

# Development of Bioinformatics Resources for Glycan-related Pathway Information using Semantic Web Technologies

2023 年 8 月

LEE SUNMYOUNG

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	The major glycan types and functions . . . . .	3
1.1.1	<i>N</i> -glycans . . . . .	4
1.1.2	<i>O</i> -GalNAc glycans (Mucin-type <i>O</i> -glycans) . . . . .	4
1.1.3	Glycosphingolipids (GSL) . . . . .	6
1.1.4	Other types of <i>O</i> -glycans . . . . .	7
1.2	Microbial glycosylation . . . . .	8
1.2.1	<i>Escherichia coli</i> ( <i>E. coli</i> ) O-antigen . . . . .	10
1.2.2	Microbial glycans and drug resistance . . . . .	12
1.2.3	Mucosal glycans and microbiota . . . . .	20
1.3	Semantic Web Technologies . . . . .	22
1.3.1	Semantic Web . . . . .	22
1.3.2	Ontologies for Data sharing . . . . .	25
1.3.3	Heterogeneity and standardization of data . . . . .	26
1.3.4	BioPAX ontology for pathway data exchange . . . . .	27
1.4	Purpose . . . . .	29
<b>2</b>	<b>Methods</b>	<b>31</b>
2.1	Inspecting semantic data . . . . .	31
2.1.1	RDFication . . . . .	31
2.1.2	ShEx for Verification . . . . .	32

2.1.3	SPARQL Protocol and RDF Query Language . . . . .	32
2.2	RDFication of Microbial glycosylation . . . . .	33
2.2.1	Preparation of O-antigen resources . . . . .	33
2.2.2	Protégé and OLS (ontology lookup service) . . . . .	33
2.2.3	Applying BioPAX ontology and controlled vocabularies . . . . .	34
2.2.4	Serialization of data in the csv file format . . . . .	34
2.2.5	Loading the generated triples into the Virtuoso . . . . .	35
2.3	RDFication of Pathway Repository . . . . .	35
2.3.1	Taxonomy . . . . .	35
2.3.2	Protein information . . . . .	36
2.3.3	Web page . . . . .	36
2.3.4	Tables for list of pathways and resources . . . . .	36
2.3.5	Visualization of the pathway data . . . . .	37
<b>3</b>	<b>Results</b>	<b>39</b>
3.1	Microbial glycosylation . . . . .	39
3.1.1	<i>Escherichia coli</i> O-antigen . . . . .	40
3.1.2	<i>Bifidobacterium bifidum</i> . . . . .	42
3.1.3	<i>Bifidobacterium longum</i> . . . . .	47
3.1.4	<i>Campylobacter jejuni</i> . . . . .	49
3.1.5	<i>Cryptococcus neoformans</i> . . . . .	53
3.1.6	<i>Mycobacteroides abscessus</i> . . . . .	56
3.1.7	<i>Mycobacterium tuberculosis</i> . . . . .	59
3.2	Pathway repository . . . . .	63
3.2.1	Inspecting Ontologies . . . . .	63
3.2.2	Biosynthetic glycosylation pathway . . . . .	66
3.2.3	Glycan related protein pathway . . . . .	70

<b>4 Discussion</b>	<b>75</b>
4.1 Semantic Data of Microbial Glycosylation . . . . .	75
4.2 Semantic Data of Glycan-related Pathway Data . . . . .	79
4.3 Future work . . . . .	82
<b>5 Conclusion</b>	<b>85</b>
<b>A Appendices</b>	<b>87</b>
A.1 SPARQL query for protein pathway table . . . . .	87
A.2 Source codes . . . . .	88
<b>Bibliography</b>	<b>95</b>





# Abstract

A lot of databases have been developed with a huge amount of glycomics data to help us understand the function and impact of glycans on cellular activity. On the other hand, many efforts have been made to integrate disparate pathway data from a variety of biological domains, leading to the development of a new knowledgebase that helps a better understanding of biological processes. However, one of the many obstacles to data integration is the diversity of biological data types and an approach to representing pathway concepts between databases. To address this challenge, Semantic Web techniques including Resource Description Framework (RDF) and ontology development have been adopted by various study groups to standardize data in a computer-readable manner with data.

It has been well-established that glycosylated molecules and their glycosylation system in microorganisms such as bacteria, viruses, fungi, and protozoa influence their interaction with their environment, including host organisms. Thus, microbial glycosylation has posed as the key factor in understanding the mechanisms of inflammatory processes in disease or cancer, antibacterial resistance, and maintaining host health by microbiota. Importantly, tremendous information is still scattered in the literature and different database. A comprehensive understanding has not been achieved until the structure and roles of glycans in microbes are provided within the context showing how the glycans contribute to interactions with the host under the microenvironment. As a starting point, the well-organized data on microbial glycosylation provided by our collaborators is modified in formal format to be ready for data integration.

I demonstrated how Semantic Web techniques can be applied to the unformatted information of microbial glycosylation for semantic description. In addition, I created a repository to save pathway information in which different kinds of resources and concepts such as catalytic activation, translocation, and modification are transformed into semantic data and saved. I expect that bench scientists who have experimental findings will be able to easily participate in the construction of new knowledgebases using our repository system based on Semantic Web technologies. Furthermore, the Semantic Web, which is made up of linked data, will enable future work in artificial intelligence and machine learning, allowing computers to infer new semantic linkages and hypotheses in the life sciences.

Keywords: Pathway Database, Repository, RDF, SPARQL, Ontology, Semantic web, Data integration, Glycoscience

# Chapter 1

## Introduction

### 1.1 The major glycan types and functions

In biology, it has been a central paradigm that the biological information goes from DNA to RNA to protein. However, the fact that the genomic differences between humans and mice are about 10% suggests that only genes and their products cannot fully explain phenotypic variation (Asif T. Chinwalla, 2002). The studies of post-transcriptional changes that regulate gene expression, including ubiquitination, phosphorylation, and glycosylation, have revealed the functions of post-transcriptional modifications in controlling biological processes (Monaco et al., 2015).

The glycans provide energy to the cell as a metabolite, determine the blood type that is decided by glycan on the red blood cells (RBC), shape into the proper tertiary structure, and direct to where to go proteins should go. The glycosylated proteins, which is estimated that over 50% of the secreted and membrane-bounded human proteins are glycosylated, which has shown their physiological roles in normal and disease states such as infection or cancer. Glycans have recently been stressed as being as universal in nature as nucleic acids, proteins, lipids, and metabolites, and as essential to the survival of all known living species (Varki, 2011).

### 1.1.1 *N*-glycans

*N*-linked glycans are covalently attached to the asparagine (Asn) side chain in glycoproteins by *N*-glycosidic linkage. All eukaryotic *N*-glycans contain a common structure,  $\text{Man}\alpha 1-3(\text{Man}\beta 1-6)\text{Man}\beta 1-4\text{GlcNAc}\beta 1-4\text{GlcNAc}\beta 1-\text{Asn-X-Ser/Thr}$  in which "Asn-X-Ser/Thr" is a minimum amino acid sequence to acquire an *N*-glycan and "X" means any amino acid with the exception of Proline. The *N*-glycan modification process on proteins takes two steps: adding the synthesized glycans to the protein in the ER (endoplasmic reticulum) and trimming the glycan branch structure on proteins by glycosidases and glycosyltransferases in the Golgi apparatus (Schachter, 2000).

Numerous experiments have revealed the functions of *N*-glycans by utilizing inhibitors (Sørensen et al., 2023) that inhibit *N*-glycan synthesis or mutations that add or delete the glycosyltransferase gene in model organisms such as yeast, *Drosophila melanogaster*, zebrafish, *Caenorhabditis elegans* and mouse (Henry, Nickels, and Edlind, 2002; Frank and Aebi, 2005; Zacchi and Schulz, 2016; Travers et al., 2000). For example, the T cell receptor, which is essential for activating T cells and initiating the adaptive immune response by recognizing antigens presented by APCs that recognize infections and aberrant cells, is extensively glycosylated (Pereira et al., 2018). TCR glycosylation defects cause aberrant T cell growth and interfere with the appropriate assembly and activation of downstream signaling complexes, resulting in decreased T cell signaling and a weakened immune response.

### 1.1.2 *O*-GalNAc glycans (Mucin-type *O*-glycans)

*O*-GalNAc glycans initiated by GalNAc are linked to serine or threonine residue and are usually elongated to form one of four common core structures (Figure 1.1). Some *O*-GalNAc glycans have Fuc and Sia, Gal, GalNAc, GlcNAc, and sulfate at their termini, which gives them antigenicity and is utilized for lectin binding. The sialylated and sulfated Lewis antigens, in especially, are ligands for selectins, which are cell surface lectins that mediate

an interaction between leucocytes and endothelial cells, and Gal-terminating structures are ligands for galectins, which are glycan-binding proteins expressed by a wide range of cells (Modenutti et al., 2019; Yago et al., 2010).

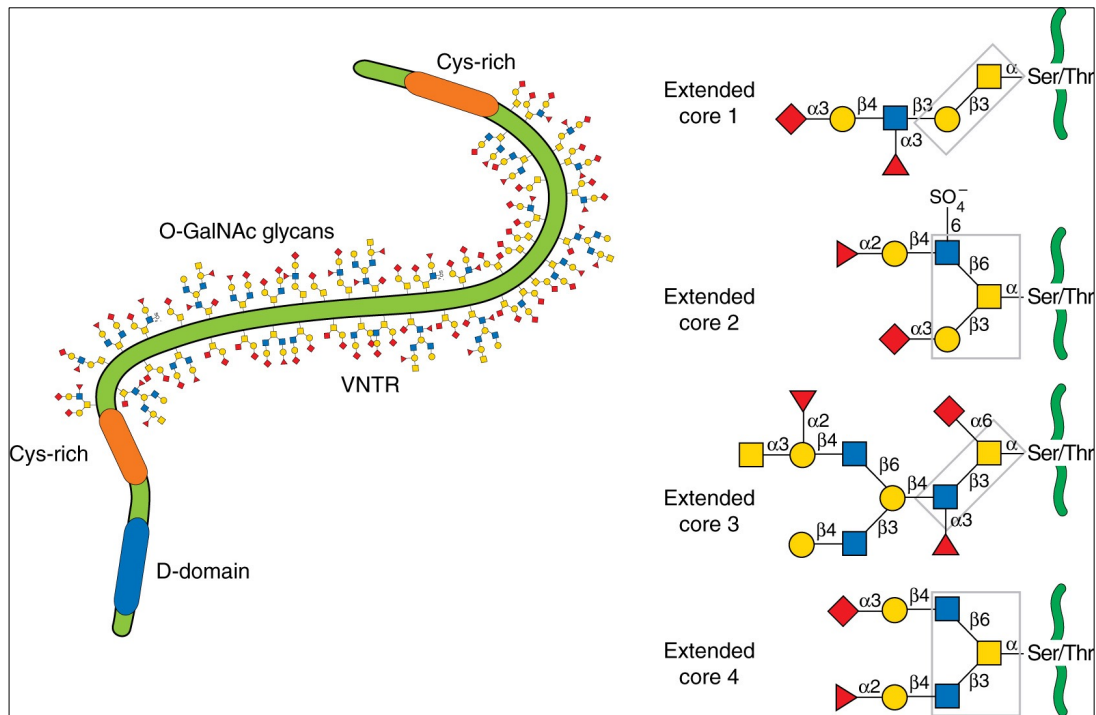


FIGURE 1.1: The highly *O*

-glycosylated mucin and *O*-GalNAc core glycans. Illustration: Essentials of Glycobiology (Varki et al., 2022)

The *O*-glycans include the A, B, and O blood group antigen glycan. Mucins are macromolecules that exist in various tissues such as the respiratory and gastrointestinal tracts that need to retain water on their surface layers and are heavily glycosylated with *O*-GalNAc glycans. As seen in Figure 1.1, sialylated *O*-glycans provide regions of a significant negative charge, allowing mucins to bind enormous amounts of water, which provide an important barrier to protect the epithelial surfaces of the salivary, gastric, intestinal, and vaginal glands against microbial invasion (Bergstrom and Xia, 2013). Also, these extensively glycosylated

proteins have been demonstrated to provide defense against protein proteases such as proteinase K (Loomes et al., 1999). Microbes have taken advantage of the mucin glycans on the surface of epithelial cells. Commensals use mucin binding to stay inside their preferred biological niche and promote biofilm growth (Takamatsu et al., 2006). Pathogens, on the other hand, have been deceived by mucin glycans as a decoy binding site, diverting them away from their intended destination in the cells (Lillehoj et al., 2013).

### 1.1.3 Glycosphingolipids (GSL)

Glycosphingolipids (GSLs) are the most frequent glycolipids in mammals, and they share a lipid component called ceramide, which is made up of long-chain amino alcohol (sphingosine) and an amide-linked fatty acid that contributes to the diversity of GSL structures (Figure 1.2) (Farwanah and Kolter, 2012). However, their basic classifications are based on glycans without bearing on ceramide variation.

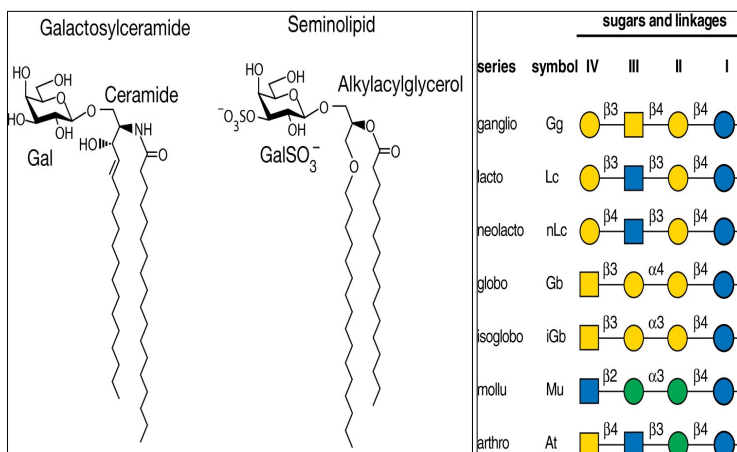


FIGURE 1.2: The core structures of GSL. Illustration: Essentials of Glycobiology (Varki et al., 2022)

Glycosphingolipids contribute to the structural integrity and stability of cell membranes. By interacting with other lipids and proteins, they help regulate membrane fluidity and organization, ensuring proper membrane function and stability (Sonnino and Prinetti, 2010).

Galactosylceramide (GalCer) is one of the most abundant molecules in the vertebrate brain, which is the first GSL to be characterized. Through mice with gene mutations in glucocerebrosidase that cause the inability to catabolize GlcCer, GlcCer was reported to act as a precursor to ceramide required to build the outermost layer of the skin.

### 1.1.4 Other types of *O*-glycans

The epidermal growth factor-like repeats (EGF-like repeats) are a common, evolutionarily conserved protein domain involved in Notch receptors, coagulation factors of blood, and various extracellular matrix proteins. It has been reported that these EGF-like repeats are modified by *O*-fucose, *O*-glucose, and *O*-GlcNAc (Takeuchi and Haltiwanger, 2014; Ma et al., 2020; Haltom and Jafar-Nejad, 2015). The EGF repeats contain six cysteine residues as in Figure 1.3, and their glycan modification on the motif has been considered an important factor in the relevant pathway because they influence signal transduction during early development, cell differentiation, and cancer progression.

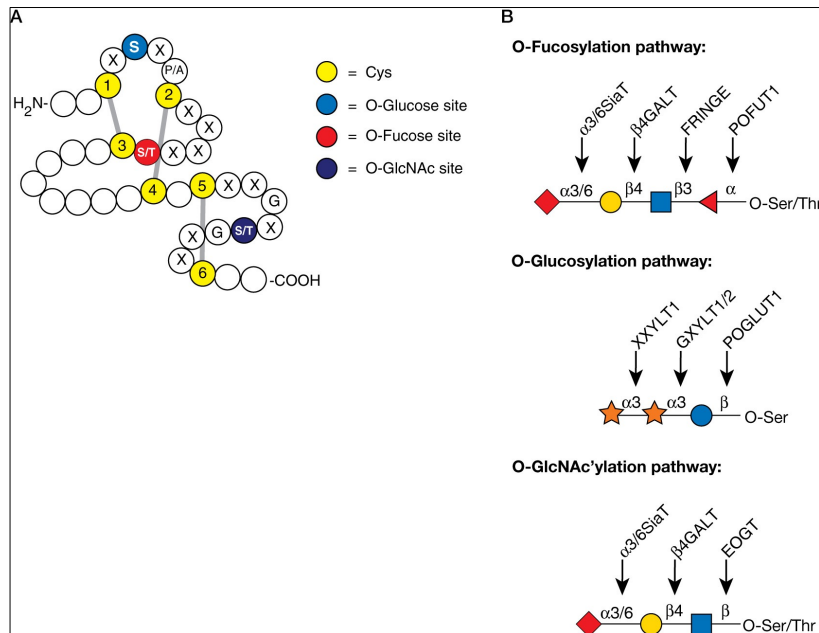


FIGURE 1.3: A schematic representation of epidermal growth factor (EGF)-like repeats. Illustration: Essentials of Glycobiology (Varki et al., 2022)



- ***O*-Fucose and *O*-Glucose Glycans**

Protein *O*-fucosyltransferases (POFUTs) typically add *O*-fucose glycans to the EGF domain through their enzymatic activity. The consensus sequence within EGF repeat for glycan binding was identified as C<sup>2</sup>-X-X-X-X-(S/T)-<sup>3</sup>. The C represents a Cysteine amino acid. The *Drosophila* homolog is Ofut1 (Okajima and Irvine, 2002). The lack of Jagged1-induced Notch signaling in fucose-deficient cells provided the first evidence that *O*-fucose is required for Notch signaling moloney2000fringe. A lot of studies suggest that POFUT1 regulates mammalian Notch signaling at the step of Notch-ligand binding, also recent research has shown that *O*-fucosylation may play a role in controlling protein quality since it alters only properly folded EGF-like repeats (Haltom and Jafar-Nejad, 2015).

The  $\beta$ -linked *O*-glucose modification occurs at the consensus sequence C<sup>1</sup>XSX(P/A)C<sup>2</sup> between the first and second conserved cysteines of EGF-like repeats (Rana et al., 2011). The first identification was from the EGF repeats of bovine blood coagulation factors VII and IX (Hase et al., 1988). The *Drosophila* gene encoding protein *O*-glucosyltransferase (POGLUT) is Rumi, human in POGLUT1. A single mutation of an *O*-glucose does not appear to impair ligand-mediated Notch1 activation but is required for optimal Notch activation. POGLUT1, like POFUT1, is found in the ER and requires a correctly folded EGF-like repeat as a substrate.

## 1.2 Microbial glycosylation

Bacterial classification is based on the results of a staining reagent on the microbial surface: because of their thick peptidoglycan, Gram-positive bacteria retain the Gram stain (crystal violet), while Gram-negative bacteria wash away the stain and take up the counterstain (Wilhelm et al., 2015). *Mycobacterium* cell walls, on the other hand, cannot be stained by

crystal violet and are instead required to be stained with an acid-fast stain (Titford, 2010). Almost all microorganisms produce a wide range of glycan structures including capsular polysaccharides, glycoproteins, and exopolysaccharides. Peptidoglycan is a crucial structure for maintaining cell shape and integrity in all microbes. They are made up of linear chains of alternately arranged N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) and are recognized by Toll-like receptor 2 (TLR2) of antigen-presenting cells in the host. As a result, peptidoglycan has been a prime target of antibiotics such as beta-lactams and carbapenems. Lipoteichoic acids and teichoic acids are distinct glycans in Gram-positive bacteria, whereas lipooligosaccharides or lipopolysaccharides are unique glycan structures in Gram-negative bacteria. The sugar structures of *Mycobacterium* species, on the other hand, are highly intricate.

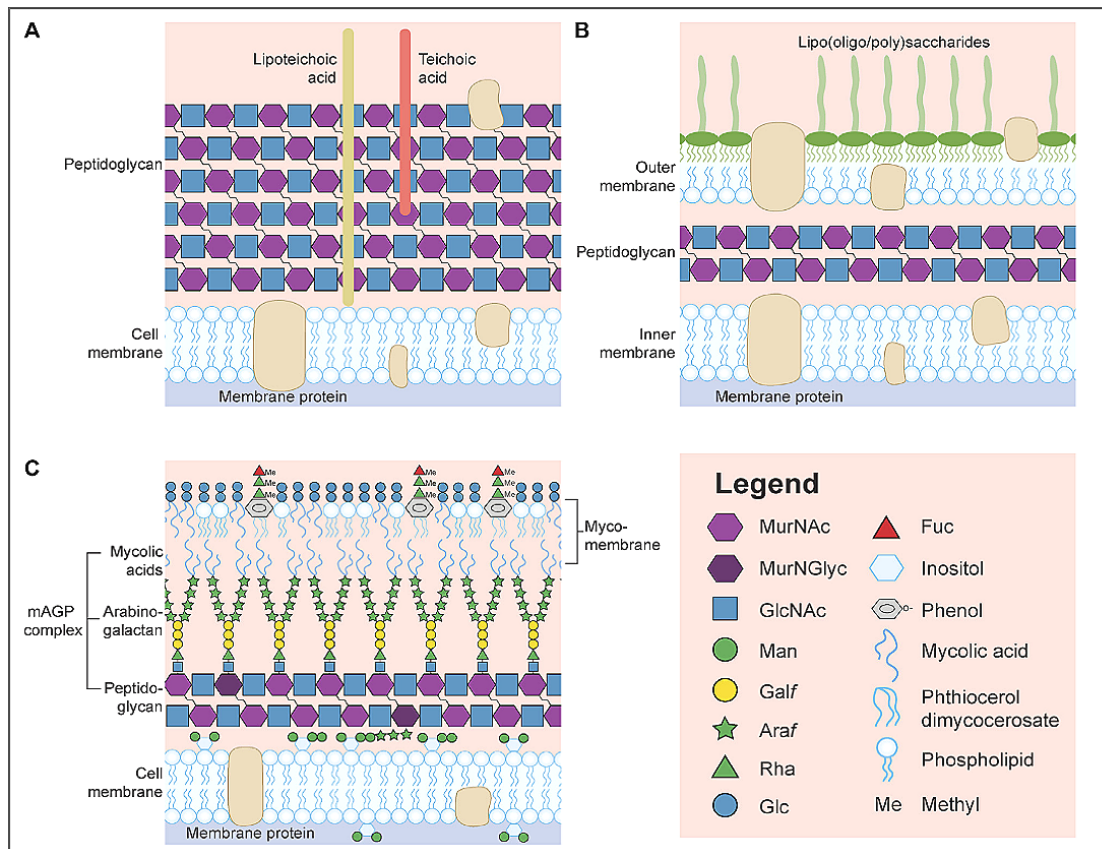


FIGURE 1.4: Structures of bacterial cell wall (Szymanski, 2022). Gram-positive (A), Gram-negative (B), and *Mycobacterium tuberculosis*

Mycobacterial cellular membranes have distinguishing features in their glycan structures. The mycolylarabinogalactan-peptidoglycan (mAGP) complex in the cell wall core consists of covalently bonded peptidoglycan (PG), arabinogalactans (AGs), and mycolic acids, which play a crucial role in immune evasion, allowing mycobacterial species to survive inside host macrophages (Catalão, Filipe, and Pimentel, 2019). As previously mentioned, bacteria have an extensive repertoire of glycoconjugates on their surfaces, and bacterial glycans play a key role in exploiting or disturbing host immune responses. Bacterial glycans found on the surfaces of pathogenic bacteria that cause disease in humans and animals are recognized by lectin in the innate immune system. After lectin attachment to the bacterial carbohydrate, it is assessed whether bacteria exploit host immunity or host immunity exerts an intrinsic defense mechanism against bacteria (Prado Acosta and Lepenies, 2019). Thus, bacterial glycans have been studied to better understand host and pathogen immune responses.

### 1.2.1 *Escherichia coli* (*E. coli*) O-antigen

*E. coli* plays a crucial role in the intestine as a commensal bacterium, aiding in digestion and competing with harmful bacteria for energy sources. While most strains of *E. coli* are harmless and even beneficial to their host, some virulence types acquire pathogenic attributes by unknown mechanisms. Once they gain a particular level of virulency, pathogens can lead to a wide range of diseases that present with severe symptoms including meningitis, cystitis, bloody diarrhea, and so on (Kaper, Nataro, and Mobley, 2004).

Lipopolysaccharide (LPS) is anchored in the outer membrane of Gram-negative bacteria and is usually assembled of three molecular components: lipid A, core oligosaccharide, and the O-antigen. Lipid A is a glycolipid anchored in LPS, core oligosaccharide contains specific glycans such as heptose and keto-deoxyoctulosonate (Kdo), and O-antigen is made up of oligosaccharide units with two to seven glycan residues (Erridge, Bennett-Guerrero, and Poxton, 2002).

- **Bioynthesis pathway of O-antigens**

The O-antigens of bacteria are important immunogens and show diversity in the glycan constituents that make up their structure. The specificity of serogroup has been utilized to determine bacterial subtypes called serotyping, which has been used as a basic tool for epidemiological surveillance to monitor the current burden and to identify outbreaks of bacterial infection (Scheutz et al., 2004).

The diversity of O-antigen means the polymorphisms present in the genes responsible for the sugar components and the glycosyl linkages. The glyco genes specific to the O-antigen and the glycosyltransferases are usually in a cluster on the chromosome known as the O-antigen gene cluster that is usually conserved within species, but some genes for sugar components are found in other loci such as *glmU* for the synthesis of UDP-GlcNAc used in metabolism. There are three biosynthesis pathways for O-antigens: the Wzx/Wzy pathway, the adenosine triphosphate-(ATP) binding cassette (ABC) transporter pathway, and the Synthase pathway, which share initiating the transfer of a sugar-phosphate to undecaprenyl phosphate (Und-P) (Greenfield and Whitfield, 2012; Raetz and Whitfield, 2002). Most O-antigens are synthesized by the Wzx/Wzy pathway, in which GTs sequentially transfer the responsive sugar nucleotide to the growing saccharides to generate the O units. The completed O-antigen units are translocated from the cytosolic face to the periplasmic side of the inner membrane by the flippase enzyme, Wzx (Hong and Reeves, 2014), and then polymerized by the polymerase enzyme, Wzy (Merino, Gonzalez, and Tomás, 2016) until the O-antigen is completed with the number of O-units regulated by Wzz protein (Guo et al., 2005) that has a main effect on the chain length of O-antigen. A few O-antigens such as E.coli O8, O9, and O9a utilize the ABC transporter pathway. In this pathway, the O-antigen structure is completed on the cytoplasmic side of the inner membrane and then translocated by ABC transporter, Wzm and Wzt (Bi et al., 2018). The synthase

pathway carries out the synthesis of O-antigens that contain homopolymers or a disaccharide unit. The *S. enterica* O54 O-antigen is the only case reported in bacteria (Keenleyside et al., 1994). Following translocation and polymerization by one of the O-antigen syntheses pathways described above, the O-polysaccharides are transferred to the lipid A-core by the ligase enzyme WaaL, and an LPS is ready for transport to the outer membrane via the LPS transport (Lpt) pathway (Okuda et al., 2016). The *wzx*, *wzy*, *wzz*, *wzm*, *wzt*, and *waaL* genes that respond to O-antigen processing are frequently identifiable from sequence alone, but assigning GTs offering a diversity of O-antigens to the specific glycosidic linkages by sequence information alone is rarely possible (Reeves and Cunneen, 2010). The database provides the currently known O-antigen structures ECODAB (<https://nevyn.org.au.se/ECODAB/>).

