BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 242–245
doi:10.1093/bioinformatics/btu609

Databases and ontologies

Advance Access publication September 12, 2014

EUROCarbDB(CCRC): a EUROCarbDB node for storing glycomics standard data

Khalifeh Al Jadda[1], Melody P. Porterfield[2], Robert Bridger[2], Christian Heiss[2], Michael Tiemeyer[2], Lance Wells[2], John A. Miller[1], William S. York[2] and Rene Ranzinger[2,*]

[1]Department of Computer Science and [2]Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: In the field of glycomics research, several different techniques are used for structure elucidation. Although multiple techniques are often used to increase confidence in structure assignments, most glycomics databases allow storing of only a single type of experimental data. In addition, the methods used to prepare a sample for analysis is seldom recorded making it harder to reproduce the analytical data and results.

Results: We have extended the freely available EUROCarbDB framework to allow the submission of experimental data and the reporting of several orthogonal experimental datasets. The features aim to increase the understandability and reproducibility of the reported data.

Availability and implementation: The installation with the glycan standards is available at http://glycomics.ccrc.uga.edu/eurocarb/. The source code of the project is available at https://code.google.com/p/ucdb/.

Contact: rene@ccrc.uga.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 19, 2014; revised on August 27, 2014; accepted on September 10, 2014

1 INTRODUCTION

In the past two and a half decades, several databases for storing glycan structures and associated meta information (e.g. biological source, species and publications) have been developed (Campbell *et al.*, 2014). Among them, the CarbBank database was the first large publicly available database of glycan structures (Doubet and Albersheim, 1992; Doubet *et al.*, 1989). Although no experimental data were stored in this database, the experimental techniques used for the identification of glycan structures were recorded. After the funding for CarbBank was discontinued, several independent new databases were created, often by importing most or all of the CarbBank data and sometimes adding experimental data. For example, GLYCOSCIENCES.de (Lütteke *et al.*, 2006) and the Bacterial Carbohydrate Structure Database (Egorova and Toukach, 2014) contain nuclear magnetic resonance (NMR) data that have been extracted from the literature for different glycan structures. The Consortium for Functional Glycomics (CFG) created its own set of databases (Raman *et al.*, 2006) to store the data generated by the consortium, including mass spectrometry (MS) profiling data and glycan array data, revealing the binding of glycans to various biomolecules. Following a similar goal, GlycoBase [developed by the NIBRT group (Campbell *et al.*, 2008)] stores the high-performance liquid chromatography (HPLC) data generated by that group. All of these specialized databases have been implemented for use by a specific research group or consortium and do not allow submission of diverse types of experimental data by outsiders.

In 2005 under the direction of Claus-Wilhelm von der Lieth, the EUROCarbDB (von der Lieth *et al.*, 2011) project was started. The aim of the project was the establishment of a publicly available database framework for the creation of a network of homogeneous databases, allowing research groups worldwide to upload annotated glycan structures and data obtained by MS, NMR and HPLC experiments. The basic idea was to enable each research group to create and populate their own database while facilitating the sharing and exchange of information among databases. The long-term goal was that by providing a free and easy-to-use framework, the heterogeneous landscape of glycan databases developed before 2005 could be systematically replaced by a set of homogeneous databases that contains experimental data with annotated glycan structures. By the end of the project in 2010, the source code of the database prototype that had been installed at the European Bioinformatics Institute (EBI) was released. This code has been subsequently used as the basis for several database projects, including the UniCarb-DB project (Hayes *et al.*, 2011) and the UniCarbKB (Campbell *et al.*, 2011) databases. Although storing experimental data along with the glycan structures, biological annotation, and their literature references was a fundamental goal of the EUROCarbDB project, approaches to storing techniques for sample preparation and methods of experimental analysis had not been developed.

Here we describe the application of different experimental techniques to establish the structures of several glycan standards provided by the CFG (Fig. 1) and the setup of a EUROCarbDB node providing public access to the produced data. We also introduce a novel way to represent the experimental data and meta data within a EuroCarbDB node. In the original implementation of EUROCarbDB, each unique structure was linked with a set of experiments. However, the information required to determine whether the separate experiments in a set were

*To whom correspondence should be addressed.

