Structural Biology

Databases of Conformations and NMR Structures of Glycan Determinants

Anita Sarkar2,4,†, Sophie Drouillard2,†, Alain Rivet2, and Serge Perez1,2,3

2Centre de Recherches sur les Macromolécules Végétales, UPR5301 CNRS-Université Grenoble Alpes, BP53, 38041 Grenoble cédex 09, France, and 3Département de Pharmacochimie Moléculaire, UMR5063 CNRS-Université Grenoble Alpes, BP53, 38041 Grenoble cédex 09, France
1To whom correspondence should be addressed: e-mail: spsergeperez@gmail.com
4Present address: The Scripps Research Institute, CA, USA.
†The authors wish it is to be known that the first two authors should be regarded as joint First Authors.

Received 15 April 2015; Revised 15 July 2015; Accepted 15 July 2015

Abstract

The present study reports a comprehensive nuclear magnetic resonance (NMR) characterization and a systematic conformational sampling of the conformational preferences of 170 glycan moieties of glycosphingolipids as produced in large-scale quantities by bacterial fermentation. These glycans span across a variety of families including the blood group antigens (A, B and O), core structures (Types 1, 2 and 4), fucosylated oligosaccharides (core and lacto-series), sialylated oligosaccharides (Types 1 and 2), Lewis antigens, GPI-anchors and globosides. A complementary set of about 100 glycan determinants occurring in glycoproteins and glycosaminoglycans has also been structurally characterized using molecular mechanics-based computation. The experimental and computational data generated are organized in two relational databases that can be queried by the user through a user-friendly search engine. The NMR (1H and 13C, COSY, TOCSY, HMQC, HMBC correlation) spectra and 3D structures are available for visualization and download in commonly used structure formats. Emphasis has been given to the use of a common nomenclature for the structural encoding of the carbohydrates and each glycan molecule is described by four different types of representations in order to cope with the different usages in chemistry and biology. These web-based databases were developed with non-proprietary software and are open access for the scientific community available at http://glyco3d.cermav.cnrs.fr.

Key words: disaccharides, molecular mechanics, NMR spectroscopy, oligosaccharides

Introduction

Carbohydrates are involved in a variety of biological functions ranging from the trivial to the crucial. They play important roles in the growth, function, development and even survival of an organism. These ubiquitous and complex molecules offer a tremendous diversity arising not only from differences in glycan composition and branching, but also from substituted components and their linkage to the aglycones. This makes glycans “bio-informational” molecules which are recognized by a variety of proteins, such as lectins, antibodies, receptors, toxins, microbial adhesins, enzymes, viruses, transporters, collectively known as glycan binding proteins. According to Cummings (2009), the number of glycan determinants likely to be important in their interactions with glycan binding proteins is estimated to be about 7000. The Consortium for Functional Glycomics (2014) (CFG) lists 7500 entries in its glycan database. Thus, considering 3000 partial glycan determinants in N- and O-glycans and glycolipids, along with the 4000 possible GAG pentasaccharides, there are, in totality, 7000 possible glycan determinants in the human glycome. This number is probably an underestimation and it may be recognized that sequencing the human glycome at present may still be unrealistic,
given the current technology for glycoproteomics, glycosaminoglycanomics or glyco-lipidomics. While the knowledge of the number of glycan determinants in the human glycome is an important issue per se, there remains the continuing challenge to obtain working quantities of well-defined glycan determinants for: (i) conjugates in vaccine development, anti-infective agents, antineurodegenerative compounds and in diagnosis, (ii) probes to be fully characterized by biophysical studies, (iii) probes to explore the molecular basis of their recognition by proteins using macromolecular crystallography or high-resolution nuclear magnetic resonance (NMR) spectroscopy.

