The Branco Weiss Fellowship Society in Science

WALLENBERG CENTRE FOR MOLECULAR AND TRANSLATIONAL MEDICINE

Team Bojar
Applied glycan data science and machine learning

Luc Thomès[1,2], Daniel Bojar[1,2]
[1]Department of Chemistry and Molecular Biology, University of Gothenburg, Gothenburg, Sweden.
[2]Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Gothenburg, Sweden.

Glycowork: a Python package for glycan data science & machine learning

Glycans are **polysaccharides** involved in many biological processes. Their roles are achieved thanks to the high combinatorial potential of chained monosaccharides that result in a large diversity of molecules with **specific properties.** Because of this inner complexity, studying glycans remains a challenge that requires **powerful automated approaches** to be tackled. Glycowork is an open-source Python package designed for **glycan-related data science** and **machine learning**. It consists of five modules containing functions to **annotate** glycan motifs, **visualize** their distribution via heatmaps and statistical enrichment, generate biosynthesis **networks** as well as to **build or use provided machine learning models**.

In glycowork, glycans are taken in the **IUPAC-condensed format** but internally processed **as graphs**. Glycan motifs are then detected as **subgraphs**, allowing the **annotation** of glycans and helping for further investigations (Figure 1). **Learned representations** of glycans obtained from a deep learning model are provided with glycowork, allowing **clustering** and **visualization**.

https://github.com/BojarLab/glycowork/

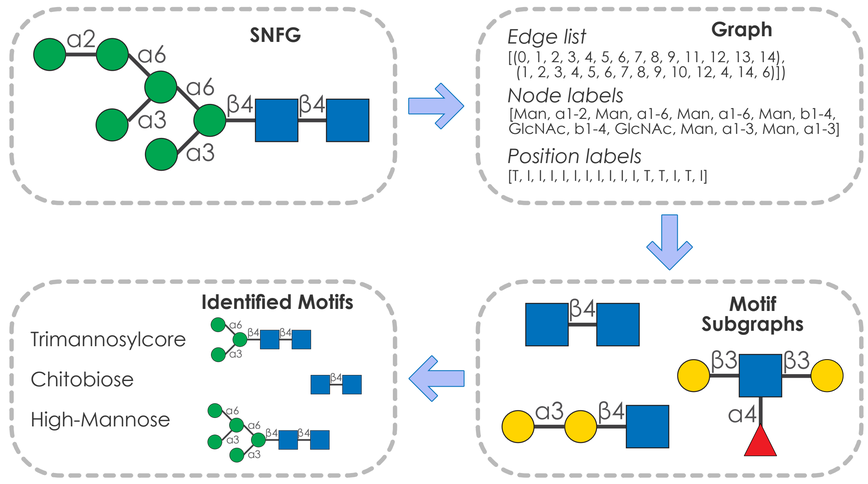

Figure 1. Annotation of glycans in glycowork

Fucose: a widespread monosaccharide involved in inter-organism communication

Fucose usage across the different domains of life

Fucose-containing glycans are found **among all kingdoms** but with **large variations in terms of their proportion**. **Animals** and **viruses** show the **highest proportions of fucose-containing glycans** while in **Fungi**, **Protista** and **Archaea**, **less than 4.5% of their glycans are fucosylated** (Figure 2).

Kingdom	Total glycans	Fucosylated glycans	Ratio (%)
Animalia	8260	3861	46.7
Bacteria	8072	1363	16.9
Fungi	3042	124	4.1
Plantae	2873	614	21.4
Protista	352	6	1.7
Virus	244	111	45.5
Archaea	45	2	4.4

Figure 2. Ratio of total versus fucosylated glycans across kingdoms

Roles of fucosylated glycans in symbiotic bacteria

Helicobacter pylori is a human-associated bacteria that produces fucosylated glycans **mimicking the Lewis blood group antigens** found at the surface of the stomach epithelium for immune evasion. With the notable exception of *H. pylori* and few other species, we noticed that **fucose-rich bacteria** are rather symbionts associated with **plant roots** (Figure 3).

