Glycan semantic model applications in substructure search

Vincenzo Daponte[1,2][0000−0003−3660−0270], Catherine Hayes[1,2], Julien Mariethoz[1,2][0000−0001−8974−4417], and Frederique Lisacek[1,2][0000−0002−0948−4537]

[1] Department of Computer Science, University of Geneva, Geneva, 1227, Switzerland
[2] Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, Geneva, 1211, Switzerland

Abstract. The complexity of glycan structures as branched tree-like molecules presents a challenge when trying to obtain a unique representation. Semantic web technologies proved to be useful in this domain, and thus the GlySTreeM knowledge base has been designed to tackle the search for complex residues substructures. This work aims to present some of the most relevant applications of this semantic platform and to prove its effectiveness through the results obtained.

Keywords: Glycan · Ontology · Knowledge Base · Substructure search.

1 Introduction

1.1 Background and context

Glycans are branched tree-like molecules composed of building blocks (monosaccharides and substituents) linked by chemical bonds (glycosidic linkages). Due to the presence of at least four bondable carbons on each monosaccharide and geometric variations (anomers) around the first carbon of the child monosaccharide, there are multiple permutations within and across the assembly of monosaccharides into oligosaccharides also called glycans. These molecules are attached to proteins (then called glycoproteins) and play vital roles in protein-protein interactions, especially at the cell surface where they are abundant. As such, they impact cell-cell or host-pathogen communication. This is clear in the ongoing COVID pandemic, which has demonstrated the importance of the glycan shield covering the spike protein in protecting the virus from attack by the human immune system [8] as well as the interaction with glycans on the human ACE2 receptor[13] .

Glycans can be decomposed into three main areas: 1. the core structure, linked to the larger carrier molecule (usually a protein), 2. the extended region, and 3. the terminal epitope/antigen (binding part). In a high number of cases, it is the terminal region that determines the interactions with receptors and, therefore, the functionality associated with the particular glycan. Therefore it is important to be able to efficiently search for these 'substructures' within a glycoprofile (the complete set of glycans attached to a protein or a tissue).

Due to the complexity of glycan structures (branching, anomericity, linkages), they do not lend themselves to unique linear representations. GlycoCT [9] has become a standard format as it portrays the glycan as a connection tree. GlyConnect [2] is a platform that brings together a number of resources and tools to characterise glycoproteins, including glycan structures which are stored in GlycoCT format. A previous study introduced the idea of using RDF as an efficient alternative to GlycoCT [3]. However, valuable as this particular RDF triple store was, it was not without its limitations, namely there was no functionality to search by monosaccharide composition, and there was a disconnect between the visual model (SNFG nomenclature) [12] and the GlycoCT knowledge model. This can be explained by taking the example of the monosaccharide N-acetyl galactosamine (GalNAc). In the SNFG cartoon model, GalNAc is depicted as a yellow square, one monosaccharide, whereas in GlycoCT format it is coded as two separate 'residues', a galactose base with an attached N-acetyl substituent. To overcome the limitations and enhance the functionalities in search of residue substructures, we have designed a unique glycan ontology as the base of the GlySTreeM [5] knowledge base developed to handle detailed research on glycan data.

2 Model design

From the requirements and the domain analysis, a structured model [5] was designed to represent the most relevant concepts of the Glycans and their structure. The main components of Glycans are represented through high level classes such as ***GlycanCore*** for the core of the glycan, ***GlycanBag*** for the undefined substructure of the glycan and the ***Glycan*** class to group all these parts to one entity (see Fig. 1).

The key choice in the design of this model is represented by the semantic decoupling of the structure from its constituent elements. This choice is made explicit through the blocks defining the glycans: the residues. In many commonly used notations such as IUPAC [10] and GlycoCT [9], residues are inseparably identified with the molecule for which it stands. In the designed model, the residues are separated from their component molecules, emphasizing the decoupling between the tree structure of the residues and the molecules. The purpose of this choice is to make navigation and therefore the search for substructures more flexible by allowing queries lacking in detail, on the molecules or their bonds.

