GNOme – Glycan Naming and Subsumption Ontology

Wenjin Zhang [1] and Nathan J. Edwards [1]

[1] *Clinical and Translational Glycoscience Research Center, Georgetown University, Washington, D.C., USA*

Abstract

The Glycan Naming and Subsumption Ontology (GNOme) is an OBOFoundry ontology that organizes the stable glycan accessions of GlyTouCan glycan sequence registry for reasoning and browsing by subsumption. The ontology enables the fast, intuitive, and interactive exploration of GlyTouCan's glycan structure accessions for glycan sequences; facilitates text-based lookup for common synonyms for structures and their GlyTouCan accessions; provides a framework for automated reasoning about glycan subsumption relationships; and annotates glycans with well-defined characterization categories. As part of the OBOFoundry, GNOme can be readily used by other ontology and standards initiatives to refer to glycans at varying degrees of characterization and is currently integrated with the GlyGen glycoinformatics resource to help users find "related glycans" and to propagate species and glycan classification annotations by subsumption.

Keywords

Glycans, ontology, subsumption

1. Introduction

Glycan sequence formats, especially GlycoCT [1] and WURCS [2], in common use by the glycobiology and glycoinformatics community for describing glycan molecules, explicitly specify, or indicate the absence of knowledge about, every detail of a glycan's structure. Experimental techniques for characterizing glycans are often unable to fully characterize glycans, and these sequence formats precisely record which details of a glycan structure are known or not known. However, the resulting long, complex, and cryptic sequences cannot be readily communicated or shared. The GlyTouCan [3] glycan (sequence) registry provides stable accessions for glycan sequences to facilitate communication and knowledge sharing. Unfortunately, the number of accessions and their glycan sequences in GlyTouCan have grown beyond 100,000, and it has become very difficult to find the corresponding GlyTouCan accessions for experimentally characterized glycans since they are not organized with respect to the degree of glycan characterization of each sequence, that is, by subsumption.

The Glycan Naming and Subsumption Ontology (GNOme) is an OBOFoundry [4] ontology that organizes the stable glycan sequence accessions of GlyTouCan for reasoning and display by subsumption. The ontology enables the fast, intuitive, and interactive exploration of GlyTouCan glycan structure accessions for experimental glycans characterized to a specific extent; facilitates text-based lookup for common synonyms for structure and their GlyTouCan accessions; provides a framework for automated reasoning about glycan subsumption relationships; and annotates glycans with well-defined characterization categories. As part of the OBOFoundry, GNOme can be readily used by other ontology and standards initiatives to refer to glycans at varying degrees of characterization, and is currently integrated with the GlyGen glycoinformatics resource [5] help users find "related glycans" and to propagate species and glycan classification annotations by subsumption.

International Conference on Biomedical Ontologies 2021, September 16–18, 2021, Bozen-Bolzano, Italy

EMAIL: nje5@georgetown.edu (A. 2)
ORCID: 0000-0001-5168-3196 (A. 2)

2. Methods

The GNOme ontology is computationally determined from the glycan sequences and accessions of GlyTouCan. The glycans, are first grouped by molecular weight, then all structures in a group aligned with each other, establishing their mutual subsumption relationships. Redundant subsumption relationships, which can be established by transitivity of the subsumption partial order, are removed. The subsumption-based characterization categories, in most characterized to least characterized order: Saccharide, Topology, Composition, Base Composition, and Molecular Weight are derived from the ROCS ontology [6] developed by the GlyTouCan project. Each subsumption category indicates the partial presence or complete absence of specific information in the structure description sequence. While the molecular weight category is not part of ROCS, it can be readily computed from the structure description, is often the result of mass-spectrometry based glycan characterization, is invariant with respect to our subsumption definition, and represents the grouping that drives our computational alignment strategy. The removal of specific classes of structural information can transform a structure to another with a subsuming characterization category, though not all such transformations will be reflected by a registered sequence in GlyTouCan. Table 1 shows the categories and the specific information whose absence defines membership in each category.

Table 1
GNOme subsumption categories and missing structure information. X indicates the complete absence of the indicated structure information.

