GlyCosmos: An Integrated Semantic Knowledge Base for Glycoscience

Sunmyoung Lee[1,*], Yasunori Yamamoto[2], Achille Zappa[1] and Kiyoko F. Aoki-Kinoshita[1,3,4]

[1]KinoshitaGlycan and Life Systems Integration Center (GaLSIC), Soka University, Hachioji, Tokyo, Japan
[2]Database Center for Life Science (DBCLS), ROIS-DS, Kashiwa, Chiba, Japan
[3]Graduate School of Science and Engineering, Soka University, Hachioji, Tokyo, Japan
[4]Institute for Glyco-core Research, Nagoya University, Nagoya, Japan

Abstract

Abstract GlyCosmos, a comprehensive web resource for glycoscience research, has been significantly enhanced to provide a unified platform for accessing and analyzing glycan structures, related genes, proteins, pathways, and diseases. This paper presents the latest developments in GlyCosmos, highlighting the adoption of a unified Resource Description Framework (RDF) schema and semantic web technologies. These advancements address previous challenges in data integration and enable seamless connection of diverse glycan-related datasets. The implementation of SPARQL endpoints allows for powerful querying capabilities, including federated queries, knowledge discovery, and inference-based queries. The enhanced data integration has improved search functionality, supporting multi-faceted searches across various parameters and nomenclatures. GlyCosmos now offers expanded resources for glycans, genes, lectins, and diseases, integrating data from multiple sources to provide a comprehensive view of glycobiology. This semantic web approach improves data interoperability, enhances knowledge discovery, and positions GlyCosmos as a crucial platform for integrating and disseminating glycan-related knowledge in the evolving field of glycoscience.

Keywords

Glycoscience, Semantic Web, RDF, Ontology, SPARQL, Bioinformatics, Data Integration, Standards, Knowledge Discovery, glycans, genes, proteins, pathways

1. Introduction

The GlyCosmos Glycoscience Portal (https://glycosmos.org) is a comprehensive web resource that integrates diverse glycan-related data and tools to support glycoscience research [1]. Developed as part of the Japanese Society for Carbohydrate Research, GlyCosmos aims to provide a unified platform for accessing and analyzing glycan structures, related genes, proteins, pathways, and diseases.

The GlyCosmos Glycoscience Portal has undergone significant updates in recent versions (e.g. we just released version 4.1 on December 9, 2024) to enhance its capabilities as a comprehensive resource for glycoscience research. A key focus of these updates has been the implementation of advanced semantic web technologies to improve data integration, accessibility, and querying capabilities.

2. Semantic Web Technologies and RDF Data Model

At the core of GlyCosmos' improvements is the adoption of a unified Resource Description Framework (RDF) schema. This new schema addresses previous challenges where imported data from external databases had disparate structures, leading to complex queries and slower performance. The unified RDF model allows for seamless integration of diverse glycan-related datasets, including genes, proteins, glycans, lectins, and diseases. The RDF data model represents information as a graph of interconnected

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
sunmyoung@soka.ac.jp (S. Lee); yy@dbcls.rois.ac.jp (Y. Yamamoto); zappa@soka.ac.jp (A. Zappa); kkiyoko@soka.ac.jp (K. F. Aoki-Kinoshita)
A0000-0002-1327-0689 (S. Lee); 0000-0002-6943-6887 (Y. Yamamoto); 0000-0003-4040-9620 (A. Zappa); 0000-0002-10002-6662-8015 (K. F. Aoki-Kinoshita)

triples (subject-predicate-object statements). This structure is particularly well-suited for describing the complex relationships in glycobiology. For example, a glycan can be linked to its associated genes, proteins it modifies, and diseases it's involved in, all within a single, queryable graph. GlyCosmos leverages several key ontologies listed below to standardize vocabulary and enable sophisticated inference capabilities:

- GlycoRDF[2] and GlycoCoO[3] for glycan-specific concepts
- Gene Ontology (GO)[4] for gene functions and processes
- Disease Ontology (DO)[5] for standardized disease terminology
- Uber-anatomy Ontology (UBERON)[6] for anatomical terms
- NCBI Taxonomy for species classification

By mapping resources to these ontologies, GlyCosmos can provide more accurate and comprehensive search results. For instance, when searching for a specific glycosyltransferase gene, the system can infer related processes and functions based on GO annotations, even if they're not explicitly stated in the original data.