### 1.2.2 Microbial glycans and drug resistance

- **The glycans in bacterial categories**

The bacterial conjugates are usually located on cellular surfaces and membranes. Bacterial membranes consist of flagella, capsular polysaccharides (CPS), exopolysaccharides (EPS), lipopolysaccharides (LPS), and peptidoglycan (PG). These membrane components carry uncommon carbohydrate residue (Schmidt, Riley, and Benz, 2003) and these glycoconjugates have been targeted for treatments and diagnostics of infectious diseases.

The majority of bacteria have been classified into three groups based on their cell wall structure: Gram-negative bacteria, Gram-positive bacteria, and mycobacteria (Figure 1.4). Gram-negative bacteria's cellular walls are composed of a thin peptidoglycan (PG) layer enclosed between the inner and outer membranes and are decorated with LPS O-antigen, core-sugars, and CPS. In contrast, Gram-positive bacteria lack an outer membrane and have a relatively thick PG layer, where the polymers such as teichoic

acids that act as an antigenic determinant, polysaccharides, and peptidoglycolipids are covalently attached (Silhavy, Kahne, and Walker, 2010).

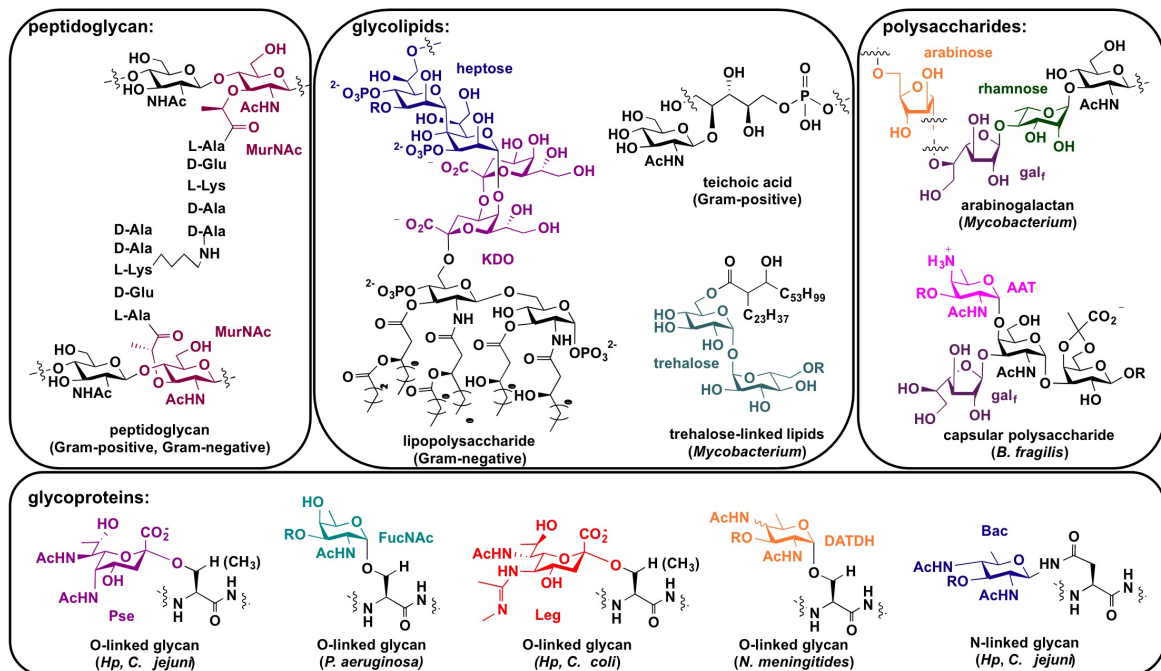


FIGURE 1.5: The representative glycans on the bacterial surfaces. Abbreviations: MurNAc = muramic acid; KDO = 3-deoxy-D-manno-oct-2-ulosonic acid; heptose = L-glycero-D-mannoheptose; gal<sub>f</sub> = galactofuranose; AAT = 2-acetamido-4-amino-2,4,6-trideoxyhexose; Pse = pseudaminic acid; Leg = legionaminic acid; DATDH = 2,4-diacetamido-2,4,6-trideoxyhexose; FucNAc = N-acetylfucosamine; Bac = bacillosamine (Tra and Dube, 2014).

The mycobacteria cell wall is very thick and has a complex structure called the mycolyl-arabinogalactan-peptidoglycan (mAGP) complex. The mAGP complex is composed of a peptidoglycan layer consisting of alternating *N*-acetylglucosamine (GlcNAc) and *N*-Glycolylmuramic acid (MurNGlyc), arabinogalactan (AG), and mycolic acids (MA) (Alderwick et al., 2015), which play an important role in antibiotic resistance avoiding antibiotics that interrupt cell wall synthesis and virulence (Maitra et al., 2019). *Mycobacterium tuberculosis* is the most notorious species of mycobacteria causing tuberculosis in humans.

- **Bacterial glycans as drug target**

All bacteria have peptidoglycan consisting of repeating units of  $\beta$ 1,4-linked *N*-acetylglucosamine (GlcNAc) and *N*-acetylmuramic acid (MurNAc). The MurNAc is a unique glycan that is not found in other prokaryotic or eukaryotic cells. Thus, this glycan has been a target to fight bacterial infections such as penicillin. LPS, on the other hand, is exclusively generated in Gram-negative bacteria and is constituted with the characteristic monosaccharides 3-deoxy-D-manno-oct-2-ulosonic acid (Kdo) and L-glycero-D-mannoheptose. Therefore, LPS has been targeted for drugs to diagnose or combat Gram-negative bacteria (Cipolla et al., 2011). In addition, Gram-positive bacteria have specific glycopolymers such as teichoic acids (TAs) that play a critical role in bacterial pathogenesis, immunological evasion, and antibiotic resistance to  $\beta$ -lactam antibiotics, including penicillin. Thus, the inhibition of TAs synthesis has been targeted for drugs due to their various effects (Pasquina, Santa Maria, and Walker, 2013). Mycobacteria have unique structures in their cell wall. Arabinogalactans (AGs) are polysaccharides that are required to produce a cell wall because they function as a linker between peptidoglycan and mycolic acids. Mycolic acids are long-chain fatty acids that have hydrophobic and waxy properties. Due to their nature, mycolic acids are highly resistant to antibiotics and small molecules derived from host immunity, such as antimicrobial peptides that defend the host from pathogenic organisms. These glycoconjugates are also linked to mycobacteria's ability to remain viable and latent in the host. Because of their crucial role in cell wall formation and function, the membrane components have been targets of drug investigation, although they have developed resistance to some medicines (Maitra et al., 2019; Batt et al., 2020).



- **Antimicrobial resistance and bacterial glycans**

Antibiotics of conventional origin are natural compounds produced by microorganisms to protect them from other microbes (McCarter, 2017) such as  $\beta$ -lactam. Numerous antibiotics have been created to target the essential enzymes responsible for membrane synthesis to penetrate the bacterial membrane that serves as a protective barrier. However, through mutations in the target enzymes or mechanism of drug activation, microorganisms have acquired strategies to resist antibiotics (Batt et al., 2020). Antimicrobial resistance means that bacteria or fungi have acquired the ability to resist the antibiotics used to kill pathogens. There are two factors contributing to antibiotic resistance. The first is the overuse of antimicrobial medicines in both animals and humans. Excessive use of antibiotics used to treat and prevent disease in animals, similar to antibiotics used in humans, has resulted in the development of antibiotic-resistant bacteria, which can be transmitted to humans who consume that poultry (Landers et al., 2012). In the case of humans, overuse of antibiotics due to unnecessary prescription that is able to cover a wide range of bacteria might result in the development of resistant organisms (Victor, 2011).

The ability of bacteria to limit the uptake of a drug comes from an intrinsic mechanism or acquired genetic material known as horizontal gene transfer (HGT). There are four main mechanisms of antimicrobial resistance: reduction of drug uptake, alteration of drug targets, inactivation of drugs, and the drug efflux system (Reygaert, 2018). As previously mentioned, a natural variation in the structure of the bacterial cell wall creates a barrier to certain drug categories. For example, *Mycoplasma* without a cell wall is resistant to medicines that inhibit peptidoglycan production, such as vancomycin, a glycopeptide antibiotic (Béb ear and Pereyre, 2005). *Mycobacterium* genus showing vancomycin resistance, on the other hand, has a thick cell wall that makes it difficult for the antibiotics to penetrate the cell (Miller, Munita, and Arias,



2014).

***Methicillin-resistant Staphylococcus aureus (MRSA)*** is another strain of *Staphylococcus aureus* (*S. aureus*) that lives on the skin and is usually harmless. However, it has acquired a multi-drug resistance to  $\beta$ -lactam medicines, including penicillin derivatives like methicillin and cephalosporins, and leads to the spread of the infection in healthcare facilities including hospitals and the community (Turner et al., 2019), which eventually leads to pneumonia (Turner et al., 2019), bacteremia meaning the presence of bacteria in the bloodstream (Siddiqui and Koirala, 2018), and endocarditis (Ruiz, Guerrero, and Tuazon, 2002). The resistance to antibiotics penicillin, methicillin, vancomycin, and daptomycin that target the cell membrane is exerted by different mechanisms (Foster, 2017). In the case of  $\beta$ -lactam antibiotics, they target the transpeptidase (TP) domain of penicillin-binding protein 2 (PBP2) (McCarter, 2017), which is responsible for the final stage to complete peptidoglycan synthesis of the bacterial cell wall. The  $\beta$ -lactam binds to the active site TP of PBP2 and prohibits peptidoglycan biosynthesis. However,  $\beta$ -lactamase (BlaZ) promotes active site regeneration by producing enzyme intermediates with higher catalytic activity than TP enzyme (Cho, Uehara, and Bernhardt, 2014). The beta-lactamase gene blaZ is transferred by transposition in a plasmid or integration into a bacterial chromosome (Jensen and Lyon, 2009). Methicillin resistance is gained via acquiring a gene that encodes PBP2a, a homolog of PBP2 (Peacock and Paterson, 2015). The active site of the TP enzyme in the PBP2 is positioned in a pocket site that is not accessible to methicillin. As a result, even when the PBP2 TP is inactive, peptidoglycan can be produced (Pinho, Lencastre, and Tomasz, 2001). Vancomycin is one of the drugs used to treat patients with antibiotic resistance to MRSA strain. Vancomycin is one of the drugs used to treat patients with antibiotic resistance to MRSA strain. It targets dipeptide D-Ala4-D-Ala5 of lipid II that consists of one GlcNAc-MurNAc-pentapeptide subunit

linked to long polyisoprenoids and prevents transglycosylation and transpeptidation for peptidoglycan (Figure 1.6) (Zeng et al., 2016).

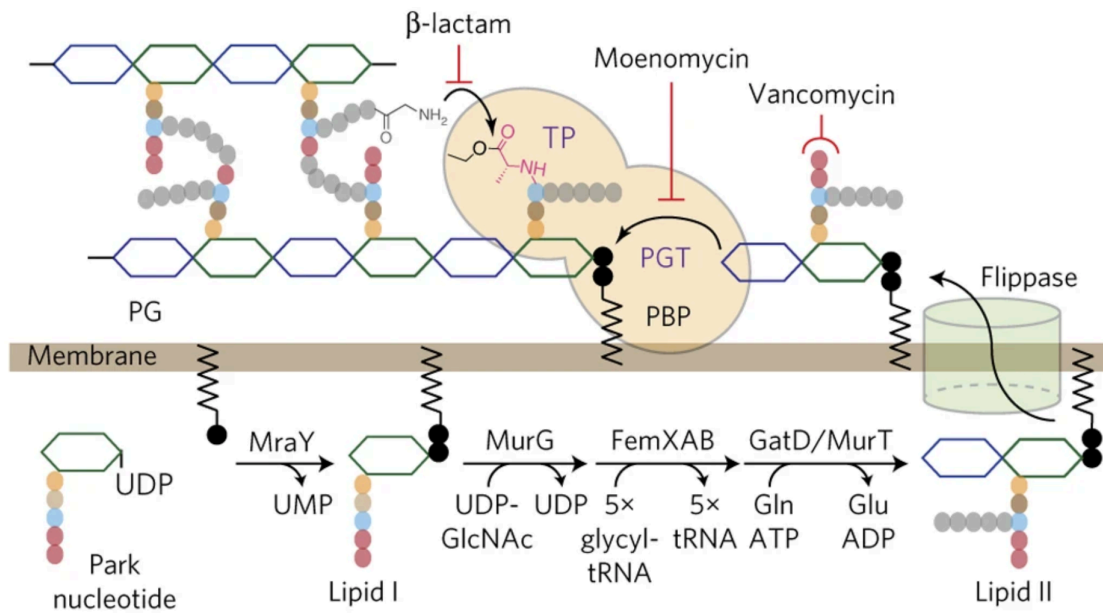


FIGURE 1.6: Inhibition mechanism of antibiotics against peptidoglycan assembly of *Staphylococcus aureus*. (Qiao et al., 2017).

Multiple mutations in chromosomal genes including *van* gene, which disturbs bacterial cell wall synthesis, lead to vancomycin resistance.

*Mycobacterium tuberculosis* (Mtb) is a notorious bacteria that has been demonstrated to be multidrug-resistant, referring to resistance to the most powerful anti-TB medicines, isoniazid, and rifampicin, in 82% of the 558,000 new cases of *Mtb* with rifampicin resistance in 2017 (Singh et al., 2020). *Mtb* generally gains antibiotic resistance through mutations in genes encoding drug targets such as RNA polymerase (Williams et al., 1998), the efflux pump system (Ghajavand et al., 2019), enzymes responding to cell wall synthesis (Brennan and Crick, 2007), etc. (Figure 1.7). As

an intrinsic drug resistance, the *Mtb* cell wall's thick lipid layer limits the passage of hydrophilic small molecules and even hydrophobic antibiotics, such as rifamycins, fluoroquinolones, and tetracyclines (Smith, Wolff, and Nguyen, 2012; Gygli et al., 2017). Moreover, the unique polymers present in the inner and outer membranes-mycolic acid, arabinogalactan, and peptidoglycan-contribute to inherent drug resistance (Nguyen, 2016). *MurA* and *MurB* are critical enzymes in the biosynthesis of peptidoglycan. The mutation of a cysteine residue into aspartic acid in the active region of *MurA* results in resistance to the antibiotic fosfomycin (Nasiri et al., 2017). One of the key elements of drug resistance mechanisms is efflux pump proteins, which transmit antimicrobials and toxins (Piddock, 2006; Blair et al., 2015). In *Mtb*, there are five superfamilies of drug efflux pumps. Antibiotic pressure, such as rifampicin or isoniazid, induces overexpression of the number of certain genes encoding each efflux pump protein, which leads to resistance to antibiotics such as isoniazid, streptomycin, etc. (Wang et al., 2013; Ghajavand et al., 2019).

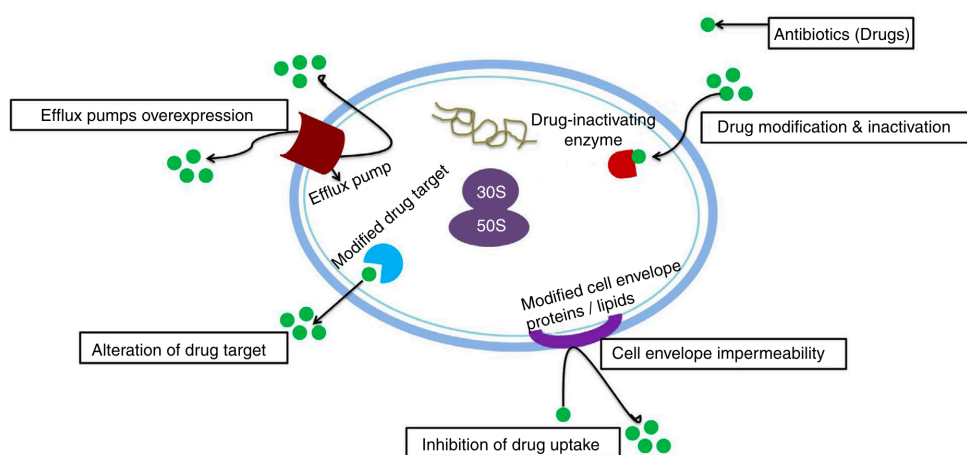


FIGURE 1.7: Diagram showing the drug resistance strategies in *M. tuberculosis* (Singh et al., 2020).

*Campylobacter jejuni* (*C. jejuni*) is one of the most common pathogens of food-borne diseases such as gastroenteritis. Food contaminated with *Campylobacter* in the animal's intestinal tract is recognized as the primary reservoir for transmission to humans. To prevent the transfer, *Campylobacter* is continuously under the pressure of antibiotics. In coping with antimicrobial stress, *Campylobacter* has established diverse resistance mechanisms such as modification of the target antibiotics by enzymes, decreased membrane permeability by efflux systems, and mutations in 23S rRNA (Luangtongkum et al., 2009; Iovine, 2013). Ciprofloxacin (Dai et al., 2020), erythromycin (Authority et al., 2022), tetracyclines (Gibreel, Wetsch, and Taylor, 2007), and ampicillin (Lin, Michel, and Zhang, 2002) are antibiotics that neutralize the ability of the drug to kill bacteria, which have been a serious threat. CmeABC is a multidrug efflux system found in *C. jejuni* that consists of CmeA, a periplasmic protein, CmeB, a multidrug transporter in the inner membrane, and CmeC, an outer membrane-anchored channel modified with an N-linked heptasaccharide glycan (GalNAc-a4-GalNAc-a4-[Glc-b3]GalNAc-a4-GalNAc-a4-GalNAc-a3-diNAcBac) (Pumbwe and Piddock, 2002; Lin, Michel, and Zhang, 2002). Abouelhadid et al. recently demonstrated the significant role of N-linked glycans in multidrug efflux pumps (Abouelhadid et al., 2020). They have already demonstrated that glycan impairment on the CmeABC pump decreases resistance to ampicillin, erythromycin, tetracycline, and ciprofloxacin (Abouelhadid et al., 2019). The study revealed that N-linked glycans have a role in antimicrobial resistance by increasing the activity of multidrug efflux pumps. They proved that an impairment of glycosylation has decreased efflux pump efficiency, which leads to increased susceptibility to antibiotics, through N-glycans enhance protein thermostability, stabilize protein complexes, and promote protein-protein interaction. The results imply that glycosylation could be a promising target for the development of antimicrobials against multidrug-resistant bacteria.

### 1.2.3 Mucosal glycans and microbiota

The gut microbiota that lives in the mucus layer covering the epithelium of the gastrointestinal tract plays an important role in human health (Bell and Juge, 2021). The gut microbiomes break down the undigestible dietary polysaccharides of the host using their carbohydrate-active enzymes (CAZymes) and transport systems, enabling the uptake of polysaccharides, thereby the microbes take the nutrient source (La Rosa et al., 2022). Because the microbiota in the gut can utilize divergent glycans continually provided by the host as an energy source, they use this advantage to colonize the mucus layer, which is connected to maintaining a healthy condition that prevents pathogenic microbes from replacing the microbiota (Marcobal et al., 2013). Also, because sustaining mucus production is critical to maintaining the barrier function of gut epithelial cells against pathogenic bacteria, the host cells benefit from effective mucin oligosaccharide recycling and the promotion of mucus formation through cytokine environments controlled by the interaction of immune cells with goblet cells (McGuckin et al., 2011).

Prebiotic effects of *O*-glycans are considerably studied along with the physical barrier function of mucin glycans (Figure 1.8) (González-Morelo, Vega-Sagardía, and Garrido, 2020). Through the experiment of feeding exogenous polysaccharides such as human milk oligosaccharides (HMOs) into mice, Pruss et al. demonstrated that mucin *O*-glycans increase the species variety of microbiota and population of resident commensal species such as *Bacillus* spp., *Bacteroides* spp. This can be carried out by the selective consumption of exogenous glycans by microorganisms that possess various carbohydrate-active enzymes or specific hydrolase activity to cleave certain glycosidic linkages. In addition, *O*-glycans aid the reconstruction of the perturbed microbiota following antibiotic therapy and retard the development of obesity caused by a diet (Pruss et al., 2021).

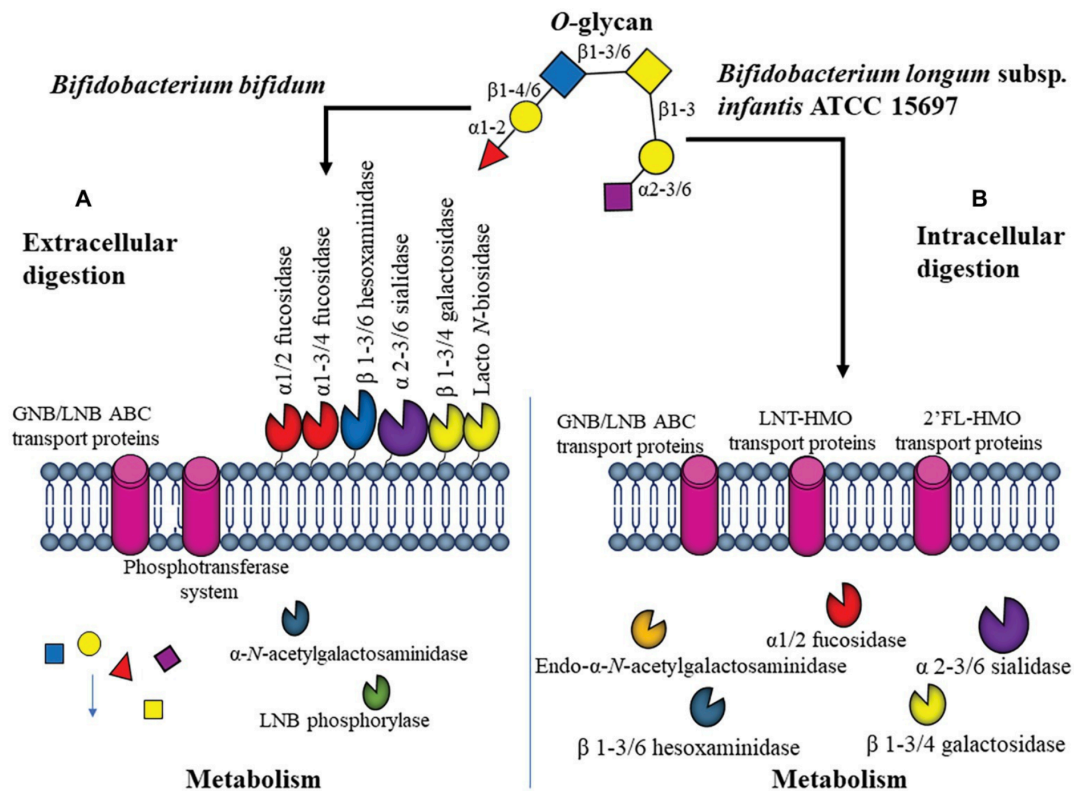


FIGURE 1.8: Representation of O-glycans utilization by (A) *Bifidobacterium bifidum*, (B) *Bifidobacterium infantis*

(González-Morelo, Vega-Sagardía, and Garrido, 2020).

A number of studies have demonstrated that prebiotics can modulate the immune response by promoting the production of immunological molecules such as anti-inflammatory cytokines or the activities of immune cells such as macrophages, natural killer (NK) cells, T cells (Shokryazdan et al., 2017; Liu, Wang, and Wu, 2022) or increasing the population of protective microbes. As previously stated, resident bacteria such as *Lactobacilli* and *Bifidobacteria* hamper the colonization of pathogenic bacteria. Mannose, for example, binds to *Salmonella* fimbriae (Oyofe et al., 1989), and dietary oligosaccharides prevent pathogens from attaching to intestinal epithelium (Jeurink et al., 2013).

## 1.3 Semantic Web Technologies

### 1.3.1 Semantic Web

- Semantic Web

The present World-Wide Web provides fixed information written in documents, where users move from one Web site to another following hyperlinks to get information. However, in the Semantic Web, computers can understand and process the information on the web pages. Furthermore, this web can enable search by inference and can become a database for knowledge discovery across different domains to contribute to the construction of a new knowledge base (Kuck, 2004). The Semantic Web cake that is referred to as a stack of technologies shows how each technical layer contributes to the goal of creating a web of data that is machine-readable and interpretable (Figure 1.9).

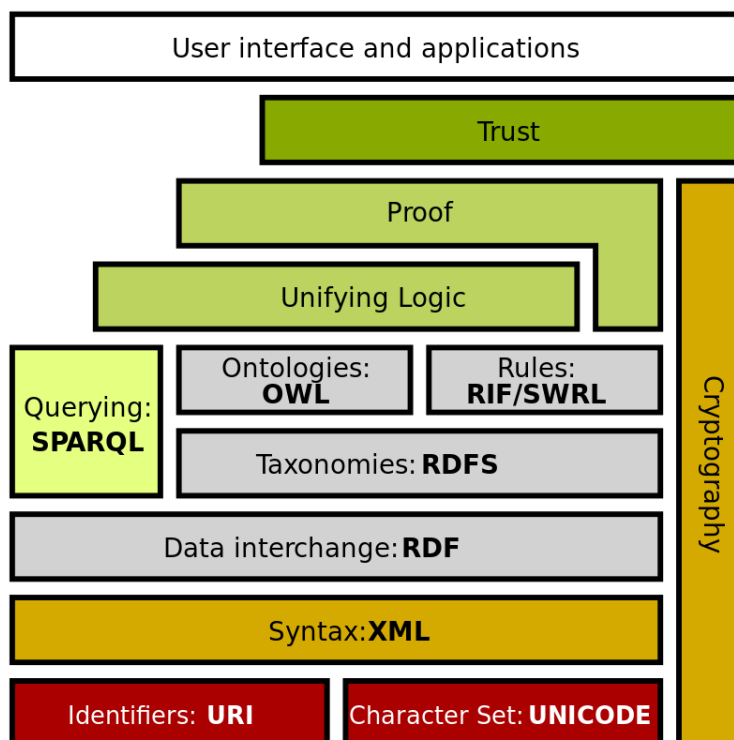


FIGURE 1.9: The Web cake diagram for the Semantic Web (Hammar, 2018).

I generated microbial glycan data, based on data obtained from expert collaborators, that can be described in a machine-readable and machine-comprehensible data format by transforming the text-based spreadsheet data provided by collaborators utilizing Semantic Web technologies such as RDF, RDFS, SPARQL, and ontologies.

- **RDF overview**

Resource Description Framework (RDF) is a data model providing a logical organization for semantic description. The model is quite simple as a triple composed of a **subject (s)**, a **predicate (p)**, and an **object (o)**. The subject is referred to as the resource to be explained on the specific predicate with a value (i.e., the object). RDFS (RDF Schema) is developed to extend the expression scope of RDF as one of the standards of the Semantic Web. Figure 1.10 (A) can represent the fact that the subject has a relationship with the Object. Triples can be chained together into complex networks, called RDF graphs, which are easier to read than textual representations when resources extend their information. There are several types of textual representations of RDF serialization such as N-triples, RDF/XML, N3, JSON-LD, and Turtle. Among them, the turtle format is a compact and human-friendly format to express triples and allow some abbreviations to assist readability such as *a* signifies **rdf:type**. Figure 1.10 (B) illustrates a directed and labeled graph of the RDF model. As depicted in the figure, a triple is represented by an edge corresponding to the predicate between two nodes; the source node and the destination node. Resources can be expanded by retrieving data from multiple distributed sources, which may result in resource duplication. To resolve the problem of the identification of common resources, the resources are presented by the use of URIs (Uniform Resource Identifiers) which is a string of characters that identifies a unique name or location for a resource on the internet. These URI references allow any data on the Web can be referred to by a unique identifier.