For this reason, the ***Residue*** class has been incorporated as the node of a semantic tree structure that allows navigating at all levels and without information on the molecules. These molecules are represented by the classes ***Base*** and ***Substituent*** and linked to the corresponding residue through the object properties *hasBase* and *hasSubstituent*. These classes include also the details on the anomeric links. As an example, the representation of the GalNac residue, which is very common in glycans, is shown in different syntaxes in table 1:

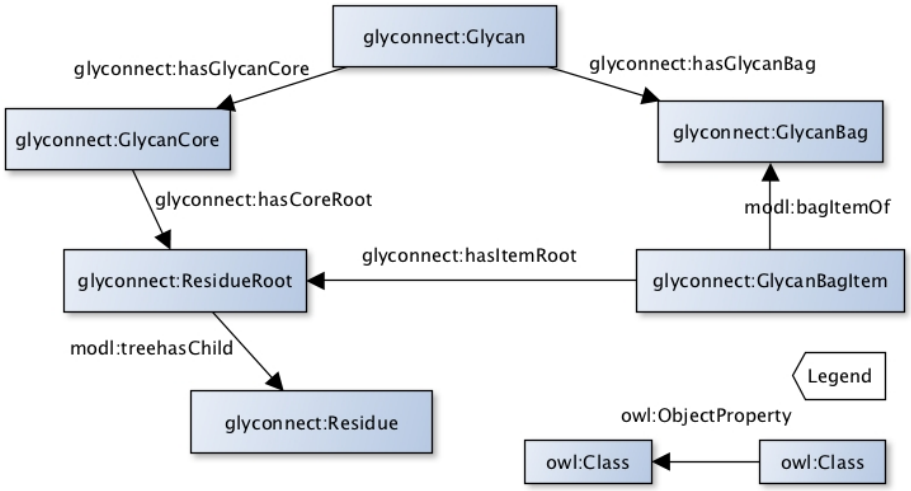

Fig. 1. The main classes representing the logical components of the glycan structure, from the Glycan to the Residue.

Table 1. Textual representation of GalNAc

IUPAC	GlycoCT
GalNAc	RES 1b:x-dgal-HEX-1:5 2s:n-acetyl LIN 1:1d(2+1)2n

In accordance with the model underlying the GlySTreeM knowledge base, this residue would be represented in Fig. 2.

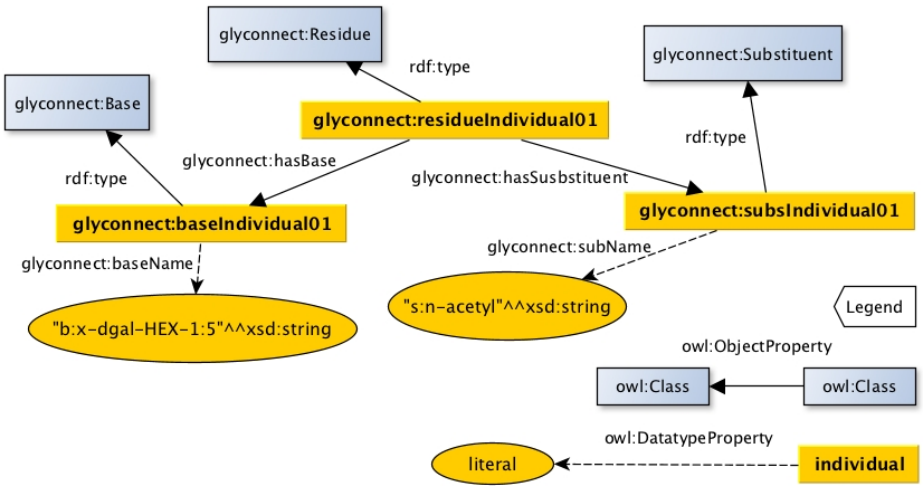

Fig. 2. The decomposition of the GalNac residue, not only in base and substituent but also separated from the *node* Residue.

Once the residue structure has been defined, it can then be used to represent the different logical components of the glycan that include it. So the core is related to the root of the residue tree (***ResidueRoot*** class) from which descend all the residues referring to the ***GlycanCore*** class. The ***GlycanBag*** class is used to group all the substructures whose link with the core is undefined. These substructures are identified as the items of the ***GlycanBag*** class, and each of them is associated with the ***ResidueRoot*** of each substructure.