3. SPARQL Endpoints and Query Capabilities

The adoption of RDF enables GlyCosmos to offer powerful SPARQL (SPARQL Protocol and RDF Query Language) endpoints. These endpoints allow users and external applications to perform complex queries across the entire integrated dataset. The SPARQL capabilities of GlyCosmos include:

1. Federated queries: Users can retrieve information from multiple databases in a single query. For example, one could search for glycans associated with a specific gene, their expression in various tissues, and related diseases in one operation.
2. Hierarchical searches: Leveraging ontology structures, queries can traverse hierarchical relationships.
3. Inference-based queries: The system can deduce implicit relationships based on ontology rules. This allows for the discovery of non-obvious connections between glycans, genes, and diseases.
4. Flexible data retrieval: Researchers can construct custom queries to extract precisely the data they need, combining information from genes, glycans, proteins, and diseases in novel ways.

The GlyCosmos Programmatic Access page (https://glycosmos.org/programmatic) provides detailed documentation on the RDF schema and available SPARQL endpoints (https://ts.glycosmos.org/sparql), enabling advanced users to construct sophisticated queries.

4. Enhanced Data Integration and Search Functionality

The new RDF schema has significantly improved search functionality across GlyCosmos. Users can now perform multi-faceted searches using a variety of parameters, including gene names, protein identifiers, glycan structures, and disease terms. The system supports multiple glycan nomenclatures (GlycoCT[7], IUPAC[8], Linear Code[9], WURCS[10]), and users can search by molecular mass or monosaccharide composition. For example, in the Glycans resource, users can search by:

- Text-based glycan names in various formats
- Monosaccharide composition, useful for mass spectrometry data analysis
- Molecular mass ranges, with results categorized by structural detail level
- Species or taxonomic information
- Glycan motifs or substructures

The Genes resource now offers a hierarchical view based on Gene Ontology terms, allowing researchers to explore glycan-related genes by function or process. The interface provides filtering options for gene names, species, and associated diseases, making it easier to navigate the over 8,000 gene entries. The Lectins resource has been expanded with data from CarboGrove[11] and UniLectin[12] databases, providing comprehensive information on lectin-glycan binding specificities. Users can explore lectins by species, recognized monosaccharides, or binding affinities. The Diseases resource now integrates data from GDGDB[13] and the Alliance of Genome Resources[14], offering a more comprehensive view of glycan-related disorders. This includes information on genetic variants, model organism data, and links to external resources like OMIM and Orphanet.

5. Key Features and Value Proposition

GlyCosmos offers several key advantages as a semantic knowledge base for glycoscience:

Integrated Data Resources: GlyCosmos aggregates data from multiple glycoscience databases and repositories, including GlyTouCan[15], UniProt[16], KEGG[17], and others. This integration allows researchers to easily access diverse glycan-related information from a single portal.

Semantic Web Technologies: The portal leverages RDF, SPARQL, and ontologies to represent glycan data in a standardized, machine-readable format. This semantic approach enables sophisticated querying and data integration capabilities.

User-Friendly Interface: GlyCosmos provides an intuitive web interface for browsing and searching glycan structures, genes, proteins, and related information. Interactive visualizations and analysis tools enhance data exploration.

Data Submission and Curation: The portal includes repositories like GlyTouCan for glycan structure registration and GlycoPOST for mass spectrometry data submission, promoting data sharing in the glycoscience community.

Interoperability: As part of the GlySpace Alliance[18], GlyCosmos supports data exchange and integration with other glycoinformatics resources like GlyGen[19] and Glycomics@ExPASy[20].

GlyCosmos employs a unified RDF schema to represent glycan-related data, enhancing interconnectivity between different resources. Key aspects of the semantic model include:

- Use of established ontologies like GlycoRDF, GlycoCoO, and SIO to standardize terminology
- Hierarchical organization of concepts using rdfs:subClassOf relationships
- Cross-linking between resources using rdfs:seeAlso properties
- Integration of external ontologies like Gene Ontology and Disease Ontology

This semantic approach enables:

- Inference-based analysis and discovery of hidden relationships
- Flexible querying across multiple data types
- Easy extension and integration of new data sources

The semantic foundation of GlyCosmos enhances user experience and knowledge discovery:

- Faceted Search: Users can filter and explore data using multiple criteria
- Visual Analytics: Interactive visualizations of glycan structures, pathways, and relationships
- Cross-Domain Queries: Researchers can easily connect glycan data to genes, proteins, and diseases
- Programmatic Access: SPARQL endpoint and API access support integration with analysis workflows

GlyCosmos supports various applications in glycobiology research and beyond:

- Glycan Structure Analysis: Researchers can search, compare, and analyze glycan structures across species and biological contexts

- Biomarker Discovery: Integration of glycan and disease data facilitates identification of potential diagnostic markers
- Drug Target Identification: Connections between glycans, proteins, and pathways can reveal novel therapeutic targets
- Glycoengineering: Knowledge of glycosylation pathways and enzymes aids in designing modified glycoproteins

Glycoengineering: Knowledge of glycosylation pathways and enzymes aids in designing modified glycoproteins GlyCosmos represents a significant advancement in glycoinformatics by providing a semantically integrated knowledge base for glycoscience. Its use of RDF, ontologies, and SPARQL enables powerful data integration and discovery capabilities. As glycomics data continue to grow, GlyCosmos will play an increasingly important role in connecting glycan information to broader biological and biomedical research.

