Development of Bioinformatics Resources for Glycan-related pathway information using Semantic Web Technologies

セマンティックウェーブ技術を活用した糖鎖関連経路情報のバイオインフォマティクスリソースの開発

19D5604　李　宣明　　　指導教員　木下　聖子

SYNOPSIS

It has been attempted to integrate distributed data in many biological disciplines, which will enable the development of new knowledge bases and provide insight into underlying biological processes. However, the diversity of biological data types and the complexity of concepts has presented an obstacle to data integration. On the other hand, Semantic Web Techniques, created to provide a standard for data sharing on the web, have been used for integrating biological information derived from various data types. I implemented the fundamental methods of Semantic Web technology features in order to standardize the glycan-related data that is gathered from public databases or co-researchers in a computer-readable manner and to build a repository for pathway information, which is described by interpreting different types of resources and concepts including catalytic activation, translocation, modification, and so on. Given the importance of glycans in pathway information, sharing data with other information from existing databases or more specific details provided by users will support in data integration.

Keywords: Pathway Database, Repository, RDF, SPARQL, Ontology, data share

Introduction

The key roles of glycans have been revealed in the normal cellular process and disease states with the tremendous growth of glycomics data to be generated by a high-throughput analysis [1]. Those data are scattered across different databases or scientific publications and the information that gives an insight into the role of resources in the cellular process and the dynamic behavior under the microenvironment is fragmented in various formats. Considering the importance of the glycan data in cell biology and the exponentially growing amount of data with the advancing analysis instruments, it would be urgent to devote effort to building the method to describe the information in a standardized and computable format in order to share and integrate with other data such as glycomics or proteomics.

For this purpose, I adopted Semantic Web technologies that have been developed to instruct computers to understand the semantics of data [2]. The Semantic Web has been constructed based on the core technologies such as RDF (Resource Description Framework), SPARQL (SPARQL Protocol and RDF Query Language), and OWL (OWL Web Ontology Language). The RDF provides a model for describing information on resources that are presented in an informal or incomputable format. In the RDF model, each resource is identified with URI (Uniform Resource Identifier) which ensures that resources are identified without confusion on the web and the information is specified in the form of subject, predicate, and object that is usually referred to as a "triple". The relationship between the subject and object is defined by the predicate. The expansion of the information is accomplished by using a predicate to explain additional information about the subject or object node. SPARQL is used to query RDF data in the triple store which is a database that contains triple data that is called a SPARQL endpoint and useful endpoints are opened to the public such as the Uniprot(http://beta.sparql.uniprot.org/), ChEMBL (http://www.ebi.ac.uk/rdf/services/chembl/sparql), Reactome (http://www.ebi.ac.uk/rdf/services/reactome/sparql), etc. The OWL is designed to help the computer interpret the meaning of the resources using vocabulary derived from the specified document that is designed by logic. The core concepts are to define the individual and the type of resources using the *<rdf:type>* property indicating ***Class*** or ***Property*** and to restrict properties representing a relationship between two ***Classes***. Ontologies that are created based on the OWL have been developed to describe the complex concept and knowledge in the Life Science domain and enable the distributed data to be shared between different databases regardless of the complexity of concepts.

The structure of microbial glycan, which has evolved over time to adapt to hosts or environments, has been reported that they are dissimilar from eukaryotes' structure and have unique modifications. Microbes also have a wide variety of glycan structures depending on species and strains. It is highly significant to reveal the beneficial role of glycan in the symbiotic enteric bacteria or pathogenic bacteria that cause diseases. The glycan-related research data of microbes were obtained in tabular format from co-researchers and RDFization was performed on that data about *Campylobacter jejuni,* *Cryptococcus neoformans, Mycobacterium tuberculosis, Mycobacterium abscessus, Bifidobacterium bifidum, and Bifidobacterium longum*. The RDFized data are provided as fundamental data to create MicroGlycoDB, a database for the glycan information of microbes.

E.coli O-antigen is a complex polysaccharide that consists of repeating units of sugar molecules in bacterial cell walls and has been known that it is an important virulence factor in pathogenesis. For that reason, the O-antigens have been targeted to biomarkers for diagnosing infectious diseases or developing glycoconjugate vaccines. The collected data from the ECODAB database [3] are organized in tabular format and the O-antigens are RDFized to describe in computer-readable format.