Recent developments in bacterial fermentation of metabolically engineered strains have opened the road to large-scale productions of bioactive glycan determinants (Priem et al. 2002). Approximately, 200 bioactive oligosaccharides are involved in key biological pathways (such as cell adhesion, immunological recognition, embryogenesis, oncogenesis, infection). They belong to classes such as human blood groups (ABH, Lewis, P) or are moieties of the major glycosphingolipids (globosides, gangliosides) and can be produced in large-scale quantities. Prior to using them for further research and applications, it was necessary to structurally characterize them. This was carried out by the complementary techniques of NMR spectroscopy, including 1H and 13C spectra (in certain cases COSY, TOCSY, HMQC, HMBC correlation spectra) and molecular modeling techniques. The low-energy conformations generated by molecular modeling supplement the deficit of structural information from X-ray crystallography on glycans or protein–glycan complexes (Imbert and Perez 2000; Perez 2007) while facilitating the investigations of protein–glycan interactions by computational methods (Perez and Tvaroska 2014). The thrust of the glycoinformatics endeavor thus far has mainly been glycan sequence-related (Perez and Mulloy 2003; Demarco and Woods 2008). Web-based tools have been developed to build preliminary 3D structures starting from a sequence as implemented in SWEET-II (Bohne et al. 1999), Glycam (2015) (Woods 2014), POLYS (Engelsen et al. 2013) carbohydrate builders. Development and improvement of high-throughput techniques and computational power have made high-throughput molecular modeling experiments possible wherein complex glycan structures can be investigated during the course of time-limited investigations (Rosen et al. 2009).

The purpose of the present work is to provide to the scientific community two interlinked open-access, web-based databases that store structural information of glycan determinants, combining results derived from experimental and computational investigations. These data sets can be used as reliable starting models for interaction studies and structural investigations involving glycans. These databases can be expanded as new glycan structures are determined and linked to other databases.

Scope of the work
A comprehensive list of 170 prominent glycan determinants involved in recognition events was established taking into account the availability of these molecules in ample quantities. Following their experimental characterizations, the 3D features of these glycans were established by molecular modeling procedures and NMR spectroscopy. A second set of 120 more glycans was established which were solely investigated by computational methods. Table I provides a classification of glycan determinants covered in the present investigation.

The management of these data (i.e. NMR and 3D coordinates of low energy conformations), in relation to glycan sequence and structure, requires the construction of large-scale repositories for storage, organization and dissemination. Additionally, various algorithms and tools have to be developed that can query (search) these repositories, interlink them and, in context, be useful in calculations and analyses of the existing data.

### Table I. Classification of glycan determinants in BioOligo

<table>
<thead>
<tr>
<th>Index</th>
<th>BioOligo category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Blood group A antigens</td>
</tr>
<tr>
<td>2</td>
<td>Blood group B antigens</td>
</tr>
<tr>
<td>3</td>
<td>Blood group H antigens (blood group O)</td>
</tr>
<tr>
<td>4</td>
<td>Blood group H antigens (blood group O) and Globo H tetrasaccharide</td>
</tr>
<tr>
<td>5</td>
<td>Core structures</td>
</tr>
<tr>
<td>6</td>
<td>Core structures (Type 1 and Type 2)</td>
</tr>
<tr>
<td>7</td>
<td>Core structures (Type 1)</td>
</tr>
<tr>
<td>8</td>
<td>Core structures (Type 2)</td>
</tr>
<tr>
<td>9</td>
<td>Core structures (Type 4)</td>
</tr>
<tr>
<td>10</td>
<td>Fucosylated oligosaccharides</td>
</tr>
<tr>
<td>11</td>
<td>Fucosylated oligosaccharides (3 Fucosyllactose core)</td>
</tr>
<tr>
<td>12</td>
<td>Fucosylated oligosaccharides (Lacto-Series)</td>
</tr>
<tr>
<td>13</td>
<td>GAAs</td>
</tr>
<tr>
<td>14</td>
<td>GaIe-3Gal oligosaccharides (Galili and xeno antigens)</td>
</tr>
<tr>
<td>15</td>
<td>GaIe-3Gal oligosaccharides (Isogloboseries)</td>
</tr>
<tr>
<td>16</td>
<td>Ganglioside sugars</td>
</tr>
<tr>
<td>17</td>
<td>Globoside sugars (P antigens) (Forsman antigens)</td>
</tr>
<tr>
<td>18</td>
<td>Globoside sugars (P antigens) (Globo series—core structure type 4)</td>
</tr>
<tr>
<td>19</td>
<td>Globoside sugars (P antigens) (P blood group antigens and analogues)</td>
</tr>
<tr>
<td>20</td>
<td>Globoside sugars (P antigens) (stage-specific embryonic antigens: SSEA-3 and SSE-A)</td>
</tr>
<tr>
<td>21</td>
<td>Glucosylated oligosaccharides</td>
</tr>
<tr>
<td>22</td>
<td>Glycosphingolipid</td>
</tr>
<tr>
<td>23</td>
<td>Lewis antigens</td>
</tr>
<tr>
<td>24</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>25</td>
<td>Miscellaneous (Blood group-related oligosaccharides)</td>
</tr>
<tr>
<td>26</td>
<td>Miscellaneous (Chitin oligosaccharides)</td>
</tr>
<tr>
<td>27</td>
<td>Miscellaneous (Fibrinogen-related oligosaccharides)</td>
</tr>
<tr>
<td>28</td>
<td>Miscellaneous (LDN-related oligosaccharides)</td>
</tr>
<tr>
<td>29</td>
<td>Miscellaneous (Lewis X-related oligosaccharides)</td>
</tr>
<tr>
<td>30</td>
<td>Miscellaneous (TF-related oligosaccharides)</td>
</tr>
<tr>
<td>31</td>
<td>Miscellaneous (TN-related oligosaccharides)</td>
</tr>
<tr>
<td>32</td>
<td>Miscellaneous (Trehalose-like sugars)</td>
</tr>
<tr>
<td>33</td>
<td>N-linked oligos</td>
</tr>
<tr>
<td>34</td>
<td>Sialylated oligosaccharide (Type 1)</td>
</tr>
<tr>
<td>35</td>
<td>Sialylated oligosaccharide (Type 2)</td>
</tr>
</tbody>
</table>

