Toward an Ontological Model of Molecular Glyco-Phenotypes

Matthew H. Brush*, Jean-Phillipe Gourdine,
Nicole Vasilevsky, Melissa A. Haendel
Oregon Health & Science University
Portland, OR, USA
*brushm@ohsu.edu

Christopher J. Mungall
Lawrence Berkeley National Laboratory
Berkeley, CA, USA

Peter Robinson, Sebastian Köhler
Institut für Medizinische Genetik und Humangenetik, Charité - Universitätsmedizin Berlin
Berlin, Germany

***Abstract*— The codification of phenotypes in ontologies has provided a level of standardization and computability that can support innovative means of data integration and analysis. The Monarch Initiative is a large scale effort that exploits ontologies for semantic integration of diverse genotype-to-phenotype data, and applies analytic methods that leverage the graph structure of ontologies to identify similarities between patients, diseases, and model systems based on the set of phenotypes with which they are annotated. A recent collaboration with the Undiagnosed Disease Program (UDP) has provided an abundance of glycomic data that is being used to characterize disease states in UDP patients. In order to apply Monarch tools toward diagnosis of patients and matching diseases to experimental models, we have begun to translate this quantitative glycomic data into qualitative phenotypes. Here we describe the development of the Molecular Glyco-Phenotype Ontology[1] (MGPO) that formalizes these phenotypes into an ontological model. We focus on methodologies for gathering requirements and translation of these requirements into an ontological model that can be integrated with existing phenotype ontologies, and which can support annotation and data analysis use cases of the Monarch Initiative and the UDP.**

Keywords—glycobiology, phenotype, ontology requirements

I. INTRODUCTION

In the arena of molecular biology research, carbohydrates have long taken a backseat to their protein and nucleic acid siblings. However, with the discovery that the human genome is comprised of only 30,000 genes, it has been widely recognized that that post-translational modification drives much of the diversity of protein function. This has sparked an increased interest in interrogating the glycan moieties that decorate much of the proteome. Emerging high-throughput technologies are now allowing researchers to characterize this incredibly complex class of macromolecules, and 'glycomics' is maturing as a first tier field that is rapidly contributing to our understanding of human biology and disease.

The term 'glycan' refers to any sugar monomer or chain produced in living organisms. The structure and composition of glycan chains is extremely diverse, and, the glycome is estimated to be up to 10^4 times larger than the proteome [1]. Glycans can be found free or covalently attached to protein and lipid macromolecules where they impact folding, stability, and molecular interactions. Glycans are highly prominent at the cell surface where they regulate cell surface interactions with matrix, neighboring cells, or microorganisms, through interactions with biological ligands. Inside of cells, glycans play roles in diverse structural and signaling functions critical to cellular development, growth, and survival. It is therefore not surprising that aberrant glycosylation has been associated with many hereditary and chronic conditions including cancers and immunological, cardiovascular, and neurological disorders [2]. Over 30 genetic disorders that alter glycan structure or synthesis have been identified in the past decade, which impact nearly all organ systems [1,3]. Patients suffering from these disorders display an overwhelmingly diverse range of clinical phenotypes that complicates diagnosis and treatment, and the molecular defects associated with these disorders are only now becoming accessible to interrogation.

As we begin to appreciate the relevance of glycosylation in human disease, informatics approaches that have been successful in harnessing genomic and proteomic data toward understanding and treating disease must be applied to the field of glycomics. One such approach is the application of ontologies for the standardized and computable representation of phenotype and disease data. Many phenotype ontologies exist [4] that encode knowledge about the molecular, physiological, and anatomical defects resulting from genetic variation, in a graph structure that formally represents the relationship between these phenotypes, and between phenotypes and related biological entities such as biological processes, cell types, anatomical entities, chemicals, and qualities. These include several taxon-specific phenotype ontologies such as the Human Phenotype Ontology (HPO) [5] and the Mammalian Phenotype Ontology (MP) [6], and the recently developed "Uberpheno" ontology that merges these and many others into a single model that allows cross-species analysis of phenotypic similarities [7]. This Uberpheno ontology is being leveraged in the work of the Monarch Initiative[2], which aims to integrate data from diverse resources

