Published online 17 October 2015

doi: 10.1093/nar/gkv1041

GlyTouCan 1.0 – The international glycan structure repository

Kiyoko Aoki-Kinoshita[1,2,*], Sanjay Agravat[3], Nobuyuki P. Aoki[1], Sena Arpinar[4], Richard D. Cummings[5], Akihiro Fujita[1], Noriaki Fujita[2], Gerald M. Hart[6], Stuart M. Haslam[7], Toshisuke Kawasaki[8], Masaaki Matsubara[9], Kelley W. Moreman[4], Shujiro Okuda[10], Michael Pierce[4], René Ranzinger[4], Toshihide Shikanai[2], Daisuke Shinmachi[1], Elena Solovieva[2], Yoshinori Suzuki[2], Shinichiro Tsuchiya[1], Issaku Yamada[9], William S. York[4], Joseph Zaia[11] and Hisashi Narimatsu[2]

[1]Faculty of Science and Engineering, Soka University, Tokyo 192-8577, Japan, [2]Glycoscience and Glycotechnology Research Group, AIST, Ibaraki 305-8568, Japan, [3]Department of Mathematics and Computer Science, Emory University, Atlanta, GA, 30322, USA, [4]Complex Carbohydrate Research Center, University of Georgia, Athens, GA, 30602, USA, [5]Department of Surgery, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, 02115,USA, [6]Department of Biological Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD, 21205, USA, [7]Department of Life Sciences, Faculty of National Sciences, Imperial College London, London, SW7 2AZ, UK, [8]Research Center for Glycobiotechnology, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan, [9]The Noguchi Institute, Tokyo 173-0003, Japan, [10]Niigata University Graduate School of Medical and Dental Sciences, Niigata 951-8510, Japan and [11]Center for Biomedical Mass Spectrometry, Dept. of Biochemistry, Boston University School of Medicine, Boston, MA, 02118, USA

Received August 10, 2015; Revised September 15, 2015; Accepted September 30, 2015

ABSTRACT

Glycans are known as the third major class of biopolymers, next to DNA and proteins. They cover the surfaces of many cells, serving as the 'face' of cells, whereby other biomolecules and viruses interact. The structure of glycans, however, differs greatly from DNA and proteins in that they are branched, as opposed to linear sequences of amino acids or nucleotides. Therefore, the storage of glycan information in databases, let alone their curation, has been a difficult problem. This has caused many duplicated efforts when integration is attempted between different databases, making an international repository for glycan structures, where unique accession numbers are assigned to every identified glycan structure, necessary. As such, an international team of developers and glycobiologists have collaborated to develop this repository, called GlyTouCan and is available at http://glytoucan.org/, to provide a centralized resource for depositing glycan structures, compositions and topologies, and to retrieve accession numbers for each of these registered entries. This will thus enable researchers to reference glycan structures simply by accession number, as opposed to by chemical structure, which has been a burden to integrate glycomics databases in the past.

INTRODUCTION

As there are already numerous publicly available glycan and carbohydrate databases, it is important to clarify how GlyTouCan is different. As reported from the ACGG-DB meeting (1) it was agreed that uniquely identifying glycan structures would be the main content of this repository. The user and time/date of registration would also be attached with the glycan structure information. Therefore the ability to register glycan structures will be the main service provided, and a unique accession number will be generated to be used for reference in any lycan-related research or publications. We note that it was decided at this ACGG-DB meeting that in order to simplify development, GlyTouCan will be responsible only for glycan structures and minimal metadata (user and registration time/date). That is, other metadata such as experimental data, aglycon information and publication would be outside the scope of this repository, and it would be the responsibility of curated databases such as UniCarbKB (2) and BCSDB (3) to cover such information. Other attributes such as mass and motifs are attached to each entry as they can be computed from the structure. Sim-

[*]To whom correspondence should be addressed. Tel: +81 42 691 4116; Fax: +81 42 691 4116; Email: kkiyoko@soka.ac.jp

ilarly, links to other database entries are computed from the structure as well. However, glycan structures need not consist purely of monosaccharides as GlyTouCan also covers substituents on the monosaccharides that are not on the reducing end of the glycan. Thus the major substituents such as sulfates, phosphates, etc. are handled in version 1.0.