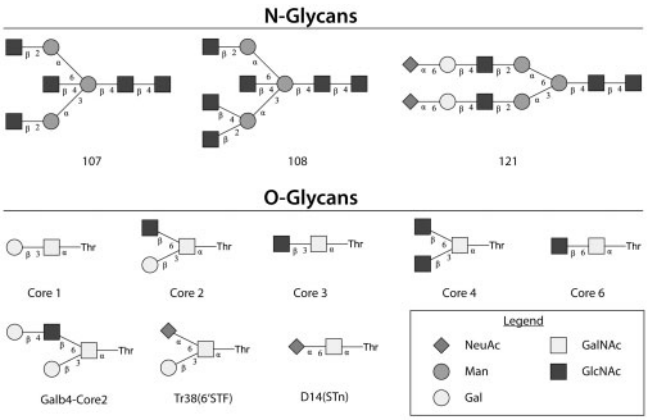

Fig. 1. Three N-glycan and eight O-glycan compounds have been provided by the CFG as analytical standards. Each is displayed using the CFG cartoon representation together with the name used to identify each structure

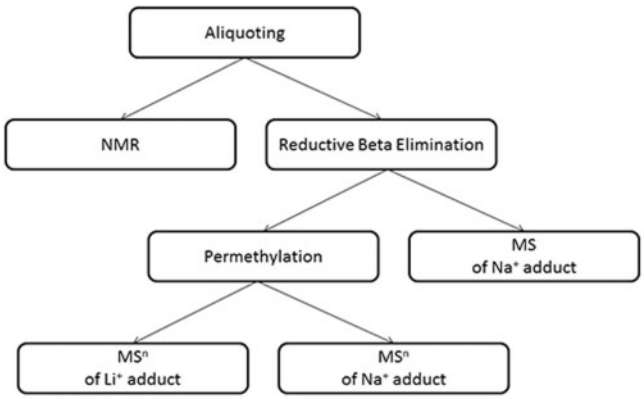

Fig. 2. Design of the O-glycan analysis experiment. Each glycan sample was divided into aliquots and analyzed by NMR without further treatment and tandem MS^n. Additional aliquots were analyzed by MS^n after being subjected to reductive β-elimination to release O-glycans as oligoglycosyl alditols, and again after permethylation and selection of ions formed by complexation with Li^+ and Na^+

performed using exactly the same sample or performed using different samples that happened to contain the same structure was unavailable. This critical information is required to evaluate the experimental data and assess the reliability of the resulting structural annotation. Confidence in a structural annotation is greater when it is based on the application of varied orthogonal methods to the same sample rather than on the analysis of several similar samples, which may contain different contaminants.

2 METHODS

In all, 100 µg of each glycan standard was prepared for MS analysis. O-linked glycan standards were released from their threonine linker by reductive beta elimination, neutralized and passed over a strong cation-exchange resin, as previously described (Orlando *et al.*, 2009). The O-glycans were permethylated (Ciucanu and Kerek, 1984), and dissolved (0.1 µg/µL) in either 1 mM sodium hydroxide or 1 mM lithium hydroxide, in 50% aqueous methanol, before direct infusion into a linear ion trap mass spectrometer equipped with an Orbitrap FT detector (LTQ Orbitrap XL, Thermo-Fisher) by nanospray ionization (NSI) at 0.4 µL/min flow rate. A full mass spectrum was continuously recorded by the instrument for 20 sec, and glycan peaks were then manually selected for fragmentation using 35% collision-induced dissociation (CID). Each resulting MS/MS spectra was also continuously detected and recorded for 20 sec and MS^3 fragmentation was performed when necessary using the same time and fragmentation energy. All scans for each measurement were averaged into a single RAW file and converted to mzXML using MSConvert (Chambers *et al.*, 2012) and imported to GlycoWorkbench to facilitate structural annotation of fragmentation pathways and thereby confirm structural assignments.

In all, 100 µg of each N-glycan standard was prepared for MS analysis as previously described (Orlando *et al.*, 2009) and dissolved (0.1 µg/µL) in either 1 mM sodium hydroxide or 1 mM lithium hydroxide, in 50% aqueous methanol, before direct infusion into a linear ion trap mass spectrometer equipped with an Orbitrap FT detector (LTQ Orbitrap XL, Thermo-Fisher) by NSI at 0.4 µL/min flow rate. A high-resolution full MS scan was recorded in FT mode, and all peaks 3-fold above noise level were fragmented in the ion trap by CID with 35% collision energy. Collected spectra were converted from RAW to mzXML format using MSConvert and imported to GlycoWorkbench (Damerell *et al.*, 2012) to facilitate structural annotation fragmentation pathways and thereby confirm structural assignments.