*The constructions of the databases are such that novel categories can be implemented for further extensions.*

**Representing and encoding complex carbohydrates**

Representation in text of the primary structure, or sequence, of complex carbohydrates was first described following the IUPAC-IUBMB terminology in its extended and condensed forms (McNaught 1997a, b). Other types of representations have been developed in glycobiology, favoring pictorial representations that facilitate the visualization of any of the constituting units of any complex carbohydrate molecule. This is adequate, given that the number of basic carbohydrate units found in mammals is limited. Extensions to the constituents found in bacterial and plant polysaccharides have also been developed.

The variety in nomenclature and structural representation of glycans makes it complex to decide the best form to use. The choice of notation is frequently based on whether the study is focused on chemistry or biology. The information content of each representation may vary or highlight a particular aspect when compared with others.
While representing a complex glycan, chemists prefer to elucidate the structure that includes information about the anomeric carbon, the chirality of the glycan, the monosaccharides present and the glycosidic linkages that connect them. For others, it is more interesting to visualize the monosaccharides present and hence a symbolic/diagrammatic notation is favored. In the present investigation, all these different types of lexical and graphical representations were used to describe the glycans under investigation. To establish databases, it is nevertheless essential to decide on a simple representation with a common/standard format that can be computationally manipulated. This would facilitate computational processing, while ensuring that the data content is nonredundant and facilitate intercommunication with distinct databases. Chemical file formats like InChI (Mcnaught 1997a, b) and SMILES (Weininger 1988) have been developed to aid storing molecule information in chemical databases like PubChem (Wang et al. 2010) or ChEBI (Degtyarenko et al. 2008). IUPAC (extended), InChI and SMILES encoding are computed from the chemical drawing (ring structure) thereby allowing auto-generation of these encodings. There are severe limitations as none of these existing structural encoding schemata are capable of coping with the full complexity to be expected for experimentally derived structural carbohydrate sequence data across all taxonomic sources. Instead, the GlycoCT encoding scheme was selected in its capacity of coping with the heterogeneous landscape of digital encoding in glycoscience (Herget et al. 2008). Finally, the 3D depiction of glycans, polysaccharides and glycoconjugates as per the accepted nomenclature and pictorial representation used in carbohydrate chemistry, biochemistry and glyco-biology has been made using the molecular visualization program Sweet Unity Mol (Perez, Tubiana et al. 2015). Figure 1 summarizes the different types of representations used to encode glycan structures.