Figure 1: Glycan-related gene list

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcterms:<http://purl.org/dc/terms/>
PREFIX glycan: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX sio:<http://semanticscience.org/resource/>
PREFIX go: <http://www.geneontology.org/formats/oboInOwl#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX ddbj: <http://ddbj.nig.ac.jp/ontologies/taxonomy/>
SELECT DISTINCT ?gene_id ?symbol ?description ?organism ?organism_id
(GROUP_CONCAT(DISTINCT ?doid ; separator = "|") AS ?doids)
FROM <http://rdf.glycosmos.org/glycogenes>
FROM <http://ddbj.nig.ac.jp/ontologies/taxonomy/>
FROM <http://rdf.glycosmos.org/disease>
FROM <http://purl.obolibrary.org/obo/doid>
WHERE {
  ?gene a glycan:Glycogene.
  BIND(STRAFTER(STR(?gene), "http://glycosmos.org/glycogene/") AS ?id)
  BIND(xsd:integer(?id) AS ?gene_id)
  ?gene rdfs:label ?symbol .
  ?gene dcterms:description ?description .
  ?gene rdfs:seeAlso ?ncbi .
  FILTER(REGEX(STR(?ncbi),"ncbigene"))
  ?gene glycan:has_taxon ?taxonomy .
  ?taxonomy dcterms:identifier ?organism_id .
  ?taxonomy rdfs:label ?organism .
  OPTIONAL{
    ?do go:id ?doid .
    ?do rdfs:label ?do_label .
    ?disease rdfs:seeAlso ?do .
    ?disease a glycan:Disease .
    ?disease sio:SIO_000255/sio:SIO_000001 ?gene .
  }
} LIMIT 10
```

6. Value and Future Directions

The semantic web approach adopted by GlyCosmos offers several key advantages:

1. Improved data interoperability: The standardized RDF format facilitates data exchange with other glycoscience resources and broader life science databases.
2. Enhanced knowledge discovery: By connecting diverse datasets through a unified schema, researchers can uncover new relationships between glycans, genes, and diseases.
3. Scalability: The flexible RDF model allows for easy integration of new data types and sources as the field of glycoscience evolves.
4. Support for AI and machine learning: The structured, machine-readable format of RDF data is well-suited for advanced analytical techniques and predictive modeling.

As glycoscience continues to grow in importance across biology and medicine, GlyCosmos is positioned to play a crucial role in integrating and disseminating glycan-related knowledge. Future developments may include further integration with other omics data types, improved tools for glycan structure prediction and analysis, and enhanced support for glycoengineering applications. By leveraging semantic web technologies, GlyCosmos version 4 provides researchers with a powerful platform for exploring the complex world of glycobiology, facilitating new discoveries and advancing our understanding of the roles of glycans in health and disease.

Acknowledgments

This work was funded by Japan Science and Technology (JST) – National Bioscience Database Center (NBDC) Grant Number JPMJND2204. This work was also partially supported by the Human Glycome Atlas Project, which is funded as a Large-scale Academic Frontier Promotion Project of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. Lee, T. Ono, S. Masaaki, A. Fujita, M. Matsubara, A. Zappa, I. Yamada, K. F. Aoki-Kinoshita, Updates implemented in version 4 of the GlyCosmos glycoscience portal, Anal. Bioanal. Chem. 417 (2025) 907–919.

[2] R. Ranzinger, K. F. Aoki-Kinoshita, M. P. Campbell, S. Kawano, T. Lütteke, S. Okuda, D. Shinmachi, T. Shikanai, H. Sawaki, P. Toukach, M. Matsubara, I. Yamada, H. Narimatsu, GlycoRDF: an ontology to standardize glycomics data in RDF, Bioinformatics 31 (2015) 919–925.

[3] I. Yamada, M. P. Campbell, N. Edwards, L. J. Castro, F. Lisacek, J. Mariethoz, T. Ono, R. Ranzinger, D. Shinmachi, K. F. Aoki-Kinoshita, The glycoconjugate ontology (GlycoCoO) for standardizing the annotation of glycoconjugate data and its application, Glycobiology 31 (2021) 741–750.