1 https://phenotype-ontologies.googlecode.com/svn/trunk/src/ontology/mgpo.owl
2 http://monarchinitiative.org

and apply semantic technologies to facilitate the characterization of disease and the identification of model systems for disease research. OWLSim [8] is a graph-based semantic similarity engine and a central component Monarch suite of tools, which allows comparisons between entities (patients, diseases, genes, and model organisms) based on the set of phenotypes related to them. OWLSim leverages the connectivity of phenotype classes in the Uberpheno ontology to quantify the similarity of any two entities based on the aggregate pair-wise similarity score of all phenotypes associated with the entities. In this way, diagnoses can be made based on a set of clinical findings, patients can be compared based on their phenotypic presentations, and model systems exhibiting phenotypes similar to a disease profile can be identified.

A major barrier to the OWLSim-based approach in the domain of glycosylation-related disorders is an under-representation of phenotypes describing defects in glycan levels, structure, and processing in existing ontologies - which we collectively refer to as "molecular glyco-phenotypes". Large amounts of glycomic data are becoming available to Monarch Initiative through its partnership with the Undiagnosed Disease Program [9], which has contracted labs to perform large-scale glycomic profiling on patients with rare and undiagnosed diseases. In order to fully leverage this data it is necessary to have rich coverage of molecular glyco-phenotypes in the suite of Monarch ontologies.

Here we present our efforts toward developing a cross-species phenotype ontology module to meet this need, which models molecular glyco-phenotypes and can be integrated into the framework of existing phenotype ontologies. Due to the highly complex and nuanced nature of data in the glycomics domain, and the need for ontology hierarchies to support specific use cases in the context of the Monarch and UDP efforts, the requirements gathering process was a critical component of this work. Accordingly, the approaches taken to extracting domain knowledge from text, data, and experts, and the conversion of this information into actionable development guidelines, will be a major focus of this report.

II. Requiremetts Gathering Methodology

Unlike other anatomical and organism-level phenotypes, molecular glyco-phenotypes are 'born' as quantitative data about metabolite levels, reporter intensities, and enzymatic activities. It is through the often subjective and variable interpretation of this data by experimentalists, authors, and curators that such measurements are translated into statements about the "normality" or "abnormality" of a particular glycan characteristic. This translation of quantitative data to qualitative phenotype assertions was an important consideration as we began to investigate the key types of glycan characteristics interrogated by biologists, and define the scope and content that would comprise our molecular glyco-phenotype module.

Several sources of data were consulted and curated to extract key elements that would inform our model. **Primary data** in the form of real datasets from different glycomics assays were provided by our collaborators from the UDP. In many cases these were annotated with their interpretations of normality of individual measurements. Analysis of raw data gave us an important view of the originating end of the data pipeline, and offered insight into the process through which researchers make qualitative evaluations on quantitative data. **Primary literature** was consulted to expand our understanding of molecular glycan defects, with a set of 20 research articles curated to extract information about glyco-phenotypes. We recorded the verbatim text reporting a phenotype (e.g. "accumulation of sialylated glycans in the serum") along with metadata describing the provenance of this claim (e.g. the assay and techniques used, tissues sampled, evaluants and analytes examined, etc). This descriptive information was critical, especially where the raw data was unavailable, as assay context can have significant implications for interpreting of phenotype assertions and translating them into ontology classes. **Literature reviews and textbooks** were also consulted to attain expert level summaries and classifications of glycosylation defects. These were primarily organized around genetic disorders, and provided additional assurance that our landscape analysis was comprehensive and spanned full diversity of glycan abnormalities. Finally, **public databases** describing disease and glycomics data were explored, including OMIM[1], the Functional Glycomics Database[2], and the Unicarb Knowledgebase[3]. These offered insights into how experts and curators organize and interpret molecular data.