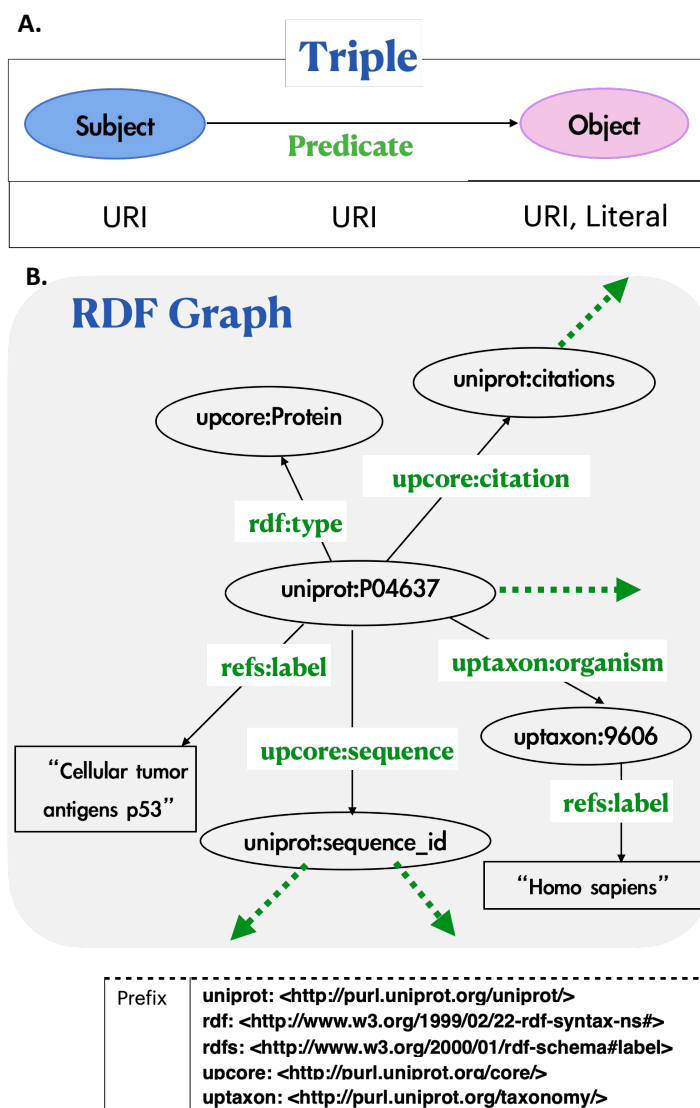


FIGURE 1.10: RDF graph model representation

By virtue of RDFS, the RDF data model can describe groups of related resources as well as the relationships between these resources. RDFS terms provide a method that assigns the resources to attributes such as the domains (*rdfs:domain*) and ranges (*rdfs:range*) of properties. The *rdfs* stands for the namespace that is defined in this URI (<http://www.w3.org/2000/01/rdf-schema>) (McBride, 2004).

### 1.3.2 Ontologies for Data sharing

With the advancement of high-throughput technologies, the complexity and volume of biological data in the life sciences have greatly increased. The data have been used to build databases of biological knowledge in their specialized disciplines. However, there is no database covering all types of resources to provide an answer to the question of researchers that includes various types of resources such as genes, proteins, and model organisms. The users must traverse many databases consisting of heterogeneous data formats and information, and it is necessary to become highly skilled at 'database surfing'. Thus, attempts have been made to integrate biological databases to allow the information that they have to be easily shared and exchanged across databases. Assigning names to biological objects across databases without causing confusion is one of the most challenging aspects of data integration.

Ontologies have been developed to capture biological concepts and objects in specialized scientific disciplines. The terms defined in the domain-specific ontology can be shared, and the existence of the shared terms facilitates data integration between multiple databases because they allow knowledge to be communicated in a standardized and machine-readable format. The main components of ontologies are classes and relations that are referred to using an identifier such as a URI. A 'Class' is a group that refers to a set of elements, such as the class "Protein" referring to the all protein set, or "Catalysis" referring to the set of all catalytic processes. The relationships that the elements belonging to a class have been defined by what are called "properties". In an ontology, an identifier for a **Class** or relationship consists of a prefix string that represents a short form of URI, a colon, and a number of digits. For example, GO:0007165 is an identifier for a **Class** with the prefix GO that will be transformed to the complete IRI ([http://purl.obolibrary.org/obo/GO\\_0007165](http://purl.obolibrary.org/obo/GO_0007165)) with the definition labeled 'signal transduction'. To resolve the ontology term written in the prefix:identifier format into the full IRI, the prefix declared in an ontology namespace is analyzed and

concatenated to the identifier by the computer.

### 1.3.3 Heterogeneity and standardization of data

- **Data format and annotation**

Biological databases have been developed independently for various research aims, and each data source is unique in terms of data format and access system. As previously mentioned, in order to gain a comprehensive understanding of biological processes, there was a need to communicate distinctively different datasets. The efforts to integrate heterogeneous data into interoperable systems have been made based on the FAIR principles standing for "Findable, Accessible, Interoperable, and Reusable" for data sharing (Wilkinson et al., 2016).

Nucleic acids and proteins are represented in the form of sequence data, and protein structures or metadata are also given in their own data types such as textual description, or tabular data. Furthermore, data exchange is hampered by conflicts of concept and inconsistencies in the annotation of objects among databases. For example, **Rad24** is a protein in the checkpoint pathway that arrests the cell cycle in *Saccharomyces cerevisiae* when DNA damage is detected. The *rad24* is a gene name associated with the checkpoint pathway of *Saccharomyces pombe*, but it is an ortholog of *rad17*. Furthermore, the *rad* gene series in *C. elegans* have no orthologs to *S. cerevisiae rad17* (Stein, 2003). Using ontologies including species information, such as PRotein Ontology (PRO) or GO terms (Botstein et al., 2000), can help to resolve the confusion of naming that exists among different communities. I introduced RDF to describe data in a formal and standardized manner to allow computers to process them.

### 1.3.4 BioPAX ontology for pathway data exchange

It is difficult to effectively utilize pathway data because pathway information is fragmented and distributed across several pathway databases in incompatible formats. In addition, each pathway data from a high throughput analysis is increasing exponentially in its own format. To collect and integrate the scattered pathway data, data have to be prepared into a computer-processable and standard format. BioPAX (Figure 1.12) is a standard language that is created to facilitate the exchange of pathway data. The BioPAX covers the following pathways: metabolic pathways, which primarily involve catalytic reactions by enzymes; molecular interactions, which primarily involve protein-protein or protein-DNA interactions; signaling pathways, which primarily involve molecules and events in the chaining of reactions; and gene regulatory networks, which comprise relationships between transcription factors and genes (Demir et al., 2010). BioPAX also includes important **Classes** for explaining data details, such as cross-references to external databases, chemical structures, linkages to controlled vocabulary from other ontologies, and protein modification (Figure 1.11).

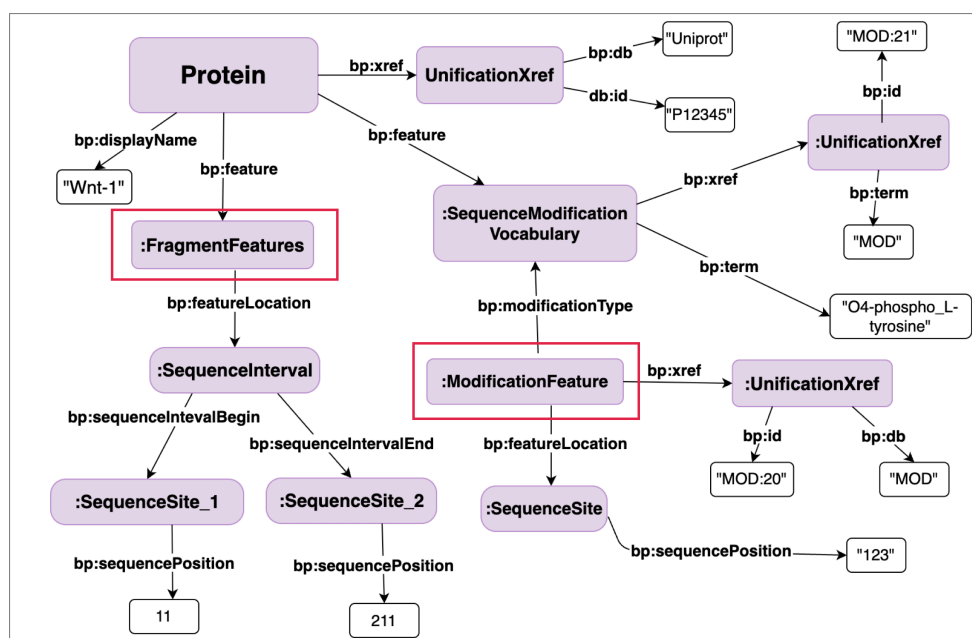


FIGURE 1.11: Introduced Class of BioPAX ontology for the description of protein modification.

The standard BioPAX format can also be used to visualize pathway data. Pathway data is eventually represented in a graphic diagram, which is quite beneficial for abstract concepts. They provide a visualizing tool, such as Cytoscape (Shannon et al., 2003), to standardize a pathway diagram.

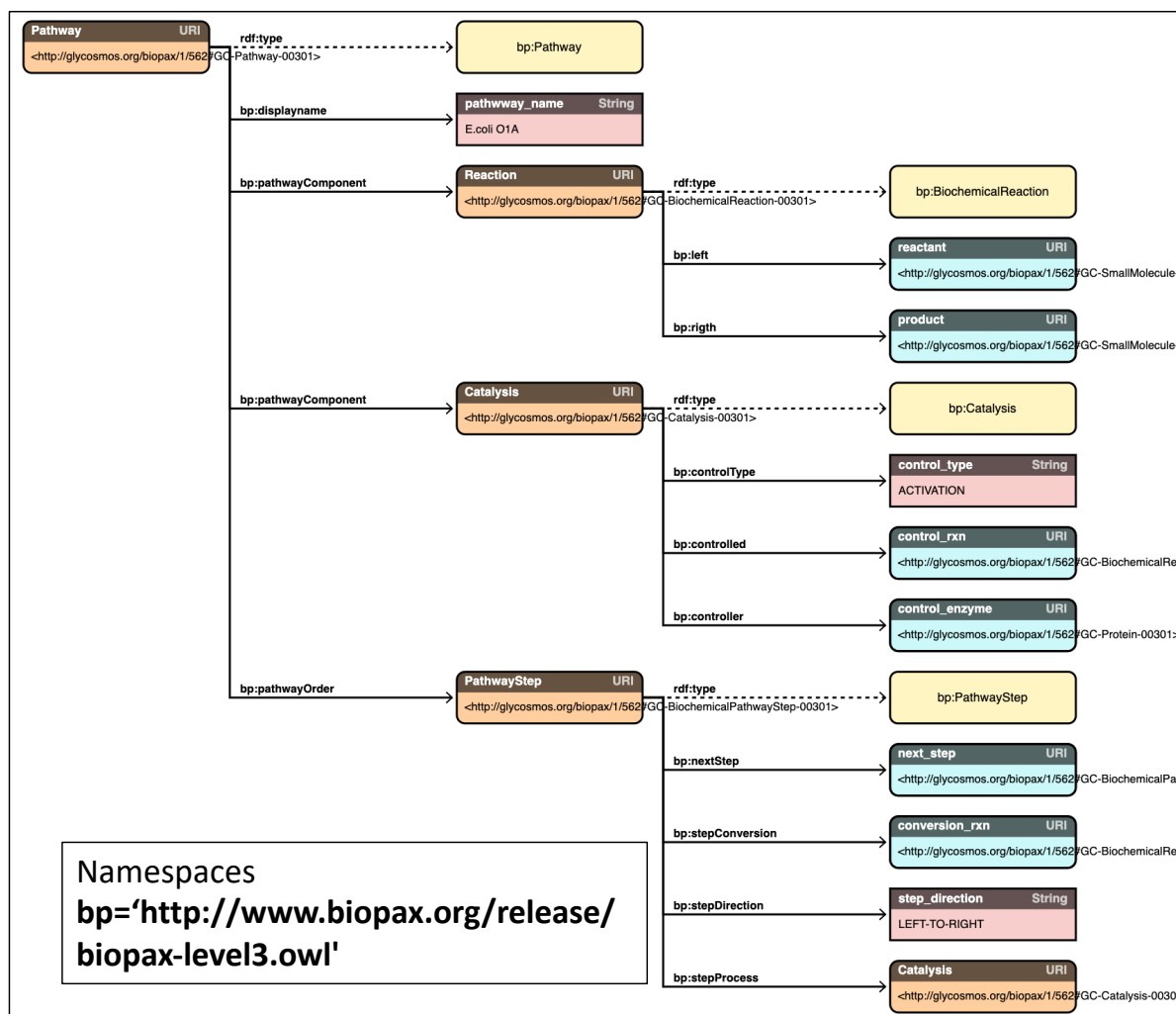


FIGURE 1.12: The RDF model for pathway representation using BioPAX.

## 1.4 Purpose

Since the N-glycosylation pathway was identified in *C. jejuni* (Szymanski et al., 1999), through extensive research on prokaryotic glycoconjugates, it has been accepted that Bacteria and Eukaryotes both have protein glycosylation systems in the scientific community. In prokaryotes, glycans are particularly abundant on the cell surface where they serve critical roles in microbe–host interactions and immune evasion, and colonization. The composition of bacterial glycans including glycolipids, glycoproteins, and polysaccharides differs from that of eukaryotic and the variety of sugars that microorganisms can create is wide (Schäffer and Messner, 2017). In addition, many strains within the same species have extra pathways for particular monosaccharides or their modifications, which contribute to the variation of carbohydrate structure from strain to strain (Szymanski, 2022). Numerous studies on identifying glycoenzymes and glycan structures in pathogens, commensals, and environmental isolates have risen for a few decades. However, the scattered information in the publications made it extremely difficult for researchers to gain a comprehensive understanding of the function of glycan or glycoenzyme in context including their cellular environment condition. Thus, I considered there is a need to make a database containing the information encompassing glycan structure, glycoenzyme, the relevant cellular architecture, and the roles in the interactions since no such database has been available, although there is a database for glycan structure, CSDB (<http://csdb.glycoscience.ru/database/>). Also, from the MicroGlycoDB, information on the glycosylation pathway will give researchers working in relevant medical microbiology a fundamental understanding of the development of new diagnostics and medications.

Massive volumes of glycome data have demonstrated that glycosylation at all levels of global cellular activity has inevitable roles to understand comprehensively the cellular function (Ohtsubo and Marth, 2006; Reily et al., 2019). It is well known that mammals have a highly complex glycan repertoire that differs from that of eukaryotes and prokaryotes. Furthermore,

structurally distinct glycans play a pivotal role in the correct development and activation of the mammalian immune system and cellular processes (Cobb and Kasper, 2005). Thus, to gain accurate knowledge and insight into the biological pathway, pathway data provided through pathway analysis, which is mostly derived from gene and protein resources, has to encompass glycan information. However, representative pathway databases like Reactome, Wikipathway, and KEGG do not provide glycan information sufficiently (Fabregat et al., 2018; Waagmeester et al., 2016; Kanehisa and Goto, 2000). To fill this gap, I intended to develop a repository that could describe glycan modifications in the relevant pathway by users. Pathway information provided by a number of pathway databases has been mapped and combined to complete pathway information as intended because pathway data vary in terms of resource types and content amounts. The acquired data through the repository must eventually be shared or integrated with other pathway data, and I designed that glycan-related data of the repository can be described in a standardized format using Semantic Web Techniques, which are designed to connect data semantically in order to make it easier for machines to interpret the data on the web and is mainly composed of RDF, SPARQL, controlled vocabularies and ontologies, etc. The RDFized data allow glycan-related pathway information to be easily shared with other public pathway data in a formal format.

# Chapter 2

## Methods

### 2.1 Inspecting semantic data

#### 2.1.1 RDFication

The spreadsheet data obtained from collaborators was reorganized based on the RDF model in the spreadsheet. With the process of RDFication, the non-structured data was transformed into a standard format. Values in every single cell were defined as instances of a particular **Class** using the *rdf:type* property. To link subject and object values, the proper vocabulary was inspected using Protégé which is the editor for ontology development where the usage or definition of **Classes** and *properties* can be identified. RDFication is a mechanical process that can be executed by simple programming code creating triples. I developed code using RDFLib which is a library to help users work with RDF. As a methodology, I proposed an approach to produce triples from table data (Lee, Ono, and Aoki-Kinoshita, 2021). The organized data in the spreadsheet were serialized in the RDF turtle format. The Python code and the generated triple files can be found in the GitLab repository (<https://gitlab.glyco.info/glycosmos/microglycodb/-/tree/master/RDFication>).



### 2.1.2 ShEx for Verification

I verified the RDF data using PyShEx (Python), version 0.8.1 as Shape Expressions (ShEx) engines that interpret based on the defined constraints (Thornton et al., 2019). Based on the validation results, I can refine the ShEx shape definition or update RDF data to address any problems identified. ShEx has been developed to model and validate an RDF model. ShEx was implemented using simple Python code to evaluate whether the triples were created correctly based on their schema. Shape definitions are contained in a ShEx schema. When an RDF node is validated against a shape, the nearby nodes are tested against the restrictions in the shape. A triple constraint defines and describes the possible object values for the corresponding predicate, for example, exactly one, one or more, zero or more. Object values can be limited to a single type, range, or list (Solbrig et al., 2017). The RDF data was passed and loaded shape to the ShEx engine, which compares the data against the shape definition. Each class in their shema was examined to see if it was connected to the correct object type, such as URI or literal. Also using the ShEx results, RDF data was able to be corrected to have the right relationships with objects. They also reported cardinality, which refers to the number of instances of an object entity and is necessary for defining the constraints that govern the relationships between entities.

### 2.1.3 SPARQL Protocol and RDF Query Language

SPARQL is a query language for RDF data (Prud'hommeaux and Seaborne, 2008). Because all RDF stores contain structured triple data, we can access and retrieve linked data over multiple RDF databases by executing federated SPARQL (Juty, Le Novere, and Laibe, 2012). Despite the lack of any information such as disease, taxon, chemicals, etc in our RDF database, the federated SPARQL query allows us to query a more complex question or expand the RDF database in a relatively easy way.

Once I defined the triple data using ShEx (Thornton et al., 2019), I uploaded these data

to the local triple storage using Virtuoso (Erling and Mikhailov, 2009). The uploaded data were examined with SPARQL queries generated by the RDF config (<https://github.com/dbcls/rdf-config>), a tool developed by DBCLS (Database Center for Life Science). I was able to check whether the identifier of entries made through the RDFication process is correct based on the retrieved results because if the entity URI of the linked nodes in the RDF graph is not identical, the connection between the resources will not be built. Figure 3.9 shows the SPARQL query to retrieve the object resources in the triples of *B. longum*.

## 2.2 RDFication of Microbial glycosylation

### 2.2.1 Preparation of O-antigen resources

The O-antigen list was collected using Beautiful Soup, a Python module, from the Escherichia coli O-antigen Database (ECODAB), which is dedicated to E.coli O-antigen structures. This O-antigen list was converted to a linear text using in-house code and the Python programming language (Python Software Foundation (<http://www.python.org/>)), version 3.7.0, and then to WURCS format using the GlycanFormatConverter API provided by the GlyCosmos Portal to register in GlyTouCan. The glycosyltransferase information required for mapping was collected from the Carbohydrate Structure Database (CSDB), a curated database established for providing glycan information in prokaryotes, plants, and fungi (Toukach and Egorova, 2016).

### 2.2.2 Protégé and OLS (ontology lookup service)

The OLS online service (<https://www.ebi.ac.uk/ols4>) is used to assign controlled vocabulary to enzyme activity and bacterial glycan localization. OLS has provided the most recent ontology versions as a repository for ontologies developed in the biomedical domain. Also, Protégé was used to inspect the usage of controlled vocabularies defined in their ontology.

For example, to convert the glycosylation reaction of microbes to RDF format, ontologies devoted to the description of the structure and important reactions of glycans and glycoconjugates, such as GlycoRDF or GlycoConjugate Ontology (GlycoCoO), were employed according to the definition. The ontology file was examined in the Protégé editor to identify the usage.

### 2.2.3 Applying BioPAX ontology and controlled vocabularies

The BioPAX ontology (Demir et al., 2010) was utilized to represent the RDF model for O-antigens. In general, an ontology is made up of instances, classes, and properties. Instances, such as O1A, O4, and O157, are items within the same class of named O-antigens; classes are used to represent concepts or objects; and properties are used to define relationships between two individuals or concepts. The BioPAX Working Group supplied documentation for the use of each Class and property in representing concepts such as biological reactions, catalytic processes, and pathways. Because the BioPAX ontology does not support glycan-related terminology, external ontology concepts such as SugarNucleotide and Saccharide were imported from ChEBI and GlycoRDF, respectively, to describe a glycan resource and information linked to it.

### 2.2.4 Serialization of data in the csv file format

The organized data in the spreadsheet from the RDF data model for microbial glycosylation have to be converted into RDF sentences. RDF documents can be stored in a variety of forms, including N3, N-Triples, RDF/XML, and Turtle. RDF statements were expressed in Turtle format in this study since it is the most human-readable and concise. We were able to save the RDF data in a compact textual form, in which a long and repeated IRI can be expressed as a short prefixed name. I utilized the RDFLib python library (Ranzinger et al., 2015) to extract RDF triples from all instances in the same table file. There are three sections

to the Python code for RDFication: importing a required API (Application Programming Interface) from the import lib module, reading a CSV file and generating nodes related to the Classes, adding instances to the RDF triple graph, and serialization. The transformed RDF triples were imported into a Virtuoso database server (Erling and Mikhailov, 2009), a graph database designed for RDF data storage, and viewed using a SPARQL endpoint (<https://ts.alpha.glycosmos.org/sparql>).

### 2.2.5 Loading the generated triples into the Virtuoso

To be processed and queried, RDF triples must be saved in an RDF store, such as the Virtuoso database server. To begin using the Virtuoso service, the open-source version of the Virtuoso server program was downloaded and installed ([virtuoso.openlinksw.com](http://virtuoso.openlinksw.com)). After installation, I can open the triple data on my personal computer to access the Virtuoso Conductor website, from which data can be queried and submitted. After logging in, navigate from "Linked Data" to "Quad Store Upload" in the Web interface to upload. I can upload an RDF data file here. The uploaded RDF data can be accessed at <https://ts.glycosmos.org/sparql>. This is called an endpoint, which is the place one can ask a question using SPARQL query language for RDF data. Integrated queries over all datasets on the Virtuoso server can be executed via this SPARQL endpoint. All query results can also be examined at <https://ts.glycosmos.org/sparql>.

## 2.3 RDFication of Pathway Repository

### 2.3.1 Taxonomy

We obtained a taxonomy dump in a Web Ontology Language (OWL) ontology file format from DDBJ (DNA Databank of Japan) (Mashima et al., 2016). Then the file is processed to contain the species name, and taxonomic identifier number in NCBI taxonomy using

SPAQRList. The result table is provided to help keyword search using Web Components of TogoStanza (<https://www.webcomponents.org/>) (Katayama et al., 2019).

### 2.3.2 Protein information

We prepared the RDF triples containing protein information such as UniProt identifier, species, protein name including the unreviewed proteins, and gene name through the queries into the SPARQL endpoint of UniProt (<https://sparql.uniprot.org/>) that is RDF data storage of UniProt. The final result data are also processed in the same way as the method used in the Taxonomy.

### 2.3.3 Web page

To develop the web services for the pathway repository ( <https://gpr.alpha.glycosmos.org/>, test version), version 6.1.0 of the 'Ruby on rails' framework (RailsGuides, [https://guides.rubyonrails.org/v6.1/getting\\_started.html](https://guides.rubyonrails.org/v6.1/getting_started.html)) was used, which is a popular web application framework written in the Ruby programming language. Also, the standard techniques for web application development such as HTML, CSS, JavaScript, and AJAX (Asynchronous JavaScript and XML) were used for providing a user-friendly interface.

### 2.3.4 Tables for list of pathways and resources

The registered pathway information was provided in a table representation, including the list of pathways, reactions, and participants. The table representation is served by web components that are developed to promote code reuse. I customized the needed elements such as UniProt identifier number, taxon name, protein name, etc and then the processed data were embedded in our pathway repository using the service package provided by DBCLS

(<https://togostanza.github.io/metastanza/pagination-table.html>). This table allows users to search pathway information using strings such as GlytouCan identifier, or species name and move to the linked page that may contain more specific information.

### 2.3.5 Visualization of the pathway data

For visualization of a glycan-related pathway, a pathway diagram was prepared using visualization tools such as Cytoscape (<http://www.cytoscape.org/>) to help simplify and identify the complex input data. Pathways are shown in the biggest compartment nodes, which are the cellular places where every reaction takes place. If multiple cellular locations exist, pathways are linked by common participants between the reactions. Edges can describe catalysis, stimulation, and inhibitory reactions, and nodes present the cellular compartment, complexes, proteins, and small molecules including labels (Figure 3.29).

To visualize a glycan biosynthesis pathway, GoJS (<https://gojs.net/latest/>), which is a JavaScript library for creating interactive diagrams and visualizations on the web was used. Each glycan structure was rendered in figure form with a GlyTouCan identifier using GoJS, and links were shown with the information on glycan-related enzymes.



# Chapter 3

## Results

### 3.1 Microbial glycosylation

Data in a tabular format containing glycan-related data was provided by expert collaborators. The information varied depending on the kinds of microorganisms: five bacterial species and one fungus. When the role of a protein that is predicted from gene sequence alignment represents enzyme activity, an additional vocabulary was prepared to describe enzyme activity.

NOTES - phase variation etc	biosynthesis	Modification/variation	Putative function
Unknown, conserved hypothetical protein	LOS	Add Man $\alpha$ 1 to inositol ring at position 2	
C4 aminotransferase specific for PseB product	flagellum	Add acyl-residue to Man at position 6	$\alpha$ -1,6-mannosyltransferase
Cytidine diphosphoramidate kinase	capsular_ polysaccharide		Chitin synthase

FIGURE 3.1: An example of the data obtained from expert collaborators.

To relate the enzyme information and glycan structure semantically for some data containing glycan information, descriptions of glycan data were also added, which can be a component



embedded in the cell wall or cytoplasmic membrane, or cellular location. A wide range of bacterial cell wall components was written in a text type (Figure 3.1). To allow that data to be described precisely by machine, the values in a single cell of the row were inspected and reorganized to adapt to ontologies.

### 3.1.1 *Escherichia coli* O-antigen

- RDF model

The *E. coli* O-antigen, which is composed of many repeats of an oligosaccharide unit, is a very variable surface polysaccharide found in Gram-negative bacteria (Russo et al., 2009). Such variation contributes to *E. coli* antigenic variability, and it is utilized to determine serogroups that allow tracking a causing bacteria when a disease suddenly rise.