The semantic model was implemented through an ontology and deployed in a triple store. The ontology was semantically validated by aligning it with the SKOO ontology [4], which is a scientific knowledge ontology model aligned and validated with reference scientific and general ontologies such as SIO [6] and Dolce [7]. Individuals obtained by automatic mapping from the GlycoCT encoded structure available in Glyconnect [1] were imported to this triple store, thus forming the GlySTreeM knowledge base. This knowledge base is used for investigations requiring specific levels of detail and exploiting the intrinsic features of the model. Some of the main features will be shown in the next section through the description of some research scenarios in which GlySTreeM was put to the test.

3 Applications

Some of the features provided by this knowledge base, such as the search with incomplete information on molecules and the flexibility in substructure queries have been used in glycan based research scenarios. In this section are presented three use cases that are considered significant in demonstrating the potential of the GlySTreeM Knowledge Base.

3.1 Use Case 1: Search for a particular structure

In GlyConnect there are a number of different types of glycans. The most common of these are N- and O-linked structures. While an experienced glycobiologist will immediately recognise the main features of these structures, it is useful to have a method to assign the types as well as the sub-types or 'cores'. The GlySTreeM knowledge base can be used to search for individual structures or types of structures. For example, does GlyConnect contain a Core 1 type O-linked glycan with two sialic acid (NeuAc) residues in line, Tab. 2?

Table 2. Use Case 1: Results

Query	Details	Results
1	O-linked structures	1233
2	O-linked and Core 1	296
3	O-linked, Core 1 and contains two sialic acids	1

3.2 Use Case 2: Fucosylation of N-linked, complex type glycans

The presence of core fucose on N-linked glycosylation of therapeutic antibodies has been shown to be of great importance [11]. To this end, it would be interesting to investigate the fucosylation status of the structures held in GlySTreeM.

If N-Linked, complex type glycans are interrogated using SPARQL queries, patterns start to emerge. This subset reflects the entire knowledgebase, in that almost two third are fucosylated (mirroring the first two results), Tab. 3. There are 714 sequenced structures (position of monosaccharides are known if not complete linkage) that are core-fucosylated.

3.3 Use Case 3: Comparison of alpha 2-3 and 2-6 sialic acid linkages

Another important terminal residue in glycans is sialic acid. It can be found in both alpha 2-3 and alpha 2-6 linkages, and this query investigates the proportion of those on bi-antennary N-linked structures. Firstly we have 109 N-linked, complex, bi-antennary sequenced structures. Using the SPARQL queries, we can pull out the proportions of these that have undetermined linkages for NeuAc (22%),

Table 3. Use Case 2: Results

Query	Details	Results
1	Structure with no Fuc	2022
2	Structures with Fuc	2757
3	N-linked, complex structures	2467
4	N-linked, complex and contain at least one Fuc	1571
5	Terminal Fuc (no core Fuc)	66
6	Terminal Fuc (no core Fuc) and no undefined residues	60
7	Only core Fuc and no undefined residues	714

all 3-linkages (22%), all 6-linkages (37%), or a mixture (29%), Tab. 4. There is a slight overlap as we have a small number of bi-antennary structures with three NeuAc residues.

Table 4. Use Case 3: Results

Query	Details	Results
1	N-Linked, Complex, no undefined structure and at least one NeuAc	541
2	N-linked, complex, no UND, bi-ant and at least one NeuAc	272
3	N-linked, complex, no UND, bi-ant and one NeuAc on each arm	117
4	N-linked, complex, no UND, bi-ant and one Neu5Ac on each arm	109
5	N-linked, complex, no UND, bi-ant, one Neu5Ac on each arm, with undefined linkage	24
6	N-linked, complex, no UND, bi-ant, one Neu5Ac on each arm, with 3 linkage	24
7	N-linked, complex, no UND, bi-ant, one Neu5Ac on each arm, with 6 linkage	40
8	N-linked, complex, no UND, bi-ant, one Neu5Ac on each arm, with 3 and 6 linkage	31

4 Conclusions

Ontologies have proven to be valuable tools in the representation of glycan structures. We present here a method to search these complex molecules for substructures which in a lot of cases, equate to interactions and/or functions of the glycan. The scenarios on which the GlySTreeM Knowledge base has been employed allows the demonstration of the flexibility and expressivity of the ontology behind it. The base will be further tested on current data for consistency checks of glycan classifications on types, cores, and structural patterns such as antennas, core-fucosylation, or bisecting GlcNAc.