III. Ontology Development

A. *Defining Ontology Scope and Axes for Classification*

With a rich set of curated data in hand, the next challenge was to synthesize this information to identify the key dimensions of glycan phenotypes on which the ontology would be classified. We began by listing each distinct glyco-phenotype curated from the sources above (roughly 100 phenotypes in total), and looked for patterns that would reveal important features for their categorization. Four essential dimensions revealed themselves through this exercise:

1. **Affected glycan characteristic.** This dimension distinguishes phenotypes based on the general type of glycan feature or process affected. Four key features are further described in Table 1 (glycan occupancy, composition, levels, and glycan-related processes).
2. **Affected glycan type**. This dimension distinguishes phenotypes based on the class of glycan exhibiting abnormal occupancy, structure, level, or processing. Glycan classes are well-defined according to a structure-based classification (e.g. N-linked glycans vs O-linked glycans vs glycosaminoglycans).
3. **Affected glycosylation target**. This dimension distinguishes phenotypes based on the aglycone target affected by a defect in glycan occupancy, composition, or processing (e.g. transferrin-conjugated glycans, phosphatidylinositol-linked glycans).

1 http://www.ncbi.nlm.nih.gov/omim
2 http://www.functionalglycomics.org/fg/
3 http://www.unicarbkb.org/

4. **Affected locus of abnormality.** This dimension distinguishes phenotypes based on the subcellular location of the defect (e.g. golgi glycans, ER glycans), and/or the tissue or fluid where it was measured (e.g. urine glycans, serum glycans).

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. Consultation with glycomic experts and Monarch developers pointed to the axis of 'affected glycan characteristic' (Table 1) as being the best fit for top-level classification. This is because phenotypes affecting a given glycan characteristic are likely to result from a common underlying biological defect. For example, abnormal glycan structures typically result from defective glycosylation or deglycosylation enzyme activity, and abnormal glycan levels often result from defective degradation or trafficking. This ensures that phenotypes with common biological bases will show high connectivity in the graph structure of the ontology, and therefore be scored highly in phenotype-based comparisons using the OWLSim algorithm. Notably, if it is determined that simultaneous classification along these axes in different orders is useful, logical definitions can be used to infer multiple inheritance and provide parallel classification schemes in the ontology (see below).

The remaining three dimensions were applied as classification axes in the order presented above to build down a hierarchy of phenotype classes that describes glycan-related abnormalities commonly reported in the literature and clinical tests. Accordingly, under each of the four 'abnormal glycan characteristic' subtypes listed in Table 1, more specific phenotypes were distinguished first by which class of glycan was found to be abnormal for the given characteristic, then the identity of any aglycone target that may have been specified, and finally by any reported anatomical or subcellular locus indicated where the abnormality was found. Figure 1 shows the hierarchical structure that results from the classification scheme outlined above, and gives examples of the a few more granular classes that result from this application of classification

TABLE 1: Description of four key glycan characteristics used to define top- level classification axes in the glyco-phenotype module.

Glycan Characteristic	Description
Occupancy	The extent of conjugation of a glycan chain to glycosylation sites on an aglycone target. This does not speak to the structure or composition of the glycan chain, just the number of glycan chains attached to a protein or lipid target.
Structure	The structural organization, sugar composition, or other modifications on a glycan chain. Typical phenotypes are abnormal branching patterns, or altered occurrence of a monosaccharide residues.
Level	The amount or concentration of a canonical glycan product present in the assayed anatomical entity.
Related Process	An event with a glycan as a participant, typically related to glycan metabolism or function (e.g. hydrolysis, glycosylation, degradation, binding)