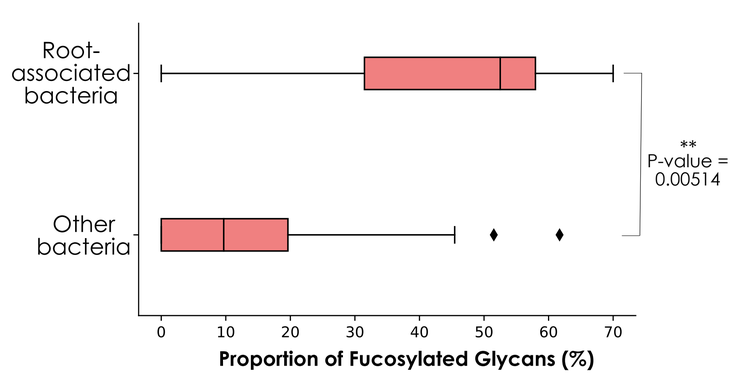

Figure 3. Fucose content across bacterial glycans

Roles of fucosylated glycans during *E. coli* infection

Under given circumstances, certain *Escherichia coli* strains infiltrate the human gastrointestinal and urinary tracts to **induce infections**. The representations of fucosylated glycans from a set of **pathogenic** or **non-pathogenic strains** showed a **segregation** of the dots according to their pathogenic potential (Figure 4). **Enriched motifs** computed in each condition showed that only the pathogenic strains **favored usage of fucose to mimic host glycans**. They displayed the **Fuc(α1-2)Gal motif** observed in **human Lewis antigens**, presumably to **evade the immune system**.

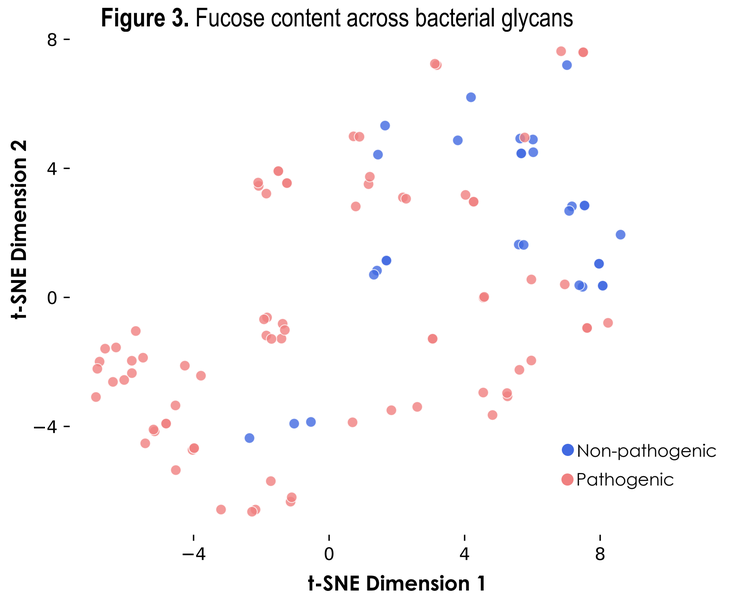

Figure 4. Representation of fucosylated glycans from pathogenic and non-pathogenic *E. coli* strains

Near-sequons, reservoirs of potential new glycosylation sites in viruses

Near-sequons are protein motifs that are **one mutation away** of being *N*-glycosylation sites called **sequons**. In several viral species, the binding proteins interacting with host cells are **enriched both in sequons and near-sequons**. Knowing the important role of glycosylations on these proteins, near-sequons constitute interesting markers to predict the apparition of potential variants.

