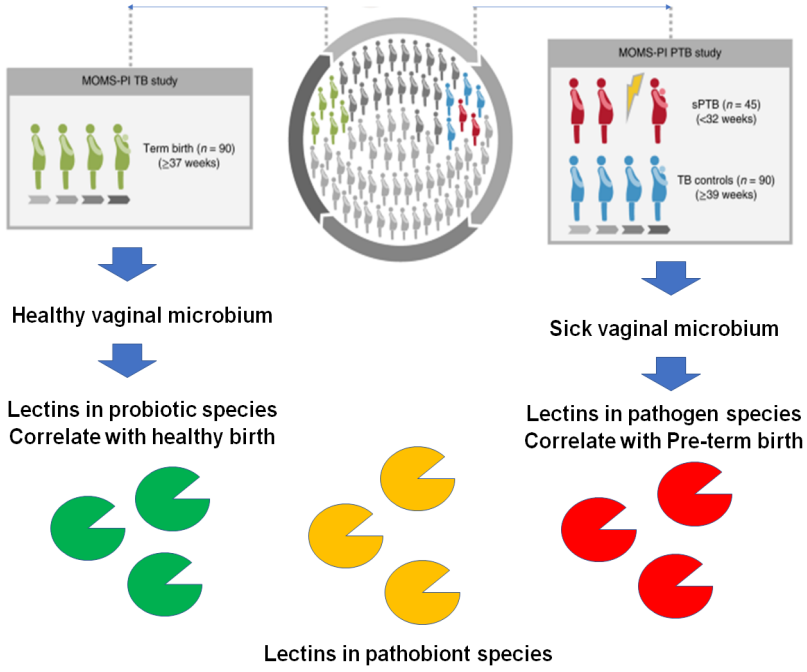
Manuscript submitted to Microbiome

Abstract

Bacteria use protein receptors called lectins to anchor to specific host surface sugars. The role of lectins in the vaginal microbiome, and their involvement in reproductive tract pathophysiology is poorly defined. Here we establish a classification system based on taxonomy and protein 3D structure to identify 109 lectin classes. Hidden Markov Model (HMM) profiles for each class were used to search bacterial genomes, resulting in the prediction of >100 000 bacterial lectins available at unilectin.eu/bacteria. Genome screening of 90 isolates from 21 vaginal bacterial species showed that potential pathogens produce a larger variety of lectins than commensals indicating increased glycan binding potential. Both the number of predicted bacterial lectins, and their specificities for carbohydrates correlated with pathogenicity. This study provides new insights into potential mechanisms of commensal and pathogen colonisation of the reproductive tract that underpin health and disease states.

Personal contributions

Discussion with the clinicians from Imperial College. Predictions of bacterial lectins, filtering and analysis of the results uploaded on the dedicated part of UniLectin. Building the vaginal microbiome proteome dataset. Predictions of lectins in the vaginal microbiome. R statistical analysis and generation of the figures.



Prédiction dans les protéomes de lectines bactériennes comme outil pour comprendre la colonisation commensale et pathogène du microbiome vaginal

François Bonnardel, Stuart M. Haslam, Anne Dell, Ten Feizi, Yan Liu, Virginia Tajadura Ortega, Yukie Akune, Lynne Sykes, Phillip R. Bennett, David A. MacIntyre, Frédérique Lisacek and Anne Imberty

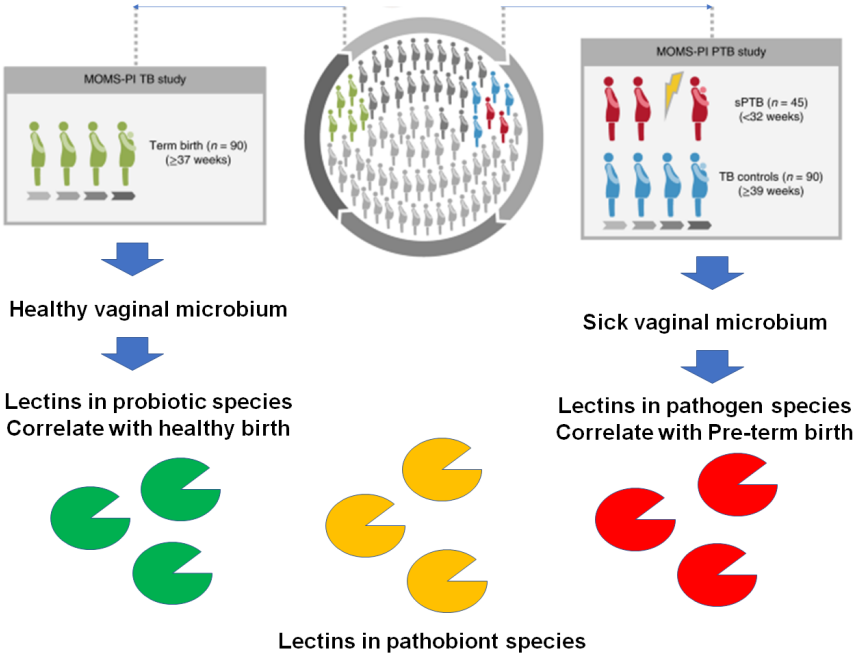
Manuscrit soumis à Microbiome

Abstract

Les bactéries utilisent des récepteurs protéiques appelés lectines pour s'ancrer à des sucres de surface spécifiques de l'hôte. Le rôle des lectines dans le microbiome vaginal et leur implication dans la physiopathologie de l'organe reproducteur est mal défini. Nous établissons ici un système de classification basé sur la taxonomie et la structure 3D des protéines pour identifier 109 classes de lectines. Les profils Hidden Markov Model (HMM) de chaque classe ont été utilisés pour analyser les génomes bactériens, ce qui a permis de prédire plus de 100 000 lectines bactériennes disponibles sur unilectin.eu/bacteria. Le criblage du génome de 90 isolats provenant de 21 espèces de bactéries vaginales a montré que les pathogènes potentiels produisent une plus grande variété de lectines que les commensales, ce qui indique un potentiel accru de liaison aux glycanes. Le nombre de lectines bactériennes prédites ainsi que leurs spécificités pour les glucides sont corrélés avec la pathogénicité. Cette étude apporte de nouvelles connaissances sur les mécanismes potentiels de colonisation de l'organe génital par des agents commensaux et des pathogènes, qui sont à la base des états de santé et des maladies.

Contributions

Discussion avec les cliniciens de Imperial College. Prédictions des lectines bactériennes, filtrage et analyse des résultats mis en ligne sur la partie dédiée d'UniLectin. Construction de l'ensemble de données sur le protéome du microbiome vaginal. Prédictions des lectines dans le microbiome vaginal. R analyse statistique et génération des figures.



8.4 Article VIII: Exploration of Fungi lectomes

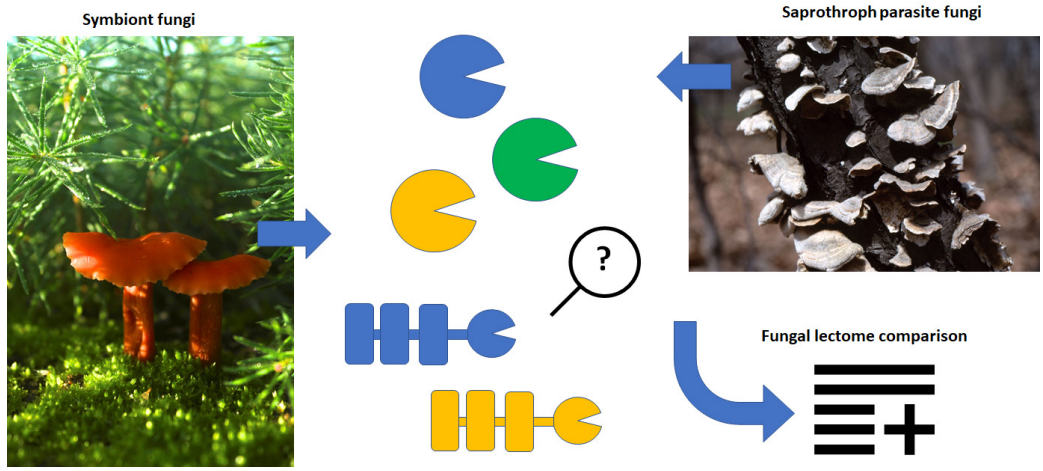
A Comprehensive Phylogenetic and Bioinformatics Survey of Lectins in the Mycota kingdom

Annie Lebreton, François Bonnardel, Yu-Cheng Dai, Anne Imberty, Francis Martin, Frédérique Lisacek

Manuscript in preparation

Abstract
Fungal lectins are a large family of glycan-binding proteins, with no enzymatic activity. They play fundamental biological roles in the interactions of fungi with their environment and are found in many different species throughout the Mycota kingdom. In particular, their contribution to defence against feeders has been emphasized and extracellular lectins may be involved in the recognition of bacteria, fungal competitors and specific host plants. Their carbohydrate specificities and quaternary structures vary widely, but evidence for an evolutionary relationship within the different classes of lectins is provided by the high degree of amino acid sequence identity shared by the different fungal lectins. The UniLectin3D database contains 193 3D structures of fungal lectins, of which 126 are characterised with their carbohydrate ligand. These lectin classes were used to construct 107 lectin motifs in 26 folding configurations and to screen 1,223 genomes deposited in the Joint Genome Institute MycoCosm database. The characterization of 33,518 lectin candidates in fungal genomes is based on systematic statistics regarding potential carbohydrate ligands, protein lengths, signal peptides, relative motif positions and amino acid compositions of fungal lectins. These results shed light on the evolution of the lectin gene families.

Personal contributions
Discussion with colleagues at INRA Nancy. Predictions of lectins in mycocosm proteomes. Python pipeline for scoring, features acquisition and loading in unilectin-mycocosm database. Development of Mycocosm database in UniLectin. Together with Annie, R statistical analysis and generation of the figures.