The first release of the International Glycan Structure Repository was recently completed to enable this registration of glycan structures through either a web browser or a programmable interface. The repository was pre-registered with structures available in major public glycan databases that provided their data: at the time of this writing, links are available for GlycomeDB (4), BCSDB (3), GlycoEpitope (5), and progress is currently being made to link with UniCarbKB (2). For entries in other databases that contained duplicate glycan structures (e.g. because of the inclusion of aglycons), these became single entries in GlyTouCan, and the Linked DB section would list all entries in the original database containing that glycan structure. For the current version, there were instances where structures could not be registered because they could not be converted to GlycoCT (6). In this case, we put these entries on hold for the next version, which will use the Web 3.0 Unique Representation of Carbohydrate Structures (WURCS) (7) as the base format and should be able to handle such cases. The site also provides a variety of methods to query and browse the relationships of glycan structures based upon logic inherent among the registered structures. Herein, we describe the functionality developed in the frontend of GlyTouCan, backend web services, and linked data endpoint. The aim of GlyTouCan is to simplify the identification of glycan structures and to help link related research within the many life sciences databases available worldwide. In order to facilitate the integration of various life sciences databases, we chose to use Semantic Web technologies, in particular Resource Description Framework (RDF) as the base technology. Linked Data refers to the data that are contained within the Semantic Web. As such, a triplestore (RDF database) is incorporated into the GlyTouCan architecture, and an endpoint is made available as a URL where the RDF data (triplestore) can be accessed using the SPARQL query language. Figure 1 illustrates the overall architecture of GlyTouCan, which consists of three major parts: the front-end web interface, the backend which also includes a relational database, and the triplestore. As GlyTouCan was an international collaboration between the US and Japan, the US developers first developed the backend system, which was used by the Japanese developers to develop the web interface, and then converted to the triplestore.

This article will cover a brief explanation of fundamental GlyTouCan functionality. It should be noted that detailed documentation is available on the user guide website (http://code.glytoucan.org). This site not only includes a detailed explanation of all functionalities, but also Linked Data information and links to source code. It will primarily cover the process by which a accession number can be searched for, registered via a graphical user interface, accessed using a program and finally queried through our publicly available Linked Data endpoint.