Pathway data in biology have been used as an essential means to obtain insight into a cellular process that is consisted of molecular components and their interactions. With the accumulation of glycan data, it is realized that glycosylation plays an important role in a wide variety of cellular pathways, and the lack of glycan information in pathway data can cause an improper pathway analysis. Also, the major pathway database such as Reactome [4], and KEGG [5] does not provide sufficient glycan information at present because considerable time and effort is a need for data curation. I designed to create a repository that allows the data to be shared with other pathway databases that lack glycan information. For this purpose, the input resources have been described in a standardized format using ontologies.

Methods

1. Data collection and processing

The list of *E.coli* O-antigens and the information related to glycosyltransferases are collected from ECODAB (*E.coli* O-antigen database) and CSDB (Carbohydrate structure database) [6], respectively. The collected data from public web pages or co-researchers are organized in tabular format to help data to be converted into RDF format. The collected O-antigen list was processed into proper text type such as Linear Code of the glycan to obtain the GlyTouCan [7] identity number, which is essential to reference a glycan structure as an IRI (Internationalized Resource Identifier) providing consistent referencing. Before conducting this step, the proper ontologies were searched because the first row represents the property that links the subject and object. The ontologies are listed in the result Figure 1. The processed data are converted using RDFLib, a library of Python.

2. Taxonomy and Proteins

A taxonomy dump file is obtained in a Web Ontology Language (OWL) ontology file format from DDBJ (DNA Databank of Japan). The file is processed to contain the species name, and taxonomic identifier number in NCBI taxonomy using SPAQRList. To gain protein information on the Uniprot, the proper query is asked to the SPARQL endpoint of the Uniprot. The protein entries are extracted along with the corresponding species, identifier number, and gene name including the unreviewed proteins. The final result data are also processed in the same with the method used in the Taxonomy.

3. Verification of RDF document by the ShEx

The created triples are inspected before being uploaded to the triple storage using the ShEx (Shape Expression) [8] which is developed to validate the schema of the RDF model.

Results

1. Schemas for RDF model and RDFication

a. Microbes

As mentioned in the introduction, The glycan-related information on the 6 species of microbes was obtained in excel file format from collaborate. As the first step of RDFization, a schema was designed depending on their information (Table 1). Because the anatomical structure of microbes as a location for the glycan and types of glycans present in that structure is described in plain text, the data was processed to make it more explicit and easier to serialize RDF sentences. The information on bacterial structure such as capsule, outer membrane, and the bacterial glycans including N-glycan, LPS (lipopolysaccharide), LOS (lipooligosaccharide), GPI (glycophospha-tidylinositol), GPL (Glycophospholipid) was presented using ontologies such as GeneOntology [9], ChEBI [10]. I made a simple code to serialize the data in tabular format into RDF sentences using RDFLib, a Python library. The created RDF files were saved to the RDF storage namely VIRTUSO [11] which is a graph database for the RDF triple store. Because the virtuoso supports the storage of graph data, users can create new data through data federation or integration. SPARQLists [12] for each microbe was created to present their own data according to their unique set of characteristics.