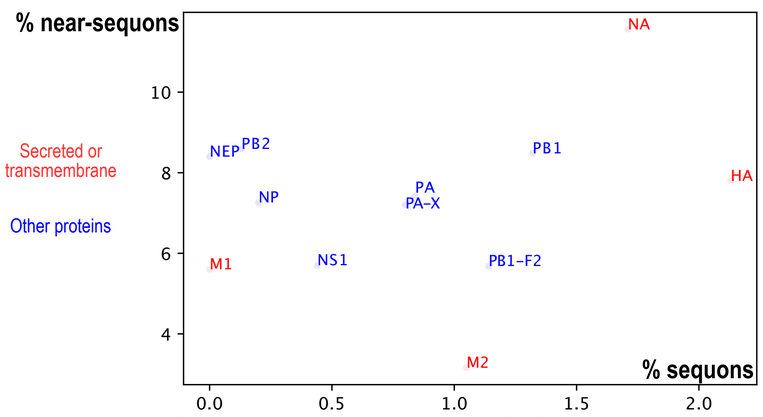

Figure 5. Sequon and near-sequon content of influenza A proteins

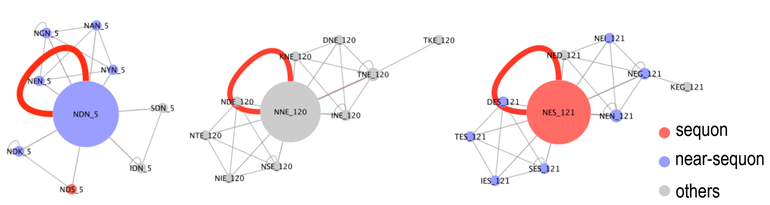

Figure 6. Near-sequon-to-sequon transitions in influenza A hemagglutinin

The role of gut glycans in ulcerative colitis and during *C. difficile* infections

Ulcerative colitis and *Clostridioides difficile* infections are pathological conditions affecting the human intestine, leading to **colon inflammation**. We analyzed different glycan types detected in patients and compared them to control samples to identify the **specific** and **conserved glycan motifs** involved. Preliminary results are consistent with publicly available RNA-seq data showing the differential expression of several glycoenzymes. They highlight **a relation between gut glycans** and the **development of these pathological conditions**.

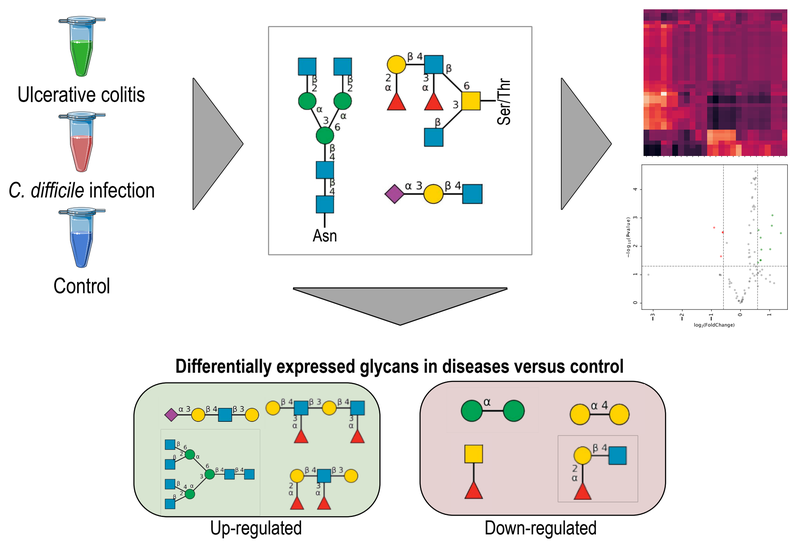

Figure 7. General workflow for the comparative analysis of glycan motifs

Present in all living organisms and able to interact with proteins, lipids and even RNAs, glycans are **involved in virtually every biological process**.

We have developed **glycowork**, a Python package aimed to facilitate the analysis of glycan data and have applied it to various situations to expand the frontiers of our knowledge in glycobiology. We have shown the importance of glycans from a monosaccharide (**fucose**) point of view as well as from a **glycoprotein side** where polysaccharides modulate the functions of viral binding proteins. Finally, we have also investigated the roles of glycans in **human diseases** with the aim to contribute to the development of curative strategies.

References:

Thomès, L., Burkholz, R., and Bojar, D. (2021). Glycowork: A Python package for glycan data science and machine learning. Glycobiology,1-5. doi:10.1093/glycob/cwab067

Thomès, L., and Bojar, D. (2021). The Role of Fucose-Containing Glycan Motifs Across Taxonomic Kingdoms. Front. Mol. Biosci. 8, 755577. doi:10.3389/fmolb.2021.755577.