[4] Gene Ontology Consortium, Gene ontology consortium: going forward, Nucleic Acids Res. 43 (2015) D1049–56.

[5] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, W. A. Kibbe, Disease ontology: a backbone for disease semantic integration, Nucleic Acids Res. 40 (2012) D940–6.

[6] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, M. A. Haendel, Uberon, an integrative multi-species anatomy ontology, Genome Biol. 13 (2012) R5.

[7] S. Herget, R. Ranzinger, K. Maass, C.-W. V. D. Lieth, GlycoCT-a unifying sequence format for carbohydrates, Carbohydr. Res. 343 (2008) 2162–2171.

[8] A. D. McNaught, Nomenclature of carbohydrates (IUPAC recommendations 1996), Pure Appl. Chem. 68 (1996) 1919–2008.

[9] E. Banin, Y. Neuberger, Y. Altshuler, A. Halevi, O. Inbar, N. Dotan, A. Dukler, A novel linear code nomenclature for complex carbohydrates., Trends in Glycoscience and Glycotechnology 14 (2002) 127–137. doi:10.4052/tigg.14.127.

[10] M. Matsubara, K. F. Aoki-Kinoshita, N. P. Aoki, I. Yamada, H. Narimatsu, WURCS 2.0 update to encapsulate ambiguous carbohydrate structures, J. Chem. Inf. Model. 57 (2017) 632–637.

[11] Z. L. Klamer, C. M. Harris, J. M. Beirne, J. E. Kelly, J. Zhang, B. B. Haab, Carbogrove: a resource of glycan-binding specificities through analyzed glycan-array datasets from all platforms, Glycobiology 32 (2022) 679–690. doi:10.1093/glycob/cwac022.

[12] B. Schnider, Y. M'Rad, J. El Ahmadie, A. G. de Brevern, A. Imberty, F. Lisacek, HumanLectome, an update of UniLectin for the annotation and prediction of human lectins, Nucleic Acids Res. 52 (2024) D1683–D1693.

[13] E. Solovieva, T. Shikanai, N. Fujita, H. Narimatsu, GGDonto ontology as a knowledge-base for genetic diseases and disorders of glycan metabolism and their causative genes, J. Biomed. Semantics 9 (2018).

[14] R. Kishore, V. Arnaboldi, C. E. Van Slyke, J. Chan, R. S. Nash, J. M. Urbano, M. E. Dolan, S. R. Engel, M. Shimoyama, P. W. Sternberg, T. A. O. Genome Resources, Automated generation of gene summaries at the alliance of genome resources, Database (Oxford) 2020 (2020).

[15] A. Fujita, N. P. Aoki, D. Shinmachi, M. Matsubara, S. Tsuchiya, M. Shiota, T. Ono, I. Yamada, K. F. Aoki-Kinoshita, The international glycan repository GlyTouCan version 3.0, Nucleic Acids Res. 49 (2021) D1529–D1533.

[16] P. Choudhary, S. Anyango, J. Berrisford, J. Tolchard, M. Varadi, S. Velankar, Unified access to up-to-date residue-level annotations from UniProtKB and other biological databases for PDB data, Sci. Data 10 (2023) 204.

[17] K. F. Aoki-Kinoshita, M. Kanehisa, Glycomic analysis using KEGG GLYCAN, Methods Mol. Biol. 1273 (2015) 97–107.

[18] F. Lisacek, M. Tiemeyer, R. Mazumder, K. F. Aoki-Kinoshita, Worldwide glycoscience informatics infrastructure: The GlySpace alliance, JACS Au 3 (2023) 4–12.

[19] W. S. York, R. Mazumder, R. Ranzinger, N. Edwards, R. Kahsay, K. F. Aoki-Kinoshita, M. P. Campbell, R. D. Cummings, T. Feizi, M. Martin, D. A. Natale, N. H. Packer, R. J. Woods, G. Agarwal, S. Arpinar, S. Bhat, J. Blake, L. J. G. Castro, B. Fochtman, J. Gildersleeve, R. Goldman, X. Holmes, V. Jain, S. Kulkarni, R. Mahadik, A. Mehta, R. Mousavi, S. Nakarakommula, R. Navelkar, N. Pattabiraman, M. J. Pierce, K. Ross, P. Vasudev, J. Vora, T. Williamson, W. Zhang, GlyGen: Computational and informatics resources for glycoscience, Glycobiology 30 (2020) 72–73.

[20] J. Mariethoz, D. Alocci, A. Gastaldello, O. Horlacher, E. Gasteiger, M. Rojas-Macias, N. G. Karlsson, N. H. Packer, F. Lisacek, Glycomics@ExPASy: Bridging the gap, Mol. Cell. Proteomics 17 (2018) 2164–2176.