Building a molecular glyco-phenotype ontology to decipher undiagnosed diseases

Jean-Philippe Gourdine*, David M. Koeller*,
Matthew H. Brush, Melissa A. Haendel*
Oregon Health and Science University
Portland, Oregon USA

Thomas O. Metz*
Pacific Northwest National Laboratory
Richland, WA USA

***Abstract*—Hundreds of rare diseases are due to mutation on genes related to glycans synthesis, degradation or recognition. These glycan-related defects are well described in the literature but largely absent in ontologies and databases of chemical entities and phenotypes, limiting the application of computational methods and ontology-driven tools for characterization and discovery of glycan related diseases. We are curating articles and textbooks in glycobiology related to genetic diseases to inform the content and the structure of an ontology of Molecular Glyco-Phenotypes (MGPO). MGPO will be applied toward use cases including disease diagnosis and disease gene candidate prioritization, using semantic similarity and pattern matching at the glycan level with glycomics data from patient of the Undiagnosed Diseases Network.**

Keywords—rare diseases, glycans, phenotypes, ontology

I. Introduction

Rare diseases and disorders affect 25 million people in the United States, 30 million people Europe and 350 million worldwide [1]. 7,000 rare diseases and disorders have been identified and 80% of them are related to genetic defects [2]. More than 100 of these genetic defects are related to glycan biology, affecting glycan synthesis, degradation or recognition [3]. Glycans, also referred to as sugars, carbohydrates, monosaccharides or polysaccharides, can be found free or attached to macromolecules to form glycosphingolipids when attached to lipids, N or O-linked glycoproteins when attached to proteins or glycophosphatidylinositol (GPI)-anchored glycoproteins when glycoproteins are attached to GPI.

Two broad categories define their biological roles: structural/modulatory function and intrinsic/extrinsic molecular recognition [4]. Glycan-related diseases can be observed by their glycan markers or glyco-phenotypes, defined as an abnormality related to glycan structure, level, activity, and processing. Many of these glycophenotypes can be detected in patients' bodily fluids or cells (e.g. urine [5]).

The Human Phenotype Ontology (HPO) is a structured clinical vocabulary that has been used for deep clinical phenotyping and allows the integration of data across sources and organisms. The ontology is the *de facto* standard for clinical phenotyping for rare disease diagnostics using phenotypic profile comparison [6]. Glycan-related defects and the molecules they affect are well described in the literature, but largely absent in ontologies such as the HPO, and from databases of chemical entities and phenotypes such as metabolomics databases like the 'Metabolomics Workbench' [7]. This limits our ability to apply computational methods and ontology-driven tools in the characterization and query of glycobiological diseases. Therefore, we are developing the Molecular Glyco-Phenotypes Ontology (MGPO).

The MGPO aims to address these gaps by providing an ontological representation of glycan defects that complements and integrates with existing phenotype ontologies, and informs improved representation of glycans and related concepts in orthogonal efforts. The MGPO will support computational tools for disease diagnosis, disease gene candidate prioritization, and model system discovery.

II. Results

A. Requirements Analysis and Curation Activities

MGPO development was largely driven by curation of glyco-phenotype data from the literature and datasets provided by collaborators such as the Undiagnosed Disease Program [8], and the Undiagnosed Disease Network. To date, we have manually curated glycol-phenotypes from 100 human diseases, 31 papers from human, and 11 papers from mouse related to abnormalities in glycobiology as showed in Table 1. For each glycophenotype captured, we indicated the source (e.g. PMID), the assay (e.g. MALDI-TOF), the subject (e.g. patient), the disease (e.g. OMIM), the evaluant (e.g. urine), the analyte (e.g. free oligosaccharides), the phenotype statement (e.g. accumulation of high mannose free oligosaccharides).

TABLE I. Glycobiology and disease curation spreadsheet - example of a glycophenotype for Gaucher disease

Source	Assay/Design	Subject Type	Subject Description	Evaluant	Analyte	Phenotype Statement
PMID: 23676310	Analysis of urinary FOS by MALDI-TOF/TOF, in diseases patients vs healthy controls. FOS permethylated prior to analysis.	human patients	GAUCHER DISEASE	urine	free oligosaccharides (FOS)	**accumulation of high-mannose FOS species in** Gaucher disease

B. Development of initial MGPO prototype.

With a rich set of curated data in hand, the next challenge is to synthesize this information to identify the key dimensions of

* Members of the Undiagnosed Disease Network, USA

UDN Metabolomics grant number: IH 1U01TR001395

glycan phenotypes on which the ontology would be classified. We begin by listing each distinct glyco-phenotype curated from the sources above, and looked for patterns that would reveal important features for their categorization. Four essential dimensions reveal themselves through this exercise:

1. Affected glycan characteristic. This dimension distinguishes phenotypes based on the general type of glycan feature or process affected: (e.g. glycan structure, occupancy, composition, levels, activity).
2. Affected glycan type. This dimension distinguishes phenotypes based on the class of glycan exhibiting abnormal characteristics.
3. Affected glycosylation target. This dimension distinguishes phenotypes based on the aglycone target affected by a defect.
4. Affected locus of abnormality. This dimension distinguishes phenotypes based on the subcellular location of the defect and/or the tissue or fluid where it was measured.