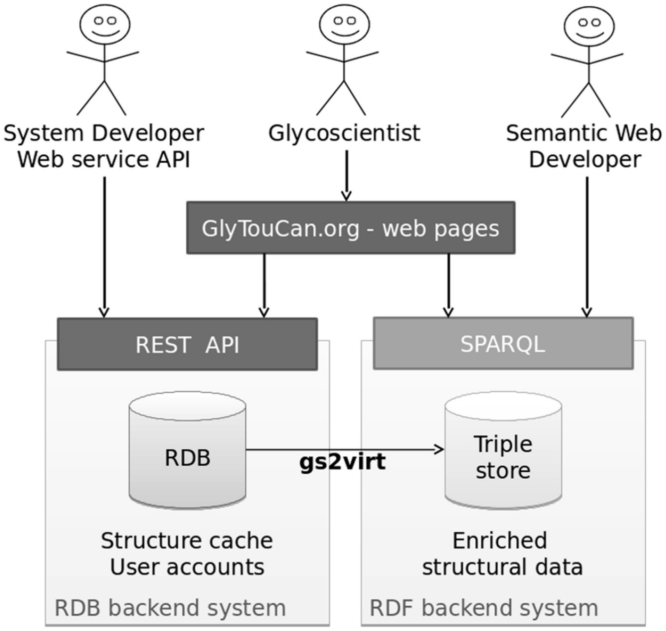
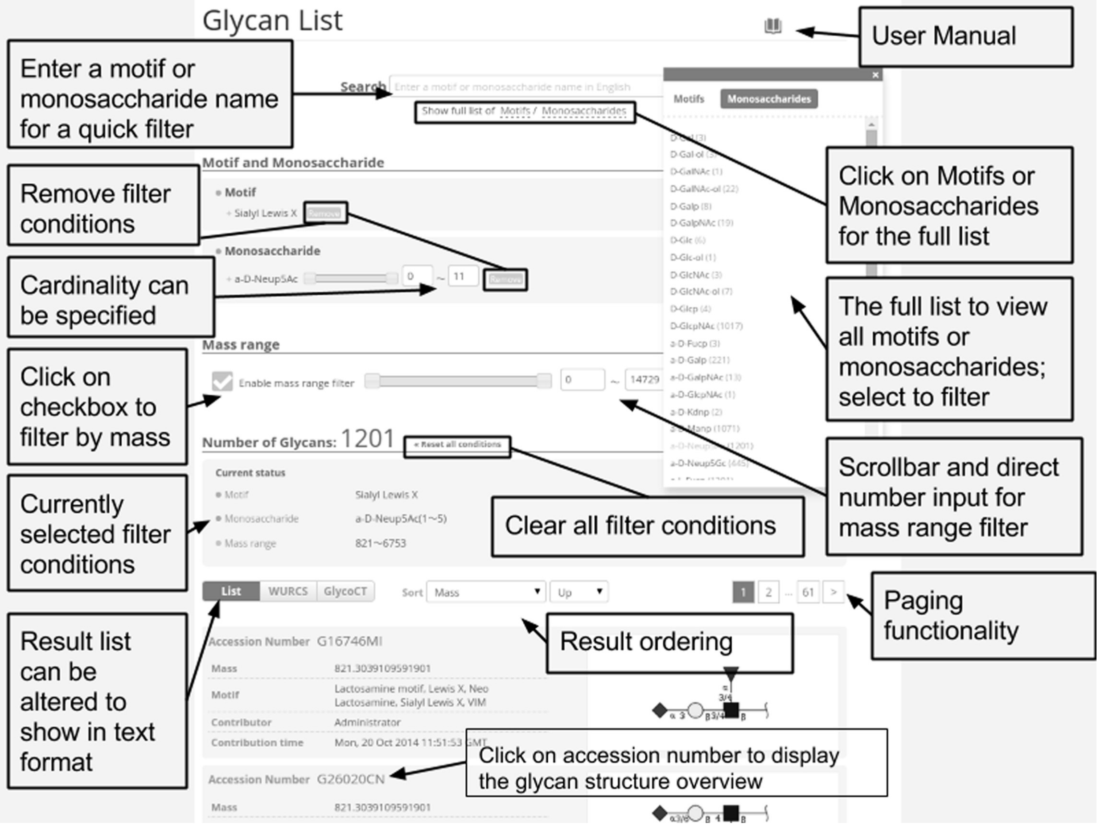


Figure 1. A schematic of the Glycan Repository architecture, which comprises of the front-end web pages, a REST API interface to the relational database, and a SPARQL interface to the triplestore.

SEARCH, BROWSE AND FILTER

GlyTouCan offers a variety of methods to search and browse through the glycan structures registered and other preloaded data available. In the cases where there are a vast number of search results, it is possible to filter based upon common structural attributes. Structures from GlycomeDB (4), BCSDB (3) and GlycoEpitope (5) are a few of the databases that were pre-registered in GlyTouCan. The browsing functionality shows the wealth of information that was gathered from the currently available data in these public glycan databases. This section will explain the browsing options available, as well as the three search methods: graphical, sequence text and motif.

Within the menu bar, the 'View All' section has a 'Glycan List' link which is a quick method to view the glycan structures contained within the entire repository. The options to filter down the results are shown in Figure 2. The top text input field can be used to screen for a specific motif or monosaccharide name. The mass range can be quantified by inputting the minimum and/or maximum values, or by altering the scrollbar. The ranges for these filtering options, such as the maximum mass or monosaccharides to choose from, are retrieved from the entire scope of the glycan structures currently registered. If a monosaccharide is chosen, it is then possible to filter for a specific cardinality of that monosaccharide. In the results list, the format can also be altered to display the glycan sequences in either GlycoCT (6) or WURCS (7). Moreover, the results can be sorted according to accession number, contributor, mass or date entered. This same browsing and filtering functionality is available in the search results of the motif search, described next.