For NMR, the samples were deuterium exchanged by dissolution in D_2O (99.9% D, Aldrich) and lyophilization. The deuterium-exchanged N- and O-glycans were each dissolved in D_2O (80 µL, 99.96% D, Cambridge Isotope Laboratories), and a small volume of 1% acetone (5 µL for N-glycans and 2 µL for O-glycans) was added as an internal standard before transferring the samples to a 3 mm Shigemi NMR tube. 1-D proton, total correlation spectroscopy (TOCSY), nuclear overhauser effect spectroscopy (NOESY) and rotating frame nuclear overhauser effect spectroscopy (ROESY) NMR spectra were recorded with water presaturation. All NMR data, including gradient enhanced Correlation spectroscopy (COSY) and Heteronuclear single-quantum correlation spectroscopy (HSQC) spectra, were acquired at a sample temperature of 25°C using a Varian Inova-600 MHz spectrometer. The number of transients was 32 for 1-D proton and 2-D TOCSY and NOESY experiments, 16 for gCOSY and 256 for gHSQC. Mixing time was 80 ms for TOCSY, 300 ms for NOESY and 200 ms for ROESY. Chemical shifts were measured relative to internal acetone (δ_H = 2.218 ppm, δ_C = 33.00 ppm) (Wishart *et al.*, 1995).

3 RESULTS

In this paper, we introduce a novel way to represent the experimental data within a EUROCarbDB database. The modified database allows the data obtained by different analytical methods using a single sample to be combined into an ***experiment***, thereby increasing confidence in structural annotations inferred by combining the orthogonal data. These modifications increase the efficacy of EUROCarbDB for scientists using diverse techniques to assign glycan structures.

The design of the ***experiment***s described in the Section 2 is shown in Figures 2 and 3. Each experiment consists of a series of laboratory tasks, indicated by boxes in the diagram. These tasks are defined as ***protocols***, which are recorded in the database by storing the protocol name, a short description and a reference to a Web page in the Glycoscience Protocol Online Database (http://jcggdb.jp/GlycoPOD/) or in our wiki (http://glycomics.ccrc.uga.edu/GlycomicsWiki/Main_Page).

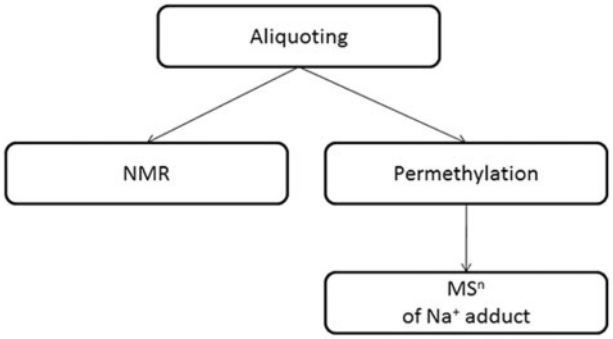

Fig. 3. Design of the N-glycan experiment. Each native glycan sample was divided into two aliquots: one was analyzed by NMR without modification, and the other was permethylated and analyzed as the Na^+ adduct by tandem MS

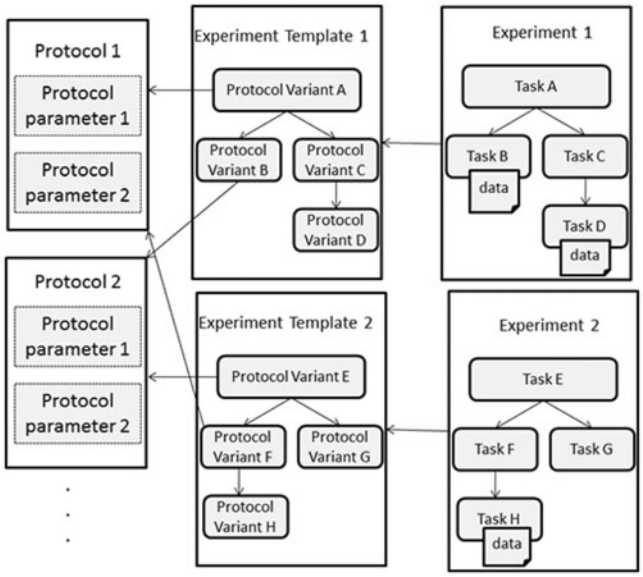

Fig. 4. The design of an **experiment** is a multistep process. It starts by defining **protocols**, which are general but detailed descriptions of laboratory tasks, including lists of any parameters that may vary. **Protocol variants** are then instantiated in the context of an **experiment template**, which specifies a linear or branched sequence of **protocol variants**. In the Figure, arrows leading from **protocol variants** to **protocols** represent instantiation. These steps involve the explicit specification of protocol parameter values. An **experiment template** thus defined can then be instantiated as an actual **experiment** (arrows from **experiment** to **experiment template** represent instantiation). Each **experiment** encapsulates all of the MS, HPLC and/or NMR data acquired during a discrete analysis of a specific sample along with the metadata describing how these data were obtained