Subsumption Category	Superclass	Stereochemistry	Linkage & Ring Information	Anomers & Carbon Bond Positions
Molecular Weight	✗	✗	✗	✗
Base Composition		✗	✗	✗
Composition			✗	✗
Topology				✗
Saccharide				

We automatically categorize and partition the glycans based on the presence or absence of these specific pieces of information and determine whether the glycan sequences in a molecular weight group are related by the removel transformations described above.

The GNOme ontology uses GlyTouCan accessions to define its primary class terms, with the required OBO Foundry structure (e.g. GNO_G00912UN). When a glycan's sequence describes a glycan composition or base composition that is consistent with a so-called composition string that succinctly describes the number of each type of monosaccharide, we associate the composition string, in various formats, with the GNOme term, as a synonym.

When GlyTouCan accessions representing structure sequences that have been published as GNOme terms in a release are subsequently archived or replaced, this is reflected in the ontology by marking it as obsolete and indicating the replacement accession if available.

In order to support glycoinformatics resources that capture a subset of GlyTouCan's structures, GNOme also releases ontology restrictions to subsets of GlyTouCan accessions, in which only the relevant GNOme terms are retained and subsumption relationships inferred by transitivity from the primary ontology. GNOme supports ontology restrictions for GlyGen, BSCDB [7], and GlyCosmos [8].

3. Results and Discussion

3.1. Construction of OBO Foundry Ontology

The GNOme ontology has been constructed as an OBO Foundry ontology with prefix GNO and released as an OWL-format ontology with automatically generated OBO and JSON derived formats using the robot tool [9]. In addition to GNOme classes for each supported GlyTouCan accession, GNOme creates molecular weight class terms, to represent the molecular weight grouping of subsumption relationships (to two decimal places), and a root Glycan class that subsumes each of the molecular weight terms. The subclass predicate is used to represent the subsumption relationship. The subclass relationships define a DAG, not a tree, since a given structure may have multiple immediately subsuming structures.

GNOme provides a variety of annotation properties that provide more information to each class term representing a GlyTouCan accession and its sequence:

- *has_glytoucan_id*, *has_glytoucan_link*: GlyTouCan accession and deep linking URL to the corresponding GlyTouCan webpage.
- *has_subsumption_category*: GNOme URI of the subsumption category.
- *has_basecomposition*; *has_composition*; *has_topology*: GNOme URI of the structure with the appropriate information removed, if it exists.
- *has_structure_browser_link*, *has_composition_browser_link*: URL of deep link to interactive GNOme browser web-applications for structures and compositions, where appropriate.
- *has_Byonic_name*, *hasExactSynonym*: Synonyms for the GNOme term, including specific predicate for composition strings as formatted by the Byonic glycopeptide identification search engine. *hasExactSynonym* is defined in the oboInOwl ontology.

In addition, GNOme uses the *definition* property from the Information Artifact Ontology, predicates from the oboInOwl ontology (*hasExactSynonym* in particular), and typical Web Ontology Language terms, such as *label*. The GNOme ontology strives to meet and abide by the OBO Foundry principles, adopting annotation properties consistent with other OBO Foundry ontologies, and as such is well-integrated with ontology support services such as OntoBee [10] and OLS [11], and can be readily referenced by other OBO Foundry ontologies.

In all, GNOme (version 1.7.2, June 16, 2021) defines 11 annotation properties, 5 named instances (subsumption categories), and 111,696 classes. The classes are made up of 1 root glycan class, 1 subsumption category class, 13,698 molecular weight classes, and 97,996 classes representing GlyTouCan accessions, of which 7,775 represent GlyTouCan obsoleted accessions.

3.2. Interactive Exploration of Glycans by Subsumption

The glycan structures represented by GlyTouCan accessions and sequences are not easily understood as human readable labels or other text that is the basis for the widely available ontology browsing and interrogation tools. Glycobiologists, GNOme’s primary target audience, typically use so-called cartoons or images to communicate and describe a glycan structure. While these images do not necessarily reveal all details of a structure’s characterization or sequence, they represent the most accessible method for interactive navigation and browsing of glycan structures. We have constructed built-for-purpose, visual, image-based web-applications for browsing and exploring glycan structures by subsumption. The web-applications rely on users to visually select from available glycan topologies and then navigate up and down the subsumption relationships by choosing visual representations of glycans that match their conceptual glycan of interest.