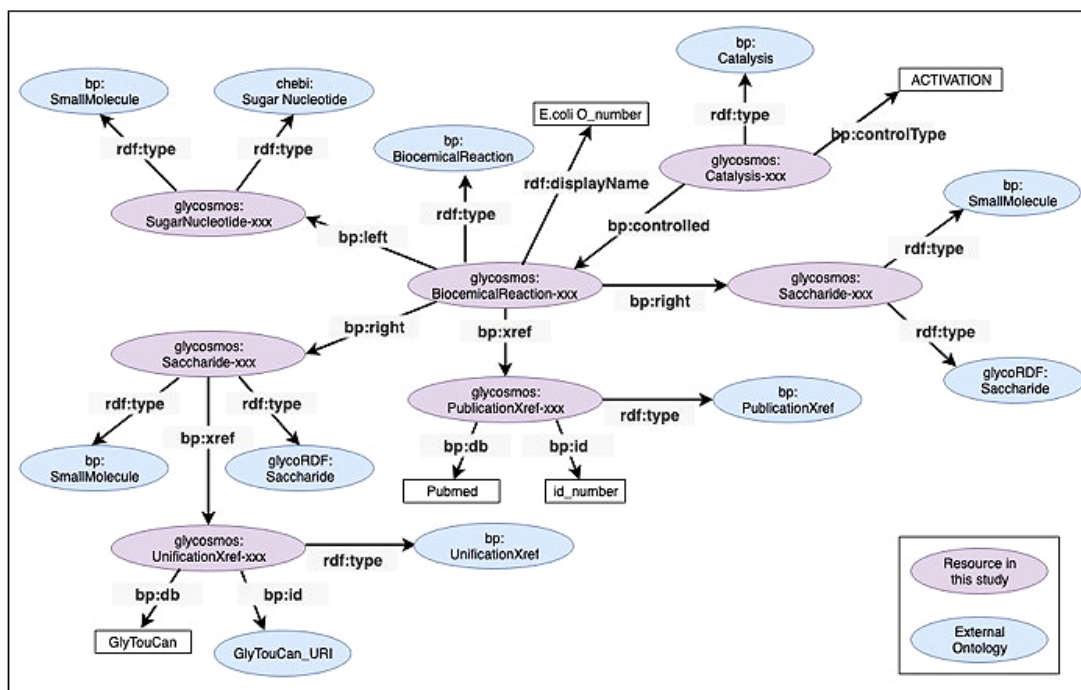


FIGURE 3.2: The schema for *E. coli* O-antigen.

A lot of studies have demonstrated that polysaccharide-based O-antigens cause a wide range of diseases including sepsis and urinary tract infections (UTI) in both people and animals (DebRoy et al., 2016; Sarkar et al., 2014). Also, with the emergence of multidrug-resistant (MDR) strains causing treatment failure, O-antigens have also become the target for effective vaccine development (Royer et al., 2022; Xing et al., 2023). Data of the O-antigens were collected from a public database, ECODAB (Liu et al., 2020) unlike the information of other microbes, which is provided by co-researchers. I prepared a list of O-antigens, which includes details on the glycan structure and the enzymes responsible for the glycosidic link. As the information on the enzymes is insufficient, more detailed information was collected from the CSDB database dedicated to microbial glycan data (Toukach and Egorova, 2019). The O-antigen glycan is intended to be represented as a series of catalytic reactions by the activity of glycosyltransferase. In this case, each reaction acquires order until the entire glycan structure is completed. To represent the structures as a series of enzyme reactions, BioPAX ontology was used (Figure 3.2). A GlycoRDF ontology was used to describe glycans because the BioPAX lacks a **Class** for glycan and glycans were also assigned unique GlyTouCan identifiers.

- **SPARQL query**

The SPARQL language's `INPUT()` method was used to load triple data to the RDF storage. Figure 3.3 represents a part of the SPARQL query to describe an RDF triple that presents an enzyme reaction extending a glycan string. Because the data (subject and object) that belongs to their **Class** group must be instantiated, the data was passed to the SPARQL query as variables that were embarrassed with "{ { } }".

```

INSERT DATA
{
  GRAPH <http://localhost:8890/testSpace>
  {
    :{{pw_id}}_GC-BiochemicalReaction{{rxn_id}} rdf:type bp:BiochemicalReaction ;
      bp:conversionDirection "LEFT-TO-RIGHT"^^xsd:string .
    {#each reactantArray_id}
      :{{this.pwid}}_GC-BiochemicalReaction{{this.id}} bp:left :{{this.pwid}}_GC-{{this.val}} .
    {/each}
    {#each productArray_id}
      :{{this.pwid}}_GC-BiochemicalReaction{{this.id}} bp:right :{{this.pwid}}_GC-{{this.val}} .
    {/each}
    :GC-Pathway{{pw_id}} bp:pathwayComponent :{{pw_id}}_GC-BiochemicalReaction{{rxn_id}};
      bp:pathwayOrder :{{pw_id}}_GC-PathwayStep{{rxn_id}}.
    :{{pw_id}}_GC-PathwayStep{{rxn_id}} rdf:type bp:PathwayStep ;
      bp:stepConversion :{{pw_id}}_GC-BiochemicalReaction{{rxn_id}} .
  }
}

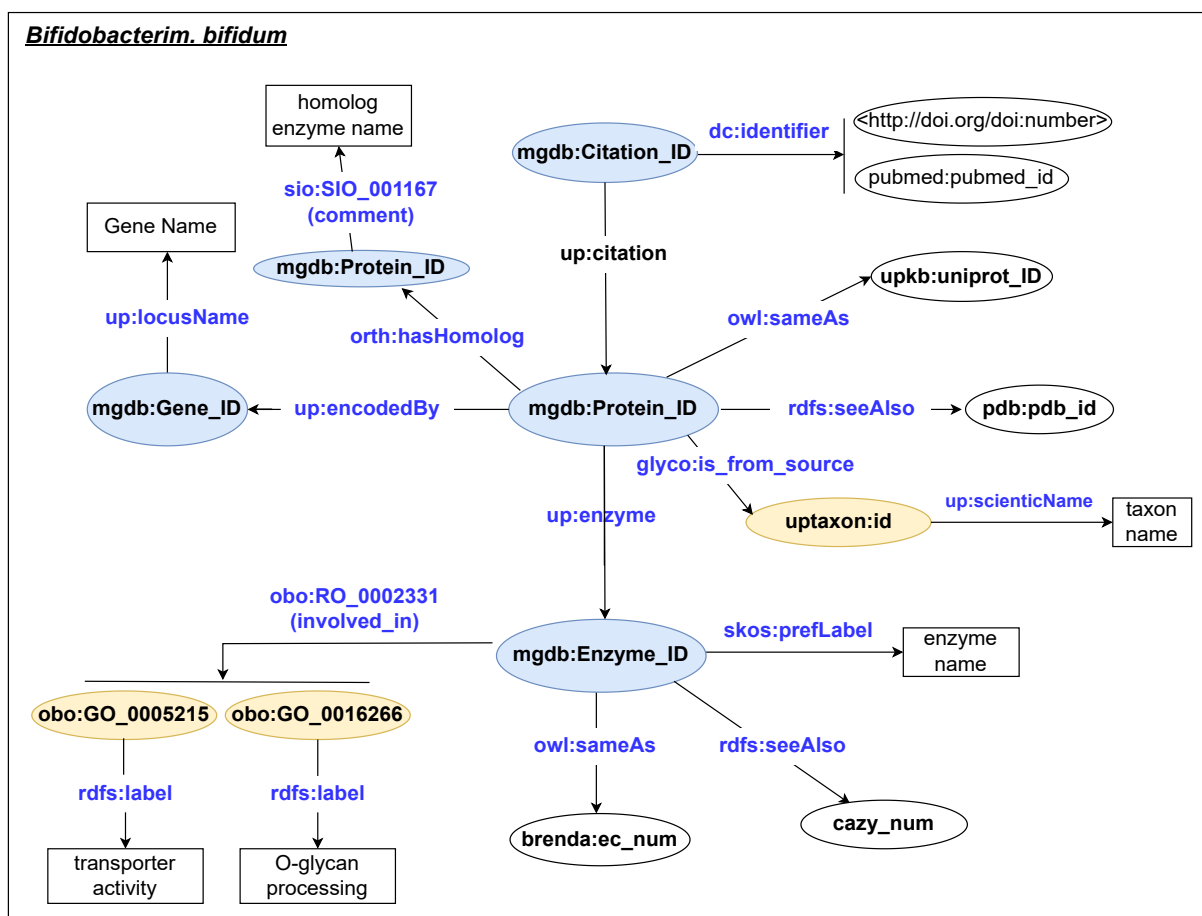
```

FIGURE 3.3: The SPARQL for input a glycosylation reaction of *E.coli* O-antigen .

### 3.1.2 *Bifidobacterium bifidum*

- RDF model

Bifidobacterium is one of the genus-producing probiotics, which are living organisms that are found in the human gut and serve as a functional food for human health. *B. bifidum* belongs to this genus and they have been investigated for production on the industrial scale. Exopolysaccharides (EPSs) produced by this bacteria have been employed in the dairy industry because they have demonstrated benefits such as a prebiotic effect, immune system modulation, and the capacity to decrease cholesterol (Ku et al., 2016). EPSs are polysaccharides affiliated with the external surface forming capsules. The data obtained from the co-researcher showed the role of enzymes that break down polysaccharides, even those that cannot be improved through experimentation and transport of the partial glycan. Also, the identifiers from external databases were supplied for related information such as proteins, enzymes, and gene locus (Figure 3.4).

FIGURE 3.4: The schema for *B. bifidum*

- SPARQL query

Figure 3.5 displays the SPARQL query that is used to obtain object instances to see if the properties link the subject and object as intended. The proteins and enzymes were described with information from an external database through the *owl:sameAs* or *rdfs:seeAlso* properties. I had in mind that the enzyme activity written rough description must be described as linked data using an ontology identifier number. The *obo:involved\_in* property given by Relational Ontology (RO) was employed to do this. With this property, the enzyme activity was represented by the molecular activity of GO (gene ontology), although the fact that the activity is not specific.

```

PREFIX up: <http://purl.uniprot.org/core/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssci: <http://semanticscience.org/resource/>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX glyrdf: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT ?protein ?uniprot_id
      (GROUP_CONCAT(DISTINCT ?pubmed_id ; separator = ",") AS ?pubmed_ids)
      (GROUP_CONCAT(DISTINCT ?pdb_id ; separator = ",") AS ?pdb_ids)
      ?homolog_name ?gene ?locus_name ?taxonomy ?species ?go_function_id
      ?enzyme ?ec_num ?cazy_num ?enzyme_label
FROM <http://localhost:8890/bifidum>
WHERE {
  ?protein a up:Protein ;
    owl:sameAs ?uniprot_id ;
    up:citation / dc:identifier ?pubmed_id ;
    up:encodedBy ?gene .
  ?gene up:locusName ?locus_name.
  ?protein glyrdf:is_from_source ?taxonomy ;
    up:enzyme ?enzyme .
  ?enzyme a up:Enzyme .
  ?taxonomy up:scientificName ?species.
  OPTIONAL {
    ?enzyme obo:RO_0002331 ?go_function_id ;
      owl:sameAs ?ec_num ;
      rdfs:seeAlso ?cazy_num ;
      skos:prefLabel ?enzyme_label .}

  OPTIONAL { ?protein rdfs:seeAlso ?pdb_id .}

  OPTIONAL {
    ?protein orth:hasHomolog ?homolog .
    ?homolog ssci:SIO_001167 ?homolog_name.
  }
}

```

FIGURE 3.5: SPARQL query for *B. bifidum*

Figure 3.6 is an example of the results retrieved from the SPARQL query. The data format is the same for every instance. The variable name `interest` is on the left of the double colon, and the value is on the right. The information connected with the protein containing ProteinDataBank ID (Berman et al., 2002) such as enzyme with EC enzyme or CAZy number, gene, taxonomy, and citation with the PubMed ID were obtained as designed in the RDF model. Also, if homolog protein information is available, it can be obtained. As intended, the data value can be presented with a URI carrying a unique identifier, allowing the data to be linked with other RDF data in the public database. The RDF sentences of other microbes also were inspected using their SPARQL query and these data will be added to the MicroGlycoDB database. All triple data can be checked in our SPARQL endpoint (<https://ts.alpha.glycosmos.org/sparql>).

```
{
  "protein_id": "http://purl.jp/bio/12/database/microglycodb/MGDP_000006",
  "pdb_id": [
    "https://www.rcsb.org/structure/3M04",           # ProteinDataBank ID
    "https://www.rcsb.org/structure/3UES",
    "https://www.rcsb.org/structure/H173UET"
  ],
  "citation_id": [
    "https://pubmed.ncbi.nlm.nih.gov/19520709",    # Reference information
    "https://pubmed.ncbi.nlm.nih.gov/22451670"
  ],
  "uni_id": "https://www.uniprot.org/uniprotkb/C5NS94",
  "homolog_name": "",
  "enzyme_id": "http://purl.jp/bio/12/database/microglycodb/MGDE_000006",
  "enz_go_name": "1,3/1,4- $\alpha$ -L-fucosidase",    # Enzyme information
  "enz_go_id": "http://purl.obolibrary.org/obo/GO_0033932",
  "ec_enzyme_num": "https://www.brenda-enzymes.org/enzyme.php?ecno=3.2.1.111",
  "cazy_num": "http://www.cazy.org/GH29.html",
  "enzyme_label": "1,3/1,4- $\alpha$ -L-fucosidase",
  "gene_id": "http://purl.jp/bio/12/database/microglycodb/MGDG_000006",
  "gene_locus": "afcB",
  "taxon_id": "http://purl.uniprot.org/taxonomy1681",
  "taxon_name": "Bifidobacterium bifidum"         # Taxon information
},
```

FIGURE 3.6: The result values of SPARQL query for *B. bifidum*

- **ShEx**

RDF triples can be checked against a given ShEx definition to whether it fits the constraints defined in the schema.

```

:Protein
{
  rdf:type [uni:Protein] ; # 100.0 %
  glyco:is_from_source IRI ?;
    # 67.64705882352942 % obj: IRI. Cardinality: {1}
  owl:sameAs IRI ?;
    # 67.64705882352942 % obj: IRI. Cardinality: {1}
  uni:enzyme @:Enzyme ?;
    # 67.64705882352942 % obj: @:Enzyme. Cardinality: {1}
  uni:citation @:Citation ?;
    # 67.64705882352942 % obj: @:Citation. Cardinality: {1}
  uni:encodedBy IRI ?;
    # 67.64705882352942 % obj: IRI. Cardinality: {1}
  seman:SI0_001167 xsd:string ?;
    # 32.35294117647059 % obj: xsd:string. Cardinality: {1}
  ns2:hasHomolog @:Protein ?;
    # 32.35294117647059 % obj: @:Protein. Cardinality: {1}
  rdfs:seeAlso IRI *
    # 26.47058823529412 % obj: IRI. Cardinality: +
    # 11.76470588235294 % obj: IRI. Cardinality: {3}
}
:Enzyme
{
  rdf:type [uni:Enzyme] ;| # 100.0 %
  obo:RO_0002331 IRI ; # 100.0 %
  skos:prefLabel xsd:string ; # 100.0 %
  owl:sameAs IRI ?;
    # 86.95652173913044 % obj: IRI. Cardinality: {1}
  rdfs:seeAlso IRI ?
    # 60.86956521739131 % obj: IRI. Cardinality: {1}
}
:Citation
{
  rdf:type [uni:Citation] ; # 100.0 %
  dc:identifier IRI + # 100.0 %
    # 52.17391304347826 % obj: IRI. Cardinality: {1}
    # 43.47826086956522 % obj: IRI. Cardinality: {2}
}

```

FIGURE 3.7: Verification of *B. bifidum* turtle file

### 3.1.3 *Bifidobacterium longum*

- RDF model

*B. longum* is also belonged to the Bifidobacterium genus. In terms of content, *B. longum* is relatively simple to compare the *B. bifidum*. The schema was designed to explain protein data with an NCBI protein identifier and enzyme data with a CAZy identifier. However, the information showing the enzyme function could not be described as linked data because they lack any information on substrate or product including a textual reference (Figure 3.8).

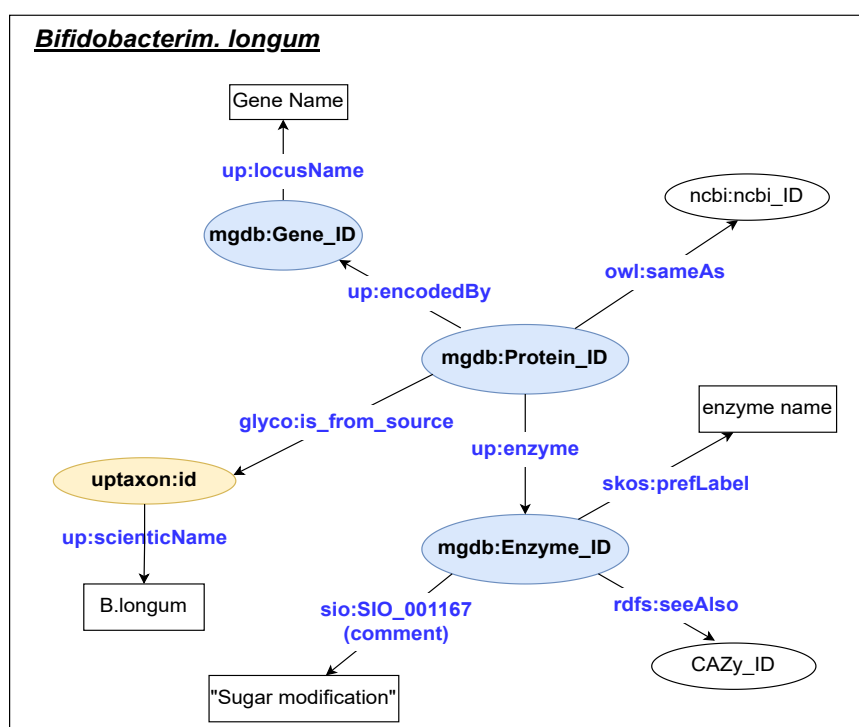


FIGURE 3.8: The schema for *B. longum*

- SPARQL query

The RDF model of *B.longum* was relatively simple. The PubMed protein identification number was identified using the owl: sameAs property for Protein instances. Enzyme



instances data was obtained using the `rdfs:seeAlso` property.

```

PREFIX up: <http://purl.uniprot.org/core/>
PREFIX glyrdf: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?ProteinEntry ?ncbi_id ?protein_label ?enzyme
?cazy_id ?enzyme_label ?taxonomy ?organism ?gene ?gene_name
FROM <http://localhost:8890/longum>
WHERE {
  ?ProteinEntry a up:Protein ;
    glyrdf:is_from_source ?taxonomy ;
    up:encodedBy ?gene ;
    up:enzyme ?enzyme ;
    owl:sameAs ?ncbi_id ;
    skos:prefLabel ?protein_label .
  ?taxonomy up:scientificName ?organism.

  ?enzyme rdfs:seeAlso ?cazy_id ;
    skos:prefLabel ?enzyme_label .

  ?gene a up:Gene ;
    up:locusName ?gene_name .
}

```

FIGURE 3.9: SPARQL query for *B. longum*

- **ShEx**

The ShEx schema was used to validate RDF triples of *B. longum*. The result shows that Classes such as **Protein**, **Enzyme**, and **Gene**, as well as instances, belonged to their Class consistent with the schema requirements. I identified that **Protein Class**

matched only **Enzyme** and **Gene Class** as designed. That is presented in "@" mark in Figure 3.10

```
:Protein
{
  rdf:type [uni:Protein] ; # 100.0 %
  uni:enzyme @:Enzyme ; # 100.0 %
  owl:sameAs IRI ; # 100.0 %
  glyco:is_from_source IRI ; # 100.0 %
  skos:prefLabel xsd:string ; # 100.0 %
  uni:encodedBy @:Gene # 100.0 %
}
:Enzyme
{
  rdf:type [uni:Enzyme] ; # 100.0 %
  rdfs:seeAlso IRI ; # 100.0 %
  skos:prefLabel xsd:string ; # 100.0 %
  seman:SI0_001167 xsd:string # 100.0 %
}
:Gene
{
  uni:locusName xsd:string ; # 100.0 %
  rdf:type [uni:Gene] # 100.0 %
}
```

FIGURE 3.10: Verification of *B. longum* turtle file

### 3.1.4 *Campylobacter jejuni*

- **RDF model**

*C.jejuni* is one of the most common bacteria responsible for gastroenteritis or diarrhea (Cain et al., 2020). Also, they contribute to immune-mediated disorders like Guillain-Barre Syndrome (GBS) or immunoproliferative small intestine disease because gangliosides on the human cell surface and the lipooligosaccharide (LOS) on the *C.jejuni* surface are similar (Goodfellow and Willison, 2016). It has been reported that *C.jejuni* is capable of significantly modifying proteins via *N*- and *O*-linked glycosylation. I obtained the enzyme data showing their substrate and product glycan, which is

able to describe the enzyme activity using ontology. However, the information about proteins with and without enzymatic activity was placed in the same column for additional explanation of data, so the commented data was organized to be separated. In addition, a Gene Ontology (GO) identifier was also assigned to the enzyme activity based on the comment such as "glycosyltransferase activity", "acyl carrier activity", and "methyltransferase activity" (Figure 3.11).

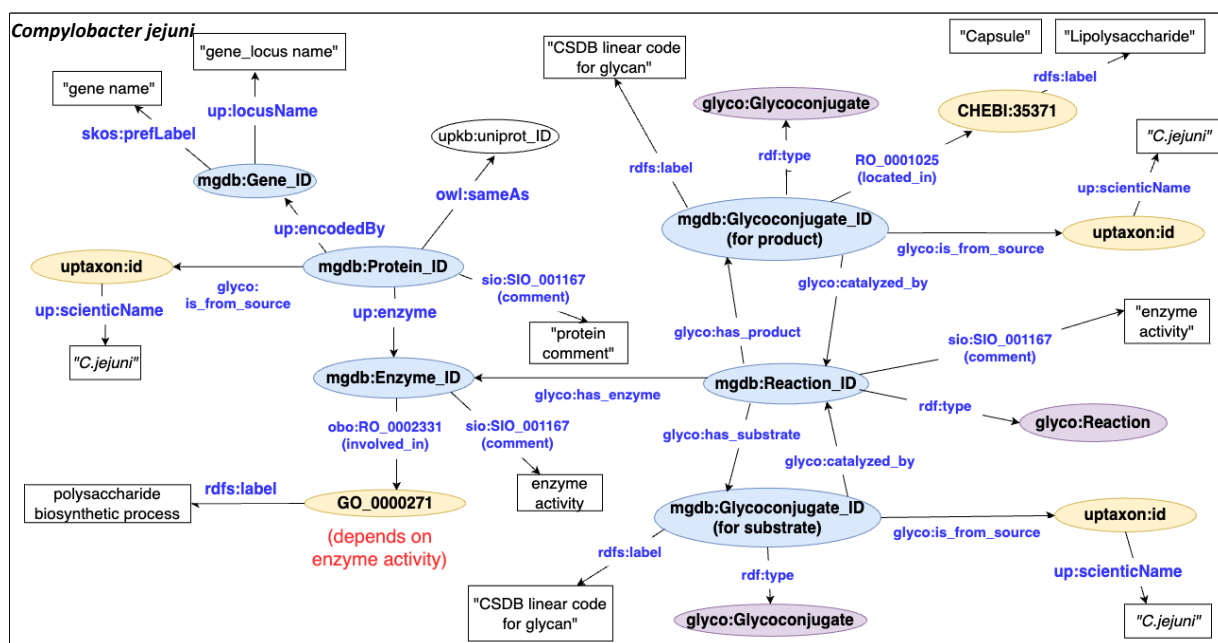


FIGURE 3.11: The schema for *C. jejuni*

- SPARQL query

The SPARQL query used to find all object resources is seen in Figure 3.11. *C.jejuni* included glycoconjugate data written in CSDB linear code. To describe the glycans as participants in enzyme reactions, the GlycoRDF ontology was used. The `:has_enzyme` property of GlycoRDF has related the **Enzyme** and **Reaction Class**. To extract the reactant and product glycan, a SPARQL query was generated like the below lines of “`#reaction`” in Figure 3.11. I got the object resource data successfully, the reactant

and product glycan data were extracted in the text type of the CSDB linear code. **Protein** and **Glycan** instances both carry taxon information that can be used to search for glycans included in the species.

```

SELECT ?protein ?protein_comment ?uniprot_id ?gene ?enzyme
      ?locus_name ?gene_label ?go_id ?enzyme_comment ?species
      ?reaction ?reactant_glycan ?reactant_glycan_text ?product_glycan
      ?product_glycan_text
FROM <http://localhost:8890/jejuni>
WHERE {
  #protein
  ?protein a up:Protein;
           owl:sameAs ?uniprot_id ;
           up:encodedBy ?gene ;
           up:enzyme ?enzyme ;
           glyrdf:is_from_source ?taxon.
  ?taxon up:scientificName ?species.
  OPTIONAL { ?protein ssci:SI0_001167 ?protein_comment . }
  ?gene a up:Gene ;
        up:locusName ?locus_name ;
        skos:prefLabel ?gene_label .

  #enzyme
  ?enzyme a up:Enzyme ;
          obo:RO_0000271 ?go_id .
  OPTIONAL { ?enzyme ssci:SI0_001167 ?enzyme_comment. }

  #reacton
  ?reaction glyrdf:has_enzyme ?enzyme.
  ?reaction glyrdf:has_product ?product_glycan.
  ?product_glycan rdfs:label ?product_glycan_text.
  OPTIONAL { ?reaction glyrdf:has_substrate ?reactant_glycan.
            ?reactant_glycan rdfs:label ?reactant_glycan_text. }
}

```

FIGURE 3.12: SPARQL query for *C.jejuni*

- **ShEx**

The Reaction node of ShEX in Figure 3.13 shows that each Reaction has one or more enzymes and is matching Enzyme node. I identified that the Glycoconjugate node was

linked to the Reaction node via the GlycoRDF ontology's `:catalyzed_by` property. It is only possible to identify intracellular localization information for 52 percent of Glycoconjugates.

```

:Protein
{
  rdf:type [uni:Protein] ; # 100.0 %
  uni:encodedBy @:Gene ; # 100.0 %
  owl:sameAs IRI ; # 100.0 %
  glyco:is_from_source IRI ; # 100.0 %
  uni:enzyme @:Enzyme ?;
    # 66.40625 % obj: @:Enzyme. Cardinality: {1}
  seman:SIO_001167 xsd:string ?
    # 46.875 % obj: xsd:string. Cardinality: {1}
}
:Enzyme
{
  rdf:type [uni:Enzyme] ; # 100.0 %
  obo:RO_0000271 IRI +; # 100.0 %
    # 98.80952380952381 % obj: IRI. Cardinality: {1}
  seman:SIO_001167 xsd:string + # 100.0 %
    # 98.80952380952381 % obj: xsd:string. Cardinality: {1}
}
:Gene
{
  rdf:type [uni:Gene] ; # 100.0 %
  skos:prefLabel xsd:string ; # 100.0 %
  uni:locusName xsd:string # 100.0 %
}
:Glycoconjugate
{
  rdf:type [glyco:Glycoconjugate] ; # 100.0 %
  glyco:is_from_source IRI ; # 100.0 %
  glyco:catalyzed_by @:Reaction +; # 100.0 %
    # 98.66666666666667 % obj: @:Reaction. Cardinality: {1}
  rdfs:label xsd:string +; # 100.0 %
    # 97.33333333333334 % obj: xsd:string. Cardinality: {1}
  obo:RO_0001025 IRI ? # 52.0 % obj: IRI. Cardinality: {1}
}
:Reaction
{
  rdf:type [glyco:Reaction] ; # 100.0 %
  glyco:has_enzyme IRI +; # 100.0 %
    # 97.5 % obj: IRI. Cardinality: {1}
    # 97.5 % obj: @:Enzyme. Cardinality: +
  glyco:has_product @:Glycoconjugate *;
    # 95.0 % obj: @:Glycoconjugate. Cardinality: +
    # 92.5 % obj: @:Glycoconjugate. Cardinality: {1}
  glyco:has_substrate @:Glycoconjugate ?
    # 92.5 % obj: @:Glycoconjugate. Cardinality: {1}
}

```

FIGURE 3.13: Verification of *C. jejuni* turtle file

### 3.1.5 *Cryptococcus neoformans*

- RDF model

*C. neoformans* is a fungus that causes meningoencephalitis by opportunistic infections that do not normally cause illness in healthy individuals but can cause disease in people with weakened immune conditions such as those with HIV/AIDS, cancer, or autoimmune disease. It has been reported that the highly mannosylated proteins on their cell wall have an important role in the pathogenicity (Thak et al., 2020; Lee et al., 2023). The protein data that is predicted to have enzyme activity, such as "mannosyltransferase", "galactosyltransferase", and "glucosyltransferase" was described using the GO ontology corresponding to the catalytic activity.