These Web pages contain step-by-step descriptions of the experimental procedures, including the necessary materials and instrumentation. In addition to the Web page references for each ***protocol***, a list of ***protocol parameters*** can be stored to capture variations of each standard procedure described in the Web pages. For example, the incubation time, temperature or pH of a chemical reaction may vary from one experiment to another. These specific ***protocol parameter*** values, together with the complete but more general description in the Web pages, fully

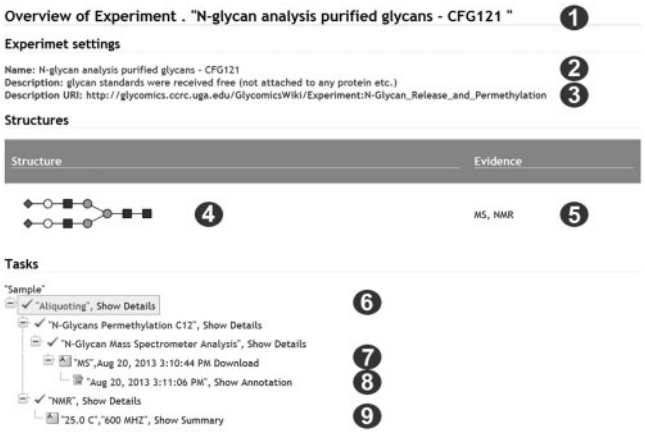

Fig. 5. Screenshot of an experiment overview containing (1) experiment name; (2) description; (3) description URI; (4) glycan structure identified in this experiment; (5) evidence types provided to identify the glycan structure; (6) protocol names; (7) evidence/data files (mzXML); (8) annotation (Glyco Workbench File); (9) NMR protocol/data

describe the experimental methods used to obtain each dataset. The availability of this detailed information increases the analyst's capacity to understand, evaluate and reproduce the experimental results.

To minimize the work necessary to create an experiment and upload experimental data, we created several template mechanisms for the creation of an experiment (Fig. 4). First, the user creates a ***protocol*** by providing its name, description, Web page and a list of parameters with their units of measurement (more details can be found in part I of the Supplementary Material). Specific values for these parameters are assigned later. The example shown as scheme in Figure 3 and as screenshot in Figure 5 'N-Glycan Mass Spectrometer Analysis' includes four ***protocols***.

In the second step, the user creates specific implementations of these ***protocols*** called ***protocol variants***, which encapsulate explicit values for each parameter (for more details see the Supplementary Material part II). In Figure 2, 'Mass spectrometer analysis Na+' and 'Mass spectrometer analysis Li+' are two ***protocol variants*** of the ***protocol*** 'Mass spectrometer analysis' specifying Li+ and Na+, respectively, as values for the protocol parameter *adduct*. The advantage of these templates is that creating a new variation of a protocol does not require the user to reenter all the information. Rather, a ***protocol variant*** is created as a specific instance of the more general ***protocol*** and distinguished by explicit protocol parameters values. The instantiated ***protocol variants*** are placed in a user-defined sequence called an ***experimental template***, which fully describes a specific series of tasks used to prepare and analyze a sample. Each actual use of the ***experiment template*** to analyze a specific sample is documented by instantiating a new ***experiment***. The data produced by an ***experiment*** are thus associated with the metadata specified by the above procedure, facilitating the upload and archival of the annotated data. Each new ***experiment*** is created with a few button clicks, which retrieve and integrate all the relevant information about the experiment and its protocols, including protocol parameter values. Then, it is easy to associate the specific dataset or datasets that were generated

Aliquoting

Description:
Dividing the sample into smaller, more reasonable, quantities.
URI: See wiki page
Parameters

Name	Description	Unit of measurement	Value
Solvent Type	Solvent used to re-suspend sample. Unit of Measure - concentration and name of solvent.	mM	1mM NaOH/50%MeOH
Sample Quantity	Amount of sample in each aliquot. Unit of Measure - ug.	mM	100

Fig. 6. Screenshot of a protocol summary. For each protocol that is part of an experiment a similar summary is available. Each summary includes the name of the protocol (e.g. Aliquoting), its description, a link to the Web page describing the protocol and a list of variable parameters, each documented with a name, description, unit of measurements and value

by the ***experiment*** with a well-defined sequence of parameterized ***protocols***.