The Structure Browser web-application has two panes - the Topology Selector and Subsumption Navigator. The Topology Selector provides buttons to select the monosaccharide composition of the glycan of interest, displaying representative topology images in real-time as the monosaccharide composition changes. The interactive update of topologies provides feedback to the user on the types of glycan structures consistent with the current state of the monosaccharide composition buttons. Selecting a topology switches focus to the Subsumption Navigator pane, where the immediate parents (subsuming) and children (subsumed by) of the current structure are shown. A double click refocuses the Subsumption Navigator on a structure higher or lower in the subsumption hierarchy. The Composition Browser web-application works similarly but is focused on glycans with subsumption category Composition or Base Composition. A pop-up menu provides the opportunity to jump to

GlyTouCan or GlyGen, place descendants or ancestor accessions on the clipboard, or switch between Structure and Composition Browsers.

Importantly, the Structure Browser and Composition Browser web-applications permit deep linking by GlyTouCan accession, synonym, or composition (button state), and can be embedded in other websites. Interactive users of the Browsers can use the Find or Align tools to explore subsumption near structures identified by accession, synonym, or GlycoCT/WURCS sequence. The GNOme Structure Browser, restricted to the GlyGen glycan set, is integrated with the GlyGen glycoscience knowledgebase as a "Related Glycans" button on each Glycan page; and GlyCosmos integrates the Structure Browser as one of its many glycan search strategies.

3.3. Automated Subsumption Reasoning

In addition to the interactive use of the GNOme Structure Browser to find related glycans, the GlyGen project also makes use of the subsumption relationships computed and provided by GNOme for annotation propagation. A long-standing issue with glycan annotation is the scattering of annotations for a single conceptual glycan structure across many accessions related by subsumption. Mass-spectrometry based inference typically annotates composition or base composition category structures, while more detailed characterization techniques might determine the location of fucosylation on complex monosaccharides and carbon bond positions of glycosidic linkage. In this case, propagation of important annotations, such as the species observed with a glycan, via the subsumption relationships of GNOme, allow information known for more well characterized structure to be associated with less well characterized structures that subsume them. In addition to species, GlyGen also propagates glycan type and subtype annotations via subsumption relationships.

3.4. Use by Other Ontology Projects

GNOme, as an OBO Foundry ontology, has been adopted by a number of other ontology projects, including ChEBI [12] and the Protein Ontology (PRO) [13], by the HUPO Protein Standards Initiative post-translational modification ontology (PSIMOD) [14], and the HUPO Protein Standards Initiative effect to support glycopeptide identifications in mzIdentML [15].

4. Conclusion

We hope GNOme helps bridge the semantic gap between glycoinformatics data-resources and the glycobiology community they serve.

5. References

[1] Herget S, Ranzinger R, Maass K, Lieth CW. GlycoCT-a unifying sequence format for carbohydrates. Carbohydr Res. 2008 Aug 11;343(12):2162-71. doi: 10.1016/j.carres.2008.03.011. Epub 2008 Mar 13. PMID: 18436199.

[2] Matsubara M, Aoki-Kinoshita KF, Aoki NP, Yamada I, Narimatsu H. WURCS 2.0 Update To Encapsulate Ambiguous Carbohydrate Structures. J Chem Inf Model. 2017 Apr 24;57(4):632-637. doi: 10.1021/acs.jcim.6b00650. Epub 2017 Mar 22. PMID: 28263066.

[3] Tiemeyer M, Aoki K, Paulson J, Cummings RD, York WS, Karlsson NG, Lisacek F, Packer NH, Campbell MP, Aoki NP, Fujita A, Matsubara M, Shinmachi D, Tsuchiya S, Yamada I, Pierce M, Ranzinger R, Narimatsu H, Aoki-Kinoshita KF. GlyTouCan: an accessible glycan structure repository. Glycobiology. 2017 Oct 1;27(10):915-919. doi: 10.1093/glycob/cwx066. PMID: 28922742; PMCID: PMC5881658.