Number of Glycans: 1201

Current status

Motif	Sialyl Lewis X
Monosaccharide	a-D-Neup5Ac(1~5)
Mass range	821~6753

Accession Number G16746MI

Mass	821.3039109591901
Motif	Lactosamine motif, Lewis X, Neo Lactosamine, Sialyl Lewis X, VIM
Contributor	Administrator
Contribution time	Mon, 20 Oct 2014 11:51:53 GMT

Accession Number G26020CN

Mass	821.3039109591901

User Manual

Enter a motif or monosaccharide name for a quick filter

Remove filter conditions

Cardinality can be specified

Click on checkbox to filter by mass

Currently selected filter conditions

Result list can be altered to show in text format

Click on Motifs or Monosaccharides for the full list

The full list to view all motifs or monosaccharides; select to filter

Scrollbar and direct number input for mass range filter

Clear all filter conditions

Paging functionality

Result ordering

Click on accession number to display the glycan structure overview

Figure 2. The browsing functionality can filter out the results based upon motifs, monosaccharide names, monosaccharide cardinality and mass range. It is also possible to sort the results in order of accession number, contributor, mass or date entered.

The Motif search method (https://glytoucan.org/Motifs/search) initially displays a list of the 61 predefined motifs, which are defined as commonly found glycan substructure patterns in the literature. The *N*-Glycan core and Sialyl-Lewis X are examples of motifs. Clicking on a motif in this list displays the listing of the glycans registered in GlyTouCan that contain the selected motif. Some motifs are very common and can thus return a very large number of glycan structures containing the selected motif. Therefore, multiple filtering options allow users to easily find their glycan structure of interest.

In the cases where a more specific glycan structure needs to be specified, it is possible to build a specific glycan substructure via the GlycanBuilder interface (8) (https://glytoucan.org/Structures/graphical) as displayed in Figure 4. Once the structure building is complete and the search button is pressed, it is converted into GlycoCT (6) format and used to search the repository. Regardless of whether the structure has been registered already, it and any superstructures (registered glycan structures that contain the input glycan structure), will be displayed in the search results. If the structure used to search for is not registered yet, it will indicate so by displaying 'not registered' instead of an accession number in the results page.

The text search method (https://glytoucan.org/Structures/structureSearch) is similar to the graphical interface, however, the input can be specified using commonly utilized sequence formats such as GlycoCT (6), KCF (9) or LinearCode® (10). After submission, the format is converted to GlycoCT (6) as necessary, after which the results screen is displayed, similar to the graphical search method. From the results listing, more details about a specific glycan structure can seen by clicking on an accession number, which will display the glycan structure overview page.

The glycan structure overview page displays all data specific to the glycan structure and any related information. Each section on this page is labeled in Figure 3. The core structural information is at the top of the page, where a graphical representation is displayed (graphical representation display options can be changed using the Preferences menu). Below the image, sections describing motifs contained and monosaccharide compositions are displayed. Further below, a listing of the public databases under the Linked DB section is shown with links to the database site where more (curated) information can be referenced.

REGISTRATION WORKFLOW

From the GlyTouCan website, there are currently three different methods to register structures into the repository: graphically, text and file upload. Text and file upload are limited to GlycoCT (6), LinearCode (10) and KCF (9) glycan structure text sequence formats. For the sake of brevity we will review the simplest method which is via the graphical interface using GlycanBuilder (8).

