GlyGen.org
Computational and Informatics Resources for Glycoscience

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Complex Carbohydrate Research Center, University of Georgia, Athens, GA, USA
The George Washington University, Washington, DC, USA

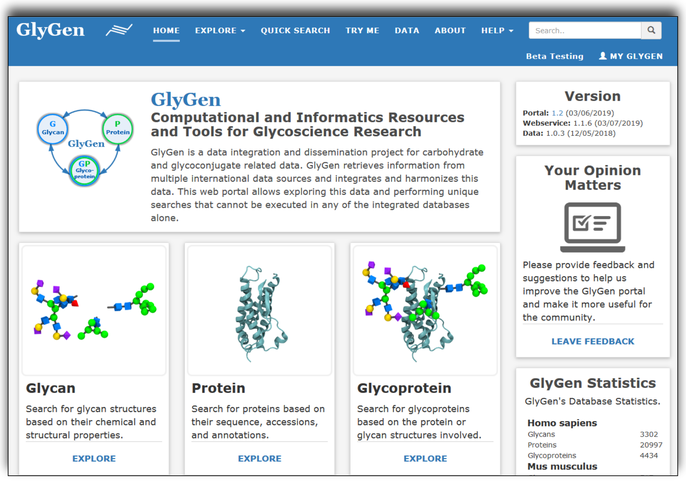
UNIVERSITY OF GEORGIA

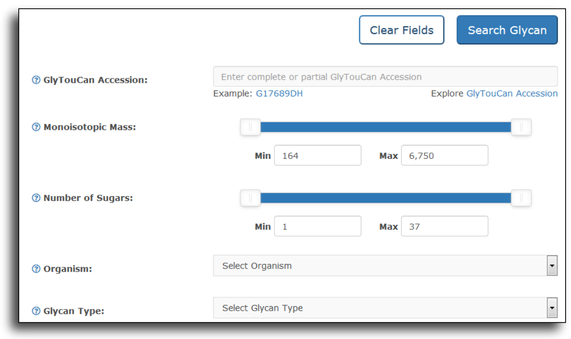
Advancing our understanding of the roles that glycosylation plays in development and disease is hindered by the diversity of the data that must be integrated to gain insight into these complex phenomena. GlyGen is a new initiative supported by the NIH Common Fund with the goal of democratizing glycoscience by implementing a comprehensive data repository that integrates diverse types of data, including glycan structures, glycan biosynthesis enzymes, glycoproteins, and three-dimensional glycoprotein structures along with genomic and proteomic knowledge. Our goal is to provide both trained and aspiring glycoscientists and informaticians an easy way to access the complex information involving glycans in biology. The GlyGen project provides both webpage based and machine readable interfaces to access the integrated data.

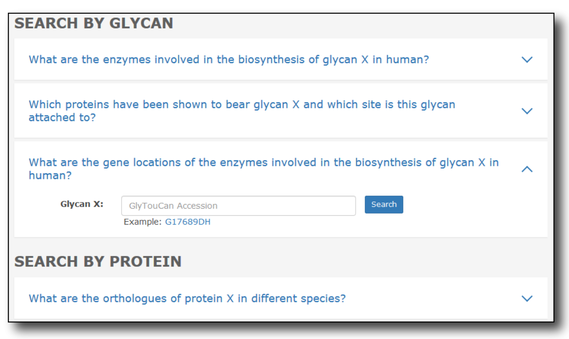
Web portal

The GlyGen portal allows browser-based access to the data integrated in the data store. The portal is divided into three information groups: glycans, proteins and glycoproteins. For each type of molecule an extensive list of information integrated from the different data sources will be displayed. Protein and glycan information pages are interlinked with each other allowing the user to navigate from a protein to its glycans and vice versa.

Extensive search interfaces have been implemented for each type of molecule allowing the user to search by molecule IDs, common namespaces, chemical properties (e.g. mass or number of residues) or relationships to other molecules (e.g. glycosylation of a protein with a certain glycan).







All information in the webpages is tagged and linked with the original data sources, allowing users to find the source of each piece of information and link out to additional data that has not been integrated into GlyGen.