The concept of an experiment allows users to upload different kinds of ***evidence*** (experimental data) used to assign a structure to the analyzed glycan. Different types of evidence for glycan samples include quantification data, MS data, HPLC data and NMR data.

Figure 5 shows a screenshot of the Web interface after uploading the experimental data generated by analysis of the N-glycan CFG 121 standard. The experiment itself (diagrammed in Figure 3) consists of ***tasks*** displayed as a hierarchical tree in the lower part of the image. These include the protocols used to prepare and analyze the sample together with the experimental data and the annotation data, which can be retrieved by the user. Expanding this tree representation provides an increasingly detailed view of the results of executing each protocol, ultimately showing the dataset produced by the final analytical protocol (e.g. a file containing mass spectral data) and the annotation file for that data. Figure 6 shows an example protocol with its name, the description, the link to the Web page and the parameters with their values.

4 DISCUSSION

The EUROCarbDB database framework is an extremely useful resource for scientists seeking to create their own database to store analytical Glycomics data. We have extended the original EUROCarbDB code to collect and associate the different types of analytical data obtained by analyzing single samples, adding confidence to the structural annotations deduced from that data. The enhanced code simplifies the generation, storage and retrieval of complete descriptions of the experiment process, facilitating the interpretation and reproduction of the experiment. We have installed a EUROCarbDB node with these enhancements on our server and populated this node with carefully annotated experimental data generated using several standard glycans provided by the CFG. This database and its contents are freely available at http://glycomics.ccrc.uga.edu/eurocarb/. In addition, we encourage users to download this easy-to-use database so they can organize and archive their own glycomics datasets and associate them with metadata describing the procedures they use to generate that data. The EUROCarbDB source code is available at https://code.google.com/p/ucdb/.

The fundamental utility of the tools described here lies in their capacity to dramatically increase the public availability of well-documented glycoanalytic data. Use of these tools will allow researchers around the globe to download and examine experimental datasets and structural annotations created by other scientists, rigorously evaluate those datasets and faithfully reproduce the experimental protocols that were originally used to generate them. To the best of our knowledge, no other publicly available database offers this kind of representation and this level of detail. To make this representation more informative, we also link each protocol to a Web page (GlycoPOD or our wiki) where the creator of the protocol can provide a more detailed description of the experimental methods used. These wiki pages can be accessed via a hyperlink in the details section of the experimental tree (Fig. 6).

ACKNOWLEDGEMENTS

The authors would like to express their deep gratitude to the EUROCarbDB team who created the freely available open-source framework, which they built upon.

Funding: This work was supported by Consortium for Functional Glycomics bridging grant [5U54GM062116-10]; and National Institute of General Medical Sciences, a part of the National Institutes of Health [8P41GM103490].

Conflict of interest: none declared.

REFERENCES

Campbell,M.P. *et al.* (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics*, **24**, 1214–1216.
Campbell,M.P. *et al.* (2011) UniCarbKB: putting the pieces together for glycomics research. *Proteomics*, **11**, 4117–4121.
Campbell,M.P. *et al.* (2014) Toolboxes for a standardised and systematic study of glycans. *BMC Bioinformatics*, **15** (**Suppl. 1**), S9.
Chambers,M.C. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.
Ciucanu,I. and Kerek,F. (1984) A simple and rapid method for the permethylation of carbohydrates. *Carbohydr. Res.*, **131**, 209–217.
Damerell,D. *et al.* (2012) The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. *Biol. Chem.*, **393**, 1357–1362.
Doubet,S. and Albersheim,P. (1992) Letter to the Glyco-Forum CarbBank. *Glycobiology*, **2**, 505–505.
Doubet,S. *et al.* (1989) The complex carbohydrate structure database. *Trends Biochem. Sci.*, **14**, 475–477.
Egorova,K.S. and Toukach,P.V. (2014) Expansion of coverage of Carbohydrate Structure Database (CSDB). *Carbohydr. Res.*, **389**, 112–114.
Hayes,C.A. *et al.* (2011) UniCarb-DB: a database resource for glycomic discovery. *Bioinformatics*, **27**, 1343–1344.
Lütteke,T. *et al.* (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology*, **16**, 71R–81R.
Orlando,R. *et al.* (2009) IDAWG: metabolic incorporation of stable isotope labels for quantitative glycomics of cultured cells. *J. Proteome Res.*, **8**, 3816–3823.
Raman,R. *et al.* (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, **16**, 82R–90R.
von der Lieth,C.W. *et al.* (2011) EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology*, **21**, 493–502.
Wishart,D.S. *et al.* (1995) 1H, 13C and 15N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR*, **6**, 135–140.