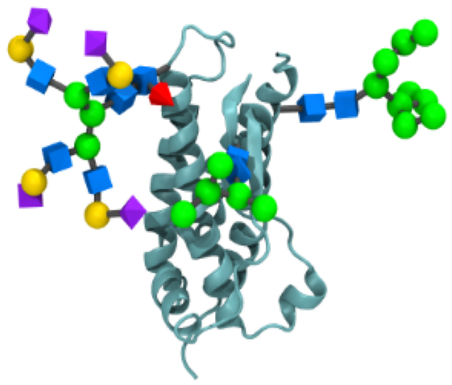
GlyGen Data Integration: Creating A Collaborative Environment For Data Generators, Bioinformatics Resources, And Users

Rahi Navelkar, GlyGen Consortium

Department of Biochemistry & Molecular Medicine, The George Washington University.

Data-Model



GlyGen is an NIH-funded interdisciplinary glycoinformatics project developed to implement a comprehensive data repository with an aim to answer crucial community-wide use cases. GlyGen utilizes novel data integration methodologies to collect and harmonize different types of data to ultimately provide a consolidated view of protein, gene, and glycan in one resource.

GlyGen has developed a comprehensible data model by integrating namespaces from existing protein and glycan ontologies which can be easily extended to accommodate diverse data types and accompanying annotations. All generated datasets feed into the GlyGen frontend interface (https://glygen.org) .For details about GlyGen frontend please visit poster presented by Will York and Rupali Mahadik, #B106.

Data Collection

GlyGen currently captures over 40 different cross-references (such as BRENDA, KEGG, REACTOME, HGNC, CarbBank, Glycosciences.de, GlyConnect, etc.) collectively for protein and glycan entities. GlyGen also captures over 5 different glycan sequence formats. (e.g. WURCS, GlycoCT, InChI, SMILES etc.). GlyGen includes data associated with Human, Mouse and Rat species. Future plans include the addition of new species (like HCV), synthesized glycan data, GAGs interaction data, etc. Currently, GlyGen has data for ~20,000 Human, ~22000 Mouse, ~21,000 Rat canonical proteins and ~5000 glycans.

GlyGen

https://glygen.org

Data Integration and Use Cases

The data integration infrastructure and protocols allow a systematic and streamlined approach for individual researchers or large molecular biology database developers to submit and integrate their data into GlyGen for further dissemination.

Collection of such diverse datasets allows GlyGen users to not just browse and download data but also perform complex queries across different domains. such as: Which protein(s) bear glycan X?, Which enzyme(s) are involved in synthesizing glycan X?, etc.

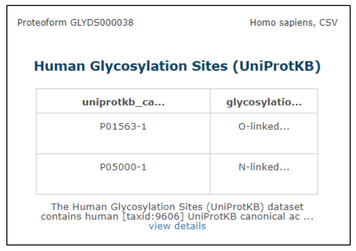
What are the glycosylation sites and the associated glycans on the Hepatocyte growth factor (HGF) protein?

Scan this QR code →

Researchers interested in integrating their data into GlyGen can reach out to us using our contact page (https://www.glygen.org/contact.html)

Quality Control and Documentation

Each data submission involves the co-development of quality control and pipeline with the data submitter(s) to generate a stable, versioned dataset assigned with a unique identifier. (https://data.glygen.org/)



GLYDS000001

The workflows along with other details (such as authors, contributors, QC steps, usability, keywords, etc) are documented in dataset-specific README's generated using BioCompute specifications. The README also provides access to the input, log (eliminated entries) and scripts to users to provide replicability.

Data Access

Users can access GlyGen data programmatically through RESTful web service-based APIs (https://api.glygen.org) and (in-future) SPARQL endpoint. All GlyGen data is under CC BY 4.0 license (free to copy, distribute, display and make commercial use, given proper citation) https://glygen.org/license.html

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.