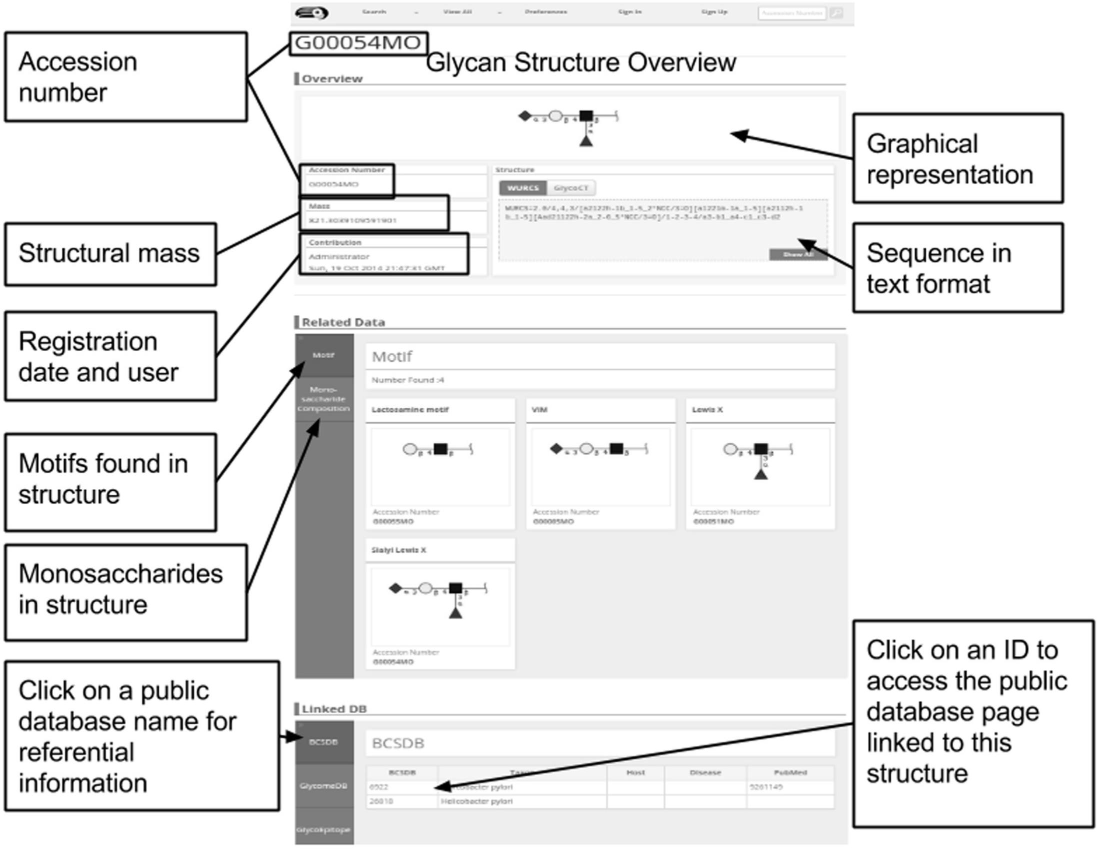


Figure 3. The glycan structure overview page displays the core structural information as well as any related data such as motifs found and monosaccharide compositions. A listing of the public databases under the Linked DB section is also displayed with links to reference the structure information directly.

In order to access the registration functionality, it is first necessary to sign in by clicking on the 'Sign In' button. Login is required in order to associate a user to the structures that are registered. For each glycan structure, a username is displayed indicating the person who registered the structure.