**Generating the primary data**

**Glycan sample preparation**

Glycosylation is carried out in bacterial cells, which offer high yields (several grams per liter of culture) of complex glycoforms at low production costs (Priem et al. 2002). The conditions for using bacterial transference gene in *Escherichia coli*, for producing such quantities of well-defined glycans have been described (Samain 2007). Large-scale synthesis of a series of glycolipids and glycoproteins epitopes was achieved using previously described methods (Dumon et al. 2001; 2006; Antoine et al. 2003; Antoine, Bosso, et al. 2005; Antoine, Heyraud, et al. 2005; Bettler et al. 2003; Randriantsoa et al. 2007; Pierfort and Samain 2008; Yavuz et al. 2008; Drouillard et al. 2010; Ilg et al. 2010; Bastide et al. 2011; Yavuz et al. 2011; Barreteau et al. 2012; Gebus et al. 2012; OligoTech, Elictyl, 2012).

**NMR experiments**

Each glycan, available in sufficient quantity, was submitted to a complete structural characterization by NMR spectroscopy. The $^1H$ and $^{13}C$-NMR assignments were based on homonuclear $^1H–^1H$ and

---

**Sialyl Lewis X (sLeX) Pentaose**

**A**  
$\alpha$-D-Neu5Ac (2→3) $\beta$-D-Galp (1→4)-[ $\alpha$-L-Fucp (1→3)]$\beta$-D-GlcNAc (1→3)$\beta$- D-Galp

**B**  
Neu5Ac a2–3 Gal b1–4 (Fuc a1–3) GlcNAc- b1–3 Gal

**C**

![Chemical structure of Sialyl Lewis X (sLeX) Pentaose](image)

**D**

![“cartoon-type” representation](image)

**E**

![3D representation using the ring blending facility of SweetUnityMol](image)

**F**

| RES | 1b: x-dgal-gal-HEX-1:5 | 2b: b-dglc-glcHEX-1:5 | 3b: n-acetyl | 4b: a-dgal-gal-HEX-5:6 | 5b: b-dgal-gal-HEX-1:5 | 6b: a-dgro-dgal-NON-2:6 | 1a: 2: keto | 3d: s: n-acetyl | LIN | 1:1o(3+1)2d | 2:2d(2+1)3n | 3:2o(3+1)4d | 4:2o(3+1)5d | 5:5o(3+2)6d | 6:6d(5+1)7n |
|-----|------------------------|------------------------|-------------|------------------------|------------------------|------------------------|-------------|------------------------|-----|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|

---

*Fig. 1. Nomenclature and structural representations commonly used for complex glycans, as exemplified by Sialyl Lewis X (sLeX) pentaose. (A) Condensed IUPAC Nomenclature; (B) Neu5Ac α2–3 Gal β1–4 (Fuc α1–3) GlcNAc β1–3 Gal; (C) chemical representation (drawn using ChemDraw®); (D) “cartoon-type” representation (Essential of Glycobiology (Varki et al. 2009); (E) 3D representation using the ring blending facility of SweetUnityMol (Perez, Tubiana et al. 2015)). This figure is available in black and white in print and in color at Glycobiology online.*
heteronuclear $^1$H-$^{13}$C correlation experiments (correlation spectroscopy COSY, heteronuclear multiple bond coherence HMBC, heteronuclear multiple quantum coherence HMQC). The assignments were conducted following well-established protocols as described by Kover et al. (2010).

3D structures—molecular mechanics
The common computational approach to 3D structure prediction is based on searching through the conformational space of the glycan in order to find low-energy regions, i.e. conformers, which the molecule is likely to populate. This can be accomplished using multiple ways by a variety of different algorithms (Fadda and Woods 2010).

Despite their inherent structural intricacies, complex carbohydrates are particularly suited for computational conformational predictions (Perez 2007; Frank and Schloissnig 2010). Speed and computational time are critical and decisive factors in high-throughput conformational analysis of complex glycans. Keeping this delicate balance in consideration, a software called Shape was developed for automatic conformation prediction of carbohydrates using a genetic algorithm (Rosen et al. 2009). Its robustness and accuracy have been tested on a series of studies on previously published conformation predictions of oligosaccharides performed using other conformational search tools. In these cases, all major local minima could be recovered with a major improvement in computational time.