Exploration des lectines dans le royaume fongique : relation avec la reconnaissance de l'hôte et l'écologie

Annie Lebreton, François Bonnardel, Anne Imberty, Francis Martin, Frédérique Lisacek

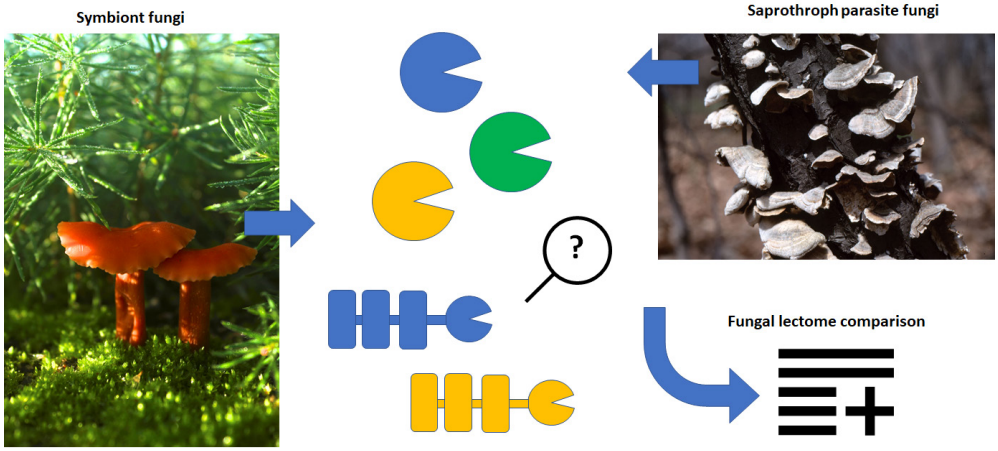
Manuscript in preparation

Abstract

Le royaume des champignons est une classe très diversifiée d'eucaryotes. Principalement présent dans les écosystèmes terrestres mais aussi dans les milieux marins, il est composé d'organismes unicellulaires et multicellulaires, libres et symbiotiques qui interagissent avec un large éventail d'autres organismes en tant que pathogènes, saprotrophes et symbiotes. Les lectines sont des protéines réversibles de liaison au glycane, sans activité enzymatique. Les lectines fongiques sont utilisées dans la défense immunitaire innée contre les bactéries, les virus et autres champignons. Elles jouent également un rôle dans la reconnaissance de l'hôte et les interactions. Les lectines ayant des structures 3D connues sont classées dans UniLectin3D avec 2225 structures 3D de lectines, dont 193 structures de lectines fongiques (126 avec un ligand glucidique). Les classes de lectines permettent de définir 107 motifs (appartenant à 26 folds) qui ont été utilisés pour identifier les lectines du règne fongique. Les protéomes de Mycocosm ont été analysés à l'aide de motifs de lectines pour identifier 33518 lectines fongiques candidates. Les lectomes des Agaricomycetes ont été comparés en fonction de leur écologie. Des lectines intéressantes de *Laccaria Bicolor* ont été mises en évidence. Les prévisions sont disponibles sur le site unilectin.eu/mycocosm.

Contributions

Discussion avec les collègues de l'INRA Nancy. Prédictions des lectines dans les protéomes de Mycocosm. Pipeline Python pour le scoring, l'acquisition de caractéristiques et le chargement dans la base de données unilectin-mycocosm. Développement de la base de données Mycocosm dans UniLectin. Analyse statistique avec R et génération des figures.



9 Discussion and Perspectives

9.1 Achievements

The UniLectin.eu portal has been developed to provide the scientific community, worldwide, with both curated and predicted lectin information. On this portal, the UniLectin3D database covers lectin 3D structures (Article I). A tutorial is available on UniLectin3D and has been used to build a tutorial chapter for the database (Article II). UniLectin3D administration for creating and modifying lectins is available in a dedicated interface. A first step towards predicting lectins in genomes has been to focus on tandem repeating lectins and in particular on the β-propeller and β-trefoil lectins, which have interesting symmetrical characteristics (see Article III). Dedicated databases were built, called PropLec for predicted β-propeller lectins using blade sequences (Article IV) and TrefLec database for predicted β-trefoil lectin using leaf sequences (Article V).

For all the other lectin classes, a different approach was needed. The identification of signature patterns requires the definition of appropriate groups sharing a similar amino acid sequence. A new classification is proposed for the lectins, the fold is more conserved than the sequence and is used as a first level for grouping lectins. Then the second level based on 20% of sequence similarity form 109 conserved classes. HMMER motifs have been defined and used to identify lectins in genomes, leading to the LectomeXplore database of predicted lectins (Article VI), dedicated to the exploration and comparison of lectomes. In collaboration with colleagues at Imperial College (London) working on vaginal microbiomes, the lectome of Lactobacilli associated with vaginal health was compared with those of bacteria associated with diseases, i.e. pathogens and pathobionts. (Article VII). Another collaboration with the "Tree-Microbe Interaction" group (INRAE, Nancy) has made it possible to analyse the lectomes of fungal species available in the Mycocosm database and to analyse them according to their ecology (Article VIII).

The UniLectin3D database and the classification of lectins can be further improved. The prediction of lectins can be improved through lectin classification and the selection of multiple protein sources. Prediction modules can be improved by adding information. The classification of lectins allows the prediction and discovery of new lectins, which can then be used to improve the classification again, forming a full cycle as shown in Figure 40.

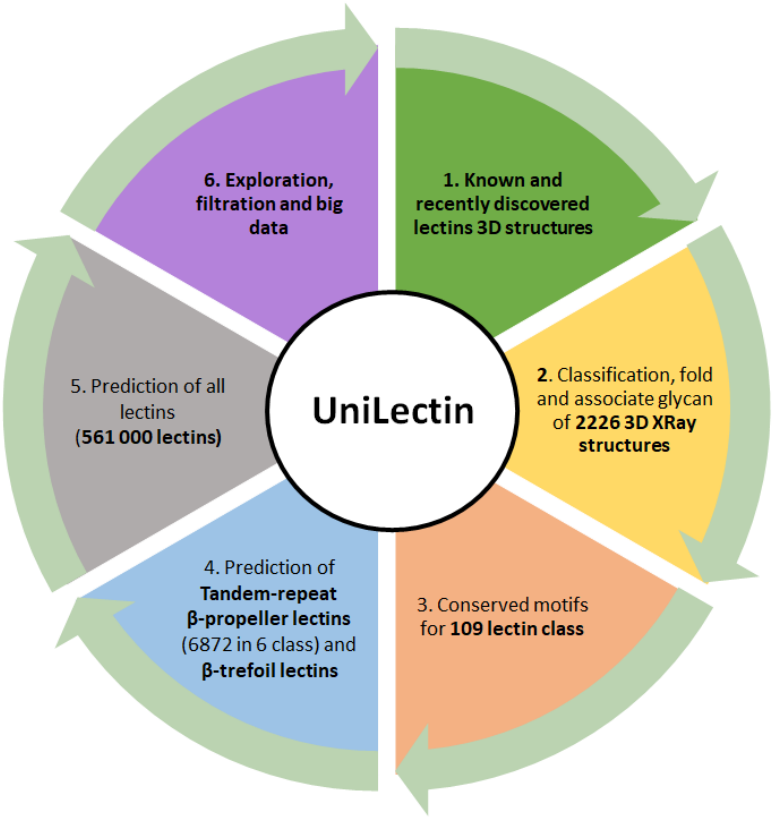


Figure 40: UniLectin platform represented as a perpetual circle, composed of the developed modules PropLec, TrefLec, LectomeXplore, MycoLec.

9.2 Evolution of Lectin classification and UniLectin3D

9.2.1 Lectin classification

The new classification of the lectins has three levels: the fold, the class and the family. The first and second levels correspond to the structural classification based on the CRD, and the third family level provides a finer grouping associated with binding pocket specificity. The classification of lectins must be maintained manually with the creation of fold levels, classes or families in the case of newly discovered 3D structures of lectins. The classification can be further improved by making better use of the protein 3D structure which contain more information to compare proteins than just the sequence, and linking the binding pocket information, using automatic 3D classification tools such as ProtNN (Dhifli and Diallo 2016).

The lectin classification must be maintained, and a future improvement would be to design and implement an automatic tool for searching 3D structures of lectins in the weekly release of the PDB. RCSB structures can be queried using lectin keywords or related names, interacting with a

carbohydrate-bound ligand. Another way is to use the predicted lectins 3D structures, and focus on the structures bound non-covalently to a glycan.

The present classification of lectins could be extended since it covers only the lectins with available 3D structures. It would be of interest to include the sequences of other lectins from manual exploration of literature. Some UniProt proteins and the Pfam and InterPro protein families are potential lectins.

9.2.2 UniLectin3D evolution

The UniLectin3D database stores curated lectin 3D structures and provides tools for their exploration and the maintenance of the database. UniLectin3D could provide additional and much needed information, such as the lectin affinity for specific glycan ligands, possibly under different conditions. Furthermore, a standardised protocol for the isolation of the 3D structure of the lectin could be very useful to biochemists.

Currently, UniLectin3D still uses a copy of Glyco3D database structure, which was the logical choice to reuse the information stored in Glyco3D. After three years of development, UniLectin3D has evolved and the database will be rebuilt as shown in the following graph in Figure 41.