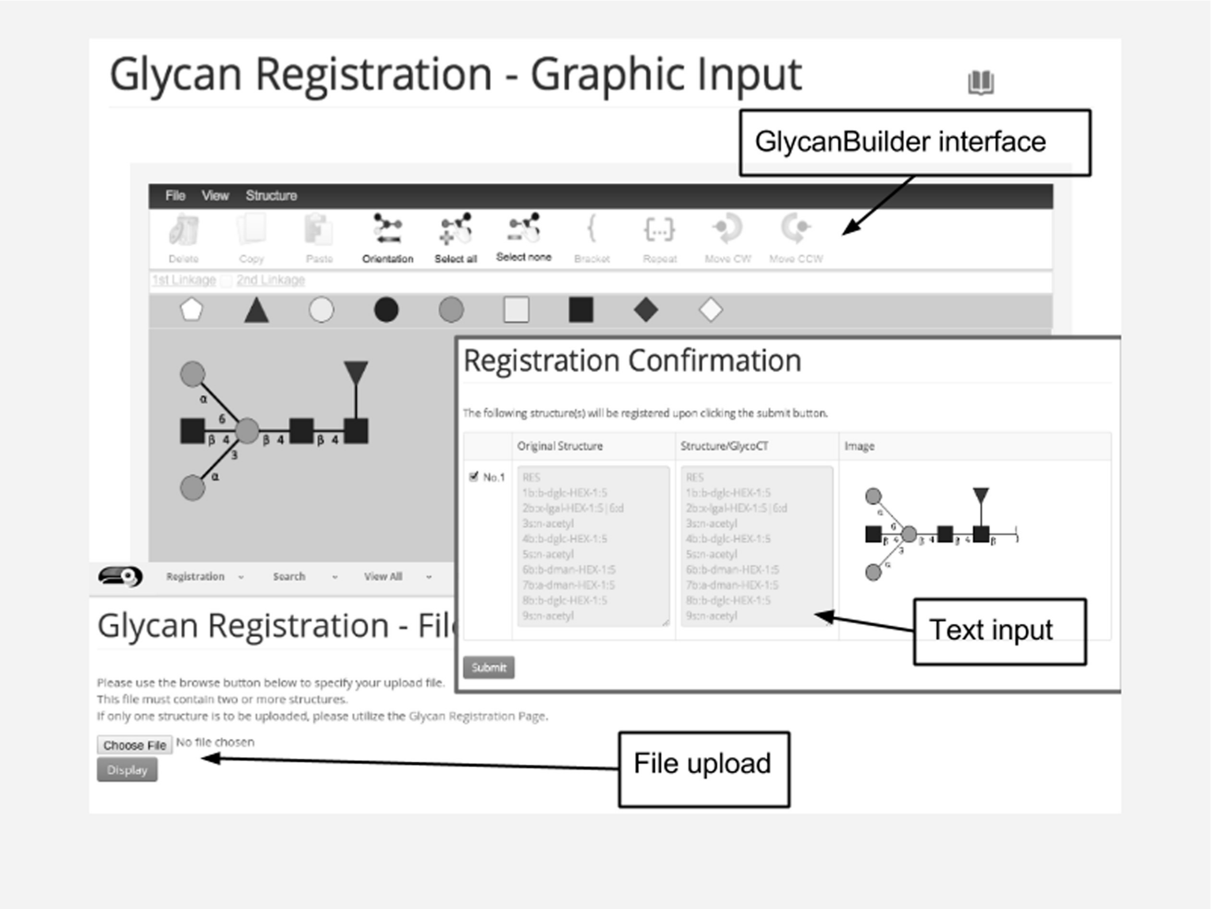
Glycan structures can be drawn on the GlycanBuilder (8) interface and once the submit button is pressed, the graphical form of the structure is converted into the GlycoCT format (6) and submitted to the backend system. The backend system checks for duplicates and generates the accession number for the newly registered glycan structure. Structures with ambiguous linkages and fragments are all accepted, however, checks such as conflicting linkages are made to ensure that chemically impossible structures are not registered. We note here that structures with ambiguous linkages will be registered if the GlycoCT (6) string comparison does not match any other registered structures, even if other ambiguous structures could potentially match it. This will become an important feature for users to be able to browse ambiguous structures and find more defined structures that match it (e.g. a structure registered as 'Gal(?1-?)GlcNAc' could match structures 'Gal(b1–4)GlcNAc' and 'Gal(b1-?)GlcNAc'). Such related (subsumed) structures will be made browseable in the next release using a visualization tool currently being developed, as all 'subsumed' structures and their relationships will be added to the triplestore, thus allowing users to find these relationships easily through a visualization interface.

It is possible to utilize the backend server for other methods of access using a standard programmable interface, which will be described later. After registering the glycan structure, multiple server-side processes are executed in order to analyze and link the structure to previously registered glycan structures or glycan motifs (a commonly known glycan substructure pattern, explained below in the Motif Search section). Lastly a confirmation page will be shown which displays the details found for the registered structure, such as the calculated mass, motif and monosaccharide content.

Following the registration process, another batch process (labeled gs2virt in Figure 1) is executed which retrieves the structure and accession number from the relational database and converts the data into RDF format. These data are stored into a separately available triplestore and linked to content such as conversion to other textual formats.

BACKEND SERVER

The GlyTouCan backend server provides a method for system or application developers to access the information stored in the repository. Using the provided library, or application programming interface (API), it is possible to search, retrieve and register new glycan structures. Methods



Registration Confirmation

The following structure(s) will be registered upon clicking the submit button.