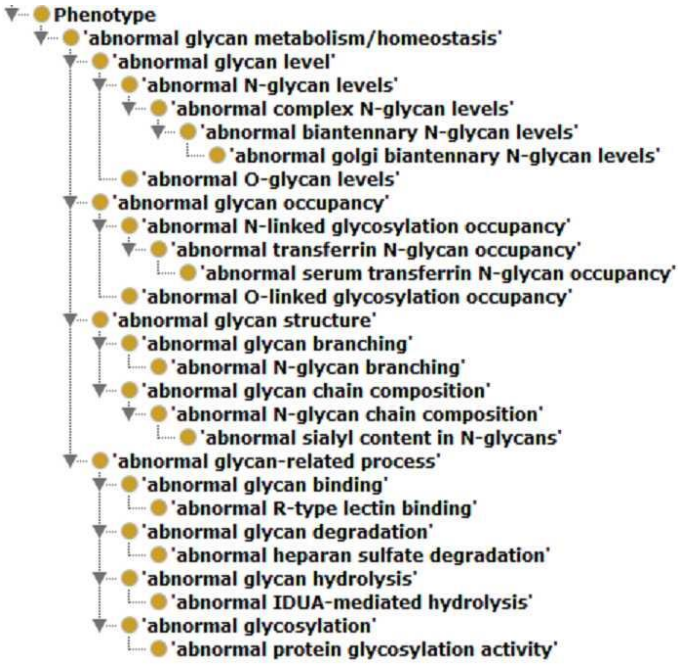

FIGURE 1: Hierarchy showing examples of the types of classes our model will distinguish based on defined classification axes.

axes. Note this is not meant to be a complete representation, but only illustrate the kinds of distinctions our model allows as to make and classes that result.

B. *Integration with Community Ontologies*

As noted above, the glyco-phenotype ontology will be built as a generic but extensible module that can be integrated with individual taxon-specific phenotype ontologies, and also fit within the larger Uberpheno framework. Uberpheno relies on logical definitions in taxon-specific member ontologies to support automatic classification of phenotype classes from these ontologies into a unified hierarchy, where "orthologous" phenotypes from different ontologies are grouped under taxon-agnostic Uberpheno nodes. This is possible given that classes in each member ontology have logical definitions based on interoperable design patterns, and built using a common set of referenced classes describing anatomy (Uberon), biological processes and activities (GO), small molecules (ChEBI), proteins (PRO), and qualities (PATO).

To similarly allow a reasoner to automatically integrate the glyco-phenotype module into the Uberpheno framework, its classes require logical definitions that are consistent with existing design patterns and referenced ontologies. To enable this to happen, the glyco-pheno module MIREOTs [10] terms from the external ontologies listed above that are needed to craft logical definitions. In many cases this means requesting that new glycan-specific terms be implemented in these referenced efforts to meet needs of our work. For example, we are working with ChEBI to add terms describing glycan moieties and molecules, and with GO to add needed glycan related biological processes and activities. In addition, new ontologies may be brought into the fold in order to avoid

duplication of modeling. For example, the Glycomics Ontology[2] (GlycO) can provide very specific glycan-related terms that may not fit in the ChEBI scope or framework.

With these referenced terms in place, logical definitions are built using them and a core set of properties from the Relations Ontology (RO[1]). These definitions follow design patterns applied in Uberpheno member ontologies, which use a "subq" approach based on entity-quality (EQ) statements describing phenotypes in terms of an affected quality (Q) and a process or material entity (E) in which it inheres [11]. The "subq" moniker refers to the fact that all descriptions begin with an encapsulating `has_part` relation to account for non-atomic phenotypes. An example of a complex logical definition that uses terms from PATO, ChEBI, GlycO, and Uberon is shown below for the class '**reduced sialyl content of serum proteins**':

```
has_part some ((has_fewer_parts_of_type and
(toward some 'sialyl residue')) and
(inheres_in some ('glycan' and (part_of some
('glycoprotein' and (located_in some
'serum'))))))
```

Implementation of such logical definitions which re-use and extend existing patterns are critical for supporting inferred classification of molecular glyco-phenotype classes within the larger Uberpheno framework.

IV. Lessons Learned and Best Practices

Our experience with the requirements-based methodology outlined above can be generalized into best-practices for similar efforts in ontology development. While a detailed discussion is beyond the scope of this report, we stress the importance of two key activities related to attaining knowledge of the domain, and consideration of specific use cases.