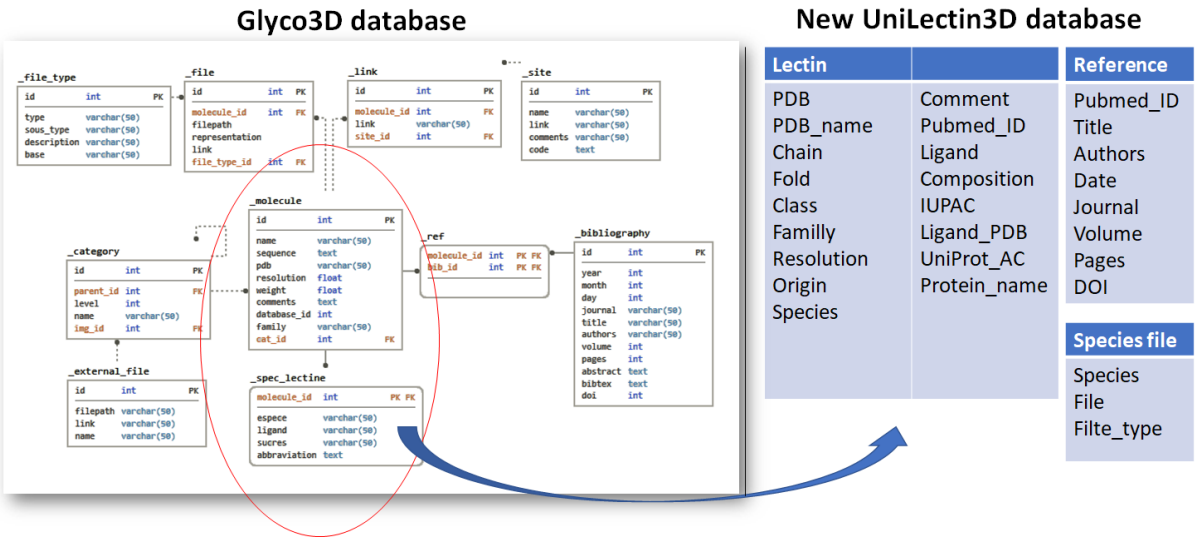


Figure 41: UML model of Glyco3D database used currently by UniLectin3D, with on the right the model of UniLectin3D future database. The information relative to lectins is currently split between two tables named molecule and spec_lectine which will be merged in a unique lectin table. Each lectin has now a fold, a class, and a family. For each lectin PDB structure, only the first publication is associated. Each species has one picture file.

Bioinformatics database are supposed to respect open science, which aims to make research data available for public access and reuse, allowing other scientists to make new discoveries by integrating

data from multiple sources. Complex and diverse data are hard to integrate and understand, as they are generated from a wide variety of experimental and computational sources, formatted by distinct teams. In 2016, the "Fair Guidelines for the Management and Stewardship of Scientific Data" was published, that stand for "Findability, Accessibility, Interoperability and Reuse of Digital Assets" (Wilkinson et al. 2016). UniLectin portal already provides open access to all information, and UniLectin3D list the PDB ids of all lectin structures, that are cross referenced with UniProt.

With FAIR data, it is possible to integrate multiple sources and make websites interact with each other. Known as the semantic web website interactions, using triple-shop RDF technology as web services, facilitate the exploration of scientific data. The UniLectin-to-GlyConnect and SugarBind interaction glycans are currently being developed using an RDF store and Sparql server available on GlyConnect, which will allow each glycan containing the epitopes recognised by the lectins to be searched. The structure of GlycoCT glycans can be translated into SPARQL to query an RDF glycan shop. Future development of UniLectin will also include RDF communication with general databases such as UniProt and PDB, as well as other glyco-bioinformatics resources found in the GlyGen or the GlyCosmos environments where RDF is a common concern (Campbell et al. 2017).

9.3 Lectin prediction: new methods and databases

Thanks to the new lectin classification, candidate lectins have been predicted across kingdoms and are provided for further exploration in LectomeXplore database on UniLectin portal.

9.3.1 Prediction method and LectomeXplore module

Among the 109 classes of lectins in UniLectin3D, two classes are for the moment excluded from LectomeXplore: the chi lectin TIM due to its enzymatic origin, and the variable lymphocyte receptors that are adaptative immune receptors from jawless fishes able to bind to a large number of antigens, causing too many false positives. The 107 classes of lectins left are used to generate conserved patterns using MUSCLE to calculate MSA alignments, which can also be improved by using more accurate methods than MUSCLE, or combining 2D protein sequences with the available 3D structures that contain much more information, based on methods such as Expresso (Armougom et al. 2006), to reliably align distantly related sequences.

To ensure the glycan binding functionality of the predicted lectin, glycan binding motifs are currently displayed for manual verification. An appropriate method for storing and representing glycan binding residues is needed. Glycan binding domains are currently stored using small patterns of five residues

to identify their positions in the sequences and display them in LectomeXplore. However, the binding sites are not always correctly identified.

Lectin prediction require protein dataset to analyse, with both UniProt and NCBI-nr protein sequence datasets used because of the simplicity of a single downloadable protein file containing all available species, instead of other platforms such as GenBank, and because of the limitations of the server available for the analysis. Sadly such databases provide protein from both verified and non-verified genomic sources. Therefore, protein from partial and fragmented genomes contain many sequencing errors and associated information, which would require a filtering step. In the current LectomeXplore, there are no direct criteria to ensure the quality of a predicted lectin source genome. The solution could be to give the user the possibility to select reference genomes, to choose the quality of predicted lectins. Reference genomes are available in other protein sources such as translated GenBank, UCSC and JGI. For example, the Ginkobylobin lectin domain that belongs to the plant kingdom was predicted on 13 proteins from the bacteria *Acinetobacter* which cannot be identified in *Acinetobacter* genomes and comes from genomic data of a *Ginkgo biloba* sequencing project. For this reason, it is highly probable that the protein has been described with the wrong species.

Statistical validation of prediction methods is required to ensure that generated results makes sense. Although the prediction of lectins has been evaluated using the SwissProt curated dataset, validation method (ie. cross-validation method) is required to 1/ ensure the results are lectins and no other protein such as enzymes; 2/ refine lectin prediction and find the optimal similarity score for each class or family of lectins. To reduce the number of false positive predictions, a more specific prediction at the family level could be used. For classes of lectins derived from lectin-enzymes denatured after mutation, the presence of this mutated site could be used as a selection criterion.

To avoid too much redundancy in the predicted lectins, the predicted lectins are currently grouped by species. However, the lectomes of distinct species strains can differ greatly. These differences may be due to variations in growth conditions, a sequencing error or the acquisition of external plasmids with additional lectins. Future developments will use the information on strain diversity to expand the possibilities for exploring living kingdoms.

The protein identified with lectin domains can have architectures composed of multiple functional domains, combination of either lectin domain(s), or functional domain(s), or other transmembrane domains and signal peptides structural domain(s). Such Architectures can be studied using the Pfam and InterPro domains, but also based on frequently found CAZy enzymes and CBM domains. The Pfam domains do not cover all known protein domains that are available in other protein databases

such as CATH-GENE3D integrated with other domain sources, including transmembrane domains and signal peptides, in InterPro. A better integration of these domains in LectomeXplore could provide a larger overview.

Researchers are interested in genomes and proteins already published and available in UniProt/NCBI databases, but they are also interested in having the possibility to analyse their own protein or protein dataset to identify candidate lectins. Currently, LectomeXplore does not offer lectome prediction as a service, with the possibility for a user to input sequences of the proteome of a specific sample/species/environment and check whether it contains lectins. Moreover, deploying and managing web computing requires permanent support, and a dedicated server with both computing power and large storage. Another solution is to provide the lectin domains to a database of protein domains such as InterPro.

9.3.2 3D Modelling of candidate lectins and prediction of their affinity

Scientist are interested in lectin 3D structures to visualize their glycan binding pocket, generate docking with diverse glycan and glycomimetic compounds for drug design. 3D models can be generated either by homology with an existing structure, homology with domains/folds, or by *ab initio* modeling. The SWISS-MODEL server provides an automated protein structure homology modeling, and integrates the PLIP interactions of NGL viewers. The Rosetta software suite includes algorithms for computer modelling and analysis of protein structures. It has enabled advances in *de novo* protein design, enzyme design, ligand anchoring and structure prediction of biological macromolecules and macromolecular complexes.

Once built, the 3D structure of candidate lectins can be analyzed and compared to the 3D structure of the reference lectin, to identify which residues are exposed in the binding pocket and potentially involved in glycan binding. Such predictions have already been developed and are progressing either specifically for glycan binding (Zhao et al. 2014, 2018) or for all protein binding pockets to aid drug design (Jiménez et al. 2017), using lectin engineering to find new affinities with other glycans.

The prediction of reliable glycan binding specificity and affinity would greatly increase the interest in the predicted lectins. However, the acquisition of glycan binding specificity data requires either direct wet lab analysis using for example glycan-lectin arrays, or *de novo* docking approaches followed by wet lab validation. For laboratory data, only a handful of lectins have been tested against glycan arrays and even those with a few hundred oligosaccharide structures do not cover all glycans in nature. For *de novo* methods, the docking tools generate possible interactions between the protein and the ligands provided. This only works after fine adjustment of the positions corresponding to the

known binding pocket and requires a lot of computing resources. Flexibility of oligosaccharides is an issue in most available docking procedures.

9.4 Exploration of lectin candidates

With the new lectin classification, lectins can be precisely identified in the proteomes of available species. Our prediction toolbox led to detect more than 500.000 lectins among bacteria, viruses, archaea and eukaryotes, including fungi, animals, nematodes, plants, including algae. Specific tools have been developed for the visualisation of the results and the exploration of lectomes in LectomeXplore that is accessible on the web. Candidate lectins may be further analysed and this raises the possibility of obtaining new 3D structures.

Predicted lectins are of great interest to research teams around the world as potential tools to identify glycans in cancer cells. They can also be used as a lead to create a new treatment against pathogenic bacteria and fungi by competing with the binding activity of the lectins to the host. To discover new lectins, the first possibility is to focus on the predicted lectin fold across the kingdoms (bacteria, fungi, plants, nematodes and algae or protists) and analyse them to get a new insight into the diversity of lectins across the kingdoms. Each kingdom has pathogenic species that are involved in many diseases. For example, lectin search can be useful for blocking emerging pathogens in mammals due to antibiotic-resistant bacterial strains. They also have ecological niches with dedicated ecological environments with specific physiology. Ecological databases are available based on cleaned ecological information and predicted ecology based on metabolic pathways encoded in the genomes of the species. The Macadam and Faprotax databases provide information on the ecology of bacteria (Louca et al. 2016; Le Boulch et al. 2019). Based on ecological data, candidate lectins can be used to explore the evolution of lectin by expansion/duplication and compression of the corresponding domain across species.