[4] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of

ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-5. doi: 10.1038/nbt1346. PMID: 17989687; PMCID: PMC2814061.

[5] York WS, Mazumder R, Ranzinger R, Edwards N, Kahsay R, Aoki-Kinoshita KF, Campbell MP, Cummings RD, Feizi T, Martin M, Natale DA, Packer NH, Woods RJ, Agarwal G, Arpinar S, Bhat S, Blake J, Castro LJG, Fochtman B, Gildersleeve J, Goldman R, Holmes X, Jain V, Kulkarni S, Mahadik R, Mehta A, Mousavi R, Nakarakommula S, Navelkar R, Pattabiraman N, Pierce MJ, Ross K, Vasudev P, Vora J, Williamson T, Zhang W. GlyGen: Computational and Informatics Resources for Glycoscience. Glycobiology. 2020 Jan 28;30(2):72-73. doi: 10.1093/glycob/cwz080. PMID: 31616925; PMCID: PMC7335483.

[6] Relation Ontology Carbohydrate Structure (ROCS). URL: https://github.com/glytoucan/rocs.

[7] Toukach PV, Egorova KS. Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. Nucleic Acids Res. 2016 Jan 4;44(D1):D1229-36. doi: 10.1093/nar/gkv840. Epub 2015 Aug 18. PMID: 26286194; PMCID: PMC4702937.

[8] Yamada I, Shiota M, Shinmachi D, Ono T, Tsuchiya S, Hosoda M, Fujita A, Aoki NP, Watanabe Y, Fujita N, Angata K, Kaji H, Narimatsu H, Okuda S, Aoki-Kinoshita KF. The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. Nat Methods. 2020 Jul;17(7):649-650. doi: 10.1038/s41592-020-0879-8. PMID: 32572234.

[9] Jackson RC, Balhoff JP, Douglass E, Harris NL, Mungall CJ, Overton JA. ROBOT: A Tool for Automating Ontology Workflows. BMC Bioinformatics. 2019 Jul 29;20(1):407. doi: 10.1186/s12859-019-3002-3. PMID: 31357927; PMCID: PMC6664714.

[10] Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, Mungall C, Courtot M, Ruttenberg A, He Y. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res. 2017 Jan 4;45(D1):D347-D352. doi: 10.1093/nar/gkw918. Epub 2016 Oct 12. PMID: 27733503; PMCID: PMC5210626.

[11] Côté RG, Jones P, Martens L, Apweiler R, Hermjakob H. The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W372-6. doi: 10.1093/nar/gkn252. Epub 2008 May 8. PMID: 18467421; PMCID: PMC2447739.

[12] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016 Jan 4;44(D1):D1214-9. doi: 10.1093/nar/gkv1031. Epub 2015 Oct 13. PMID: 26467479; PMCID: PMC4702775.

[13] Chen C, Huang H, Ross KE, Cowart JE, Arighi CN, Wu CH, Natale DA. Protein ontology on the semantic web for knowledge discovery. Sci Data. 2020 Oct 12;7(1):337. doi: 10.1038/s41597-020-00679-9. PMID: 33046717; PMCID: PMC7550340.

[14] Montecchi-Palazzi L, Beavis R, Binz PA, Chalkley RJ, Cottrell J, Creasy D, Shofstahl J, Seymour SL, Garavelli JS. The PSI-MOD community standard for representation of protein modification data. Nat Biotechnol. 2008 Aug;26(8):864-6. doi: 10.1038/nbt0808-864. PMID: 18688235.

[15] Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, Selley JN, Searle BC, Shofstahl J, Seymour SL, Julian R, Binz PA, Deutsch EW, Hermjakob H, Reisinger F, Griss J, Vizcaíno JA, Chambers M, Pizarro A, Creasy D. The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics. 2012 Jul;11(7):M111.014381. doi: 10.1074/mcp.M111.014381. Epub 2012 Feb 27. PMID: 22375074; PMCID: PMC3394945.