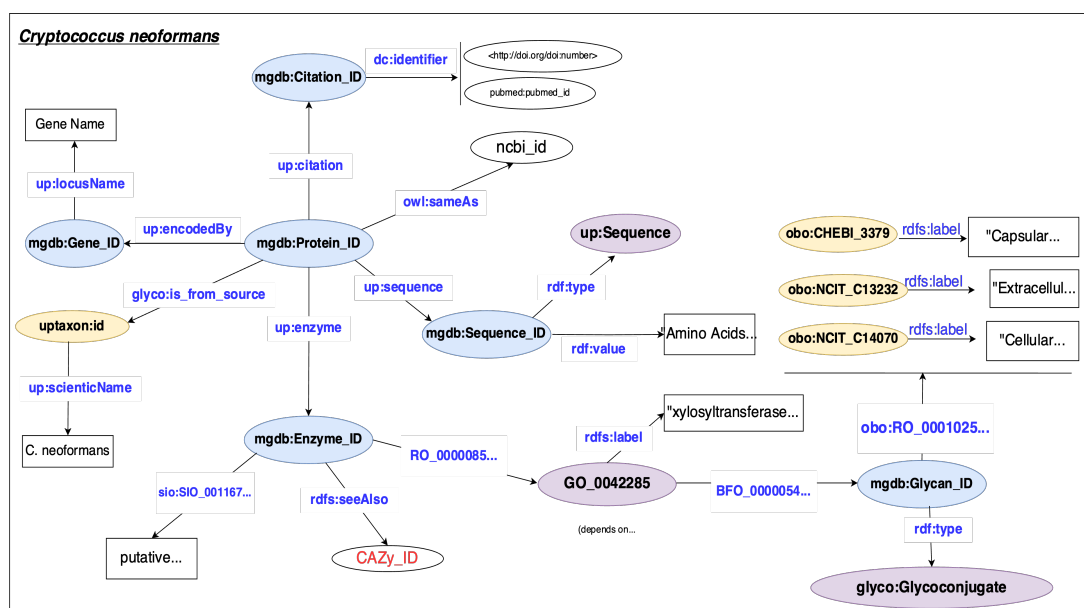


FIGURE 3.14: The schema for *C. neoformans*

However, there is no information showing the participant involved in the enzyme reaction and there is data indicating the anatomical position of glycan created by enzyme activity. I structured the enzyme information and anatomical data so that they could

be linked without misunderstanding by introducing a glycan node connected to the bacterial anatomy (Figure 3.14).

- **SPARQL query**

A SPARQL query was created to identify the intracellular location information of glycans that did not have any information.

```

SELECT ?protein ?ncbi_id ?gene ?fungi_db ?enzyme ?go_function
       ?cazy_uri ?enzyme_comment ?pubmed_id ?taxon_id ?taxon_name
       ?go_loc_id ?location ?seq_id ?aa_seq
FROM <http://localhost:8890/neoformans>
WHERE {
  #protein
  ?protein a up:Protein ;
    owl:sameAs ?ncbi_id ;
    up:encodedBy ?gene ;
    up:enzyme ?enzyme ;
    up:sequence ?seq_id .
  OPTIONAL { ?protein up:citation/dc_term:identifier ?pubmed_id .}
  ?protein glyrdf:is_from_source ?taxon_id.
  ?taxon_id up:scientificName ?taxon_name.
  #enzyme
  ?enzyme a up:Enzyme .
  OPTIONAL { ?enzyme obo:R0_0000085 ?go_function;
    ssci:SI0_001167 ?enzyme_comment.
    ?go_function obo:BF0_0000054 ?glycan.}
  OPTIONAL { ?glycan obo:R0_0001025 ?go_loc_id.
    ?go_loc_id rdfs:label ?location }
  ?enzyme rdfs:seeAlso ?cazy_uri.
  #sequence
  ?seq_id a up:Sequence;
    rdf:value ?aa_seq .
  #gene
  ?gene a up:Gene ;
    owl:sameAs ?fungi_db .
}

```

FIGURE 3.15: SPARQL query for *C. neoformans*

According to the RDF model, the GO molecular activity that is connected to **Enzyme** through the RO ontology's "*has\_function*" property is linked to the **Glycan** resource. To reach the localization data that glycans placed from this **Clycan** node, "*located\_id*" property were used. Based on this RDF schema, the intracellular location of glycans shown in the row data was extracted.

- **ShEx**

```

:Protein
{
  rdf:type [uni:Protein] ; # 100.0 %
  uni:sequence IRI ; # 100.0 %
  glyco:is_from_source IRI ; # 100.0 %
  owl:sameAs IRI ; # 100.0 %
  uni:encodedBy @:Gene ; # 100.0 %
  uni:enzyme @:Enzyme ; # 100.0 %
  uni:citation IRI ?;
      # 20.0 % obj: IRI. Cardinality: {1}
  rdfs:label xsd:string ?
      # 10.0 % obj: xsd:string. Cardinality: {1}
}
:Enzyme
{
  rdf:type [uni:Enzyme] ; # 100.0 %
  rdfs:seeAlso IRI ; # 100.0 %
  seman:SI0_001167 xsd:string ; # 100.0 %
  obo:RO_0000085 IRI ?
      # 78.57142857142857 % obj: IRI. Cardinality: {1}
}
:Gene
{
  rdf:type [uni:Gene] ; # 100.0 %
  owl:sameAs IRI # 100.0 %
}
:Glycoconjugate
{
  rdf:type [glyco:Glycoconjugate] ; # 100.0 %
  obo:RO_0001025 IRI ?
      # 74.60317460317461 % obj: IRI. Cardinality: {1}
}

```

FIGURE 3.16: Verification of *C. neoformans* turtle file

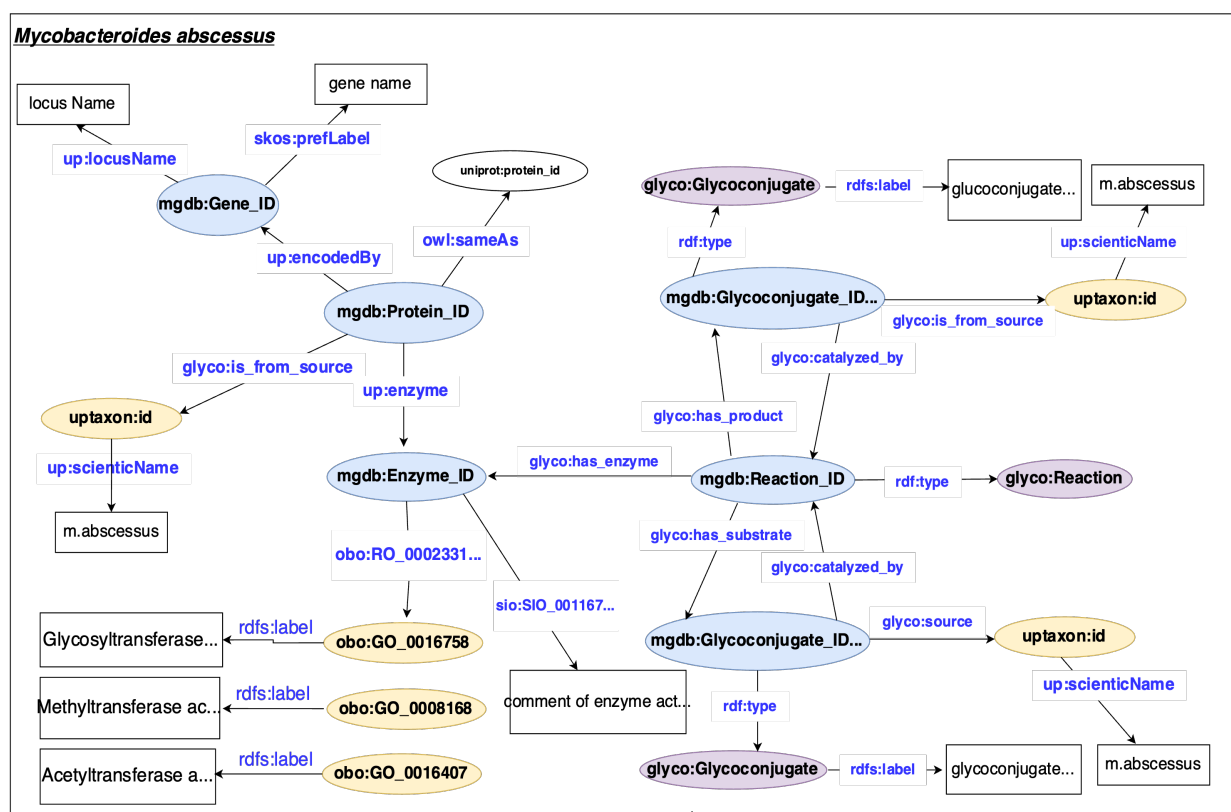


I identified that about 74 percent of **Glycoconjugate** node has a value, which means that not every glycan has localization information. Figure 3.16 shows that in the **Protein** data, only 20 percent has reference information, while in the enzyme data, 78 percent have **Enzyme** activity.

### 3.1.6 *Mycobacteroides abscessus*

- **RDF model**

*M. abscessus* is a non-tuberculosis mycobacterium that causes a variety of illnesses, including lung disease that is accompanied by cystic fibrosis in pulmonary (Esther Jr et al., 2005). This bacteria has demonstrated drug resistance to numerous classes of antibiotics due to the unique structure of the mycobacterial cell wall that consists of peptidoglycan (PG), arabinogalactan, and mycolic acids (Akusobi et al., 2022). I received data about glycopeptidolipids (GPLs) in the outer layer of the cell wall. The loss of GPLs cause the morphologic change of bacteria from smooth to rough and this transition correlated to the virulence (Viljoen et al., 2020). Each enzyme had the glycan information about substrate and product that is represented with CSDB linear code that is one of the text types for glycan nomenclature. Enzyme activity is presented using GO ontology: Glycosyltransferase, Methyltransferase, Acetyltransferase (Figure 3.17).

FIGURE 3.17: The RDF schema for *M. abscessus*

- SPARQL query

*M. abscessus* has enzyme function as well as glycan data. To identify the enzyme activity, *obo:involved\_in* property RO ontology was used. The obtained object values were shown in URI of GO ontology term for example, '[http://purl.obolibrary.org/obo/GO\\_0016758](http://purl.obolibrary.org/obo/GO_0016758)>', and the meaning was retrieved using *rdfs:label*. Reaction information that is prepared for the description of glycan data was queried using *glyco:has\_enzyme* property like *C. jejuni*. Basically, all resources including **Protein**, **Taxon**, **Enzyme**, **Gene** also were queried. Overall, the search results are consistent with RDF data.

```

SELECT ?protein ?uniprot_id ?species ?gene ?locus_name ?gene_label
       ?enzyme ?involved_activity ?enzyme_comment ?reaction ?reactant_glycan
       ?reactant_glycan_text ?product_glycan ?product_glycan_text
FROM <http://localhost:8890/abscessus>
WHERE {
  ?protein a up:Protein ;
    owl:sameAs ?uniprot_id ;
    up:encodedBy ?gene ;
    up:enzyme ?enzyme;
    glyrdf:is_from_source ?taxon.
  ?taxon up:scientificName ?species.
  ?enzyme a up:Enzyme ;
    obo:RO_0002331 ?involved_activity ;
    ssci:SIO_001167 ?enzyme_comment .
  ?gene a up:Gene ;
    up:locusName ?locus_name ;
    skos:prefLabel ?gene_label .
  ?reaction glyrdf:has_enzyme ?enzyme.
  ?reaction glyrdf:has_product ?product_glycan.
  ?product_glycan rdfs:label ?product_glycan_text.
  ?reaction glyrdf:has_substrate ?reactant_glycan.
  ?reactant_glycan rdfs:label ?reactant_glycan_text.
}

```

FIGURE 3.18: SPARQL query for *M. abscessus*

- **ShEx**

In Figure 3.19, we can see that the Reaction resources are connected to the **Glycoconjugate** and **Enzyme Class** using *glyco: has\_substrate* or *glyco: has\_product* and *glyco: has\_enzyme* property, respectively. Additionally, it can be shown that every instance of a glycoconjugate took part in the reaction. All **Protein Class** instances are linked to the **Enzyme** and **Gene Class**. The verification results agreed with the triple data and I loaded the RDF data into the RDF storage.

```

:Protein
{
  rdf:type [uni:Protein] ; # 100.0 %
  glyco:is_from_source IRI ; # 100.0 %
  uni:enzyme @:Enzyme ; # 100.0 %
  uni:encodedBy @:Gene ; # 100.0 %
  owl:sameAs IRI # 100.0 %
}
:Enzyme
{
  rdf:type [uni:Enzyme] ; # 100.0 %
  seman:SIO_001167 xsd:string ; # 100.0 %
  obo:RO_0002331 IRI # 100.0 %
}
:Gene
{
  rdf:type [uni:Gene] ; # 100.0 %
  uni:locusName xsd:string ; # 100.0 %
  skos:prefLabel xsd:string # 100.0 %
}
:Glycoconjugate
{
  rdf:type [glyco:Glycoconjugate] ; # 100.0 %
  glyco:catalyzed_by @:Reaction ; # 100.0 %
  rdfs:label xsd:string # 100.0 %
}
:Reaction
{
  rdf:type [glyco:Reaction] ; # 100.0 %
  glyco:has_substrate @:Glycoconjugate ; # 100.0 %
  glyco:is_from_source IRI ; # 100.0 %
  glyco:has_product @:Glycoconjugate ; # 100.0 %
  glyco:has_enzyme @:Enzyme # 100.0 %
}

```

FIGURE 3.19: Verification of *M. abscessus* turtle file

### 3.1.7 *Mycobacterium tuberculosis*

- RDF model

*M. tuberculosis* (Mtb) is a causative agent for tuberculosis (TB) and has shown multidrug resistance. Unlike ordinary bacteria, Mtb does not possess virulence factor so

they can survive in the host without causing severe illness during the host is non-immunocompromised state (Chai, Zhang, and Liu, 2018). Mtb cell envelope possesses numerous glycolipids such as lipoarabinomannan (LAM), phosphatidylinositol-containing mannosidase (PIMs), lipomannan (LM), etc., which contributes to the pathogenicity of Mtb by promoting the binding to the mannose receptor and entry of the mycobacteria into antigen-presenting cells (Berg et al., 2007). The information about the enzyme was extracted from text values in the column for modification of glycolipids. To present the glycolipid information, a **Reaction** node was inoculated and is described as a participant in the biosynthesis reaction by the enzyme. The glycolipids are presented in CSDB linear code without GlyTouCan identifier (Figure 3.20).

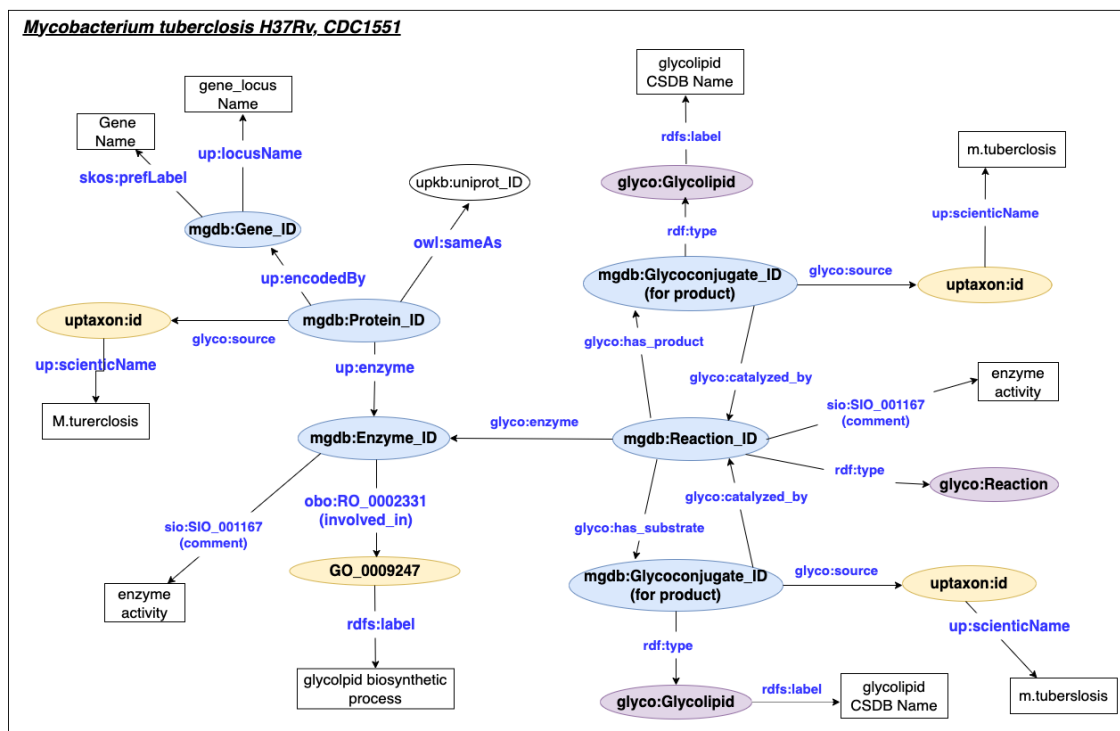


FIGURE 3.20: The schema for *M. tuberculosis*

- **SPARQL query**

SPARQL scripts were generated to query the information of all resources. Figure 3.20 shows a way to reach the interest data from a certain node resource in the graph data. In terms of Reaction node, *M.tuberculosis* and *C. jejuni* use the same RDF schema to describe **Glycoconjugate**. I can therefore identify the object resources including glycans using the same SPARQL query. The OPTIONAL method was used to obtain information from reaction nodes without reactant glycan data, even if the property in the optional pattern, such as *glyrdf:has\_substrate*, does not match. The SPARQL query's results will be supplied to the MicroGlycoDB, a database devoted to microbial glycosylation.

```
SELECT ?protein ?uniprot_id ?taxon ?gene ?locus_name ?gene_label
       ?enzyme ?involved_activity ?enzyme_comment ?reaction ?reactant_glycan
       ?reactant_glycan_text ?product_glycan ?product_glycan_text
FROM <http://localhost:8890/mtb>
WHERE {
  #protein, taxon_id
  ?protein a up:Protein .
  OPTIONAL { ?protein owl:sameAs ?uniprot_id .
             ?protein glyrdf:is_from_source ?taxon. }
  ?protein up:encodedBy ?gene ;
           up:enzyme ?enzyme .

  #enzyme
  ?enzyme a up:Enzyme .
  OPTIONAL { ?enzyme obo:R0_0002331 ?involved_activity .
             ?enzyme ssci:SI0_001167 ?enzyme_comment .}

  #gene
  ?gene a up:Gene .
  OPTIONAL { ?gene up:locusName ?locus_name .
             ?gene skos:prefLabel ?gene_label . }

  #reacton
  ?reaction glyrdf:has_enzyme ?enzyme.
  ?reaction glyrdf:has_product ?product_glycan.
  ?product_glycan rdfs:label ?product_glycan_text.
  OPTIONAL { ?reaction glyrdf:has_substrate ?reactant_glycan.
             ?reactant_glycan rdfs:label ?reactant_glycan_text. }
}
```

FIGURE 3.21: SPARQL query for *M. tuberculosis*

- ShEx

```

:Protein
{
  rdf:type [uni:Protein] ; # 100.0 %
  uni:enzyme @:Enzyme ; # 100.0 %
  uni:encodedBy @:Gene ; # 100.0 %
  glyco:is_from_source IRI ?;
    # 61.53846153846154 % obj: IRI. Cardinality: {1}
  owl:sameAs IRI ?
    # 61.53846153846154 % obj: IRI. Cardinality: {1}
}
:Enzyme
{
  rdf:type [uni:Enzyme] ; # 100.0 %
  obo:RO_0002331 IRI ?;
    # 42.30769230769231 % obj: IRI. Cardinality: {1}
  seman:SI0_001167 xsd:string ?
    # 42.30769230769231 % obj: xsd:string. Cardinality: {1}
}
:Gene
{
  rdf:type [uni:Gene] ; # 100.0 %
  uni:locusName xsd:string ?;
    # 88.46153846153845 % obj: xsd:string. Cardinality: {1}
  skos:prefLabel xsd:string ?
    # 61.53846153846154 % obj: xsd:string. Cardinality: {1}
}
:Glycolipid
{
  rdf:type [glyco:Glycolipid] ; # 100.0 %
  glyco:catalyzeed_by @:Reaction +; # 100.0 %
    # 97.5 % obj: @:Reaction. Cardinality: {1}
  glyco:is_from_source IRI ?;
    # 92.5 % obj: IRI. Cardinality: {1}
  rdfs:label xsd:string *
    # 92.5 % obj: xsd:string. Cardinality: +
    # 90.0 % obj: xsd:string. Cardinality: {1}
}
:Reaction
{
  rdf:type [glyco:Reaction] ; # 100.0 %
  glyco:has_product @:Glycolipid ; # 100.0 %
  glyco:has_enzyme @:Enzyme ; # 100.0 %
  seman:SI0_001167 xsd:string ; # 100.0 %
  glyco:has_substrate @:Glycolipid ?
    # 57.692307692307686 % obj: @:Glycolipid. Cardinality: {1}
}

```

FIGURE 3.22: Verification of *M. tuberculosis* turtle file

RDF triples of *M.tuberculosis* were tested against the ShEx schema. The result demonstrates that Classes like Protein, Enzyme, Reaction, and Gene, as well as instances, belonged to their Class in accordance with the requirements in the schema. Following verification, the RDF data were uploaded to the endpoint (<https://ts.alpha.glycosmos.org/sparql>) that contains all of our study group's RDF graph data.

## 3.2 Pathway repository

### 3.2.1 Inspecting Ontologies

- Gene Ontology (GO)

As mentioned in the introduction part, ontology makes complex concepts to be communicated unambiguously. Ontology development has been proliferated as one of the major strategies in the context of data integration in biology (Blake and Bult, 2006). GO was developed to provide a consistent description of sequences and gene products for data integration between different databases (Consortium, 2004). Almost enzyme information from the raw data was recorded in an unstructured text that had to be transformed using ontologies into standardized data. The enzyme information was provided in the text type of the comment column in the spreadsheet and they contained the function of enzymes. Each enzyme activity was represented using GO ontology based on their comment. GO ontology could not cover all kinds of enzyme activity, the terminology showing representative activity was used. For example  $\beta$ -1,4-mannosyltransferase and  $\alpha$ -1,6-mannosyltransferase were assigned to **mannosyltransferase activity** with GO identifier, GO\_0000030. The glycan that is created, modified, or transferred by enzymes is placed in a bacterial cell envelope composed of a capsule, peptidoglycan layer, outer membrane, etc. Depending on the species, bacterial membrane structures, and their constituent parts differ.



TABLE 3.1: The used terms of ontologies for the glycan-related resources of Microbes

	Ontology label	Identifier ( <a href="http://purl.obolibrary.org/obo/">http://purl.obolibrary.org/obo/</a> )
1	mannosyltransferase activity	GO_0000030
2	chin synthase activity	GO_0004100
3	galactosyltransferase activity	GO_0008378
4	xylosyltransferase activity	GO_0042285
5	glucosyltransferase activity	GO_0046527
6	trehalose synthase activity	GO_0102986
7	dolichol-phosphate mannosyltransferase(yeast)	PR_P14020
8	catalase activity	GO_0004096
9	kinase activity	GO_0016301
10	flippase activity	GO_0140327
11	Kdo transferase activity	GO_0043842
12	methyltransferse activity	GO_0008168
13	phosphoramidate-hexose phosphotransferase activity	GO_0047329
14	isomerase activity	GO_0016853
15	lipopolysaccharide core heptosyltransferase activity	GO_0071967
16	ligase activity	GO_0016874
17	lipopolysaccharide	CHEBI_35371
18	Integral membrane protein	NCIT_C16747
19	Flagellum	BTO_0002292
20	capsular polysaccharide	CHEBI_3379
21	peptidoglycan	CHEBI_8005

The localization of glycan was presented using ontology. Table 3.1 shows the list of

terminology of ontology used in this study. I introduced a **Reaction** Class into GlycoRDF (Ranzinger et al., 2015) to represent the enzyme reaction containing the data of substrate and product glycan. All bacteria contain protein, enzyme, and gene information, and ontology created for the development of the RDF model of the UniProt Database was used.

- Protégé

Protégé is software used for developing and maintaining ontologies (Musen, 2015). It provides a user-friendly interface for users to learn how to use controlled vocabulary that defines the **Classes**, *properties*, and relationships.

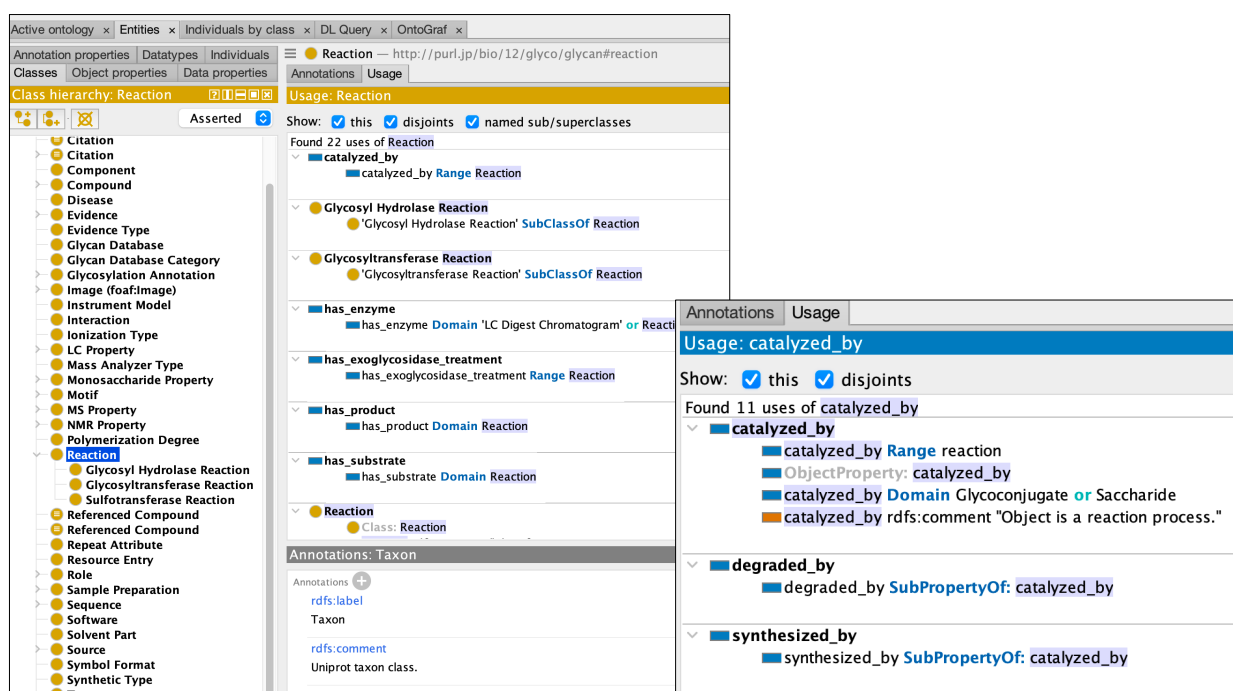


FIGURE 3.23: A screenshot of the Protégé software displaying the definition of controlled vocabulary. In this example, the Reaction class is displayed, showing that it can be used as the Subject of the *catalyzed\_by* predicate, among many others.