An operon is a genome localized clusters of co-regulated genes with related functions, belonging generally in the same functional pathway. Bacterial operons can be used to find groups of glycan-related genes encoding enzymes or possible lectins. A web database for the exploration of operon containing glycol enzymes has been developed by Cazy team, named PULDB (Terrapon et al. 2018). Operon exploration allowed to define enzymes candidate validated with combination to wide wet lab screening the discovery of new enzymes (Helbert et al. 2019).

10 References

Abbott S, Iudin A, Korir PK, et al (2018) EMDB Web Resources. Curr Protoc Bioinforma 61:5.10.1-5.10.12. doi: 10.1002/cpbi.48

Abdellah Z, Ahmadi A, Ahmed S, et al (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945. doi: 10.1038/nature03001

Abhinav K V., Samuel E, Vijayan M (2016) Archeal lectins: An identification through a genomic search. Proteins Struct Funct Bioinforma 84:21–30. doi: 10.1002/prot.24949

Abhinav K V., Sharma A, Vijayan M (2013) Identification of mycobacterial lectins from genomic data. Proteins Struct Funct Bioinforma 81:644–657. doi: 10.1002/prot.24219

Abrahams JL, Taherzadeh G, Jarvas G, et al (2020) Recent advances in glycoinformatic platforms for glycomics and glycoproteomics. Curr. Opin. Struct. Biol. 62:56–69

Agirre J, Iglesias-Fernández J, Rovira C, et al (2015) Privateer: Software for the conformational validation of carbohydrate structures. Nat. Struct. Mol. Biol.

Aithal A, Sharma A, Joshi S, et al (2012) PolysacDB: A database of microbial polysaccharide antigens and their antibodies. PLoS One 7:. doi: 10.1371/journal.pone.0034613

Akune Y, Hosoda M, Kaiya S, et al (2010) The RINGS resource for glycome informatics analysis and data mining on the web. Omi A J Integr Biol. doi: 10.1089/omi.2009.0129

Al Jadda K, Porterfield MP, Bridger R, et al (2015) EUROCarbDB(CCRC): A EUROCarbDB node for storing glycomics standard data. Bioinformatics 31:242–245. doi: 10.1093/bioinformatics/btu609

Alocci D, Ghraichy M, Barletta E, et al (2018a) Understanding the glycome: An interactive view of glycosylation from glycocompositions to glycoepitopes. Glycobiology 28:349–362

Alocci D, Mariethoz J, Gastaldello A, et al (2019) GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. J Proteome Res. doi: 10.1021/acs.jproteome.8b00766

Alocci D, Suchánková P, Costa R, et al (2018b) Sugarsketcher: Quick and intuitive online glycan drawing. Molecules. doi: 10.3390/molecules23123206

Ambrosi M, Cameron NR, Davis BG (2005) Lectins: Tools for the molecular understanding of the glycocode. Org. Biomol. Chem. 3:1593–1608

Aoki- Kinoshita KF, Kanehisa M (2015) Glycomic analysis using KEGG glycan. Methods Mol Biol 1273:97–107. doi: 10.1007/978-1-4939-2343-4_7

Aoki-Kinoshita KF (2015) Analyzing glycan structure synthesis with the glycan pathway predictor (GPP) tool. Methods Mol Biol. doi: 10.1007/978-1-4939-2343-4_10

Armougom F, Moretti S, Poirot O, et al (2006) Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. doi: 10.1093/nar/gkl092

Armstrong DR, Berrisford JM, Conroy MJ, et al (2020) PDBe: Improved findability of macromolecular structure data in the PDB. Nucleic Acids Res 48:D335–D343. doi: 10.1093/nar/gkz990

Attwood TK, Coletta A, Muirhead G, et al (2012) The PRINTS database: A fine-grained protein sequence annotation and analysis resource-its status in 2012. Database 2012:. doi: 10.1093/database/bas019

Bao B, Kellman BP, Chiang AWT, et al (2020) Correcting for sparsity and non-independence in glycomic data through a systems biology framework. FASEB J. doi: 10.1096/fasebj.2020.34.s1.03742

Bateman A (2019) UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515. doi: 10.1093/nar/gky1049

Bauters L, Naalden D, Gheysen G (2017) The distribution of lectins across the phylum nematoda: A genome-wide search. Int J Mol Sci 18:. doi: 10.3390/ijms18010091

Baycin-Hizal D, Tian Y, Akan I, et al (2011a) GlycoFish: A database of zebrafish N -linked glycoproteins identified using SPEG method coupled with LC/MS. Anal Chem. doi: 10.1021/ac200726q

Baycin-Hizal D, Tian Y, Akan I, et al (2011b) GlycoFly: A database of Drosophila N-linked Glycoproteins identified using SPEG-MS techniques. J Proteome Res. doi: 10.1021/pr200004t

Beck A, Goetsch L, Dumontet C, Corvaïa N (2017) Strategies and challenges for the next generation of antibody-drug conjugates. Nat. Rev. Drug Discov. 16:315–337

Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2012) The protein data bank at 40: Reflecting on the past to prepare for the future. In: Structure. Structure, pp 391–396

BHENDE YM, DESHPANDE CK, BHATIA HM, et al (1951) A “new” blood group character related to the ABO system. Lancet

Birch J, Van Calsteren MR, Pérez S, Svensson B (2019) The exopolysaccharide properties and structures database: EPS-DB. Application to bacterial exopolysaccharides. Carbohydr Polym 205:565–570. doi: 10.1016/j.carbpol.2018.10.063

Böhm M, Bohne-Lang A, Frank M, et al (2019) Glycosciences.db: An annotated data collection linking glycomics and proteomics data (2018 update). Nucleic Acids Res 47:D1195–D1201. doi: 10.1093/nar/gky994

Bohne-Lang A, Lang E, Förster T, Von der Lieth CW (2001) LINUCS: LInear Notation for Unique description of Carbohydrate Sequences. Carbohydr Res 336:1–11. doi: 10.1016/S0008-6215(01)00230-0

Bohne-Lang A, Von der Lieth CW (2005) GlyProt: In silico glycosylation of proteins. Nucleic Acids Res. doi: 10.1093/nar/gki385

Boyd WC, Almodóvar LR, Boyd LG (1966) Agglutinins in Marine Algae for Human Erythrocytes. Transfusion. doi: 10.1111/j.1537-2995.1966.tb04699.x

Brinda KV, Mitra N, Surolia A, Vishveshwara S (2004) Determinants of quaternary association in legume lectins. Protein Sci 13:1735–1749. doi: 10.1110/ps.04651004

Brown RG, Marchesi JR, Lee YS, et al (2018) Vaginal dysbiosis increases risk of preterm fetal membrane rupture, neonatal sepsis and is exacerbated by erythromycin. BMC Med 16:. doi: 10.1186/s12916-017-0999-x

Burley SK, Berman HM, Christie C, et al (2018) RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. Protein

Sci 27:316–330. doi: 10.1002/pro.3331

Burley SK, Berman HM, Kleywegt GJ, et al (2017) Protein Data Bank (PDB): The single global macromolecular structure archive. In: Methods in Molecular Biology. Humana Press Inc., pp 627–641

Campbell MP, Aoki-Kinoshita KF, Lisacek F, et al (2017) Glycoinformatics. doi: 10.1101/GLYCOBIOLOGY.3E.052

Campbell MP, Packer NH (2016) UniCarbKB: New database features for integrating glycan structure abundance, compositional glycoproteomics data, and disease associations. Biochim Biophys Acta - Gen Subj. doi: 10.1016/j.bbagen.2016.02.016

Cao Y, Park SJ, Mehta AY, et al (2020) GlyMDB: Glycan Microarray Database and analysis toolset. Bioinformatics 36:2438–2442. doi: 10.1093/bioinformatics/btz934

CASP14 (2021) Critical assessment of methods of protein structure prediction (CASP)—Round XIV. https://predictioncenter.org/casp14/index.cgi. Accessed 13 Dec 2020

Caspi R, Billington R, Keseler IM, et al (2020) The MetaCyc database of metabolic pathways and enzymes-a 2019 update. Nucleic Acids Res 48:D455–D453. doi: 10.1093/nar/gkz862

Chandonia JM, Fox NK, Brenner SE (2019) SCOPe: Classification of large macromolecular structures in the structural classification of proteins - Extended database. Nucleic Acids Res 47:D475–D481. doi: 10.1093/nar/gky1134

Chandra NR, Kumar N, Jeyakani J, et al (2006) Lectindb: A plant lectin database. Glycobiology 16:938–946. doi: 10.1093/glycob/cwl012

Cheng H, Schaeffer RD, Liao Y, et al (2014) ECOD: An Evolutionary Classification of Protein Domains. PLoS Comput Biol 10:. doi: 10.1371/journal.pcbi.1003926

Cheng K, Zhou Y, Neelamegham S (2017) DrawGlycan-SNFG: A robust tool to render glycans and glycopeptides with fragmentation information. Glycobiology. doi: 10.1093/glycob/cww115

Chothia C (1984) Principles that determine the structure of proteins. Annu. Rev. Biochem. 53:537–572

Clerc O, Deniaud M, Vallet SD, et al (2019a) MatrixDB: Integration of new data with a focus on glycosaminoglycan interactions. Nucleic Acids Res 47:D376–D381. doi: 10.1093/nar/gky1035

Clerc O, Mariethoz J, Rivet A, et al (2019b) A pipeline to translate glycosaminoglycan sequences into 3D models. Application to the exploration of glycosaminoglycan conformational space. Glycobiology. doi: 10.1093/glycob/cwy084