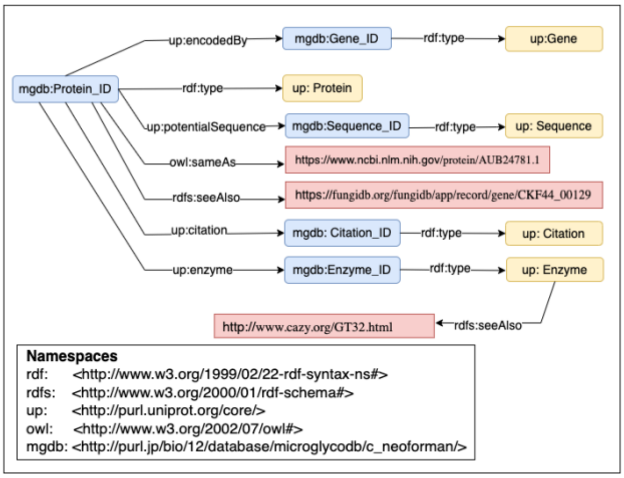

up:encodedBy
mgdb:Gene_ID
rdf:type
up:Gene
mgdb:Protein_ID
rdf:type
up: Protein
up:potentialSequence
mgdb:Sequence_ID
rdf:type
up: Sequence
owl:sameAs
https://www.ncbi.nlm.nih.gov/protein/AUB24781.1
rdfs:seeAlso
https://fungidb.org/fungidb/app/record/gene/CKF44_00129
up:citation
mgdb: Citation_ID
rdf:type
up: Citation
up:enzyme
mgdb:Enzyme_ID
rdf:type
up: Enzyme
http://www.cazy.org/GT32.html
rdfs:seeAlso
Namespaces
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
up: <http://purl.uniprot.org/core/>
owl: <http://www.w3.org/2002/07/owl#>
mgdb: <http://purl.jp/bio/12/database/microglycodb/c_neoforman/>

Figure 1. An example of the RDF graph schema and the used ontologies

The blue-colored node means a URI of MicroglycoDB. The yellow-colored node presents Uniprot ontology and the red square presents a URI of an external database.

b. E. coli O-antigens

E. coli O-antigen can act as an immunogen and has been used to classify *E. coli* into different serogroups. The O-antigen is formed through the biosynthesis process that a donor glycan binds to a substrate glycan by the glycosyl transferase. I designed the O-glycan structure to be described as a series of glycan extension reactions that form a single pathway. For the semantic representation of the O-antigen polysaccharides, the collected data from the ECODAB database were processed. The glycan nomenclature in ECODAB was transformed to the linear code format using the in-house converter which is a simple python code for organizing the data and registered into the GlyTouCan repository to get a unique identifier. The information about the glycosyltransferase enzymes was searched from the CSDB database. The cleaned data of O-antigens were organized in table format to be converted to triple sentences. To represent the structure of O-antigen as a pathway, BioPAX ontology was used which is a community-based and standardized ontology designed for biological pathway description to facilitate data sharing between pathway databases []. As the RDFizaton of the microbes, the O-antigens were serialized to triple sentences using the same python code and saved to the storage.

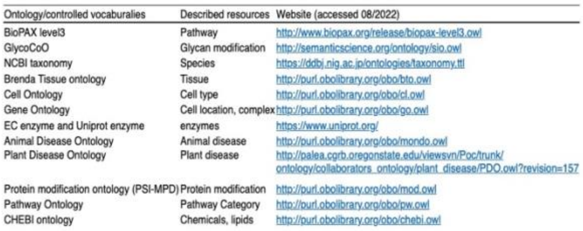
2. SPARQL and Triple store

To generate and upload triple sentences into a triple store, I have used the SPARQList application which provides a user interface to implement the SPARQL query supporting a trace for debugging and transforming the results of the query into JASON format by JavaScript. Depending on the type of data being retrieved, many SPARQList were created. The source code of the SPARQList is available at https://gpr-sparqlist.alpha.glycosmos.org/sparqlist/ To add new triples (i.e., subject, predicate, object) to an RDF graph that is created for input data in the VIRTUOSO database, the *INSERT* statement is used. Also, a new SPARQList is made to inspect whether the added triples are correct or not. I confirmed that the RDF data of microbes, *E. coli* O-antigen, and pathway information for the repository was successfully generated and visualized in the MicroGlycoDB and GlyCosmos portal sites [13].

3Repository for glycan-related pathway data

a. Ontologies for the resources

A biological pathway is very complex that consists of a series of interactions including biochemical reactions. Also, all kinds of molecules participating in a reaction become an entity of each reaction. The entities are resources that include concepts like cellular interactions and physical entities such as cellular anatomy or proteins. I inspected the existing ontologies and the controlled vocabularies to represent a wide range of resources (Table 1) semantically.