In addition, the Protégé has visualization tools that allow users to explore the **Classes** hierarchies and *properties* across the whole structure of its ontology. The owl file

containing the relevant ontology was loaded into Protégé to investigate the definition of vocabularies defined on ontology specification, such as the Uniport ontology for the UniProt knowledge base, BioPAX ontology, and GlycoRDF ontology. I reviewed the definition of the concepts and relationships that exist within each ontology, and the ontologies were applied to the description of the microbial information as indicated in the specification. Figure 3.23 shows the part of the GlycoRDF ontology. The property *catalyzed\_by* specifies that the subject is either the **Glycoconjugate** or **Saccharide** Class and the object (value) is the **Reaction** Class of the RDF triple.

### 3.2.2 Biosynthetic glycosylation pathway

- **Resources for the glycan synthesis process**

The repository's input part for the glycan synthesis process is simple to register. Users can enter reaction information such as glycans, nucleotide sugars, and enzymes, as well as the intracellular location where the reaction takes place. Glycans (Fujita et al., 2021) can be registered using the GlyTouCan identifier number, which was developed to address the difficulties in referencing or nomenclature caused by the glycan structure's complexity. Also linear code, the extended or condensed IUPAC notation, as the standardized notations for glycan structure, can be input for user convenience, and then a GlyTouCan number is assigned to the notation of the text type using the API that is developed by GlyCosmos. The ChEBI (Chemical Entities of Biological Interest) ontology (Hastings et al., 2016) is used to represent the nucleotide sugars that are in charge of the glycan structure's extension. Enzyme vocabularies are provided based on EC numbers (Enzyme commission number), which is a hierarchical classification system for enzymes based on the type of chemical reaction they catalyze, and the UniProtKB enzyme list, which includes proteins expected to have enzymatic activity.

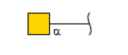
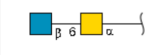

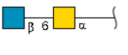
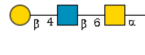




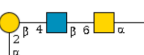
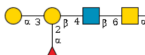

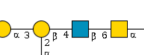
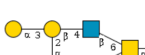

Reaction Table						
RXN ID	Reactant	Enzyme	Sugar Nucleotide	Product	Cell Location	Action
1	 G57321FI	N-Acetyl-D-glucosamine		 G00041MO	extracellular region	
2	 G00041MO	D-Galactose		 G61730RY	extracellular region	
3	 G61730RY	D-Xylose		 G48440GN	extracellular region	
4	 G48440GN	D-Galactose		 G76362LL	extracellular region	
5	 G76362LL	D-Galactose		 G94595ZZ	extracellular region	

FIGURE 3.24: The input table for a description of glycan synthesis pathway (Core 2 O-glycan of human)

According to the instruction in the BioPAX ontology, the pathway is described by chaining the reactions that are linked to the next reaction, which contains the product of the previous reaction as a reactant. To prevent input errors, a modal window is provided with the reaction number and reactant information pre-filled. After completing the input reaction, users can identify the resources they input, including the glycan structures, which are displayed in the Symbol Nomenclature For Glycans (SNFG) format, which was developed to aid in the recognition of the complex glycan structure.

Users can also correct or remove their inaccurate info (Figure 3.24).

- **Visulaization of a glycan synthesis pathway**

I used GoJS, a JavaScript library for creating diagrams, to display the extension, which is a process of the glycan string by enzymes that add a sugar nucleotide to an acceptor glycan. The required data for the graphic view was obtained from a relational database, which was created to store input data. From this view, users can confirm which enzyme is responsible for glycosidic linkage and get a quick overview of the glycosylation pathway of the glycan structure (Figure 3.25). Following confirmation, the user can choose the next step: update the input data or change it into RDF data.

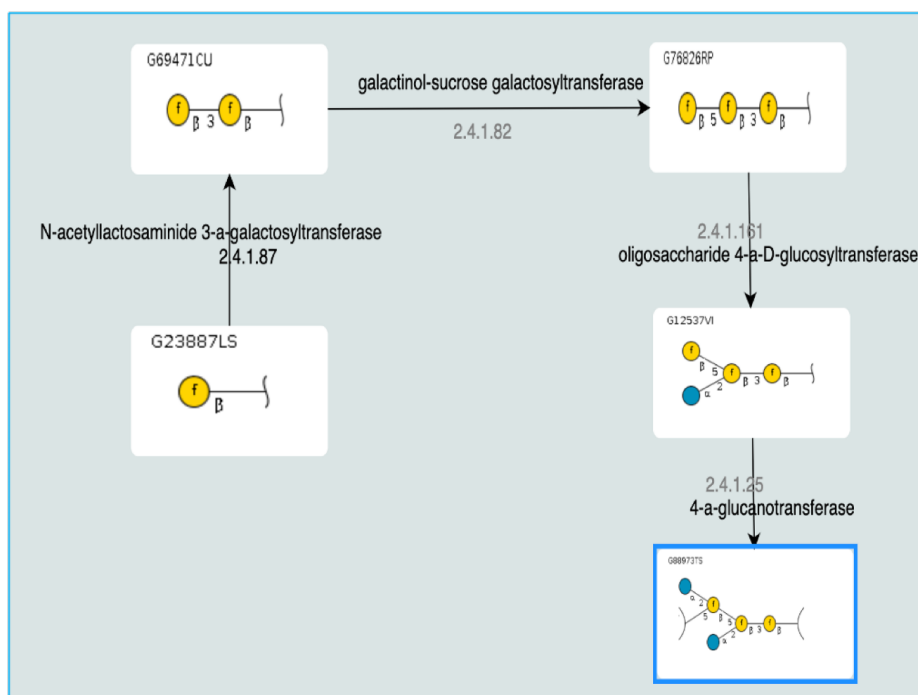


FIGURE 3.25: The graphic view for input data of glycan synthesis pathway

- **Search table for glycan synthesis pathway**

After confirming their input data in the pathway view page by clicking the RDFication button, users proceed to the table page that displays the pathway list, which is formed of processes involving participant glycans and enzymes. Each item of pathway and glycan in the table page has a link to the corresponding detail page in this repository and the GlyCosmos, respectively. Users can look for enzymes that contribute to glycan structure or diseases in which glycan is involved (Figure 3.26).

**Pathway Search**

Show 10 entries

To search, enter a keyword in the text box and press Enter (return). [Show hidden columns](#)

Pathway Name	Date	Species	Involved Glycans	Involved Enzymes	Related Disease
<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>			<input type="text" value="Search"/>
<a href="#">E.coli 0-157</a>	No data	Escherichia coli	<a href="#">G00056M0</a> <a href="#">G22979LQ</a> <a href="#">G83866MJ</a>	No data	No data
No data	No data	No data	No data	No data	No data
<a href="#">Core-2 antigen</a>	No data	Homo sapiens	<a href="#">G00041M0</a> <a href="#">G48440GN</a> <a href="#">G61730RY</a> <a href="#">G76362LL</a> <a href="#">G94595ZZ</a>	No data	No data
<a href="#">02 antigen</a>	No data	Escherichia coli	<a href="#">G04602LA</a> <a href="#">G49814WC</a> <a href="#">G77753YK</a>	No data	No data

Showing 1 to 4 of 4 entries

[Download the displayed table \(.tsv\)](#) [Download all \(.tsv\)](#)

Previous **1** Next

Line breaks are removed from cells.

FIGURE 3.26: The search table of glycan synthesis pathway

The table is created with Web components provided by GlyCosmos developers, which allow a SPARQL query result to be visualized in an HTML web page in a customized table format. I generated a SPARQL query to extract the pathway information from the triplestore. The table provides the function of keyword search.

### 3.2.3 Glycan related protein pathway

I developed a semantic web technique-based repository, GlycoPathwayRepo. The repository was designed to provide easy-to-use tools that encourage end users to participate in the creation of glycan-related pathway data without requiring technical knowledge. Depending on the number of resources, two input types are provided: one is the glycan synthesis pathway, which focuses on the enzyme contributed to the glycan linkage, and the other is the protein pathway, which is devoted to the protein pathway, which involves many different types of resources such as protein, lipids, glycans, and so on. Basic information for describing pathway data, such as species, tissue, pathway category, name, and comments, is common in both pathway branches. The ontology that was utilized to represent common resources is displayed in the table 3.2.

TABLE 3.2: The used ontologies for background information in Pathway data

Ontology	Kind of resources (URI)
Pathway ontology (PO)	Pathway category ( <a href="https://bioportal.bioontology.org/ontologies/PW">https://bioportal.bioontology.org/ontologies/PW</a> )
Tissue ontology (BTO)	Tissue ( <a href="https://bioportal.bioontology.org/ontologies/BTO">https://bioportal.bioontology.org/ontologies/BTO</a> )
NCBI Organismal Classification	Species ( <a href="https://bioportal.bioontology.org/ontologies/NCBITAXON">https://bioportal.bioontology.org/ontologies/NCBITAXON</a> )
Mondo Disease Ontology (MONDO)	Disease ( <a href="https://bioportal.bioontology.org/ontologies/MONDO">https://bioportal.bioontology.org/ontologies/MONDO</a> )

- **User interface for the resource inputs**

In comparison to glycan synthesis data, inputting protein pathway data requires more information for entity items such as proteins, lipids, complexes, and so on. To cover diverse resources, the needed vocabularies were prepared from ontologies such as Cell

ontology (Meehan et al., 2011), GO complex (Botstein et al., 2000), ChEBI (Hastings et al., 2016) and so on. The provided resources are different in accordance with reaction components such as reactant, controller, and product. For example, there are two types of resources available to controllers controlling reactivity, such as enzymes and proteins. Proteins and complexes, on the other hand, are resources for products because they are the result of specific modification, binding, or cleavage reactions, and the same product participates in the next reaction as a reactant or controller, which is executed automatically by the system, so the user does not have to enter it manually.

Whenever the input of the entity resource of reaction participants is completed, the information is converted into RDF data and saved to the triple storage under the specific graph name by SPARQL. Massive SPARQLLists were generated to transform the various types of resources, reactions, and pathways into triple data (<https://gpr-sparqlist.alpha.glycosmos.org/sparqlist/>). The figure shows an example of a SPARQL query to generate triple sentences (Figure 3.27). The generated triples are saved into the database using the **"INSERT DATA"** function of the SPARQL language.

```
INSERT DATA
{
  GRAPH <http://gpr.pathway.org/glycan_synthesis_repo_test>
  {
    pathway:GC-RXN-{{reaction_id}} a bp:BiochemicalReaction .
    {{#if reactant_toid}}
      pathway:GC-RXN-{{reaction_id}} bp:left pathway:GC-Saccharide-{{reactant_toid}}.
      pathway:GC-Saccharide-{{reactant_toid}} a bp:SmallMolecule ;
      rdf:type glycordf:Saccharide ;
      bp:xref pathway:GC-UnificationXref-{{reactant_toid}} ;
      bp:displayName "{{reactant_name}}"^^xsd:string .

      pathway:GC-UnificationXref-{{reactant_toid}} a bp:UnificationXref;
      bp:db "GlyTouCan"^^xsd:string;
      bp:id glytoucan:{{reactant_toid}}.
    {{/if}}
  }
}
```

FIGURE 3.27: A part of SPARQL query for RDFication of protein data gained by user input



Depending on the input information of the reactant, product, or controller, which are regulatory molecules responding to enzyme activity, the biochemical reaction can be displayed as a binding reaction, breakdown reaction, or enzymatic reaction and each reaction must be entered into the following order: reactant, controller (if available), and product.

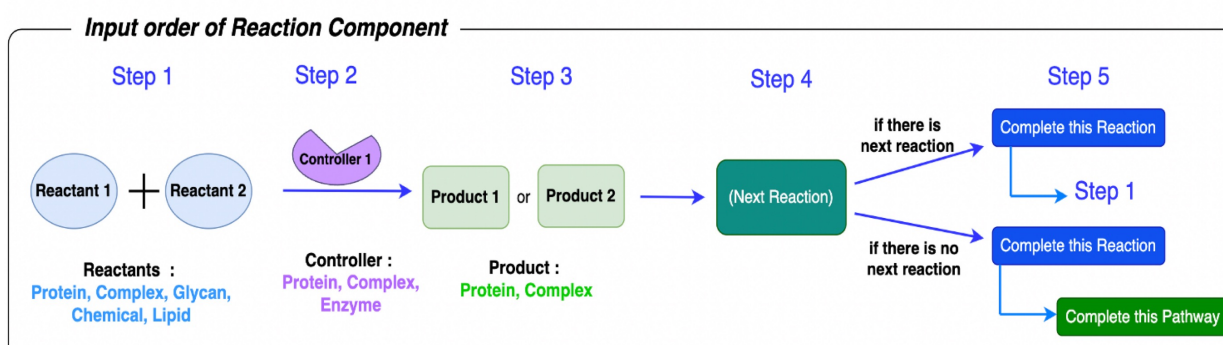


FIGURE 3.28: The input order for reaction information

To reduce the input error, a select button is provided in the final step of each reaction that indicates whether the product of the current reaction will be the reactant or controller of the next reaction, so the user does not need to enter information about the reactant or controller of the next reaction, with the exception of the initial reaction. Although it may not be convenient for the user at the present stage of development, the provided diagram will guide the order of information input for accurate data input. Users can confirm the input data for each reaction on the right side of the input page after completing the input data for one reaction (Figure 3.28).

- **Visualization of pathway data**

Whenever the users finish inputting information about each reaction set of the pathway, the input result appears in the confirmation table on the right side of the input table and accumulated until the last reaction is completed. Once the users complete entering the information about all reactions within the pathway, users will be directed to the pathway visualization page, and the pathway result is displayed on this page. Simple images make it easier to visualize pathway data that includes complex concepts such as receptor-ligand binding, protein complex formation, phosphorylation, and so on. I used Cytoscape.js, a JavaScript visualization library, to illustrate the pathway data. Among the numerous forms of graphs, I made use of SBGN (Systems Biology Graphical Notation) graphs intended for depicting biological processes. Proteins, complexes, and small molecules are presented in differentiated shapes, and the progression of reactions was delineated with lines and arrows based on reaction types such as stimulation, inhibition, or catalytic reaction by the enzyme (Figure 3.29).

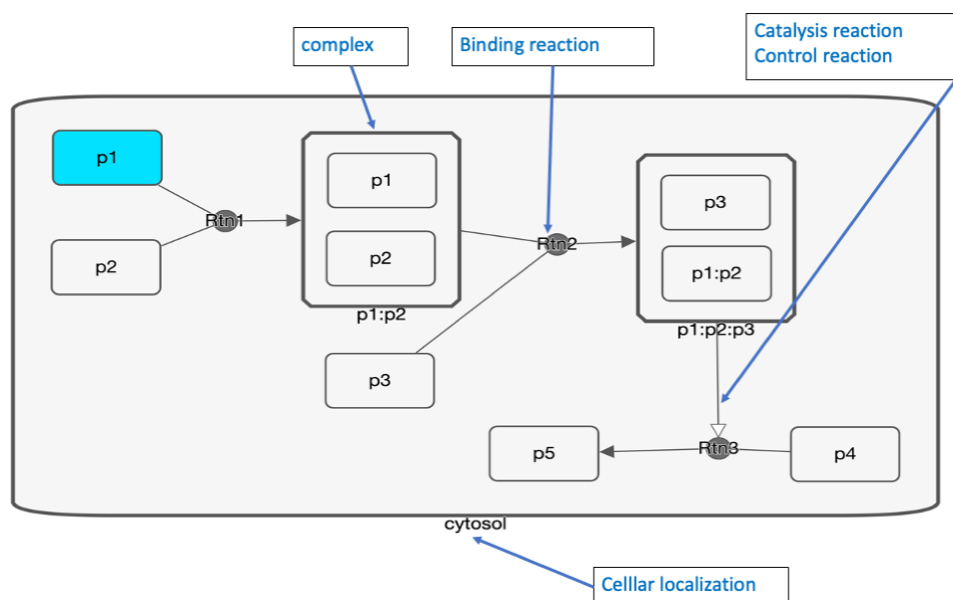


FIGURE 3.29: The display of pathway data input information

- Search Table for protein pathway data

The SPARQL query was used to get information about the pathway list containing the reaction information showing participant molecules (Appendix A).

The screenshot shows a web interface titled "Pathway Search". At the top, there is a search bar and pagination controls showing "Page 1 of 4". Below the search bar is a table with the following columns: Pathway ID, Pathway Name, Taxon, Reaction Number, Reactants, Controller, and Products. The table contains several rows of data. Two rows are highlighted with pink and blue boxes and labeled "Pathway components" and "Reaction components" respectively.

Pathway ID	Pathway Name	Taxon	Reaction Number	Reactants	Controller	Products
GC-Pathway1			1_GC-BiochemicalReaction1			
GC-Pathway2	8_23_test	Homo sapiens	2_GC-BiochemicalReaction1	p50_1,p65_1		
GC-Pathway2	8_23_test	Homo sapiens	2_GC-BiochemicalReaction2	p50_1,p65_1,p65_1	2.4.1.5	p50_1,p65_1,p50_1,p65_1,p65_1
GC-Pathway318	spargl test		318_GC-BiochemicalReaction1	p50,p65		p50,p65
GC-Pathway318	spargl test		318_GC-BiochemicalReaction2	lkb		p50,p65,p50,p65;lkb
GC-Pathway318			318_GC-BiochemicalReaction4	p50		p65
GC-Pathway332			332_GC-BiochemicalReaction1	CSK,PAG1		PAG1-CSK
GC-Pathway332	TCR test	Homo sapiens	332_GC-BiochemicalReaction2	CD4		PAG1-CSK;PAG1-CSK;CD4
GC-Pathway332	TCR test	Homo sapiens	332_GC-BiochemicalReaction3	TCR complex		TCR complex
GC-Pathway333	cyto test	Homona magnanima	333_GC-BiochemicalReaction1	p50,p65		p50,p65

FIGURE 3.30: The search table for glycan-related pathway data

The registered pathway data in the RDF triple storage are retrieved using the SPARQL query and presented in tabular format. The pagination table is prepared using Web components of Metastanza developed by DBCLS. Each column has a search field with autocomplete or sorting features. The user who wants to see specific information about a pathway can jump to the corresponding pathway view page by clicking the pathway identifier in the row field. Users are able to identify which reactions are contained in which pathways, as well as which resources are involved in each reaction (Figure 3.30).

# Chapter 4

## Discussion

### 4.1 Semantic Data of Microbial Glycosylation

The glycosylation data, which includes information regarding the genes, proteins, enzymes, and glycan structures associated with bacteria, has been transformed into a standardized format, based on the data provided by co-researchers. The RDFized resources were used as data for the MicroGlycoDB database.

Since the discovery of glycosylation machinery for *N*-linked protein glycosylation in *C. jejuni* has been reported, it has been known that glycosylation is ubiquitous in microbes including the archaea domain. (Szymanski et al., 1999). The pathogenic bacteria including *Neisseria meningitides* (Stimson et al., 1995), *Haemophilus influenzae* (Grass et al., 2003), *Campylobacter jejuni*, and *Mycobacterium tuberculosis* (Dobos et al., 1996) have evolved several strategies to invade host cell and survive. As one of the mechanisms, bacteria have developed numerous structural factors on their surface such as capsules, and cell wall containing peptidoglycan and teichoic acids that play crucial roles in host colonization and resistance to  $\beta$ -lactam drugs in Gram-positive bacteria such as *Bacillus subtilis* and *Staphylococcus aureus* (D'Elia et al., 2006; Brown et al., 2012). *S. aureus* is a Gram-positive bacterium causing infections in the skin and soft tissue.

The evidence showing the importance of the bacterial polysaccharides in interactions with the host not only in pathogenic bacteria but also in non-pathogenic microorganisms has been accumulated (Latousakis and Juge, 2018; Yakovlieva, Fülleborn, and Walvoort, 2021; Khan et al., 2022). The accumulated information has led to the development of a number of microbes-oriented databases to provide a better understanding of the pathology or physiology of microorganisms.

There are databases devoted to bacterial information such as MicrobeDB.jp (<https://microbedb.jp/>) and BacDive (Reimer et al., 2022), which provide morphologies, culture conditions, metabolism, genetic information, etc., and CSDB (Toukach and Egorova, 2019) is a representative database for prokaryotic and fungal glycans at the present time that offers glycosyltransferase activity of enzymes, structures, and NMR-spectroscopic information of microbes. These databases have not been designed to take into account a semantic description of the data using ontologies and lack information on glycosylation.

The MicroGlycoDB (<https://microglycodb.alpha.glycosmos.org/>) database has been developed to organize fragmented data from unpublished glycan research or unstructured data in a publicly accessible database and to facilitate data integration with other relevant data such as illness, route, or phenotype. The information about microbial glycosylation gained from co-researchers ranges from bacteria to fungi. When compared to mammalian glycan, the structures of glycans in bacteria and fungi displayed great diversity and distinctive composition in the structure such as Kdo (Lipopolysaccharide 3-Deoxy-d-mannooctulosonic Acid), Leg (Legionaminic acid), and Neuraminic acid (5-amino-3,5-dideoxy-D-glycero-D-galacto-non-2-ulosonic acid) (Khan et al., 2022), which have evolved by interacting with the host or environment over time. These distinctive glycan structures made it difficult to adequately describe enzyme activity that does not have information about substrate or product glycans using the controlled vocabulary. Because the enzyme information that we have obtained is not included in the enzyme list of enzymatic reactions classified by the

IUBMB (International Union of Biochemistry and Molecular Biology). To assign a URI to enzyme activity, the enzyme-related activity labeled the term that is attached to the name of the molecule with the function of the molecule such as "Pse5Ac7Am transferase", "UDP-N-acetyl-alpha-D-glucosamine C6 dehydratase", or "D-alanine-D-alanine ligase" were described using the GO ontology. However, because those particular molecules in microorganisms cannot be covered with the molecular function of GO ontology, a new ontology will need to be created that can describe the enzyme activity regardless of the types of microbial molecules. The anatomy information describing the locations of glycans in *C. jejuni* and *C. neoformans* such as flagellum, the capsule was represented using the cellular component of GO ontology. The various glycoconjugates such as phosphatidylinositol mannosidase (PIMs), lipomannan (LM), and lipoarabinomannan (LAM) were provided in a linear notation of CSDB database in the case of *M. tuberculosis*. Although it is desirable for glycan resources to represent by using an identifier of GlyTouCan repository because glycan structures can be referred to across various databases or literature without confusing (Fujita et al., 2021), glycoconjugates possessing lipids cannot be supported by this repository. As a result, the information on glycan structures was unavoidably presented in the text type.

To standardize the information on microbial glycosylation, several ontologies were used ranging from objects such as proteins, enzymes, and glycans to concepts including molecular activity. The needed ontologies can be extensively searched using Ontology Lookup Service (OLS) (Jupp et al., 2015) and Ontobee to find out the appropriate usage of vocabularies (Xiang et al., 2011). However, it was challenging to obtain a GlyTouCan identifier from the CSDB linear code of glycolipids and to describe enzyme activity expressed in a molecular function that did not contain information about a substrate and product, such as "Heptosyltransferase II." It was also not easy to describe microorganisms' distinctive architecture and glycan layers using the existing ontologies. To accommodate increasing amounts and types

of resources saved in the MicroGlycoDB database, it is considered that a new ontology tailored to our demands is required. The standardized resources eventually allow researchers to obtain the answer to more complicated queries, through data integration with other semantic data. I expect that the MicroGlycoDB will accumulate reliable information demonstrating the critical role of glycans between hosts and microbes. Also, this will be able to provide a comprehensive resource for researchers working on microbial glycosylation with the results that were gained from the inference that generates new knowledge across organisms, to assist in hypothesis testing. Inference could also be used to generate new knowledge across organisms, to assist in hypothesis testing.

Antibiotic resistance mechanisms have evolved in a large number of bacteria, including the transfer of resistance genes via plasmids, bacteriophages, and free DNA from dead bacteria; active efflux system of pumping out of the cell; change in outer membrane permeability; modification of drug target such as peptidoglycan structure, and alteration of protein synthesis via RNA polymerase (Van Hoek et al., 2011). Antibiotics that target common bacterial components, such as the efflux system or RNA polymerase, are not specific to bacteria species so they even get rid of a potentially beneficial commensal. On the other hand, the unique glycan generated by a distinct glycosylation system in a strain-specific way plays an important role in infection as a virulence factor (Yakovlieva, Fülleborn, and Walvoort, 2021). This suggests that various glycosylation products and associated enzymes may be possible targets for tackling antibiotic resistance challenges. For example, negatively charged LPS due to phosphoryl groups, heptose, and carboxyl groups on the Kdo can decrease antibiotic resistance through a change of charge density via glycan modification such as introducing positively-charged moieties (Imperiali, 2019).

Information on the biosynthetic pathway of unusual bacterial glycans, such as pseudaminic acid (Pse), legionaminic acid (Leg), Rhamnose (Rha), 3-deoxy-D-manno-oct-2ulosonic acid (Kdo), etc., will be critical to understanding their physical role and will be helpful in the

development of tools such as the bacterial glycan array to investigate pathogen-host interaction.

## 4.2 Semantic Data of Glycan-related Pathway Data

Advanced high-throughput techniques allow us to investigate the function of molecules such as DNA, RNA, proteins, and glycans at a global scale under the given condition (Eichler, 2019; Tang et al., 2009; Cui, Cheng, and Zhang, 2022; Ruhaak et al., 2018). However, a long list of differentially expressed genes or proteins cannot provide insights into the biological processes to understand the cellular behavior relevant to the change of cellular phenotype (Bauer-Mehren, Furlong, and Sanz, 2009; Rodchenkov et al., 2020). Thus numerous pathway databases have been developed to comprehend the complex interactions, relationships between biomolecules, and regulatory mechanisms within the diverse cellular processes (Raman and Chandra, 2009; Croft et al., 2014). However, pathway knowledge is dispersed in the scientific literature as well as databases; also, each database has its own naming conventions, contents, data format, and database schema for saving data based on its research focus. To integrate biological data scattered in scientific publications and the individual database with multiple formats, the biology community has applied Semantic Web Technologies to realize the Semantic Web, which is designed for computers to understand and is based on standards such as RDF and SPARQL (Kuck, 2004).

One of the most significant aspects of data integration is to standardize the data (Lapatas et al., 2015). In biology, there are multiple ways to represent similar data, where the same entity can be described by different names in several pathways databases, or vice versa. For example, *Mus musculus* is the scientific name for mouse and some oncogenes such as c-Myc, Ras, and HER2 are often referred to by their protein product names. Also, unlike DNA and proteins, the complex structure of glycans may have numerous different text representations for it (Aoki-Kinoshita, 2019). Because pathway data is composed of a variety of



resource types from diverse sources, the application of ontologies that define vocabularies in a formal way to represent resources is essential for standardizing data and proliferating it across communities (Bard and Rhee, 2004). As the efforts for dissemination of use and development of ontologies, important initiatives such as the NCBO (National Center for Biomedical Ontology) BioPortal and the OLS (Ontology Lookup Service) are providing Web services that facilitate the adoption and implementation of ontologies (Jonquet et al., 2011; Jupp et al., 2015). The data format is another significant factor in the standardization of pathway data. The data in this glycan pathway repository were represented using the Semantic Web standard. It is thought that data structured into standardized formats will be valuable in discovering new knowledge. For example, using semantic web technologies, the Bio2RDF project demonstrated how to create an RDF data model that enables interoperability and integration between different biological datasets from UniProt (Consortium, 2007), OMIM (Hamosh et al., 2005), Entrez Gene (Maglott et al., 2007), KEGG (Kanehisa and Goto, 2000), Gene Ontology (Botstein et al., 2000), and constructed a new knowledge base (Belleau et al., 2008).