Connolly ML (1983) Solvent-accessible surfaces of proteins and nucleic acids. Science (80-.). 221:709–713

Copoiu L, Torres PHM, Ascher DB, et al (2020) ProCarbDB: A database of carbohydrate-binding proteins. Nucleic Acids Res 48:D368–D375. doi: 10.1093/nar/gkz860

Corolleur F, Level A, Matt M, Perez S (2020) Innovation potentials triggered by glycoscience research. Carbohydr. Polym. 233

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. Genome Res 14:1188–1190. doi: 10.1101/gr.849004

Cummings RD (2019) “Stuck on sugars – how carbohydrates regulate cell adhesion, recognition, and

signaling." Glycoconj J 36:241–257. doi: 10.1007/s10719-019-09876-0

Cummings RD, McEver RP (2017) C-Type Lectins. doi: 10.1101/GLYCOBIOLOGY.3E.034

Cunningham F, Achuthan P, Akanni W, et al (2019) Ensembl 2019. Nucleic Acids Res 47:D745–D751. doi: 10.1093/nar/gky1113

Damodaran D, Jeyakani J, Chauhan A, et al (2008) CancerLectinDB: A database of lectins relevant to cancer. Glycoconj J 25:191–198. doi: 10.1007/s10719-007-9085-5

Dang L, Van Damme EJM (2016) Genome-wide identification and domain organization of lectin domains in cucumber. Plant Physiol Biochem 108:165–176. doi: 10.1016/j.plaphy.2016.07.009

Das R, Baker D (2008) Macromolecular modeling with Rosetta. Annu. Rev. Biochem.

Davide Alocci, Frederique Lisacek SP (2019) A Traveler's Guide to Complex Carbohydrates in the Cyber Space. http://www.glycopedia.eu/e-chapters/a-traveler-s-guide-to-complex-carbohydrates-in-the-cyber-space/article/glyco-cyber-space. Accessed 13 Dec 2020

Dawson NL, Sillitoe I, Lees JG, et al (2017) CATH-Gene3D: Generation of the resource and its use in obtaining structural and functional annotations for protein sequences. In: Methods in Molecular Biology. Humana Press Inc., pp 79–110

Dhifli W, Diallo AB (2016) ProtNN: Fast and accurate protein 3D-structure classification in structural and topological space. BioData Min 9:. doi: 10.1186/s13040-016-0108-2

Di Tommaso P, Moretti S, Xenarios I, et al (2011) T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. doi: 10.1093/nar/gkr245

Drickamer K (1989) Multiple subfamilies of carbohydrate recognition domains in animal lectins. Ciba Found Symp 145:. doi: 10.1002/9780470513828.ch4

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics. doi: 10.1093/bioinformatics/btq461

Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. doi: 10.1093/nar/gkh340

El-Gebali S, Mistry J, Bateman A, et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. doi: 10.1093/nar/gky995

El-Gebali S, Richardson L, Finn R (2018) Pfam Database: Creating Protein Families. doi: 10.6019/TOL.PFAM_FAMS-T.2018.00001.1

Engelsen SB, Hansen PI, Pérez S (2014) POLYS 2.0: An open source software package for building three-dimensional structures of polysaccharides. Biopolymers. doi: 10.1002/bip.22449

Escobar-Zepeda A, De León AVP, Sanchez-Flores A (2015) The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. Front. Genet. 6

Fu L, Niu B, Zhu Z, et al (2012) CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. doi: 10.1093/bioinformatics/bts565

Fujimoto Z, Tateno H, Hirabayashi J (2014) Lectin structures: Classification based on the 3-D structures. Methods Mol Biol 1200:579–606. doi: 10.1007/978-1-4939-1292-6_46

Fujita A, Aoki NP, Shinmachi D, et al (2020) The international glycan repository GlyTouCan version 3.0. Nucleic Acids Res. doi: 10.1093/nar/gkaa947

Gabius HJ (1997) Animal lectins. Eur. J. Biochem. 243:543–576

Gelly JC, Joseph AP, Srinivasan N, De Brevern AG (2011) IPBA: A tool for protein structure comparison using sequence alignment strategies. Nucleic Acids Res 39:. doi: 10.1093/nar/gkr333

GoldBio 15 Great biological discoveries that revolutionized life science | GoldBio. https://www.goldbio.com/articles/article/15-Great-biological-discoveries-that-revolutionized-life-science. Accessed 13 Dec 2020

Gow N, Latge J, Munro C (2017) The Fungal Cell Wall: Structure, Biosynthesis, and Function. In: The Fungal Kingdom. American Society of Microbiology, pp 267–292

Gray CJ, Migas LG, Barran PE, et al (2019) Advancing Solutions to the Carbohydrate Sequencing Challenge. J. Am. Chem. Soc.

Grigoriev I V., Nikitin R, Haridas S, et al (2014) MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Res 42:. doi: 10.1093/nar/gkt1183

Gupta R, Birch H, Rapacki K, et al (1999) O-GLYCBASE version 4.0: A revised database of O-glycosylated proteins. Nucleic Acids Res.

Haeussler M, Zweig AS, Tyner C, et al (2019) The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 47:D853–D858. doi: 10.1093/nar/gky1095

Hampton RY, Holz RW, Goldstein IJ (1980) Phospholipid, glycolipid, and ion dependencies of concanavalin A- and Ricinus communis agglutinin I-induced agglutination of lipid vesicles. J Biol Chem

Hanson RM, Lu XJ (2017) DSSR-enhanced visualization of nucleic acid structures in Jmol. Nucleic Acids Res 45:W528–W533. doi: 10.1093/nar/gkx365

Hart GW, Copeland RJ (2010) Glycomics hits the big time. Cell 143:672–676. doi: 10.1016/j.cell.2010.11.008

Hashimoto K, Goto S, Kawano S, et al (2006) KEGG as a glycome informatics resource. Glycobiology 16

Hayes CA, Karlsson NG, Struwe WB, et al (2011) UniCarb-DB: A database resource for glycomic discovery. Bioinformatics. doi: 10.1093/bioinformatics/btr137

Heger A, Holm L (2000) Rapid automatic detection and alignment of repeats in protein sequences. Proteins Struct Funct Genet 41:224–237. doi: 10.1002/1097-0134(20001101)41:2<224::AID-PROT70>3.0.CO;2-Z

Helbert W, Poulet L, Drouillard S, et al (2019) Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. Proc Natl Acad Sci U S A. doi: 10.1073/pnas.1815791116

Hirabayashi J (2004) Lectin-based structural glycomics: Glycoproteomics and glycan profiling. In: Glycoconjugate Journal. Glycoconj J, pp 35–40

Hirabayashi J, Arai R (2019) Lectin engineering: The possible and the actual. Interface Focus 9

Hirabayashi J, Tateno H, Shikanai T, et al (2015) The lectin frontier database (LfDB), and data

generation based on frontal affinity chromatography. Molecules 20:951–973

Holm L, Laakso LM (2016) Dali server update. Nucleic Acids Res. doi: 10.1093/nar/gkw357

Hosoda M, Takahashi Y, Shiota M, et al (2018) MCAW-DB: A glycan profile database capturing the ambiguity of glycan recognition patterns. Carbohydr Res 464:44–56. doi: 10.1016/j.carres.2018.05.003

Houzelstein D, Gonçalves IR, Fadden AJ, et al (2004) Phylogenetic analysis of the vertebrate galectin family. Mol Biol Evol 21:1177–1187. doi: 10.1093/molbev/msh082

Howe KL, Contreras-Moreira B, De Silva N, et al (2020) Ensembl Genomes 2020-enabling non-vertebrate genomic research. Nucleic Acids Res 48:D689–D695. doi: 10.1093/nar/gkz890

Hu D, Tateno H, Hirabayashi J (2015) Lectin engineering, a molecular evolutionary approach to expanding the lectin utilities. Molecules 20:7637–7656

Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. J Mol Graph 14:33–38. doi: 10.1016/0263-7855(96)00018-5

Imberty A, Mitchell EP, Wimmerová M (2005) Structural basis of high-affinity glycan recognition by bacterial and fungal lectins. Curr. Opin. Struct. Biol. 15:525–534

Jassal B, Matthews L, Viteri G, et al (2020) The reactome pathway knowledgebase. Nucleic Acids Res 48:D498–D503. doi: 10.1093/nar/gkz1031

Jendele L, Krivak R, Skoda P, et al (2019) PrankWeb: a web server for ligand binding site prediction and visualization. Nucleic Acids Res 47:W345–W349. doi: 10.1093/nar/gkz424

Jeske L, Placzek S, Schomburg I, et al (2019) BRENDA in 2019: A European ELIXIR core data resource. Nucleic Acids Res 47:D542–D549. doi: 10.1093/nar/gky1048

Jiménez J, Doerr S, Martínez-Rosell G, et al (2017) DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics 33:3036–3042. doi: 10.1093/bioinformatics/btx350

Johnson GT, Hertig S (2014) A guide to the visual analysis and communication of biomolecular structural data. Nat. Rev. Mol. Cell Biol. 15:690–698

Johnson M, Zaretskaya I, Raytselis Y, et al (2008) NCBI BLAST: a better web interface. Nucleic Acids Res. doi: 10.1093/nar/gkn201

Kaltner H, Gabius HJ (2001) Animal lectins: From initial description to elaborated structural and functional classification. In: Advances in Experimental Medicine and Biology. Kluwer Academic/Plenum Publishers, pp 79–94

Kanehisa M, Furumichi M, Tanabe M, et al (2017) KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361. doi: 10.1093/nar/gkw1092

Karp PD, Billington R, Caspi R, et al (2018) The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform 20:1085–1093. doi: 10.1093/bib/bbx085

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. doi: 10.1093/molbev/mst010