The next step was applying these dimensions as classification axes to build a hierarchical model. Careful consideration was given to the order in which the dimensions should be applied to confer the optimal structure to the ontology. An initial straw man was generated for testing and eliciting expert feedback for iterative improvement, and is available at http://bit.ly/23kdoyj

C. *Evaluation and feedback provided to related databases and ontologies such as ChEBI and GO*

We next evaluated glycan-related coverage of orthogonal efforts including ChEBI, GO, HPO, and various glycan databases. Here we noted an underrepresentation of concepts needed to describe glyco-phenotypes like those we had curated form the literature. In GO, grouping categories are missing, for instance, glycan binding proteins (R-type lectin, L-type lectin, C-type lectin, P-type lectin, I-type lectin, galectin). Also, glycan-related synonyms are not present (e.g. "*core 1" = T antigen)*.

In ChEBI, labels are not always aligned with common domain use - for instance ChEBI's use of 'glycans' as synonym for 'carbohydrate'. There are key gaps in representation, for instance, no representation of N- and O-glycans as groups, absence of many molecular species reported as diseases markers, absence of synonyms commonly used in glycobiology, etc. In HPO/MP, glyco-phenotypes are also underrepresented. For instance, a search for terms "oligosacchariduria" shows only sialidated and disaccharides excretion while more than fifty glycophenotypes were detected in the urine from ten different diseases [5]. In metabolomics databases, glycans with masses under 1,500 da are severely underrepresented with mostly only mono or di-saccharides and their derivatives. For instance, a search of the term "glycan" in the 'Metabolomics Workbench' database [7] leads to only 6 molecules, while more than 30 N-glycan precursors can be identified in the Endoplasmic Reticulum [4].

This contrasts with the abundance of chemical characteristics of glycans in glycobiologist databases such as UniCarbKB, GlycomeDB and Japan Consortium For Glycobiology and Glycotechnology Database (JCGGDB) [9]. Nevertheless, links between glycophenotypes and diseases are still lacking in these glycan databases.

III. Applications And Future Directions

The MGPO will be integrated with existing phenotype ontologies such as the Human Phenotype Ontology (HPO), and Mammalian Phenotype Ontology (MP), to support research and clinical application of ontology-driven methods. MGPO will serve as a bridge between existing glycan related databases and metabolomics, databases, between glycan databases and diseases databases. As part of the Monarch Initiative [6], the MGPO will be applied in semantic similarity approaches for the purposes of (1) disease diagnosis, (2) identification of patient cohorts for clinical studies, (3) discovery of model organisms to researching rare disease, and (4) identifying candidate genes underlying undiagnosed diseases. The MGPO will also be applied to annotate glycomics data from our partners in the Undiagnosed Disease Network, and support its use in semantic similarity and pattern matching analyses that can facilitate disease diagnosis and characterization.

Acknowledgment

We thank the Ontology Development Group (ODG) at the OHSU's library for their feedback.

References

[1] https://globalgenes.org/rare-diseases-facts-statistics/

[2] https://www.genome.gov/27531963/faq-about-rare-diseases/.

[3] Freeze HH. "Genetic defects in the human glycome." Nat Rev Genet. 2006 Jul;7(7):537-51.

[4] Varki A, Cummings RD, Esko JD, et al., editors. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009. "Chapter 6, Biological Roles of Glycans, Essentials of Glycobiology." 2nd edition

[5] Xia B., Asif G., Arthur L., Pervaiz M.A., Li X., Liu R., et al. "Oligosaccharide analysis in urine by maldi-tof mass spectrometry for the diagnosis of lysosomal storage diseases." Clin Chem. 2013 Sep;59(9):1357-68.

[6] Haendel MA, Vasilevsky N, Brush M, Hochheiser HS, Jacobsen J,Oellrich A, *et al.* "Disease insights through cross-species phenotype comparisons." Mamm Genome. 2015 Oct;26(9-10):548-55.

[7] Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, *et al.* "Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools." Nucleic Acids Res. 2016 Jan 4;44(D1):D463-70.

[8] Ng B.G., Wolfe L.A., Ichikawa M., Markello T., He M., Tifft C.J., *et al.* "Biallelic mutations in CAD, impair de novo pyrimidine biosynthesis and decrease glycosylation precursors." Hum Mol Genet. 2015 Jun 1;24(11):3050-7.

[9] Aoki-Kinoshita KF, Bolleman J, Campbell MP, Kawano S, Kim JD, Lütteke T, *et al.* "Introducing glycomics data into the Semantic Web" J. Biomed Semantics. 2013 Nov 26;4(1):39.