	Original Structure	Structure/GlycoCT	Image
☑ No.1	RES 1b:b-dglc-HEX-1:5 2b:x-lgal-HEX-1:5|6:d 3s:n-acetyl 4b:b-dglc-HEX-1:5 5s:n-acetyl 6b:b-dman-HEX-1:5 7b:a-dman-HEX-1:5 8b:b-dglc-HEX-1:5 9s:n-acetyl	RES 1b:b-dglc-HEX-1:5 2b:x-lgal-HEX-1:5|6:d 3s:n-acetyl 4b:b-dglc-HEX-1:5 5s:n-acetyl 6b:b-dman-HEX-1:5 7b:a-dman-HEX-1:5 8b:b-dglc-HEX-1:5 9s:n-acetyl	

Glycan Registration - File

Please use the browse button below to specify your upload file.
This file must contain two or more structures.
If only one structure is to be uploaded, please utilize the Glycan Registration Page.

Figure 4. The methods to register a structure are by building one graphically from GlycanBuilder (8), pasting a sequence text into a form or uploading a file. The uppermost image shows the GlycanBuilder user interface. The registration confirmation screen shows an example of registering a GlycoCT (6) text sequence. The file upload method allows for registering multiple structures in a file.

provided include structure registration, data retrieval and search services. The complete specification is available online and describes in detail the functionality of each method (http://api.glytoucan.org/documentation/apidoc.html).

As an example, we will request the GlycoCT (6) format of an already registered glycan structure utilizing the accession number. This is done with a simple call to a web location (uniform resource locator; URL) using the accession number as a parameter by embedding it within the URL. The accession number for an '*N*-Glycan core' (Man3GlcNAc2) structure is G00026MO. As the API is based on the REST (REpresentational State Transfer) protocol, a request in the following format can be made via a URL to retrieve the sequence (and other information) of the glycan to retrieve an XML-formatted response: http://api.glytoucan.org/glycans/G00026MO.

This request will return the following response:

```
<glycan accessionNumber="G00026MO" dateEntered="2014-10-20T06:40:24.086Z"
mass="910.3277800379999" structureLength="194">
  <contributor>Administrator</contributor>
  <structure>
    RES 1b:x-dglc-HEX-x:x 2s:n-acetyl 3b:b-dglc-HEX-1:5 4s:n-acetyl 5b:b-dman-HEX-
    1:5 6b:a-dman-HEX-1:5 7b:a-dman-HEX-1:5 LIN 1:1d(2+1)2n 2:1o(4+1)3d
    3:3d(2+1)4n 4:3o(4+1)5d 5:5o(3+1)6d 6:5o(6+1)7d
  </structure>
</glycan>
```

The xml-formatted response of a glycan structure request, where the structure is represented in GlycoCT (6) format in the <structure> tag. The response can also be returned in JSON format by simply adding '.json' to the end of the http request.

In a similar fashion, registration of a glycan structure can be completed by inputting the information in GlycoCT (6), and the XML output will indicate the relevant accession number generated. It is thus possible to create separate applications or websites utilizing this API to register structures and extract information in the repository.

LINKED DATA ENDPOINT

As described in the previous section, the registered structures are batch-processed in order to generate and enrich the RDF data. These data are then made freely available through the Linked Data endpoint at http://ts.glytoucan.org/sparql and can be accessed with standard SPARQL queries. For readers interested in using the SPARQL interface, we provide an explanation and examples of how to retrieve/download the data in GlyTouCan in the Supplementary Materials.

CONCLUSION AND FUTURE WORK

In this manuscript, we describe the first version of the Glycan Structure Repository, called GlyTouCan, which allows users to obtain unique accession numbers for any glycan structure. It is possible to register a variety of glycan structure information: glycan compositions, glycans with ambiguous linkages, ambiguous fragments and fully-defined structures. Based on the accession numbers in GlyTouCan, existing databases can simply use these numbers when referring to glycan structures. It will then be very simple for GlyTouCan to link back to the databases that have stored these same accession numbers, and display the linkages in search results. Comparison of glycan structures with specific sequence formats is no longer needed when comparing glycan data across databases or when searching for glycans in publications.