Kawabata T (2003) MATRAS: A program for protein 3D structure comparison. Nucleic Acids Res 31:3367–3369. doi: 10.1093/nar/gkg581

Kendrew JC, Bodo G, Dintzis HM, et al (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181:662–666. doi: 10.1038/181662a0

Kilpatrick DC (2002) Animal lectins: A historical introduction and overview. Biochim. Biophys. Acta - Gen. Subj. 1572:187–197

Kinjo AR, Bekker GJ, Suzuki H, et al (2017) Protein Data Bank Japan (PDBj): Updated user interfaces, resource description framework, analysis tools for large structures. Nucleic Acids Res. doi: 10.1093/nar/gkw962

Koromyslova AD, Leuthold MM, Bowler MW, Hansman GS (2015) The sweet quartet: Binding of fucose to the norovirus capsid. Virology 483:203–208. doi: 10.1016/j.virol.2015.04.006

Kumar D, Mittal Y (2011) AnimalLectinDb: An integrated animal lectin database. Bioinformation 6:134–136. doi: 10.6026/97320630006134

Kumar D, Mittal Y (2012) BacterialLectinDb: An integrated bacterial lectin database. Bioinformation 8:281–283. doi: 10.6026/97320630008281

Kumar S, Lütteke T, Schwartz-albiez R (2012) GlycoCD: A repository for carbohydrate-related CD antigens. Bioinformatics 28:2553–2555. doi: 10.1093/bioinformatics/bts481

Kunduru BR, Nair SA, Rathinavelan T (2016) EK3D: An E. coliK antigen 3-dimensional structure database. Nucleic Acids Res 44:D675–D681. doi: 10.1093/nar/gkv1313

Lassmann T, Frings O, Sonnhammer ELL (2009) Kalign2: High-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res. doi: 10.1093/nar/gkn1006

Lawson CL, Berman HM, Chiu W (2020) Evolving data standards for cryo-EM structures. Struct Dyn 7:. doi: 10.1063/1.5138589

Le Boulch M, Déhais P, Combes S, Pascal G (2019) The MACADAM database: A MetAboliC pAthways DAtabase for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups. Database 2019:. doi: 10.1093/database/baz049

Léonard S, Joseph AP, Srinivasan N, et al (2014) MulPBA: An efficient multiple protein structure alignment method based on a structural alphabet. J Biomol Struct Dyn 32:661–668. doi: 10.1080/07391102.2013.787026

Lesk AM, Hardman KD (1982) Computer-generated schematic diagrams of protein structures. Science (80-) 216:539–540. doi: 10.1126/science.7071602

Letunic I, Bork P (2018) 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 46:D493–D496. doi: 10.1093/nar/gkx922

Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261:552–558. doi: 10.1038/261552a0

Li X, Xu Z, Hong X, et al (2020) Databases and bioinformatic tools for glycobiology and glycoproteomics. Int. J. Mol. Sci. 21:1–29

Liao WR, Lin JY, Shieh WY, et al (2003) Antibiotic activity of lectins from marine algae against marine vibrios. J Ind Microbiol Biotechnol 30:433–439. doi: 10.1007/s10295-003-0068-7

Lis H, Sharon N (1986) Lectins as molecules and as tools. Annu Rev Biochem VOL. 55:35–67. doi: 10.1146/annurev.bi.55.070186.000343

Lis H, Sharon N (1998) Lectins: Carbohydrate-specific proteins that mediate cellular recognition. Chem Rev 98:637–674. doi: 10.1021/cr940413g

Liu G, Puri A, Neelamegham S (2013) Glycosylation Network Analysis Toolbox: A MATLAB-based environment for systems glycobiology. Bioinformatics. doi: 10.1093/bioinformatics/bts703

Lo WC, Lee CY, Lee CC, Lyu PC (2009) iSARST: An integrated SARST web server for rapid protein structural similarity searches. Nucleic Acids Res. doi: 10.1093/nar/gkp291

Loix C, Huybrechts M, Vangronsveld J, et al (2018) Corrigendum: Reciprocal interactions between cadmium-induced cell wall responses and oxidative stress in plants (Front. Plant Sci. 8, 1867, 10.3389/fpls.2017.01867). Front. Plant Sci. 9

Lombard V, Golaconda Ramulu H, Drula E, et al (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:. doi: 10.1093/nar/gkt1178

Louca S, Parfrey LW, Doebeli M (2016) Decoupling function and taxonomy in the global ocean microbiome. Science (80-) 353:1272–1277. doi: 10.1126/science.aaf4507

Lu S, Wang J, Chitsaz F, et al (2020) CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Res 48:D265–D268. doi: 10.1093/nar/gkz991

Lütteke T, von der Lieth CW (2004) pdb-care (PDB CArbohydrate REsidue check): A program to support annotation of complex carbohydrate structures in PDB files. BMC Bioinformatics. doi: 10.1186/1471-2105-5-69

Lv Z, Tek A, Da Silva F, et al (2013) Game On, Science - How Video Game Technology May Help Biologists Tackle Visualization Challenges. PLoS One 8:. doi: 10.1371/journal.pone.0057990

MacIntyre DA, Chandiramani M, Lee YS, et al (2015) The vaginal microbiome during pregnancy and the postpartum period in a European population. Sci Rep 5:. doi: 10.1038/srep08988

Madeira F, Park YM, Lee J, et al (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636–W641. doi: 10.1093/nar/gkz268

Mariethoz J, Alocci D, Gastaldello A, et al (2018) Glycomics@ExPASy: Bridging the gap. Mol Cell Proteomics 17:2164–2176. doi: 10.1074/mcp.RA118.000799

Mariethoz J, Khatib K, Alocci D, et al (2016) SugarBindDB, a resource of glycan-mediated host-pathogen interactions. Nucleic Acids Res 44:D1243–D1250. doi: 10.1093/nar/gkv1247

Martens M, Ammar A, Riutta A, et al (2020) WikiPathways: connecting communities. Nucleic Acids Res. doi: 10.1093/nar/gkaa1024

Martinez X, Krone M, Alharbi N, et al (2019) Molecular Graphics: Bridging Structural Biologists and Computer Scientists. Structure 27:1617–1623

McDonald AG, Boyce S, Tipton KF (2009) ExplorEnz: The primary source of the IUBMB enzyme list. Nucleic Acids Res 37:. doi: 10.1093/nar/gkn582

McDonald AG, Tipton KF, Stroop CJM, Davey GP (2010) GlycoForm and Glycologue: Two software applications for the rapid construction and display of N-glycans from mammalian sources. BMC Res Notes. doi: 10.1186/1756-0500-3-173

McNaught AD (1997) Nomenclature of carbohydrates. Adv. Carbohydr. Chem. Biochem. 52:47–177

Mehta AY, Cummings RD (2020) GlycoGlyph: A glycan visualizing, drawing and naming application. Bioinformatics. doi: 10.1093/bioinformatics/btaa190

Meiers J, Siebs E, Zahorska E, Titz A (2019) Lectin antagonists in infection, immunity, and inflammation. Curr. Opin. Chem. Biol. 53:51–67

Mende DR, Letunic I, Maistrenko OM, et al (2020) ProGenomes2: An improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. Nucleic Acids Res 48:D621–D625. doi: 10.1093/nar/gkz1002

Mi H, Muruganujan A, Huang X, et al (2019) Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc 14:703–721. doi: 10.1038/s41596-019-0128-8

Mitchell AL, Attwood TK, Babbitt PC, et al (2019) InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47:D351–D360. doi: 10.1093/nar/gky1100

Mitra A, MacIntyre DA, Ntritsos G, et al (2020) The vaginal microbiota associates with the regression of untreated cervical intraepithelial neoplasia 2 lesions. Nat Commun 11:. doi: 10.1038/s41467-020-15856-y

Moore AD, Björklund ÅK, Ekman D, et al (2008) Arrangements in the modular evolution of proteins. Trends Biochem. Sci. 33:444–451

Mylemans B, Laier I, Kamata K, et al (2020) Structural plasticity of a designer protein sheds light on β-propeller protein evolution. FEBS J. doi: 10.1111/febs.15347

Nacher JC, Hayashida M, Akutsu T (2010) The role of internal duplication in the evolution of multi-domain proteins. BioSystems 101:127–135. doi: 10.1016/j.biosystems.2010.05.005

Nagai-Okatani C, Aoki-Kinoshita KF, Kakuda S, et al (2019) LM-Glycomeatlas ver. 1.0: A novel visualization tool for lectin microarray-based glycomic profiles of mouse tissue sections. Molecules 24:. doi: 10.3390/molecules24162962

Narimatsu Y, Joshi HJ, Nason R, et al (2019) An Atlas of Human Glycosylation Pathways Enables Display of the Human Glycome by Gene Engineered Cells. Mol Cell. doi: 10.1016/j.molcel.2019.05.017

Neelamegham S, Aoki-Kinoshita K, Bolton E, et al (2019) Updates to the Symbol Nomenclature for Glycans guidelines. Glycobiology 29:620–624. doi: 10.1093/glycob/cwz045

Nguyen NH, Song Z, Bates ST, et al (2016) FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecol 20:241–248. doi: 10.1016/j.funeco.2015.06.006

Nikolskaya AN, Arighi CN, Huang H, et al (2006) PIRSF Family Classification System for Protein Functional and Evolutionary Analysis. Evol Bioinforma. doi: 10.1177/117693430600200033

Nordberg H, Cantor M, Dusheyko S, et al (2014) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res 42:. doi: 10.1093/nar/gkt1069

O'Leary NA, Wright MW, Brister JR, et al (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. doi: 10.1093/nar/gkv1189

Okuda S, Nakao H, Kawasaki T (2015) Glycoepitope: Database for carbohydrate antigen and antibody. In: Glycoscience: Biology and Medicine

Olson SA (2002) EMBOSS opens up sequence analysis. European Molecular Biology Open Software