The BioPAX project (Demir et al., 2010) was developed to allow pathway data to be shared and exchanged. The current level 3 can support the description of various types of pathways such as metabolic, signaling pathways, gene regulatory networks, and so on. Not only the signaling pathways initiated by chemicals or proteins, but the gene regulatory networks eventually include an event of translocation of transcription factors and DNA binding. To describe template-directed reactions or transcription factor activation, **DNA** and **RNA** *classes* are required. In the future, these classes will be defined for use in my pathway repository, which will allow users to define signaling pathways including information about the signal initiator to the final phenotype of the signaling pathway.

This protein pathway repository was created to contribute to data integration with other pathway data and reflect the more specific information of glycans in pathways because the

most popular pathway databases including the Reactome (Fabregat et al., 2018), WikiPathways (Martens et al., 2021), and KEGG (Kanehisa et al., 2021) have not provided information indicating the roles that the glycans or glycoconjugates play in pathways. Thanks to the unique accession numbers assigned by the GlyTouCan repository system, researchers can reference glycan structures without ambiguity, and glycan data can be linked across numerous databases (Fujita et al., 2021). Also, using the glycoconjugate ontology (GlycoCoO), it is possible to describe glycoconjugate structures, which are composed of covalently bonded oligosaccharides with proteins and lipids (Yamada et al., 2021). Leveraging the GlyTouCan and GlycoCoO ontologies, it will be possible to describe glycan modifications on molecules, and this pathway repository could contribute to making up for the deficiency of the glycan information such as epidermal growth factor receptor (EGFR) (Kaszuba et al., 2015; Azimzadeh Irani, Kannan, and Verma, 2017), T cell receptor (TCR) (Pereira et al., 2018; Zhang et al., 2017) in the major pathway databases.

The existence of biological pathway data in a formal format enables easy integration with data from other pathway databases using SPARQL queries. I created a pathway repository to capture the pathway process and the participant resources in a formal format. These semantic data make it easy to visualize and manipulate pathway data.

The pathways in Wikipathways are accumulating based on a collaborative platform, which means they benefit from the collective knowledge and expertise of a diverse user base (Martens et al., 2021). They also support integration with a variety of other databases and resources, including Gene Ontology, PubMed, and other route databases. However, they may not provide comprehensive pathways covering all kinds of pathways because of the lack of uniform standards for resource representation including annotations and visualization.

There is a need to encourage general biologists to participate in creating a new knowledge base because pathway data is constantly revised with new information, additional updates,

and corrections of inappropriate content. Also, our pathway repository has been designed to allow users to enter data without previous experience in bioinformatics, and this will lower the barrier to adding their pathway knowledge.

### 4.3 Future work

- **Introducing *Classes* for data extension**

Through the glycan biosynthesis branch menu in the GlycoPathwayRepo, users can register information about the glycan synthesis process that can be described using sugar nucleotide, substrate glycan, and enzyme. A **Modulation** Class will be introduced from BioPAX to express the influence of the drug or chemical on the glycosylation reaction that is processed by an enzyme activity such as inhibition or activation. In consequence, the enzyme resources in this repository can be linked to or expanded with information about pharmacological data.

- **Preparing user interface for enzyme kinetics**

The enzyme reaction has been represented with the Michaelis-Menten equation, which represents enzyme activity using  $V_m$ , the maximum reaction rate constant, and  $K_m$ , which is a Michaelis-Menten constant showing the affinity for a particular substrate including glycans. These reaction constants are determined through experiments that measure the reaction rate versus substrate concentration and are significant factors in reaction prediction. However, in the case of enzymes by *in silico* analysis, there is a lack of information on experimentally proven enzyme activity, so research has been undertaken to predict kinetic parameters for enzyme processes. To provide quantitative information on the catalysis event of enzymes, relevant ontologies, and user interfaces will be prepared.

- **Extension of ontology for inference**

In the Semantic Web, inferencing is a powerful tool for deriving new information from existing data using reasoning. The Semantic Web provides a formal specification that defines the meaning of the terms. For example, "**A** is a *subClassOf* **B**" in RDF Schema means that "every member of Class **A** is also a member of Class **B**". Also, because to define **Class** by describing the individuals the **Class** can contain and *rdfs:domain* or *rdfs:range properties* make implicit relationships and connections, which data will be regarded as if it had been linked even if data aren't directly connected in a graph. With this feature, I will make an ontology that allows us to infer the species or strain to which a glycan belongs.

- **Update of O-antigen data and upload to the MicroGlycoDB**

The ECODAB database containing well-organized information on *E. coli* O-antigen, including commensal and pathogenic strains, provides information on the structure of the O-antigen, glycosyl transferase genes, glycoenzymes, references, etc on their individual pages. This information was RDFized and then provided for the pathway data of the GlyCosmos portal site (<https://glycosmos.org/>). However, information on 18 O-antigen is missing from the database in the present year 2023 because the O-antigen list in the ECODB database has not been updated since 2017. I am preparing the information about the omitted O-antigens to be collected from the literature (Liu et al., 2020) and then RDFized. When this process is completed I will add the RDFized O-antigen data to MicroGlycoDB (<https://microglycodb.alpha.glycosmos.org/>) database. Also, The previously RDFized data will include additional information such as PubMed identifier numbers for references and URIs for individual view pages of O-antigen that is including more specific information.

- **Data sharing with WikiPathways**

GlycoPathwayRepo is a repository that is possible to increase the amount of information through the participation of users. Because our collection focuses on the glycosylation of lipids or proteins, when the backbone processes of well-known biological pathways are available, users can easily enter glycosylation modification information. The wikiPathways is continuously accumulating pathway data through community participation and the data that is described with WP ontology can be obtained in their endpoint (<https://sparql.wikipathways.org/>) with a SPARQL query using common resources such as proteins.

## Chapter 5

### Conclusion

I presented a method for RDFication of unstructured data in spreadsheets, by which triple sentences were generated to describe glycan-related data in microbes such as *B. bifidum*, *B. longum*, *C. jejuni*, *M. abscessus*, *M. tuberculosis*, *C. neoformans*, and *E. coli*. The method will enable biological researchers to transform their data into semantic data without having any knowledge of bioinformatics, which contributes to data integration by the community. The triple results were provided for the MicroGlycoDB database.

For enzyme activity lacking a reference identifier number from an external database, it was looked for utilizing an ontology search service such as the Ontology Lookup Service (OLS). However, it was difficult to precisely describe all of the enzyme activity using the identifier number of ontology because there are no currently available ontologies that can describe the enzymes involved in complex microbial glycoproteins, glycolipids, glycoconjugates, and so on. Thus, the development of ontology allowing particular glycan structures or enzyme activity of microbes to be described in formalized manner is required to express a more detailed description that is specialized in microbial glycosylation such as bacterial anatomy, and glycan structure.

I have developed a repository named GlycoPathwayRepo, where users can register their pathway data in a formal format without installing any software or learning about entry methods until they reached the results view page. In addition, the registered data was evaluated for the transformation into semantic data through inspection of (i) the transformation into RDF from the input text, and (ii) the access and retrieval of the pathway data.

Pathway databases will be convincingly accepted as a knowledgebase after a lot of molecular components including glycans or lipids contained in the pathway data of interest, the environmental factors such as cell type or tissue, and interactions between them can be expressed. In this regard, the integration of the fragmented pathway information from different pathway databases is a crucial and inevitable task. As one of the ways to integrate pathway data, I unified the data format and standardized the naming and description of resources using Semantic Web techniques, which will facilitate the construction of a system for collecting and combining the data. I would like to emphasize that semantic data generated by standardization can accelerate data integration between pathway databases.

# Appendix A

## Appendices

### A.1 SPARQL query for protein pathway table

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bp: <http://www.biopax.org/release/biopax-level3.owl#>
PREFIX glyrepo: <http://pathway.org/biopax/pathway#>
SELECT DISTINCT ?pw ?pw_name ?taxon_name ?reaction
              (GROUP_CONCAT(DISTINCT ?re_name ; separator = ",") AS ?reactant_names)
              ?enzyme
              (GROUP_CONCAT(DISTINCT ?prod_name ; separator = ",") AS ?product_names)
FROM <http://gpr.pathway.org/pathway_repo_test>
WHERE {
    ?pw rdfs:type bp:Pathway.
    ?pw bp:displayName ?pw_name .
    ## taxon
    OPTIONAL {
        ?pw bp:organism ?biosource .
        ?biosource bp:displayName ?taxon_name. }
    ## reaction components
    ?pw bp:pathwayOrder ?pathStep .
    ?pathStep bp:stepConversion ?reaction .
    ?reaction bp:left ?reactant;
              bp:right ?product.
    OPTIONAL {
        ?reactant bp:displayName ?re_name .
        ?product bp:displayName ?prod_name . }
    ## controller
    OPTIONAL {
        ?pathStep bp:stepProcess ?catalysis .
        ?catalysis bp:controller ?control_node .
        ?control_node bp:displayName ?enzyme. }
}
ORDER BY ?pw ?reaction
{{limit}}
OFFSET {{offset}}

```



The SPARQL query upper is an example of one of the lists of SPARQL queries I created for this study. Other SPARQL queries can be identified on this site (<https://gpr-sparqlist.alpha.glycosmos.org/sparqlist/>).

## A.2 Source codes

This is the source code for RDFization of *M. tuberculosis*. The Python code for other microbes is uploaded to this site (<https://gitlab.glyco.info/glycosmos/microglycodb/-/tree/master/RDFication>).

LISTING A.1: The python source code for RDFization of *M. tuberculosis*

```
1 #import pandas as pd
2 import csv #for handling csv and csv contents
3 from rdflib import Graph, Literal, RDF, URIRef, Namespace #basic RDF handling
4 from rdflib.namespace import XSD #most common namespaces
5 import urllib.parse #for parsing strings to URI's
6
7 # read in csv file
8 f = open('m_tuberculosis_lsm.csv', 'rt')
9 data = csv.reader(f, delimiter=',') #quotechar='', quoting=csv.QUOTE_MINIMAL)
10 header = next(data)
11 # define the graph 'g' and namespaces
12 g = Graph()
13 rdfs = Namespace('http://www.w3.org/2000/01/rdf-schema#')
14 owl = Namespace('http://www.w3.org/2002/07/owl#')
15 skos = Namespace('http://www.w3.org/2004/02/skos/core#')
16 obo = Namespace('http://purl.obolibrary.org/obo/')
17 dcterms = Namespace('http://purl.org/dc/terms/')
18 taxon = Namespace('http://purl.uniprot.org/taxonomy')
19 glyco = Namespace('http://purl.jp/bio/12/glyco/glycan#')
20 sio = Namespace('http://semanticscience.org/resource/')
21 orth = Namespace('http://purl.org/net/orth')
22 up = Namespace('http://purl.uniprot.org/core/')
23 upkb = Namespace('https://www.uniprot.org/uniprotkb/')
24 brenda = Namespace('https://www.brenda-enzymes.org/enzyme.php?ecno=')
25 pdb = Namespace('https://www.rcsb.org/structure/')
26 pubmed = Namespace('https://pubmed.ncbi.nlm.nih.gov/')
```

```
27 mgdb = Namespace('http://purl.jp/bio/12/database/microglycodb/')
28
29 # create the triples and add them to the graph 'g'
30 for row in data:
31     # Protein class
32     g.add((URIRef(mgdb+row[0]), RDF.type, URIRef(up+'Protein')))
33     if row[1]!="":
34         g.add((URIRef(mgdb+row[0]), URIRef(owl+'sameAs'), URIRef(upkb+row[1])))
35         g.add((URIRef(mgdb+row[0]), URIRef(up+'encodedBy'), URIRef(mgdb+row[2])))
36     # gene
37     g.add((URIRef(mgdb+row[2]), RDF.type, URIRef(up+'Gene')))
38     if row[3]!="":
39         g.add((URIRef(mgdb+row[2]), URIRef(up+'locusName'), Literal(row[3], datatype=XSD.string)
40             ))
41         if row[4]!="":
42             g.add((URIRef(mgdb+row[2]), URIRef(skos+'prefLabel'), Literal(row[4], datatype=XSD.
43                 string)))
44     # enzyme
45     g.add((URIRef(mgdb+row[0]), URIRef(up+'enzyme'), URIRef(mgdb+row[5])))
46     g.add((URIRef(mgdb+row[5]), RDF.type, URIRef(up+'Enzyme')))
47     if row[6]!="":
48         g.add((URIRef(mgdb+row[5]), URIRef(sio+'SIO_001167'), Literal(row[6], datatype=XSD.
49             string)))
50         g.add((URIRef(mgdb+row[5]), URIRef(obo+'R0_0002331'), URIRef(obo+'GO_0009247')))
51         g.add((URIRef(sio+'SIO_001167'), URIRef(rdfs+'label'), Literal('comment', datatype=XSD.
52             string)))
53         g.add((URIRef(sio+'R0_0002331'), URIRef(rdfs+'label'), Literal('involved_in', datatype=
54             XSD.string)))
55         g.add((URIRef(obo+'GO_0009247'), URIRef(rdfs+'label'), Literal('glycolipid biosynthetic
56             process', datatype=XSD.string)))
57     #reaction
58     g.add((URIRef(mgdb+row[10]), URIRef(glyco+'has_enzyme'), URIRef(mgdb+row[5])))
59     g.add((URIRef(mgdb+row[10]), RDF.type, URIRef(glyco+'Reaction')))
60     g.add((URIRef(mgdb+row[10]), URIRef(sio+'SIO_001167'), Literal(row[9], datatype=XSD.string
61         )))
62     g.add((URIRef(mgdb+row[11]), URIRef(glyco+'catalyzeed_by'), URIRef(mgdb+row[10])))
63     g.add((URIRef(mgdb+row[13]), URIRef(glyco+'catalyzeed_by'), URIRef(mgdb+row[10])))
64     if row[12]!="":
65         g.add((URIRef(mgdb+row[10]), URIRef(glyco+'has_substrate'), URIRef(mgdb+row[11])))
66         g.add((URIRef(mgdb+row[11]), RDF.type, URIRef(glyco+'Glycolipid')))
```

```
60     g.add((URIRef(mgdb+row[11]), URIRef(rdfs+'label'), Literal(row[12], datatype=XSD.string)
61         ))
62     g.add((URIRef(mgdb+row[11]), URIRef(glyco+'is_from_source'), URIRef(taxon+'1773'))))
63     g.add((URIRef(mgdb+row[10]), URIRef(glyco+'has_product'), URIRef(mgdb+row[13]))))
64     g.add((URIRef(mgdb+row[13]), RDF.type, URIRef(glyco+'Glycolipid'))))
65     if row[14]!="":
66         g.add((URIRef(mgdb+row[13]), URIRef(rdfs+'label'), Literal(row[14], datatype=XSD.string)
67             ))
68         g.add((URIRef(mgdb+row[13]), URIRef(glyco+'is_from_source'), URIRef(taxon+'1773'))))
69     # taxon
70     if row[4]!="":
71         g.add((URIRef(mgdb+row[0]), URIRef(glyco+'is_from_source'), URIRef(taxon+row[8]))))
72         g.add((URIRef(mgdb+row[8]), URIRef(up+'scientificName'), Literal(row[7], datatype=XSD.
73             string)))
74
75     # check the results
76     #print(g.serialize(format='turtle').decode('UTF-8'))
77
78     # save the results to rdf file
79     g.serialize('m_tuberculosis_lsm.ttl', format='turtle')
```

LISTING A.2: The python source code for verification of RDF triples of *M.**tuberculosis*

