GlycoTree: Infrastructure Supporting Semantic Annotation of Glycan Structures

Will York[1], Rene Ranzinger[1], Nathan Edwards[2], Wenjin Zhang[2] and Michael Tiemeyer[1]

[1]Complex Carbohydrate Research Center, The University of Georgia, Athens, GA, USA

[2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, Washington, DC, US

Summary: GlycoTree is just one aspect of the semantic technology being developed by the *the National Institutes of Health Common Fund* **GlyGen** initiative (https://www.glygen.org/home/). GlycoTree enables semantic annotation of glycans by specifying glycan structures as collections of prototypes (canonical residues) representing their chemical components. GlycoTree provides a means of relating glycan structures to their biosynthesis, degradation, and biological activities. So far, the GlycoTree infrastructure has allowed us to infer and publish the identities of enzymes involved in the biosynthesis of more than 3850 glycans. At present, we disseminate this type of information via the GlyGen Sandbox web site: https://glygen.ccrc.uga.edu/sandbox/

We are continuing to extend the GlycoTree approach, for example, by adding new residues to the canonical trees, annotating many of these residues with extensive biosynthetic information, and leveraging this information as a basis to automate the generation and representation of complete biosynthetic pathways for glycans.

Due to the compexity of the information that can be inferred from GlycoTree, the development of a graphical user interface (GUI) to explore and represent this information is challenging, especially considering our aim to provide relevant information that can be easily retrieved by both novice and expert glycoscientists. Our approach to this task has been to first develop independent but fully functional interfaces that are hyperlinked to the main GlyGen web portal, and completely integrate these prototypes into the portal only after thorough testing and evaluation of community feedback. A prototype GlycoTree web application is now sufficiently mature for testing and can be accessed at https://glygen.ccrc.uga.edu/sandbox/

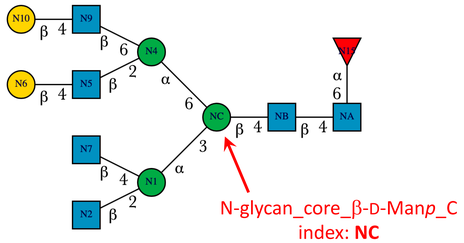

Figure 1. GlycoTree defines *prototype residues* by specifying the identity (e.g., β-D-Man*p*) and chemical context of each of these within a comprehensive *canonical tree.* Each prototype is named (e.g., "N-glycan_core_β-D-Man*p*_C"), given a concise index (e.g., "NC") and mapped to a specific node in the canonical tree. The structure of a particular glycan can be completely and unambiguously represented as a collection of prototype residues (labeled with their indices in the Figure). This semantic annotation not only provides an extremely concise way to represent glycan structure, but also allows humans and computer programs to "talk" to each other about each specific residue in the glycan and to process information associated with it. Specifically, the canonical representation provides a means to infer (explicit and implicit) information related to the atomic structure and chemical context of each residue. For example, human experts in glycobiology know what is meant by the "N-glycan core β-mannose", and can infer from this name the chemical context of this residue within the glycan and a significant amount of other detailed information regarding its biosynthesis, enzyme-catalyzed addition to the nascent glycan structure, biodegradation, and participation in molecular interactions. GlycoTree provides a formal way to represent this kind of information *for all of the residues in a glycan*, even those that have not been named by human experts.

Although the complexity of this information makes it impossible for a human to recall in its entirety, it is readily saved, retrieved, processed and visualized by the digital methods that we have developed.

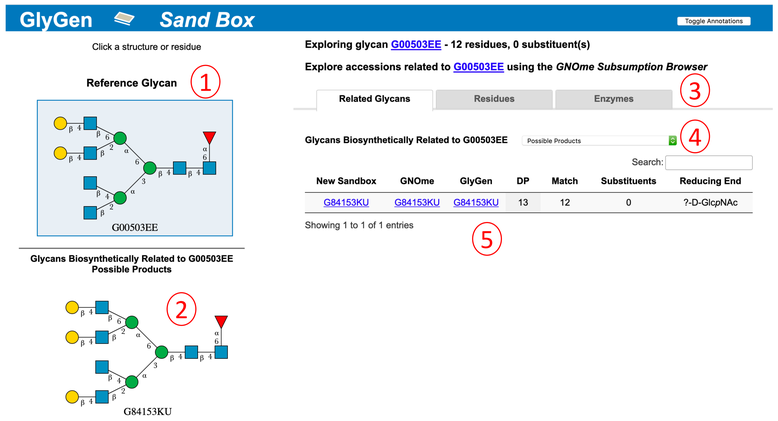

Figure 2A. Screen shot of the GlycoTree "Sandbox" Interface, illustrating some of its main features, which are annotated using red, circled numbers in the Figure:

1. A specific glycan is selected as the **Reference Glycan**, which is the focus of the active page
2. **Biosynthetically Related Gycans** are rendered below the reference glycan.
3. Three **tabs** allow different types of information (*Related Glycans*, *Residues* and *Biosynthetic Enzymes*) to be displayed.
4. A **filter** in the *Related Glycans* tab allows selection of specific classes of biosynthetically related glycans, such as possible products or possible precursors. In this example, *possible biosynthetic products are* selected and rendered in the column on the left.
5. A **list** of (filtered) Biosynthetically Related Glycans is provided. For each glycan in the list, hyperlinks to the GlyGen Portal and the GNOme Subsumption Browser are also provided, along with a link to a new Sandbox page that focuses on the related glycan as its *Reference Glycan*. Limited structural information for each glycan in the list includes its DP, the number of canonical residues that it shares with the *Reference Glycan*, and the number of non-glycosyl substituents.

These features allow the biosynthetic pathway for a particular glycan to be traversed manually (using features 4 and 5). Automated protocols to generate such pathways by leveraging the GlycoTree infrastucture are under development.

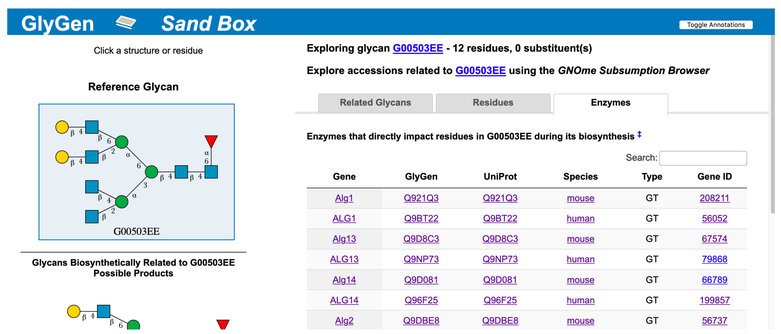

Figure 2B. Screen shot of the GlycoTree "Sandbox" Interface, illustrating the **Enzymes** tab, which contains a list of all enzymes that have been mapped to the *Reference Glycan*. This list includes links to different informatics resources that provide extensive information about each enzyme in the list.

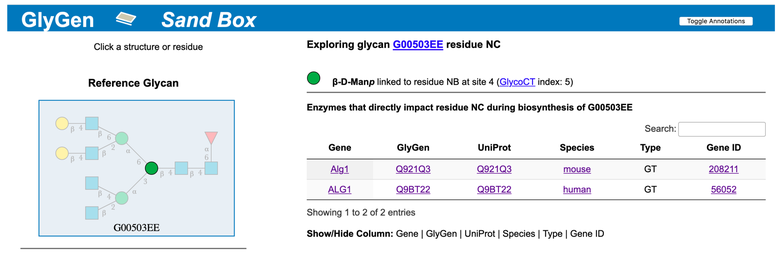

Figure 2C. Screen shot of the GlycoTree "Sandbox" Interface, illustrating the results obtained when *one* of the residues in the glycan is clicked. All other residues are "greyed out" and a new list containing enzymes that have been specifically mapped to the clicked residue in the *Reference Glycan* is shown.

Further annotation of the GlycoTree is underway, allowing this list will be extended to include other glyco-enzymes such as glycosyl hydrolases and enzymes involved in the biosynthesis of nucleotide sugars.

GlycoTree is based on concepts originally introduced by Takahashi and Kato (2003, https://www.jstage.jst.go.jp/article/tigg1989/15/84/15_84_235/_pdf/-char/ja), and extended by implementing a formal description logics approach (York, et al. 2004, https://unit.aist.go.jp/brd/jp/GTRC/HGPI/ws1/pro.html). A rigorous implementation of glycoTree was developed as the GlycO ontology (https://bioportal.bioontology.org/ontologies/GLYCO) and leveraged to enable curation of glycan structure records using the Qrator software (Eavenson, et al., 2015, PMID: 25165068).

We gratefully acknowledge funding of the GlyGen project by the United States National Institutes of Health (NIH) Common Fund Glycoscience Program (grant 1U01GM125267-01) through the National Institute of General Medical Sciences (NIGMS).