Suite. Brief Bioinform. doi: 10.1093/bib/3.1.87

Paladin L, Hirsh L, Piovesan D, et al (2017) RepeatsDB 2.0: Improved annotation, classification, search and visualization of repeat protein structures. Nucleic Acids Res. doi: 10.1093/nar/gkw1136

Pandurangan AP, Stahlhacke J, Oates ME, et al (2019) The SUPERFAMILY 2.0 database: A significant proteome update and a new webserver. Nucleic Acids Res 47:D490–D494. doi: 10.1093/nar/gky1130

Pedruzzi I, Rivoire C, Auchincloss AH, et al (2015) HAMAP in 2015: Updates to the protein family classification and annotation system. Nucleic Acids Res 43:D1064–D1070. doi: 10.1093/nar/gku1002

Pei J, Grishin N V. (2007) PROMALS: Towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics. doi: 10.1093/bioinformatics/btm017

Pei J, Kim BH, Grishin N V. (2008) PROMALS3D: A tool for multiple protein sequence and structure alignments. Nucleic Acids Res. doi: 10.1093/nar/gkn072

Pellegrini M (2015) Tandem repeats in proteins: Prediction algorithms and biological role. Front. Bioeng. Biotechnol. 3

Pennisi E (2001) The human genome. Science (80-.). 291:1177–1180

Pérez S, Sarkar A, Rivet A, et al (2015a) Glyco3d: A portal for structural glycosciences. Methods Mol Biol 1273:241–258. doi: 10.1007/978-1-4939-2343-4_18

Pérez S, Tubiana T, Imberty A, Baaden M (2015b) Three-dimensional representations of complex carbohydrates and polysaccharides - SweetUnityMol: A video game-based computer graphic software. Glycobiology. doi: 10.1093/glycob/cwu133

Perutz MF, Rossmann MG, Cullis AF, et al (1960) Structure of Hæmoglobin: A three-dimensional fourier synthesis at 5.5-. resolution, obtained by X-ray analysis. Nature 185:416–422

Peumans WJ, Van Damme EJ (1995) Lectins as plant defense proteins. Plant Physiol. 109:347–352

Peumans WJ, Van Damme EJM, Barre A, Rougé P (2001) Classification of plant lectins in families of structurally and evolutionary related proteins. Adv Exp Med Biol 491:27–54. doi: 10.1007/978-1-4615-1267-7_3

Poleksic A (2009) Algorithms for optimal protein structure alignment. Bioinformatics 25:2751–2756. doi: 10.1093/bioinformatics/btp530

Polito L, Bortolotti M, Battelli MG, et al (2019) Ricin: An ancient story for a timeless plant toxin. Toxins (Basel).

Potter SC, Luciani A, Eddy SR, et al (2018) HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204. doi: 10.1093/nar/gky448

Raman R, Venkataraman M, Ramakrishnan S, et al (2006) Advancing glycomics: Implementation strategies at the consortium for functional glycomics. Glycobiology 16

Richardson DC, Richardson JS (1992) The kinemage: A tool for scientific communication. Protein Sci 1:3–9. doi: 10.1002/pro.5560010102

Richardson JS (1977) β-Sheet topology and the relatedness of proteins. Nature 268:495–500. doi: 10.1038/268495a0

Roberts EC, Legrand C, Steinke M, Wootton EC (2011) Mechanisms underlying chemical interactions between predatory planktonic protists and their prey. J Plankton Res. doi: 10.1093/plankt/fbr005

Rodchenkov I, Babur O, Luna A, et al (2020) Pathway Commons 2019 Update: Integration, analysis and exploration of pathway data. Nucleic Acids Res 48:D489–D497. doi: 10.1093/nar/gkz946

Rose AS, Hildebrand PW (2015) NGL Viewer: A web application for molecular visualization. Nucleic Acids Res 43:W576–W579. doi: 10.1093/nar/gkv402

Russell RB, Saqi MAS, Sayle RA, et al (1997) Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. J Mol Biol 269:423–439. doi: 10.1006/jmbi.1997.1019

Salentin S, Schreiber S, Haupt VJ, et al (2015) PLIP: Fully automated protein-ligand interaction profiler. Nucleic Acids Res 43:W443–W447. doi: 10.1093/nar/gkv315

Saqi MAS, Sayle R (1994) Pdbmotif - a tool for the automatic identification and display of motifs in protein structures. Bioinformatics 10:545–546. doi: 10.1093/bioinformatics/10.5.545

Sarkar A, Pérez S (2012) PolySac3DB: an annotated data base of 3 dimensional structures of polysaccharides. BMC Bioinformatics 13:. doi: 10.1186/1471-2105-13-302

Sattin S, Bernardi A (2016) Glycoconjugates and Glycomimetics as Microbial Anti-Adhesives. Trends Biotechnol. 34:483–495

Sayers EW, Cavanaugh M, Clark K, et al (2020) GenBank. Nucleic Acids Res 48:D84–D86. doi: 10.1093/nar/gkz956

Scherbinina SI, Toukach P V. (2020) Three-dimensional structures of carbohydrates and where to find them. Int. J. Mol. Sci. 21:1–46

Schjoldager KT, Narimatsu Y, Joshi HJ, Clausen H (2020) Global view of human protein glycosylation pathways and functions. Nat Rev Mol Cell Biol 21:. doi: 10.1038/s41580-020-00294-x

Sehnal D, Deshpande M, Vařeková RS, et al (2017) LiteMol suite: Interactive web-based visualization of large-scale macromolecular structure data. Nat. Methods 14:1121–1122

Sharon N (1993) Lectin-carbohydrate complexes of plants and animals: an atomic view. Trends Biochem Sci 18:221–226. doi: 10.1016/0968-0004(93)90193-Q

Sharon N (2008) Lectins: Past, present and future. In: Biochemical Society Transactions. Biochem Soc Trans, pp 1457–1460

Sharon N (1987) Bacterial lectins, cell-cell recognition and infectious disease. FEBS Lett 217:145–157. doi: 10.1016/0014-5793(87)80654-3

Shridhar S, Chattopadhyay D, Yadav G (2009) PLecDom: A program for identification and analysis of plant lectin domains. Nucleic Acids Res 37:. doi: 10.1093/nar/gkp409

Sievers F, Higgins DG (2014) Clustal Omega. Curr Protoc Bioinforma. doi: 10.1002/0471250953.bi0313s48

Sigrist CJA, Cerutti L, De Castro E, et al (2009) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38:. doi: 10.1093/nar/gkp885

Sillitoe I, Dawson N, Lewis TE, et al (2019) CATH: Expanding the horizons of structure-based

functional annotations for genome sequences. Nucleic Acids Res 47:D280–D284. doi: 10.1093/nar/gky1097

Singh RS, Thakur SR, Bansal P (2015) Algal lectins as promising biomolecules for biomedical research. Crit. Rev. Microbiol. 41:77–88

Siva Shanmugam NR, Jino Blessy J, Veluraja K, Michael Gromiha M (2020) ProCaff: Protein-carbohydrate complex binding affinity database. Bioinformatics 36:3615–3617. doi: 10.1093/bioinformatics/btaa141

Sternberg MJE, Thornton JM (1976) On the conformation of proteins: The handedness of the β-strand-α-helix-β-strand unit. J Mol Biol 105:367–382. doi: 10.1016/0022-2836(76)90099-1

Subramanian AR, Kaufmann M, Morgenstern B (2008) DIALIGN-TX: Greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol. doi: 10.1186/1748-7188-3-6

Sýkorová P, Novotná J, Demo G, et al (2020) Characterization of novel lectins from Burkholderia pseudomallei and Chromobacterium violaceum with seven-bladed β-propeller fold. Int J Biol Macromol 152:1113–1124. doi: 10.1016/j.ijbiomac.2019.10.200

Taherzadeh G, Zhou Y, Liew AWC, Yang Y (2016) Sequence-Based Prediction of Protein-Carbohydrate Binding Sites Using Support Vector Machines. J Chem Inf Model 56:2115–2122. doi: 10.1021/acs.jcim.6b00320

Tanaka K, Aoki-Kinoshita KF, Kotera M, et al (2014) WURCS: The Web3 unique representation of carbohydrate structures. J Chem Inf Model 54:1558–1566. doi: 10.1021/ci400571e

Taylor ME, Drickamer K, Schnaar RL, et al (2017) Discovery and Classification of Glycan-Binding Proteins. doi: 10.1101/GLYCOBIOLOGY.3E.028

Terrapon N, Lombard V, Drula E, et al (2017) The CAZy Database/the Carbohydrate-Active Enzyme (CAZy) Database: Principles and Usage Guidelines. In: A Practical Guide to Using Glycomics Databases. Springer Japan, pp 117–131

Terrapon N, Lombard V, Drula É, et al (2018) PULDB: The expanded database of Polysaccharide Utilization Loci. Nucleic Acids Res. doi: 10.1093/nar/gkx1022

Tiemeyer M, Aoki K, Paulson J, et al (2017) GlyTouCan: An accessible glycan structure repository. Glycobiology. doi: 10.1093/glycob/cwx066

Toukach P V., Egorova KS (2016) Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. Nucleic Acids Res. doi: 10.1093/nar/gkv840

Tsaneva M, Van Damme EJM (2020) 130 years of Plant Lectin Research. Glycoconj. J. 37:533–551

Tsuchiya S, Yamada I, Aoki-Kinoshita KF (2019) GlycanFormatConverter: A conversion tool for translating the complexities of glycans. Bioinformatics 35:2434–2440. doi: 10.1093/bioinformatics/bty990

Ulrich EL, Akutsu H, Doreleijers JF, et al (2008) BioMagResBank. Nucleic Acids Res. doi: 10.1093/nar/gkm957

Van Damme EJM (2014) History of plant lectin research. Methods Mol. Biol. 1200:3–13

Van Holle S, De Schutter K, Eggermont L, et al (2017) Comparative study of lectin domains in model species: New insights into evolutionary dynamics. Int J Mol Sci 18:. doi: 10.3390/ijms18061136