Ontology/controlled vocaburalies	Described resources	Website (accessed 08/2022)
BioPAX level3	Pathway	http://www.biopax.org/release/biopax-level3.owl
GlycoCoO	Glycan modification	http://semanticscience.org/ontology/sio.owl
NCBI taxonomy	Species	https://ddbj.nig.ac.jp/ontologies/taxonomy.ttl
Brenda Tissue ontology	Tissue	http://purl.obolibrary.org/obo/bto.owl
Cell Ontology	Cell type	http://purl.obolibrary.org/obo/cl.owl
Gene Ontology	Cell location, complex	http://purl.obolibrary.org/obo/go.owl
EC enzyme and Uniprot enzyme	enzymes	https://www.uniprot.org/
Animal Disease Ontology	Animal disease	http://purl.obolibrary.org/obo/mondo.owl
Plant Disease Ontology	Plant disease	http://palea.cgrb.oregonstate.edu/viewsvn/Poc/trunk/ ontology/collaborators_ontology/plant_disease/PDO.owl?revision=157
Protein modification ontology (PSI-MPD)	Protein modification	http://purl.obolibrary.org/obo/mod.owl
Pathway Ontology	Pathway Category	http://purl.obolibrary.org/obo/pw.owl
CHEBI ontology	Chemicals, lipids	http://purl.obolibrary.org/obo/chebi.owl

Table 1. The ontologies for resource annotation

The process that includes catalytic reaction and the process of a pathway was described according to the guide documentation of the BioPAX ontology. The RDFized data based on the BioPAX ontology will facilitate the data link or integration with the major pathway database such as Reactome, and KEGG.

b. User interface

The repository has two different input paths: one is glycan biosynthesis, and the other is a glycan-related pathway. Both are designed based on BioPAX ontology. The former pathway focuses on the glycosyl enzymes and glycans, while the latter concerns the related glycan information on a biological pathway. Given the inherent complexity of the pathway, I have created a number of devices to make registering easier when the user input data. For example, to show a diagram of the input sequence in the input view or to provide a select option that the product of a previous reaction can be used as a reactant or controller in the next reaction without requiring the user to type it in. In addition, a confirmation table was prepared to help users to see what they entered after completing the input of each reaction information. Because the resources that participated in the reaction such as proteins, complexes, glycans, chemicals, and lipids were prepared from ontologies with a URI, they consist of a huge list. The input field for resources having to autocomplete function that suggests words or phrases based on the user's previous inputs is supplied to provide convenience.

For the web service, the repository has been developed using a web application framework called Ruby on Rails, SPARQL, Javascript libraries including the Cytoscape for visualization of the pathway, and the Web Components such as the Pagination table of MetaStanza or the Glycosmos-table that was embedded into a web page (Figure 2). The table that is created by Web Components was devoted to displaying a data list registered in the repository or the outcomes of a keyword search. Also, to visualize glycan-related pathway data, each pathway is visualized in its own view page using the Cytoscape and the information obtained from the SPARQL query.

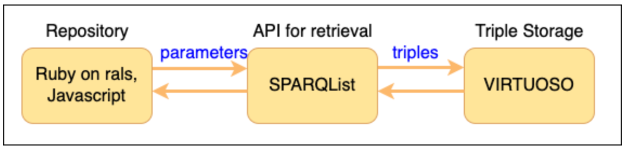


Figure 2. The architecture for the repository

The repository presents the index table in tabular form, where each row responds to a single reaction. In the glycan synthesis, the following things will be seen by users: the constituent parts of the glycan structure with GlyTouCan identifier, associated glycosyl enzymes, a species in which the glycan structure was identified, and a related disease if it is registered. Meanwhile, the reaction of the row in the glycan-related pathway index table presents biochemical reactions consisting of reactant, product, and controller that accelerate the catalytic reaction if it exists. Users can see the reaction components participating in the biochemical reaction, the reactions involved in the pathway, and the species information. The user can move to the detailed view containing visualization information by clicking on the identifier number of the pathway.

c. Input test

The EGFR (epidermal growth factor receptor) domain is heavily glycosylated, and glycosylation on the receptor regulates a number of important cellular processes in both normal and cancer cells. However, the representative pathway database does not fully provide glycosylation information. The first step is to represent glycans that are attached to the amino acid sequence. For this, the sequence number and kind of amino acids in the binding site for the glycan modification can be described using the input field with autocomplete function. The participant resources were chosen from the UniProt protein list containing the species information. Complex information that is combined by two reactant proteins was entered by selecting from complex ontology from GO complex ontology, however, When the precise complex name does not exist in the list, the complex is entered with a name linked by clone, such as A:B. Using Cytoscape, the pathway was visualized after the input was completed, showing the types of resources, cellular localization, and reaction order with symbols, lines, and arrows. The user interface is being developed to adequately represent the information of the relevant resources.