References

1. Alocci, D., Mariethoz, J., Gastaldello, A., Gasteiger, E., Karlsson, N.G., Kolarich, D., Packer, N.H., Lisacek, F.: GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. Journal of proteome research **18**(2), 664–677 (2018)
2. Alocci, D., Mariethoz, J., Gastaldello, A., Gasteiger, E., Karlsson, N.G., Kolarich, D., Packer, N.H., Lisacek, F.: GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. Journal of proteome research **18**, 664–677 (Feb 2019). https://doi.org/10.1021/acs.jproteome.8b00766
3. Alocci, D., Mariethoz, J., Horlacher, O., Bolleman, J.T., Campbell, M.P., Lisacek, F.: Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search. PloS one **10**, e0144578 (2015). https://doi.org/10.1371/journal.pone.0144578
4. Daponte, V., Falquet, G.: An ontology for the formalization and visualization of scientific knowledge (07 2021)
5. Daponte, V., Hayes, C., Mariethoz, J., Lisacek, F.: Dealing with the ambiguity of glycan substructure search. Molecules **27**(1) (2022). https://doi.org/10.3390/molecules27010065, https://www.mdpi.com/1420-3049/27/1/65
6. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., et al.: The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. Journal of biomedical semantics **5**(1), 1–11 (2014)
7. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 166–181. Springer (2002)
8. Gstöttner, C., Zhang, T., Resemann, A., Ruben, S., Pengelley, S., Suckau, D., Welsink, T., Wuhrer, M., Domínguez-Vega, E.: Structural and Functional Characterization of SARS-CoV-2 RBD Domains Produced in Mammalian Cells. Analytical chemistry **93**, 6839–6847 (May 2021). https://doi.org/10.1021/acs.analchem.1c00893
9. Herget, S., Ranzinger, R., Maass, K., Lieth, C.W.V.D.: GlycoCT-a unifying sequence format for carbohydrates. Carbohydrate research **343**, 2162–2171 (Aug 2008). https://doi.org/10.1016/j.carres.2008.03.011
10. Moss, G.P.: Basic terminology of stereochemistry (IUPAC Recommendations 1996). Pure and applied chemistry **68**(12), 2193–2222 (1996)
11. Pereira, N.A., Chan, K.F., Lin, P.C., Song, Z.: The "less-is-more" in therapeutic antibodies: Afucosylated anti-cancer antibodies with enhanced antibody-dependent cellular cytotoxicity. mAbs **10**, 693–711 (Jul 2018). https://doi.org/10.1080/19420862.2018.1466767
12. Varki, A., Cummings, R.D., Aebi, M., Packer, N.H., Seeberger, P.H., Esko, J.D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., Prestegard, J.J., Schnaar, R.L., Freeze, H.H., Marth, J.D., Bertozzi, C.R., Etzler, M.E., Frank, M., Vliegenthart, J.F., Lütteke, T., Perez, S., Bolton, E., Rudd, P., Paulson, J., Kanehisa, M., Toukach, P., Aoki-Kinoshita, K.F., Dell, A., Narimatsu, H., York, W., Taniguchi, N., Kornfeld, S.: Symbol Nomenclature for Graphical Representations of Glycans. Glycobiology **25**, 1323–1324 (Dec 2015). https://doi.org/10.1093/glycob/cwv091
13. Zhao, P., Praissman, J.L., Grant, O.C., Cai, Y., Xiao, T., Rosenbalm, K.E., Aoki, K., Kellman, B.P., Bridger, R., Barouch, D.H., Brindley, M.A., Lewis, N.E., Tiemeyer, M., Chen, B., Woods, R.J., Wells, L.: Virus-Receptor Interactions of

Glycosylated SARS-CoV-2 Spike and Human ACE2 Receptor. Cell host & microbe **28**, 586–601.e6 (Oct 2020). https://doi.org/10.1016/j.chom.2020.08.004