Fig. 2. The distinct conformations after a complete conformational sampling of the lacto-N-fucopentaose V structure [Gal $\beta$1–4 GlcNAc $\beta$1–3 Gal $\beta$1–4 (Fuc $\alpha$1–3) Glc] using Shape (Rosen et al. 2009). Chemical representation (drawn using ChemDraw®), "cartoon-type" representation (Essential of Glycobiology (Varki et al. 2009)); 3D representation using the ring blending facility of SweetUnityMol (Perez, Tubiana et al. 2015). This figure is available in black and white in print and in color at Glycobiology online.
For establishing the databases, the starting 3D structures of each constituent glycan were generated using a combination of the available carbohydrate molecular builders (Lütteke et al. 2006; Engelsen et al. 2013; Woods 2014; Sybyl (Tripos Inc.) and Chimera (Pettersen et al. 2004), followed by systematic conformational sampling to determine their conformational preferences, using Shape. In such cases, several low-energy conformations (between 1 and 5) were shortlisted and made available for each entry. As a typical example, the results of the exploration of the potential energy hypersurface of lacto N-fucopentaose are shown in Figure 2.

Classification of glycans
The classifications of the glycan determinants present in the two databases follow the main classes of glycolipids, glycoproteins and glycosaminoglycans. Most of the oligosaccharides, which have been investigated by NMR spectroscopy, are moieties of complex glycosphingolipids. As such, key representatives of the different series: isoglob, glob, neo-lacto, lacto, ganglio, besides others, have been characterized, both by NMR spectroscopy and molecular modeling. Other molecules such as N- and O-linked glycans, or small fractions of glycosaminoglycans, were not investigated by NMR, but the 3D structures in their low-energy conformations were assessed. The complete list of glycans presents in the database, along with the sequence and the commonly used names is given in the Annex, as Supplementary data.

Database construction
The databases run on an Apache web server with the application program Hypertext Preprocessor (PHP 2013) (http://www.php.net/). It has been implemented using the open source MySQL database (http://www.mysql.com/) (Vaswani 2005). They are based on a combination of three layers. The underlying layer is the MySQL database system, a relational database management system that stores all the structure-related information in the back-end and provides the facility to link two or more tables in the database. An intermediate layer is an Apache-PHP application [Apache 2, PHP] that receives the query from the user and connects to the database to fetch data from the upper layer, which comprises populated HTML pages, to the web browser client. The PHP and Java scripts are embedded in the HTML web pages for this effect and are used as application programs for integrating the back-end (MySQL database) to the web pages (HTML). Apache has been used as the web server for building the interface between the web browser and the application programs. PHP was used for writing scripts to query the database, and the Java Script (with JQuery plugin) was used to design the auto-complete function for the user interface. The graphical user interface was developed with HTML (version 5) and CSS (version 3).

Database query and results
Search and GUI features
The search engines of the two databases follow a similar logic, of simple and advanced searches, accessible via the search pages, as shown in Figure 3.

Simple search
The user types in the text box provided, based upon which a result prompt appears to guide the user in selecting from the “hits” found.
in the database. An accordion function was developed to display a preview of the results. This can be used to expand or minimize the preview of the listed results of the user query for a first glance into the entries matching the request to the database. The preview provides the glycan name, sequence, category and molecular weight to the user to make an informed choice.

**Advanced search**

This is a multicriteria search that can be used together for querying or in various combinations as best suits the user’s requirement. Four search modes are provided, namely, trivial name, type of constituent, category and molecular weight. A slider is provided for assigning a range of values to be queried in the molecular weight of the database entries. It consists of two cursors that can navigate on a bar for specifying the minimum and maximum limit of the search. Two text fields display the values of the cursor position on the slider bar. The slider cursors auto-adjust themselves when values are entered directly in the text boxes. This feature was developed by modifying a JQuery plug-in.

Both the simple and advanced search options are equipped with an “auto-complete” function, as shown in Figure 3. This is one of the prime features of result refinement provided in the GUI. It guides the user while querying the database. It comprises two parts (a) single field of entered text, and (b) the auto-prompt when the data are entered, through which the desired hit in the database can be selected either by scrolling down with the mouse or using the arrow keys on the keyboard.

**Results: BiOligo database**

The detailed results are organized under two tabs as shown in Figure 4, namely Molecule information and View and Download.