In preparation of the project an extensive user survey was performed, leading to ~250 use cases from more than 50 international scientists. These use cases are questions that cannot easily be answered using the established databases alone. Being able to answer these questions is one of the overarching goals of GlyGen. A user-friendly interface has been implemented called “Quick Searches” to address these questions that require data and querying across multiple domains.

URL: http://glygen.org

API

{ REST }

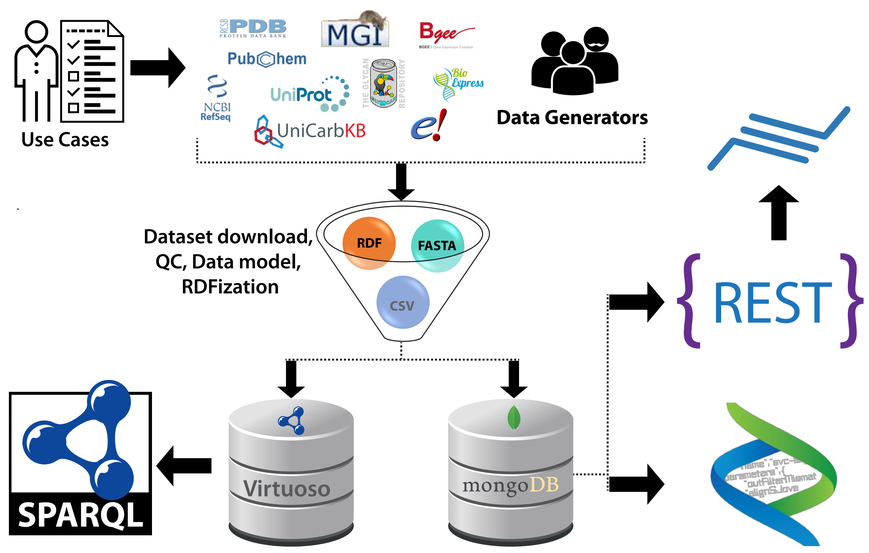
A RESTful webservice API was implemented to allow easy web-based access to the data stored in the backend. The web services use JSON (JavaScript Object Notation) to accept requests and provide data.

URL: http://api.glygen.org

JSON

Data

Data from major genomics, proteomics, and glycomics resources and data generators is collected and downloaded. All information is integrated into our datamodel. As part of this integration workflow, datasets are harmonized by translation into **consistent name spaces** and have to pass through **intensive quality control** workflows. Problems or inconsistencies found in the data are reported back to the original data providers for verification and correction as needed. The integrated information is stored in a mongoDB database.



The information is also converted to RDF format and stored in our Virtuoso triplestore which will also provide a publicly accessible **SPARQL endpoint**. Data stored in our database are publicly available as BioCompute Objects on our data website and as JSON objects provided by our webservice API. The GlyGen portal uses these webservices to retrieve the data displayed on the webpages. The data integration workflow is an iterative process that is directed by the **use cases** collected from users. These use cases also drive the selection of databases and datasets to be integrated in GlyGen.

Access

All integrated information is stored in **CSV files** annotated with README files using the **BioCompute Object** specification to store metadata describing the datasets: origin, creation date, general description, CSV column defintion, and tools/programs/scripts used to process the data. All files, including previous versions of the files can be accessed on our data page.

URL: http://data.glygen.org

Ultimately, a **triplestore** will provide access to the integrated GlyGen data. The RDF generation is based primarily on existing ontologies such as the UniProt RDF schema ontology for proteomics data, the Glycoconjugate Ontology (GlycoCoO) for Glycoproteomics data, and the Feature Annotation Location Description Ontology (FALDO) for feature positions.

RDF

URL: Coming soon

Acknowledgement: The GlyGen project is supported by an NIH Common Fund grant (U01 GM125267-01). The project is an international collaboration with many groups involved: William York, Raja Mazumder, Nathan Edwards, Radoslav Goldman, Judith Blake, Michael Pierce, Robel Kahsay, Maria Martin, Kiyoko Aoki-Kinoshita, Matthew Campbell, Evan Bolton, Darren Natale, Karen Ross, Rob Woods, Ten Feizi, Jeff Gildersleeve, Richard Cummings and many people in their groups.