A major challenge for building ontologies describing complex and nuanced domains such as molecular glyco-phenotypes is attaining an appropriate understanding of the domain to be modeled. Developers should use a variety of resources to do so, and develop a standardized curation strategy that will enable extraction and organization of knowledge from these resources. For example, given the origination of molecular glyco-phenotypes as quantitative data, it was very important to understand and document the experimental provenance of phenotype assertions made in the literature, as the contextualized data may tell us more than author's summary conclusions. Having access to this metadata provided valuable guidance for defining critical concepts and classification axes in our modeling, and was of great utility in extracting knowledge from domain experts. These domain experts were a critical resource, as they provided key insights that steered the direction of our modeling. For example, there were many nuances to classifying glyco-phenotypes based on anatomical locus of the defect that required hands on understanding of experimental assays to appreciate, and impacted the way we modeled based on this dimension.

A second core challenge for building ontologies is understanding the use cases toward which the ontology will be applied, and using this knowledge to guide development. Three use cases are critical in guiding ongoing development of the glyco-phenotype ontology module. The first is supporting the automatic integration of the module into the Uberpheno framework. Here we have to ensure that logical definitions for module classes at key integration points use consistent design patterns, and also that bridge files are created to hold OWL subclass axioms placing glyco-phenotype classes appropriately in the context of Uberpheno ontologies. A second use cases is to support annotation of phenotypes reported in patient data and the model organism literature. Monarch curators are back-annotating over 200 papers from the mouse literature, and UDP curators will annotate clinical assay results. This will provide the initial dataset on which the OWLSim algorithm can operate to identify matches between patients, diseases, and model systems. To support this use case, coverage was scoped based on outcomes of our requirements gathering curation of diverse information sources as described above. The granularity of classification is being iteratively refined to provide the appropriate specificity of terms to support consistency and precision. Notably, an added benefit of a thoughtful and well-documented requirements gathering approach is that this knowledge can inform guidelines for use of the ontology in annotation. Finally, a third use case for our work is the consumption of the ontology by the OWLSim algorithm. As discussed above, this was a major consideration in how we applied classification axes, to ensure that the connectivity of classes in the graph would reflect most biologically relevant distinctions for phenotype-based similarity comparisons.

1 http://purl.obolibrary.org/obo/ro.owl

2 http://glycomics.ccrc.uga.edu/core4/ontology/GlycO/GlycO.owl

Acknowledgments

We thank the following experts for their insights and feedback: Miao He, Hudson Freeze, Charles Hoppel, Cornelius Boerkoel, Megan Kane, Rick Cummings, and Mariska Davids. This work was funded by NIH # 1R24OD011883.

References

[1] Freeze, Hudson H. "Genetic defects in the human glycome." *Nature Reviews Genetics* 7, no. 7 (2006): 537-551.

[2] Tian, Yuan, and Hui Zhang. "Glycoproteomics and clinical applications." *PROTEOMICS-Clinical Applications* 4, no. 2 (2010): 124-132.

[3] Ajit Varki et al, Essentials of Glycobiology, 2nd edition, Ch 41-42, Cold Spring Harbor (NY), 2009.

[4] Mungall, Christopher J. et al. "Integrating phenotype ontologies across multiple species." Genome biology 11, no. 1 (2010): R2.

[5] Köhler, Sebastian, et al. "The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data." *Nucleic acids research* (2013): gkt1026.

[6] Smith, Cynthia L., and Janan T. Eppig. "The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data." *Mammalian Genome* 23, no. 9-10 (2012): 653-668.

[7] Köhler, Sebastian et al. "Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research." *F1000Research* 2 (2013).

[8] http://OWLSim.org

[9] Gahl, William A et al. "The national institutes of health undiagnosed diseases program: insights into rare diseases." *Genetics in Medicine* 14, no. 1 (2011): 51-59.

[10] Courtot, Mélanie et al. "MIREOT: The minimum information to reference an external ontology term."*Applied Ontology* 6, (2011): 23-33.

[11] Loebe, Frank, Frank Stumpf, Robert Hoehndorf, and Heinrich Herre. "Towards improving phenotype representation in OWL." *Journal of biomedical semantics* 3, no. 2 (2012): 1-17.