Varki A (2017) Biological roles of glycans. Glycobiology 27:3–49. doi: 10.1093/glycob/cww086

Varrot A, Basheer SM, Imberty A (2013) Fungal lectins: Structure, function and potential applications. Curr. Opin. Struct. Biol. 23:678–685

Vasta GR, Ahmed H, Bianchet MA, et al (2012) Diversity in recognition of glycans by F-type lectins and galectins: molecular, structural, and biophysical aspects. Ann. N. Y. Acad. Sci. 1253:E14–E26

Vasta GR, Mario Amzel L, Bianchet MA, et al (2017) F-Type Lectins: A highly diversified family of fucose-binding proteins with a unique sequence motif and structural fold, involved in self/non-self-recognition. Front. Immunol. 8

Venkataraman M, Sasisekharan R, Raman R (2015) Glycan array data management at consortium for functional glycomics. Methods Mol Biol 1273:181–190. doi: 10.1007/978-1-4939-2343-4_13

Wan S, Zou Q (2017) HAlign-II: Efficient ultra-large multiple sequence alignment and phylogenetic tree reconstruction with distributed and parallel computing. Algorithms Mol Biol. doi: 10.1186/s13015-017-0116-x

Watanabe Y, Aoki-Kinoshita KF, Ishihama Y, Okuda S (2020) GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. Nucleic Acids Res. doi: 10.1093/nar/gkaa1012

Whitfield C, Szymanski CM, Aebi M (2017) Eubacteria. doi: 10.1101/GLYCOBIOLOGY.3E.021

Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al (2016) Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:. doi: 10.1038/sdata.2016.18

Wu AM, Lisowska E, Duk M, Yang Z (2009) Lectins as tools in glycoconjugate research. Glycoconj. J. 26:899–913

Xu K, Liu N, Xu J, et al (2020) VRmol: an Integrative Web-Based Virtual Reality System to Explore Macromolecular Structure. Bioinformatics. doi: 10.1093/bioinformatics/btaa696

Yadid I, Tawfik DS (2007) Reconstruction of Functional β-Propeller Lectins via Homo-oligomeric Assembly of Shorter Fragments. J Mol Biol 365:10–17. doi: 10.1016/j.jmb.2006.09.055

Yadid I, Tawfik DS (2011) Functional -propeller lectins by tandem duplications of repetitive units. Protein Eng Des Sel 24:185–195. doi: 10.1093/protein/gzq053

Yamada I, Shiota M, Shinmachi D, et al (2020) The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. Nat. Methods

York WS, Mazumder R, Ranzinger R, et al (2020) GlyGen: Computational and Informatics Resources for Glycoscience. Glycobiology

Zhao H, Taherzadeh G, Zhou Y, Yang Y (2018a) Computational Prediction of Carbohydrate-Binding Proteins and Binding Sites. Curr Protoc Protein Sci 94:. doi: 10.1002/cpps.75

Zhao H, Yang Y, Von Itzstein M, Zhou Y (2014) Carbohydrate-binding protein identification by coupling structural similarity searching with binding affinity prediction. J Comput Chem 35:2177–2183. doi: 10.1002/jcc.23730

Zhao S, Walsh I, Abrahams JL, et al (2018b) GlycoStore: a database of retention properties for glycan analysis. Bioinformatics. doi: 10.1093/bioinformatics/bty319

Zhou H, Zhou Y (2005) SPEM: Improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics. doi: 10.1093/bioinformatics/bti582

11 Annex

I have been associated as coauthor in three articles that are not in direct line of this thesis and their abstract are listed in this annex.

Article IX: Le Mercier, P., Mariethoz, J., Lascano-Maillard, J., Bonnardel, F., Imberty, A., Ricard-Blum, S., & Lisacek, F. (2019). A bioinformatics view of glycan–virus interactions. *Viruses*, 11(4), 374.

Abstract. Evidence of the mediation of glycan molecules in the interaction between viruses and their hosts is accumulating and is now partially reflected in several online databases. Bioinformatics provides convenient and efficient means of searching, visualizing, comparing, and sometimes predicting, interactions in numerous and diverse molecular biology applications related to the -omics fields. As viromics is gaining momentum, bioinformatics support is increasingly needed. We propose a survey of the current resources for searching, visualizing, comparing, and possibly predicting host-virus interactions that integrate the presence and role of glycans. To the best of our knowledge, we have mapped the specialized and general-purpose databases with the appropriate focus. With an illustration of their potential usage, we also discuss the strong and weak points of the current bioinformatics landscape in the context of understanding viral infection and the immune response to it.

Article X: Julien Mariethoz, François Bonnardel, Anne Imberty, Frédérique Lisacek. (2021) Glycoinformatics for virology. In Press.

Abstract. The 2020 COVID-19 pandemic has demonstrated the limitations of scientific insight into virology. Among many other features, the glycan shield covering spike proteins was revealed as a missing piece of the larger picture. It confirms that the study of host-virus interactions also depends on the knowledge of glycan-protein interactions. Progress in this area can be made through collecting and organising related data in databases as well as developing user-friendly search, visualising, comparative, and predictive tools. This chapter highlights some of these resources with various examples in an attempt to bring out glyco-virology knowledge.

Article XI: Pérez, S.; Bonnardel, F.; Lisacek, F.; Imberty, A.; Ricard Blum, S.; Makshakova, O. (2020) GAG-DB, the New Interface of the Three-Dimensional Landscape of Glycosaminoglycans. *Biomolecules*, 10, 1660.

Abstract. Glycosaminoglycans (GAGs) are complex linear polysaccharides. GAG-DB is a curated database that classifies the three-dimensional features of the six mammalian GAGs (chondroitin sulfate, dermatan sulfate, heparin, heparan sulfate, hyaluronan, and keratan sulfate) and their oligosaccharides complexed with proteins. The entries are structures of GAG and GAG-protein complexes determined by X-ray single-crystal diffraction methods, X-ray fiber diffractometry, solution NMR spectroscopy, and scattering data often associated with molecular modeling. We designed the database architecture and the navigation tools to query the database with the Protein Data Bank (PDB), UniProtKB, and GlyTouCan (universal glycan repository) identifiers. Special attention was devoted to the description of the bound glycan ligands using simple graphical representation and numerical format for cross-referencing to other databases in glycoscience and functional data. GAG-DB provides detailed information on GAGs, their bound protein ligands, and features their interactions using several open access applications. Binding covers interactions between monosaccharides and protein monosaccharide units and the evaluation of quaternary structure. GAG-DB is freely available.

Résumé

Les domaines de la bioinformatique ont pour objectif de démêler les connaissances dans les données biologiques. Lorsque la bioinformatique est appliquée aux glycanes et à la glycobiologie, elle est appelée glyco-informatique. Les nouvelles technologies permettent le séquençage massif des génomes de nouvelles espèces. Mais tous les génomes et protéines découverts ne sont que partiellement annotés d'une fonction biologique, récupérée par similarité à partir des organismes de référence.

La glycobiologie est consacré à l'étude des glycanes/glucides, composés d'un ou de plusieurs monosaccharides. Les lectines sont des protéines capables de se lier de manière réversible aux glycanes, et sans fonctions enzymatiques, sont des outils puissants pour la reconnaissance des glycanes dans les échantillons, et elles sont des cibles pour les composés thérapeutiques en raison de leur implication dans le cancer, l'immunologie et les infections. Cette thèse vise à développer de nouveaux outils in-silico pour l'étude des lectines. Elle a pour objectif de fournir, dans une nouvelle base de données en ligne, des informations sur les lectines pour les génomes d'autres organismes.

Pour fournir une classification des structures 3D des lectines et leur annotation dans les génomes, un portail web dédié a été développé, appelé UniLectin. Le module UniLectin3D fournit des structures 3D classées et stockées manuellement, ainsi que leurs glycanes en interaction. En raison de la difficulté d'identifier les lectines répétées en tandem dans les génomes, une méthode spécifique a été mise au point pour permettre la prédiction de ses lectines particulières, maintenant disponibles dans les modules PropLec et TrefLec. Enfin, le module LectomeXplore fournit des lectines prédites basées sur les 107 classes de UniLectin3D, dans les génomes disponibles du NCBI et d'UniProt. Cela a permis l'étude des lectomes de différents environnements par le biais de la collaboration décrite dans la dernière partie de la thèse.

Summary

Bioinformatics uses mathematical concepts and informatics tools to unravel the knowledge hidden in biological data. When bioinformatics is applied to glycans and glycobiology, it is called glyco-informatics. New technologies allow mass sequencing of new species genomes and of environmental samples metagenomes. But all newly discovered genomes and encoded proteins are only partially annotated with biological function assessed by similarities to reference organisms.

Glycobiology is the research field dedicated to the study of glycan/carbohydrate compounds, composed of one or multiple monosaccharides. Lectins are proteins able to bind reversibly to glycans, and without enzymatic functions. Lectins are powerful tools for the recognition of glycans in samples, and they are also targets for therapeutic compounds due to their involvement in cancer, immunology and infections.

This thesis aims to use bioinformatics for developing new *in silico* tools for the study of lectins. More specifically, it addresses the need for a new online database covering curated information on lectins for both reference organisms and newly sequenced genomes belonging to other organisms.

To provide a curated classification of lectin 3D structures and their annotation in genomes, a dedicated web portal called UniLectin, was developed and includes several modules. The UniLectin3D module provides manually curated and classified 3D structures together with their interacting glycans. Due to the difficulty of identifying tandem repeated lectins in genomes, a specific method has been developed for the prediction of those particular lectins, now available in the PropLec and TrefLec modules. Finally, the LectomeXplore module includes lectin predictions based on 107 classes defined on the basis of UniLectin3D content, and resulting from screening available sequences stored in the reference protein databases NCBI-nr and UniProt. This made the study of lectomes in different environments possible as collaborative work described in the last part of the thesis.