Discussion

The ontologies were used to annotate the concepts and biomolecules to facilitate the semantic linkage between the pathway data. Each pathway database is complementary in terms of completeness and accuracy of data because the pathway data is intrinsically interlinked with various resources, the number of which has been growing rapidly with the advancement of high throughput techniques. Therefore, the annotation for the resources of biomolecules and reactions will be the most important issue for the effective integration of pathway data.

I intended that the pathway data that is unpublished or left in literature is attended to pathway knowledge integration by engaging biologists who do not have knowledge about bioinformatics. However, after the information is contributed, data have to be updated with new discoveries. To facilitate edition of pathway data, A strategic approach is being devised.

References

1. Varki, A.: Biological roles of glycans. Glycobiology. 27, 3–49 (2017). https://doi.org/10.1093/glycob/cww086
2. Wu, H., Yamaguchi, A.: Semantic web technologies for the big data in life sciences. Biosci. Trends. 8, 192–201 (2014). https://doi.org/10.5582/bst.2014.01048
3. Rojas-Macias, M.A., Ståhle, J., Lütteke, T., Widmalm, G.: Development of the ECODAB into a relational database for Escherichia coli O-antigens and other bacterial polysaccharides. Glycobiology. 25, 341–347 (2014). https://doi.org/10.1093/glycob/cwu116
4. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The Reactome Pathway Knowledgebase. Nucleic Acids Res. 46, D649–D655 (2018). https://doi.org/10.1093/nar/gkx1132
5. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. 4
6. Herget, S., Toukach, P.V., Ranzinger, R., Hull, W.E., Knirel, Y.A., Von Der Lieth, C.W.: Statistical analysis of the bacterial carbohydrate structure data base (BCSDB): Characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. BMC Struct. Biol. 8, (2008). https://doi.org/10.1186/1472-6807-8-35
7. Fujita, A., Aoki, N.P., Shinmachi, D., Matsubara, M., Tsuchiya, S., Shiota, M., Ono, T., Yamada, I., Aoki-Kinoshita, K.F.: The international glycan repository GlyTouCan version 3.0. Nucleic Acids Res. 49, D1529–D1533 (2021). https://doi.org/10.1093/nar/gkaa947
8. Thornton, K., Solbrig, H., Stupp, G.S., Labra Gayo, J.E., Mietchen, D., Prud'hommeaux, E., Waagmeester, A.: Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation. In: Hitzler, P., Fernández, M., Janowicz, K., Zaveri, A., Gray, A.J.G., Lopez, V., Haller, A., and Hammar, K. (eds.) The Semantic Web. pp. 606–620. Springer International Publishing, Cham (2019)
9. Hill, D.P., Smith, B., McAndrews-Hill, M.S., Blake, J.A.: Gene Ontology annotations: What they mean and where they come from. BMC Bioinformatics. 9, 1–9 (2008). https://doi.org/10.1186/1471-2105-9-S5-S2
10. de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C.: Chemical Entities of Biological Interest: an update. Nucleic Acids Res. 38, D249–D254 (2010). https://doi.org/10.1093/nar/gkp886
11. Erling, O., Mikhailov, I.: RDF Support in the Virtuoso DBMS.
12. Katayama, T., Kawashima, S.: SPARQList: Markdown-based highly configurable REST API hosting server for SPARQL. 2
13. Yamada, I., Shiota, M., Shinmachi, D., Ono, T., Tsuchiya, S., Hosoda, M., Fujita, A., Aoki, N.P., Watanabe, Y., Fujita, N., Angata, K., Kaji, H., Narimatsu, H., Okuda, S., Aoki-Kinoshita, K.F.: The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. Nat. Methods. 17, 649–650 (2020). https://doi.org/10.1038/s41592-020-0879-8