```
1 from shexer.shaper import Shaper
2 from shexer.consts import NT, SHEXC, TURTLE
3
4 target_classes = [
5     "http://purl.uniprot.org/core/Protein",
6     "http://purl.uniprot.org/core/Enzyme",
7     "http://purl.uniprot.org/core/Gene",
8     "http://purl.jp/bio/12/glyco/glycan#Glycolipid",
9     "http://purl.jp/bio/12/glyco/glycan#Reaction"
10 ]
11 namespaces_dict = {"http://www.w3.org/1999/02/22-rdf-syntax-ns#": "rdf",
12                   "http://www.w3.org/2001/XMLSchema/": "xsd",
13                   "http://www.w3.org/2002/07/owl#": "owl",
14                   "http://www.w3.org/2000/01/rdf-schema#": "rdfs",
15                   "http://www.w3.org/XML/1998/namespace/": "xml",
16                   "http://purl.jp/bio/12/glyco/glycan#": "glyco",
17                   "http://purl.org/dc/terms/": "dc",
18                   "http://purl.org/net/": "perl",
19                   "http://purl.uniprot.org/core/": "uni",
20                   "http://purl.obolibrary.org/obo/": "obo",
21                   "http://semanticscience.org/resource/": "seman",
22                   "http://www.w3.org/2004/02/skos/core#": "skos"
23                 }
24
25 graph_file_input = "m_tuberculosis_lsm.ttl" # file assign
26 shaper = Shaper(target_classes=target_classes,
27                 graph_file_input=graph_file_input ,
28                 input_format=TURTLE,
29                 namespaces_dict=namespaces_dict, # Default: no prefixes
30                 instantiation_property="http://www.w3.org/1999/02/22-rdf-syntax-ns#type",
31                 track_classes_for_entities_at_last_depth_level=True) # Default rdf:type
32 output_file = "tuberculosis.shex"
33 shaper.shex_graph(output_file=output_file,
34                  acceptance_threshold=0.1
35                  )
36
37 print("Done!")
```

The upper is Python code for Shex to evaluate the RDF file of *M. tuberculosis*. The Python code for verifying RDF triples of various microorganisms, such as *E. coli*, in the MicroGlycoDB database has been uploaded to this website (<https://gitlab.glyco.info/glycosmos/microglycodb/-/tree/master/RDFication>).

## List of Abbreviations

<b>AGs</b>	Arabinogalactans
<b>AIDS</b>	Acquired ImmunoDeficiency Syndrome
<b>AR</b>	Antimicrobial resistance
<b><i>B. longum</i></b>	<i>Bifidobacterium longum</i>
<b><i>B. bifidum</i></b>	<i>Bifidobacterium bifidum</i>
<b>BlaZ</b>	$\beta$ -lactamase
<b><i>C. neopormans</i></b>	<i>Cryptococcus neopormans</i>
<b><i>C. jejuni</i></b>	<i>Campylobacter jejuni</i>
<b>CAZymes</b>	carbohydrate-active enzymes
<b>CSDB</b>	Carbohydrate Structure Database
<b>DBCLS</b>	Database Center for Life Science
<b>EGF</b>	Epidermal Growth Factor
<b><i>E. coli</i></b>	<i>Escherichia coli</i>
<b>GalCer</b>	Galactosylceramid
<b>GlcNAc</b>	<i>N</i> -acetyl-glucosamine
<b>GO</b>	Gene Ontology
<b>GSLs</b>	Glycosphingolipids
<b>GPLs</b>	GlycoPeptidolipids
<b>HIV</b>	Human Immunodeficiency Virus
<b>Kdo</b>	monosaccharides 3-deoxy-D-manno-oct-2-ulosonic acid
<b>KEGG</b>	Kyoto Encyclopedia of Genes and Genomes

<b>LAM</b>	Lipoarabinomannan
<b>LM</b>	lipomannan
<b>LPS</b>	Lipopolysaccharide
<b><i>M. abscessus</i></b>	<i>Mycobacteroides abscessus</i>
<b><i>M. tuberculosis</i></b>	<i>Mycobacterium tuberculosis</i>
<b>MRSA</b>	<i>Methicillin-resistant Staphylococcus aureus</i>
<b>MurNAc</b>	<i>N</i> -acetylmuramic acid
<b>PBP2</b>	penicillin-binding protein 2
<b>PD</b>	Peptidoglycan
<b>PIMs</b>	Phosphatidylinositol- containing mannoside
<b>POFUTs</b>	Protein O-fucosyltransferas
<b>POGLUT</b>	Protein O-glucosyltransferase
<b>RDF</b>	Resource Description Framework
<b>RO</b>	Relational Ontology
<b>SPARQL</b>	SPARQL Protocol and RDF Query Language
<b>TAs</b>	teichoic acids
<b>TP</b>	transpeptidase
<b>W3C</b>	World Wide Web Consortium

# Bibliography

- Abouelhadid, Sherif et al. (2019). “Quantitative analyses reveal novel roles for N-glycosylation in a major enteric bacterial pathogen”. In: *MBio* 10.2, pp. 10–1128.
- Abouelhadid, Sherif et al. (2020). “Characterization of posttranslationally modified multidrug efflux pumps reveals an unexpected link between glycosylation and antimicrobial resistance”. In: *MBio* 11.6, pp. 10–1128.
- Akusobi, Chidiebere et al. (2022). “Transposon mutagenesis in *Mycobacterium abscessus* identifies an essential penicillin-binding protein involved in septal peptidoglycan synthesis and antibiotic sensitivity”. In: *Elife* 11, e71947.
- Alderwick, Luke J et al. (2015). “The mycobacterial cell wall—peptidoglycan and arabinogalactan”. In: *Cold Spring Harbor perspectives in medicine* 5.8, a021113.
- Aoki-Kinoshita, Kiyoko F (2019). “Glycan Nomenclature and Summary of Glycan-related Resources”. In: *Glycoforum* 22, A2.
- Asif T. Chinwalla Lucinda A. Fulton, LaDeana W. Hillier (2002). “Initial sequencing and comparative analysis of the mouse genome”. In: *Nature* 420.6915, pp. 520–562.
- Authority, European Food Safety et al. (2022). “The European Union Summary Report on Antimicrobial Resistance in zoonotic and indicator bacteria from humans, animals and food in 2019–2020”. In: *EFSA Journal* 20.3.
- Azimzadeh Irani, Maryam, Srinivasaraghavan Kannan, and Chandra Verma (2017). “Role of N-glycosylation in EGFR ectodomain ligand binding”. In: *Proteins: Structure, Function, and Bioinformatics* 85.8, pp. 1529–1549.



- Bard, Jonathan BL and Seung Y Rhee (2004). “Ontologies in biology: design, applications and future challenges”. In: *nature reviews genetics* 5.3, pp. 213–222.
- Batt, Sarah M et al. (2020). “Antibiotics and resistance: the two-sided coin of the mycobacterial cell wall”. In: *The Cell Surface* 6, p. 100044.
- Bauer-Mehren, Anna, Laura I Furlong, and Ferran Sanz (2009). “Pathway databases and tools for their exploitation: benefits, current limitations and challenges”. In: *Molecular systems biology* 5.1, p. 290.
- Bébéar, CM and S Pereyre (2005). “Mechanisms of drug resistance in *Mycoplasma pneumoniae*”. In: *Current Drug Targets-Infectious Disorders* 5.3, pp. 263–271.
- Bell, Andrew and Nathalie Juge (2021). “Mucosal glycan degradation of the host by the gut microbiota”. In: *Glycobiology* 31.6, pp. 691–696.
- Belleau, François et al. (2008). “Bio2RDF: towards a mashup to build bioinformatics knowledge systems”. In: *Journal of biomedical informatics* 41.5, pp. 706–716.
- Berg, Stefan et al. (2007). “The glycosyltransferases of *Mycobacterium tuberculosis*—roles in the synthesis of arabinogalactan, lipoarabinomannan, and other glycoconjugates”. In: *Glycobiology* 17.6, 35R–56R.
- Bergstrom, Kirk SB and Lijun Xia (2013). “Mucin-type O-glycans and their roles in intestinal homeostasis”. In: *Glycobiology* 23.9, pp. 1026–1037.
- Berman, Helen M et al. (2002). “The protein data bank”. In: *Acta Crystallographica Section D: Biological Crystallography* 58.6, pp. 899–907.
- Bi, Yunchen et al. (2018). “Architecture of a channel-forming O-antigen polysaccharide ABC transporter”. In: *Nature* 553.7688, pp. 361–365.
- Blair, Jessica MA et al. (2015). “Molecular mechanisms of antibiotic resistance”. In: *Nature reviews microbiology* 13.1, pp. 42–51.
- Blake, Judith A and Carol J Bult (2006). “Beyond the data deluge: data integration and bio-ontologies”. In: *Journal of biomedical informatics* 39.3, pp. 314–320.

- Botstein, David et al. (2000). “Gene Ontology: tool for the unification of biology”. In: *Nat genet* 25.1, pp. 25–29.
- Brennan, Patrick J and Dean C Crick (2007). “The cell-wall core of *Mycobacterium tuberculosis* in the context of drug discovery.” In: *Current topics in medicinal chemistry* 7.5, pp. 475–488.
- Brown, Stephanie et al. (2012). “Methicillin resistance in *Staphylococcus aureus* requires glycosylated wall teichoic acids”. In: *Proceedings of the national academy of sciences* 109.46, pp. 18909–18914.
- Cain, Joel A et al. (2020). “Identifying the targets and functions of N-linked protein glycosylation in *Campylobacter jejuni*”. In: *Molecular omics* 16.4, pp. 287–304.
- Catalão, Maria João, Sérgio R Filipe, and Madalena Pimentel (2019). “Revisiting anti-tuberculosis therapeutic strategies that target the peptidoglycan structure and synthesis”. In: *Frontiers in Microbiology*, p. 190.
- Chai, Qiyao, Yong Zhang, and Cui Hua Liu (2018). “*Mycobacterium tuberculosis*: an adaptable pathogen associated with multiple human diseases”. In: *Frontiers in cellular and infection microbiology* 8, p. 158.
- Cho, Hongbaek, Tsuyoshi Uehara, and Thomas G Bernhardt (2014). “Beta-lactam antibiotics induce a lethal malfunctioning of the bacterial cell wall synthesis machinery”. In: *Cell* 159.6, pp. 1300–1311.
- Cipolla, L et al. (2011). “New targets for antibacterial design: Kdo biosynthesis and LPS machinery transport to the cell surface”. In: *Current medicinal chemistry* 18.6, pp. 830–852.
- Cobb, Brian A and Dennis L Kasper (2005). “Coming of age: carbohydrates and immunity”. In: *European journal of immunology* 35.2, pp. 352–356.
- Consortium, Gene Ontology (2004). “The Gene Ontology (GO) database and informatics resource”. In: *Nucleic acids research* 32.suppl\_1, pp. D258–D261.

- Consortium, UniProt (2007). “The universal protein resource (UniProt)”. In: *Nucleic acids research* 36.suppl\_1, pp. D190–D195.
- Croft, David et al. (2014). “The Reactome pathway knowledgebase”. In: *Nucleic acids research* 42.D1, pp. D472–D477.
- Cui, Miao, Chao Cheng, and Lanjing Zhang (2022). “High-throughput proteomics: a methodological mini-review”. In: *Laboratory Investigation* 102.11, pp. 1170–1181.
- Dai, Lei et al. (2020). “New and alternative strategies for the prevention, control, and treatment of antibiotic-resistant *Campylobacter*”. In: *Translational Research* 223, pp. 76–88.
- DebRoy, Chitrita et al. (2016). “Comparison of O-antigen gene clusters of all O-serogroups of *Escherichia coli* and proposal for adopting a new nomenclature for O-typing”. In: *PLoS One* 11.1, e0147434.
- D’Elia, Michael A et al. (2006). “Wall teichoic acid polymers are dispensable for cell viability in *Bacillus subtilis*”. In: *Journal of bacteriology* 188.23, pp. 8313–8316.
- Demir, Emek et al. (2010). “The BioPAX community standard for pathway data sharing”. In: *Nature biotechnology* 28.9, pp. 935–942.
- Dobos, Karen M et al. (1996). “Definition of the full extent of glycosylation of the 45-kilodalton glycoprotein of *Mycobacterium tuberculosis*”. In: *Journal of bacteriology* 178.9, pp. 2498–2506.
- Eichler, Evan E (2019). “Genetic variation, comparative genomics, and the diagnosis of disease”. In: *New England Journal of Medicine* 381.1, pp. 64–74.
- Erling, Orri and Ivan Mikhailov (2009). “RDF Support in the Virtuoso DBMS”. In: *Networked Knowledge-Networked Media: Integrating Knowledge Management, New Media Technologies and Semantic Systems*, pp. 7–24.
- Erridge, Clett, Elliott Bennett-Guerrero, and Ian R Poxton (2002). “Structure and function of lipopolysaccharides”. In: *Microbes and infection* 4.8, pp. 837–851.
- Esther Jr, Charles R et al. (2005). “Nontuberculous mycobacterial infection in young children with cystic fibrosis”. In: *Pediatric pulmonology* 40.1, pp. 39–44.

- Fabregat, Antonio et al. (2018). “The reactome pathway knowledgebase”. In: *Nucleic acids research* 46.D1, pp. D649–D655.
- Farwanah, Hany and Thomas Kolter (2012). “Lipidomics of glycosphingolipids”. In: *Metabolites* 2.1, pp. 134–164.
- Foster, Timothy J (2017). “Antibiotic resistance in *Staphylococcus aureus*. Current status and future prospects”. In: *FEMS microbiology reviews* 41.3, pp. 430–449.
- Frank, Christian G and Markus Aebi (2005). “ALG9 mannosyltransferase is involved in two different steps of lipid-linked oligosaccharide biosynthesis”. In: *Glycobiology* 15.11, pp. 1156–1163.
- Fujita, Akihiro et al. (2021). “The international glycan repository GlyTouCan version 3.0”. In: *Nucleic Acids Research* 49.D1, pp. D1529–D1533.
- Ghajavand, Hasan et al. (2019). “Scrutinizing the drug resistance mechanism of multi-and extensively-drug resistant *Mycobacterium tuberculosis*: mutations versus efflux pumps”. In: *Antimicrobial Resistance & Infection Control* 8, pp. 1–8.
- Gibreel, Amara, Nicole M Wetsch, and Diane E Taylor (2007). “Contribution of the Cme-ABC efflux pump to macrolide and tetracycline resistance in *Campylobacter jejuni*”. In: *Antimicrobial agents and chemotherapy* 51.9, pp. 3212–3216.
- González-Morelo, Kevin J, Marco Vega-Sagardía, and Daniel Garrido (2020). “Molecular insights into O-linked glycan utilization by gut microbes”. In: *Frontiers in Microbiology* 11, p. 591568.
- Goodfellow, John A and Hugh J Willison (2016). “Guillain–Barré syndrome: a century of progress”. In: *Nature Reviews Neurology* 12.12, pp. 723–731.
- Grass, Susan et al. (2003). “The *Haemophilus influenzae* HMW1 adhesin is glycosylated in a process that requires HMW1C and phosphoglucomutase, an enzyme involved in lipooligosaccharide biosynthesis”. In: *Molecular microbiology* 48.3, pp. 737–751.
- Greenfield, Laura K and Chris Whitfield (2012). “Synthesis of lipopolysaccharide O-antigens by ABC transporter-dependent pathways”. In: *Carbohydrate research* 356, pp. 12–24.

- Guo, Hongjie et al. (2005). “Molecular analysis of the O-antigen gene cluster of *Escherichia coli* O86: B7 and characterization of the chain length determinant gene (*wzz*)”. In: *Applied and Environmental Microbiology* 71.12, pp. 7995–8001.
- Gygli, Sebastian M et al. (2017). “Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives”. In: *FEMS microbiology reviews* 41.3, pp. 354–373.
- Haltom, Amanda R and Hamed Jafar-Nejad (2015). “The multiple roles of epidermal growth factor repeat O-glycans in animal development”. In: *Glycobiology* 25.10, pp. 1027–1042.
- Hammar, Yassine (2018). *Insight into Semantic Web and why is it important today*. URL: <https://medium.com/@yassine.hammar1/insight-into-semantic-web-and-why-its-the-next-technological-revolution-24521cec4459> (visited on 08/19/2018).
- Hamosh, Ada et al. (2005). “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders”. In: *Nucleic acids research* 33.suppl\_1, pp. D514–D517.
- Hase, Sumihiro et al. (1988). “A new trisaccharide sugar chain linked to a serine residue in bovine blood coagulation factors VII and IX”. In: *The Journal of Biochemistry* 104.6, pp. 867–868.
- Hastings, Janna et al. (2016). “ChEBI in 2016: Improved services and an expanding collection of metabolites”. In: *Nucleic acids research* 44.D1, pp. D1214–D1219.
- Henry, Karl W, Joseph T Nickels, and Thomas D Edlind (2002). “ROX1 and ERG regulation in *Saccharomyces cerevisiae*: implications for antifungal susceptibility”. In: *Eukaryotic Cell* 1.6, pp. 1041–1044.
- Hong, Yaoqin and Peter R Reeves (2014). “Diversity of O-antigen repeat unit structures can account for the substantial sequence variation of *Wzx* translocases”. In: *Journal of bacteriology* 196.9, pp. 1713–1722.
- Imperiali, Barbara (2019). “Bacterial carbohydrate diversity—A brave new world”. In: *Current opinion in chemical biology* 53, pp. 1–8.

- Iovine, Nicole M (2013). “Resistance mechanisms in *Campylobacter jejuni*”. In: *Virulence* 4.3, pp. 230–240.
- Jensen, Slade O and Bruce R Lyon (2009). “Genetics of antimicrobial resistance in *Staphylococcus aureus*”. In: *Future microbiology* 4.5, pp. 565–582.
- Jeurink, Prescilla V et al. (2013). “Mechanisms underlying immune effects of dietary oligosaccharides”. In: *The American journal of clinical nutrition* 98.2, 572S–577S.
- Jonquet, Clement et al. (2011). “NCBO Resource Index: Ontology-based search and mining of biomedical resources”. In: *Journal of Web Semantics* 9.3, pp. 316–324.
- Jupp, Simon et al. (2015). “A new Ontology Lookup Service at EMBL-EBI.” In: *SWAT4LS* 2, pp. 118–119.
- Juty, Nick, Nicolas Le Novere, and Camille Laibe (2012). “Identifiers. org and MIRIAM Registry: community resources to provide persistent identification”. In: *Nucleic acids research* 40.D1, pp. D580–D586.
- Kanehisa, Minoru and Susumu Goto (2000). “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1, pp. 27–30.
- Kanehisa, Minoru et al. (2021). “KEGG: integrating viruses and cellular organisms”. In: *Nucleic acids research* 49.D1, pp. D545–D551.
- Kaper, James B, James P Nataro, and Harry LT Mobley (2004). “Pathogenic *Escherichia coli*”. In: *Nature reviews microbiology* 2.2, pp. 123–140.
- Kaszuba, Karol et al. (2015). “N-Glycosylation as determinant of epidermal growth factor receptor conformation in membranes”. In: *Proceedings of the National Academy of Sciences* 112.14, pp. 4334–4339.
- Katayama, Toshiaki et al. (2019). “TogoGenome/TogoStanza: modularized Semantic Web genome database”. In: *Database* 2019.
- Keenleyside, Wendy J et al. (1994). “A plasmid-encoded rfbO: 54 gene cluster is required for biosynthesis of the O: 54 antigen in *Salmonella enterica* serovar Borreze”. In: *Molecular microbiology* 11.3, pp. 437–448.

- Khan, Raees et al. (2022). “Bacterial polysaccharides—a big source for prebiotics and therapeutics”. In: *Frontiers in Nutrition*, p. 2631.
- Ku, Seockmo et al. (2016). “Review on Bifidobacterium bifidum BGN4: functionality and nutraceutical applications as a probiotic microorganism”. In: *International journal of molecular sciences* 17.9, p. 1544.
- Kuck, Gregk (2004). “Tim Berners-Lee’s Semantic Web”. In: *SA Journal of Information Management* 6.1.
- La Rosa, Sabina Leanti et al. (2022). “Glycan processing in gut microbiomes”. In: *Current Opinion in Microbiology* 67, p. 102143.
- Landers, Timothy F et al. (2012). “A review of antibiotic use in food animals: perspective, policy, and potential”. In: *Public health reports* 127.1, pp. 4–22.
- Lapatas, Vasileios et al. (2015). “Data integration in biological research: an overview”. In: *Journal of Biological Research-Thessaloniki* 22, pp. 1–16.
- Latousakis, Dimitrios and Nathalie Juge (2018). “How sweet are our gut beneficial bacteria? A focus on protein glycosylation in *Lactobacillus*”. In: *International journal of molecular sciences* 19.1, p. 136.
- Lee, Su-Bin et al. (2023). “Effects of altered N-glycan structures of *Cryptococcus neoformans* mannoproteins, MP98 (Cda2) and MP84 (Cda3), on interaction with host cells”. In: *Scientific Reports* 13.1, p. 1175.
- Lee, Sunmyoung, Tamiko Ono, and Kiyoko Aoki-Kinoshita (2021). “RDFizing the biosynthetic pathway of *E. coli* O-antigen to enable semantic sharing of microbiology data”. In: *BMC microbiology* 21.1, pp. 1–12.
- Lillehoj, Erik P et al. (2013). “Cellular and molecular biology of airway mucins”. In: *International review of cell and molecular biology* 303, pp. 139–202.
- Lin, Jun, Linda Overbye Michel, and Qijing Zhang (2002). “CmeABC functions as a multidrug efflux system in *Campylobacter jejuni*”. In: *Antimicrobial agents and chemotherapy* 46.7, pp. 2124–2131.

- Liu, Bin et al. (2020). “Structure and genetics of Escherichia coli O antigens”. In: *FEMS microbiology reviews* 44.6, pp. 655–683.
- Liu, Yue, Jiaqi Wang, and Changxin Wu (2022). “Modulation of gut microbiota and immune system by probiotics, pre-biotics, and post-biotics”. In: *Frontiers in nutrition* 8, p. 634897.
- Loomes, Kerry M et al. (1999). “Functional protective role for mucin glycosylated repetitive domains”. In: *European journal of biochemistry* 266.1, pp. 105–111.
- Luangtongkum, Taradon et al. (2009). “Antibiotic resistance in Campylobacter: emergence, transmission and persistence”. In.
- Ma, Cheng et al. (2020). “Comprehensive N-and O-glycosylation mapping of human coagulation factor V”. In: *Journal of Thrombosis and Haemostasis* 18.8, pp. 1884–1892.
- Maglott, Donna et al. (2007). “Entrez Gene: gene-centered information at NCBI”. In: *Nucleic acids research* 35.suppl\_1, pp. D26–D31.
- Maitra, Arundhati et al. (2019). “Cell wall peptidoglycan in Mycobacterium tuberculosis: An Achilles’ heel for the TB-causing pathogen”. In: *FEMS Microbiology Reviews* 43.5, pp. 548–575.
- Marcobal, Angela et al. (2013). “A refined palate: bacterial consumption of host glycans in the gut”. In: *Glycobiology* 23.9, pp. 1038–1046.
- Martens, Marvin et al. (2021). “WikiPathways: connecting communities”. In: *Nucleic acids research* 49.D1, pp. D613–D621.
- Mashima, Jun et al. (2016). “DNA data bank of Japan”. In: *Nucleic Acids Research*, gkw1001.
- McBride, Brian (2004). “The resource description framework (RDF) and its vocabulary description language RDFS”. In: *Handbook on ontologies*, pp. 51–65.
- McCarter, Yvette S (2017). *Antibiotics: Challenges, Mechanisms, Opportunities Written by Christopher Walsh, PhD and Timothy Wencewicz, PhD*.
- McGuckin, Michael A et al. (2011). “Mucin dynamics and enteric pathogens”. In: *Nature Reviews Microbiology* 9.4, pp. 265–278.



- Meehan, Terrence F et al. (2011). “Logical development of the cell ontology”. In: *BMC bioinformatics* 12.1, pp. 1–12.
- Merino, Susana, Victor Gonzalez, and Juan M Tomás (2016). “The first sugar of the repeat units is essential for the Wzy polymerase activity and elongation of the O-antigen lipopolysaccharide”. In: *Future microbiology* 11.7, pp. 903–918.
- Miller, William R, Jose M Munita, and Cesar A Arias (2014). “Mechanisms of antibiotic resistance in enterococci”. In: *Expert review of anti-infective therapy* 12.10, pp. 1221–1236.
- Modenutti, Carlos P et al. (2019). “The structural biology of galectin-ligand recognition: current advances in modeling tools, protein engineering, and inhibitor design”. In: *Frontiers in Chemistry* 7, p. 823.
- Monaco, Gianni et al. (2015). “A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels”. In: *BMC evolutionary biology* 15.1, pp. 1–14.
- Musen, Mark A (2015). “The protégé project: a look back and a look forward”. In: *AI matters* 1.4, pp. 4–12.
- Nasiri, MJ et al. (2017). *New insights in to the intrinsic and acquired drug resistance mechanisms in mycobacteria. Front Microbiol* 8: 681.
- Nguyen, Liem (2016). “Antibiotic resistance mechanisms in M. tuberculosis: an update”. In: *Archives of toxicology* 90, pp. 1585–1604.
- Ohtsubo, Kazuaki and Jamey D Marth (2006). “Glycosylation in cellular mechanisms of health and disease”. In: *Cell* 126.5, pp. 855–867.
- Okajima, Tetsuya and Kenneth D Irvine (2002). “Regulation of notch signaling by o-linked fucose”. In: *Cell* 111.6, pp. 893–904.
- Okuda, Suguru et al. (2016). “Lipopolysaccharide transport and assembly at the outer membrane: the PEZ model”. In: *Nature Reviews Microbiology* 14.6, pp. 337–345.

- Oyofa, BA et al. (1989). “Prevention of *Salmonella typhimurium* colonization of broilers with D-mannose”. In: *Poultry science* 68.10, pp. 1357–1360.
- Pasquina, Lincoln W, John P Santa Maria, and Suzanne Walker (2013). “Teichoic acid biosynthesis as an antibiotic target”. In: *Current opinion in microbiology* 16.5, pp. 531–537.
- Peacock, Sharon J and Gavin K Paterson (2015). “Mechanisms of methicillin resistance in *Staphylococcus aureus*”. In: *Annual review of biochemistry* 84, pp. 577–601.
- Pereira, Márcia S et al. (2018). “Glycans as key checkpoints of T cell activity and function”. In: *Frontiers in immunology* 9, p. 2754.
- Piddock, Laura JV (2006). “Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria”. In: *Clinical microbiology reviews* 19.2, pp. 382–402.
- Pinho, Mariana G, Hermínia de Lencastre, and Alexander Tomasz (2001). “An acquired and a native penicillin-binding protein cooperate in building the cell wall of drug-resistant staphylococci”. In: *Proceedings of the National Academy of Sciences* 98.19, pp. 10886–10891.
- Prado Acosta, Mariano and Bernd Lepenies (2019). “Bacterial glycans and their interactions with lectins in the innate immune system”. In: *Biochemical Society Transactions* 47.6, pp. 1569–1579.
- Prud’hommeaux, Eric and Andy Seaborne (2008). “SPARQL query language for RDF. W3C recommendation, W3C”. In: URL: <http://www.w3.org/TR/rdf-sparql-query>.
- Pruss, KM et al. (2021). “Mucin-derived O-glycans supplemented to diet mitigate diverse microbiota perturbations”. In: *The ISME Journal* 15.2, pp. 577–591.
- Pumbwe, Lilian and Laura JV Piddock (2002). “Identification and molecular characterisation of CmeB, a *Campylobacter jejuni* multidrug efflux pump”. In: *FEMS Microbiology Letters* 206.2, pp. 185–189.
- Qiao, Yuan et al. (2017). “Lipid II overproduction allows direct assay of transpeptidase inhibition by  $\beta$ -lactams”. In: *Nature chemical biology* 13.7, pp. 793–798.

- Raetz, Christian RH and Chris Whitfield (2002). “Lipopolysaccharide endotoxins”. In: *Annual review of biochemistry* 71.1, pp. 635–700.
- Raman, Karthik and Nagasuma Chandra (2009). “Flux balance analysis of biological systems: applications and challenges”. In: *Briefings in bioinformatics* 10.4, pp. 435–449.
- Rana, Nadia A et al. (2011). “O-glucose trisaccharide is present at high but variable stoichiometry at multiple sites on mouse Notch1”. In: *Journal of Biological Chemistry* 286.36, pp. 31623–31637.
- Ranzinger, Rene et al. (2015). “GlycoRDF: an ontology to standardize glycomics data in RDF”. In: *Bioinformatics* 31.6, pp. 919–925.
- Reeves, Peter R and Monica M Cunneen (2010). “Biosynthesis of O-antigen chains and assembly”. In: *Microbial Glycobiology*, pp. 319–335.
- Reily, Colin et al. (2019). “Glycosylation in health and disease”. In: *Nature Reviews Nephrology* 15.6, pp. 346–366.
- Reimer, Lorenz Christian et al. (2022). “Bac Dive in 2022: the knowledge base for standardized bacterial and archaeal data”. In: *Nucleic Acids Research* 50.D1, pp. D741–D746.
- Reygaert, Wanda C (2018). “An overview of the antimicrobial resistance mechanisms of bacteria”. In: *AIMS microbiology* 4.3, p. 482.
- Rodchenkov, Igor et al. (2020). “Pathway Commons 2019 Update: integration, analysis and exploration of pathway data”. In: *Nucleic acids research* 48.D1, pp. D489–D497.
- Royer, Guilhem et al. (2022). “O-antigen targeted vaccines against Escherichia coli may be useful in reducing morbidity, mortality, and antimicrobial resistance”. In: *Clinical Infectious Diseases* 74.2, pp. 364–366.
- Ruhaak, L Renee et al. (2018). “Mass spectrometry approaches to glycomic and glycoproteomic analyses”. In: *Chemical reviews* 118.17, pp. 7886–7930.
- Ruiz, Maria E, Isabel C Guerrero, and Carmelita U Tuazon (2002). “Endocarditis caused by methicillin-resistant Staphylococcus aureus: treatment failure with linezolid”. In: *Clinical infectious diseases* 35.8, pp. 1018–1020.

- Russo, Thomas A et al. (2009). “Capsular polysaccharide and the O-specific antigen impede antibody binding: a potential obstacle for the successful development of an extraintestinal pathogenic *Escherichia coli* vaccine”. In: *Vaccine* 27.3, pp. 388–395.
- Sarkar, Sohinee et al. (2014). “Role of capsule and O antigen in the virulence of uropathogenic *Escherichia coli*”. In: *PloS one* 9.4, e94786.
- Schachter, Harry (2000). “The joys of HexNAc. The synthesis and function of N-andO-glycan branches”. In: *Glycoconjugate journal* 17, pp. 465–483.
- Schäffer, Christina and Paul Messner (2017). “Emerging facets of prokaryotic glycosylation”. In: *FEMS microbiology reviews* 41.1, pp. 49–91.
- Scheutz, Flemming et al. (2004). “Designation of O174 and O175 to temporary O groups OX3 and OX7, and six new *E. coli* O groups that include Verocytotoxin-producing *E. coli* (VTEC): O176, O177, O178, O179, O180 and O181”. In: *Apmis* 112.9, pp. 569–84.
- Schmidt, M Alexander, Lee W Riley, and Inga Benz (2003). “Sweet new world: glycoproteins in bacterial pathogens”. In: *TRENDS in Microbiology* 11.12, pp. 554–561.
- Shannon, Paul et al. (2003). “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. In: *Genome research* 13.11, pp. 2498–2504.
- Shokryazdan, Parisa et al. (2017). “Effects of prebiotics on immune system and cytokine expression”. In: *Medical microbiology and immunology* 206, pp. 1–9.
- Siddiqui, Abdul H and Janak Koirala (2018). “Methicillin resistant *Staphylococcus aureus*”. In.
- Silhavy, Thomas J, Daniel Kahne, and Suzanne Walker (2010). “The bacterial cell envelope”. In: *Cold Spring Harbor perspectives in biology* 2.5, a000414.
- Singh, Richa et al. (2020). “Recent updates on drug resistance in *Mycobacterium tuberculosis*”. In: *Journal of applied microbiology* 128.6, pp. 1547–1567.
- Smith, Tasha, Kerstin A Wolff, and Liem Nguyen (2012). “Molecular biology of drug resistance in *Mycobacterium tuberculosis*”. In: *Pathogenesis of Mycobacterium tuberculosis and its Interaction with the Host Organism*, pp. 53–80.

- Solbrig, Harold R et al. (2017). “Modeling and validating HL7 FHIR profiles using semantic web Shape Expressions (ShEx)”. In: *Journal of biomedical informatics* 67, pp. 90–100.
- Sonnino, Sandro and Alessandro Prinetti (2010). “Gangliosides as regulators of cell membrane organization and functions”. In: *Sphingolipids as Signaling and Regulatory Molecules*, pp. 165–184.
- Sørensen, Daniel Madriz et al. (2023). “Identification of global inhibitors of cellular glycosylation”. In: *Nature Communications* 14.1, p. 948.
- Stein, Lincoln D (2003). “Integrating biological databases”. In: *Nature Reviews Genetics* 4.5, pp. 337–345.
- Stimson, Elaine et al. (1995). “Meningococcal pilin: a glycoprotein substituted with digalactosyl 2, 4-diacetamido-2, 4, 6-trideoxyhexose”. In: *Molecular microbiology* 17.6, pp. 1201–1214.
- Szymanski, Christine M (2022). “Bacterial glycosylation, it’s complicated”. In: *Frontiers in Molecular Biosciences*, p. 1089.
- Szymanski, Christine M et al. (1999). “Evidence for a system of general protein glycosylation in *Campylobacter jejuni*”. In: *Molecular microbiology* 32.5, pp. 1022–1030.
- Takamatsu, Daisuke et al. (2006). “Binding of the streptococcal surface glycoproteins GspB and Hsa to human salivary proteins”. In: *Infection and immunity* 74.3, pp. 1933–1940.
- Takeuchi, Hideyuki and Robert S Haltiwanger (2014). “Significance of glycosylation in Notch signaling”. In: *Biochemical and biophysical research communications* 453.2, pp. 235–242.
- Tang, Fuchou et al. (2009). “mRNA-Seq whole-transcriptome analysis of a single cell”. In: *Nature methods* 6.5, pp. 377–382.
- Thak, Eun Jung et al. (2020). “Core N-glycan structures are critical for the pathogenicity of *Cryptococcus neoformans* by modulating host cell death”. In: *MBio* 11.3, e00711–20.
- Thornton, Katherine et al. (2019). “Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation”. In: *European Semantic Web Conference*. Springer, pp. 606–620.

- Titford, Michael (2010). “Paul Ehrlich: histological staining, immunology, chemotherapy”. In: *Laboratory Medicine* 41.8, pp. 497–498.
- Toukach, Philip V and Ksenia S Egorova (2016). “Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts”. In: *Nucleic acids research* 44.D1, pp. D1229–D1236.
- (2019). “New features of Carbohydrate Structure Database notation (CSDB Linear), as compared to other carbohydrate notations”. In: *Journal of Chemical Information and Modeling* 60.3, pp. 1276–1289.
- Tra, Van N and Danielle H Dube (2014). “Glycans in pathogenic bacteria—potential for targeted covalent therapeutics and imaging agents”. In: *Chemical Communications* 50.36, pp. 4659–4673.
- Travers, Kevin J et al. (2000). “Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation”. In: *Cell* 101.3, pp. 249–258.
- Turner, Nicholas A et al. (2019). “Methicillin-resistant *Staphylococcus aureus*: an overview of basic and clinical research”. In: *Nature Reviews Microbiology* 17.4, pp. 203–218.
- Van Hoek, Angela HAM et al. (2011). “Acquired antibiotic resistance genes: an overview”. In: *Frontiers in microbiology* 2, p. 203.
- Varki, Ajit (2011). “Evolutionary forces shaping the Golgi glycosylation machinery: why cell surface glycans are universal to living cells”. In: *Cold Spring Harbor perspectives in biology* 3.6, a005462.
- Varki, Ajit et al. (2022). “Essentials of Glycobiology [internet]”. In.
- Victor, L Yu (2011). “Guidelines for hospital-acquired pneumonia and health-care-associated pneumonia: a vulnerability, a pitfall, and a fatal flaw”. In: *The lancet infectious diseases* 11.3, pp. 248–252.

- Viljoen, Albertus et al. (2020). “Fast chemical force microscopy demonstrates that glycopeptidolipids define nanodomains of varying hydrophobicity on mycobacteria”. In: *Nanoscale Horizons* 5.6, pp. 944–953.
- Waagmeester, Andra et al. (2016). “Using the semantic web for rapid integration of WikiPathways with other biological online data resources”. In: *PLoS computational biology* 12.6, e1004989.
- Wang, Ke et al. (2013). “The expression of ABC efflux pump, Rv1217c–Rv1218c, and its association with multidrug resistance of *Mycobacterium tuberculosis* in China”. In: *Current microbiology* 66, pp. 222–226.
- Wilhelm, Michael J et al. (2015). “Gram’s stain does not cross the bacterial cytoplasmic membrane”. In: *ACS Chemical Biology* 10.7, pp. 1711–1717.
- Wilkinson, Mark D et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1, pp. 1–9.
- Williams, DL et al. (1998). “Contribution of rpoB mutations to development of rifamycin cross-resistance in *Mycobacterium tuberculosis*”. In: *Antimicrobial agents and chemotherapy* 42.7, pp. 1853–1857.
- Xiang, Zuoshuang et al. (2011). “Ontobee: A linked data server and browser for ontology terms.” In: *ICBO*.
- Xing, Yikun et al. (2023). “Broad protective vaccination against systemic *Escherichia coli* with autotransporter antigens”. In: *PLoS Pathogens* 19.2, e1011082.
- Yago, Tadayuki et al. (2010). “Core 1-derived O-glycans are essential E-selectin ligands on neutrophils”. In: *Proceedings of the National Academy of Sciences* 107.20, pp. 9204–9209.
- Yakovlieva, Liubov, Julius A Fülleborn, and Marthe TC Walvoort (2021). “Opportunities and challenges of bacterial glycosylation for the development of novel antibacterial strategies”. In: *Frontiers in Microbiology*, p. 2737.

- 
- Yamada, Issaku et al. (2021). “The glycoconjugate ontology (GlycoCoO) for standardizing the annotation of glycoconjugate data and its application”. In: *Glycobiology* 31.7, pp. 741–750.
- Zacchi, Lucia F and Benjamin L Schulz (2016). “SWATH-MS glycoproteomics reveals consequences of defects in the glycosylation machinery”. In: *Molecular & Cellular Proteomics* 15.7, pp. 2435–2447.
- Zeng, Daina et al. (2016). “Approved glycopeptide antibacterial drugs: mechanism of action and resistance”. In: *Cold Spring Harbor perspectives in medicine* 6.12.
- Zhang, Kai et al. (2017). “Glycosylation Profiling of  $\alpha/\beta$  T Cell Receptor Constant Domains Expressed in Mammalian Cells”. In: *Synthetic Antibodies: Methods and Protocols*, pp. 197–213.