In preparation for the next release, one of the major items proposed is to have more user-editable content, such as the ability to make comments and rate entries. Thus, nearly all of the data in GlyTouCan such as the structures, motifs and monosaccharides would be able to be commented on or rated by registered users. The contents of entries themselves will not be editable, however, a community-driven framework for data quality check would improve content quality and give direction to new functionality requirements.

Adding relationships between glycan structures also adds value to the data set. Enrichment with sub/super-structure, subsumed structures and isomeric relationships is simply a matter of creating algorithms that analyze the structures already registered; however, performance analysis will be necessary as volume could be highly resource-intensive. Once this is ready, it will be possible to show ambiguous siblings or sub/super structure content while browsing through search results.

Finally, the post-registration processes are currently run at a batched interval in order to process newly submitted glycan structures, thus there may be some delay between the time when a glycan structure is registered and when its data are viewable from a browser. Ideally this should be a real-time process, however, the complexity involved and potential resource bottlenecks require more evaluation on how this is to be implemented. This is a high-priority item that will be covered after the first release.

ACKNOWLEDGEMENT

The authors would like to acknowledge Mitsuko Yamazaki of Soka University for her assistance in supporting various tasks behind the scenes, including web page design and localization support.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Bioscience Database Center (NBDC), Japan Science and Technology Agency (JST) and National Institute of General Medical Sciences [8P41GM103490]; Biotechnology and Biological Sciences Research Council [BBF0083091 and BB/K016164/1 to S.M.H]. Funding for open access charge: Japan Science and Technology Agency (JST) Grant Code TG2608 and Soka University.

Conflict of interest statement. None declared.

REFERENCES

1. Aoki-Kinoshita,K.F., Sawaki,H., An,H.J., Campbell,M., Okuda,S., Hsu,D., Kato,M., Yamada,I., Kawasaki,T., Khoo,K.H. *et al.* (2013) The Fifth ACGG-DB Meeting Report: Towards an International Glycan Structure Repository. *Glycobiology*, **23**, 1422–1424.
2. Campbell,M., Peterson,R., Mariethoz,J., Gasteiger,E., Akune,Y., Aoki-Kinoshita,K.F., Lisacek,F. and Packer,N.H. (2014) UniCarbkb: building a knowledge platform for glycoproteomics. *Nucleic Acids Res.*, **42**, D215–D221.
3. Toukach,Ph.V. (2011) Bacterial Carbohydrate Structure Database 3: Principles and Realization. *J. Chem. Inf. Model.*, **51**, 159–170.
4. Ranzinger,R., Herget,S., von der Lieth,C.W. and Frank,M. (2010) GlycomeDB–a unified database for carbohydrate structures. *Nucleic Acids Res.*, **39**, D373–D376.
5. Okuda,S., Nakao,H. and Kawasaki,T. (2015) GlycoEpitope: Database for Carbohydrate Antigen and Antibody. In: Taniguchi,N, Endo,T, Hart,GW, Seeberger,PH and Wong,C-H (eds). *Glycosci. Biol. Med*. Springer, Japan.
6. Herget,S., Ranzinger,R., Maass,K. and von der Lieth,C.W. (2008) GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr. Res.*, **343**, 2162–2171.
7. Tanaka,K., Aoki-Kinoshita,K.F., Kotera,M., Sawaki,H., Tsuchiya,S., Fujita,N., Shikanai,T., Kato,M., Kawano,S., Yamada,I. *et al.* (2014) WURCS: the Web3 unique representation of carbohydrate structures. *J. Chem. Inf. Model.*, **54**, 1558–1566.
8. Damerell,D., Ceroni,A., Maass,K., Ranzinger,R., Dell,A. and Haslam,S.M. (2012) The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. *Biol. Chem.*, **393**, 1357–1362.
9. Aoki-Kinoshita,K.F. (2009) *Glycome Informatics: Methods and Applications*. CRC Press, London.
10. Banin,E., Neuberger,Y., Altshuler,Y., Halevi,A., Inbar,O., Nir,D. and Dukler,A. (2002) A Novel Linear Code(r) Nomenclature for Complex Carbohydrates. *Trends Glycosci. Glyc.*, **14**, 127–137.