**Molecule information**

This includes the trivial name of the glycan, its sequence, the chemical (ring) and CFG (consortium of functional glycomics) cartoon representation, molecular weight, the glycan category or family into which it has been classified in BiOligo, glycan composition, i.e. the comprising glycan type and number of each such glycan in the BiOligo entry, glycosidic linkages present in it and occasionally additional comments. Each entry is associated with a reference that identifies it as a glycan determinant, and from which it has been sourced into BiOligo. The illustrative representations of the glycans can be viewed through the “Zoombox” feature that was developed by modifying an existing JQuery plug-in that allows the selected image to be zoomed and highlighted.
View and download
This tab incorporates the best representatives of the families of the most-probable low-energy conformational families from the results that have passed the filtering step. The molecules are displayed using Jmol (Hanson 2010) applet windows that also enable basic viewing and measurements under the right-click options. For each of the conformations, the atomic coordinates (at the PDB format) can be downloaded from this section for further use.

Results: glyco-NMR database
The detailed results are organized under two tabs as shown in Figure 5, namely molecule information and display and download.

Molecule information
Molecule information provides the trivial name, the sequence, the graphical representation of the stereo-chemical configuration, the symbol notation for carbohydrates of the CFG, the type of constituent and the glycan category. The experimental conditions used to record the NMR spectra are provided, i.e. temperature, solvent, frequency and concentration.

Display and download
This incorporates the representations of the chemical repeat unit, and in many cases, the $^1$H and $^{13}$C spectra, along with COSY, TOCSY, HMQC, HMBC correlations spectra (for about 50% of the glycans).

Conclusions
The present study reports a comprehensive NMR characterization of 170 glycan moieties of glycosphingolipids produced in large-scale quantities by bacterial fermentation. Glycosphingolipids are important components of the outer leaflet of the limiting plasma membrane, where the glycan moieties face the external milieu and modulate the signaling activity at the level of cell–cell interaction and modulate activities of proteins in the same plasma membrane. They constitute a large and diverse family of glycans having important roles in physiology and pathology. It is therefore important to establish their 3D structures as well as their conformational flexibility. Equally important is the characterization of their lateral associations leading to their clustering in the form of “lipid rafts”. To this end, glycan moieties of the glycosphingolipids were submitted to the application of high-throughput molecular modeling making use of an accurate...
molecular mechanics force field coupled to genetic algorithm. The ensemble of 3D data provides the basis for constructing complete structures of glycosphingolipids embedded in the plasma membrane. Incidentally, the computational characterizations were extended to a series of about 100 other biologically important glycan determinants as found in glycoproteins (N- and O-linked glycans) and glycosaminoglycans. As a result, the 3D structures of about 300 glycan determinants have been established.

The experimental and computational data generated are stored in two relational databases, which are open access and can be queried by the user through the web-interfaced search engine. It categorizes the structural information into logical sections for the user to access using precustomized searching techniques. The databases shall be maintained and regularly updated to complete the conformational explorations of some of the missing high molecular weight glycans.

Particular emphasis was given to the use of a common nomenclature for the structural encoding of the carbohydrates. Each glycan has been described using four different types of representations in order to cope with the different usages in chemistry and biology, encompassing all aspects of information contained in various representations, suited to the user’s field of research. These are important features in the context of a general insertion of these databases within a portal called Glyco3D (2015) (glyco3d.cermav.cnrs.fr), which includes other structural databases of carbohydrates [monosaccharides, disaccharides, polysaccharides (Sarkar and Perez 2012)] and proteins interacting with glycans [glycosyl-transferases, monoclonal antibodies, glycosaminoglycan-binding proteins and lectins (Perez et al. 2014; Perez, Sarkar et al. 2015)]. Each of these databases has been set up to account for the specific features of the class of molecules covered and a logical network has been established that links all these databases together. A search engine has been developed that scans the full content of all the databases for queries related to sequential information of the glycans or other related descriptors.

Future directions of this work would be to extend the network of information available by connecting to other related databases present freely over the web for the academic community to enhance and assimilate functional aspects of complex carbohydrates that are involved in intra-, inter- and extra-cellular interactions.

Methods

Nuclear magnetic resonance

$^{13}$C and $^1$H NMR spectra were recorded with a BRUKER Avance 400 spectrometer operating at a frequency of 100.618 MHz for the $^{13}$C and 400.13 MHz for $^1$H. All the experiments were performed at a pH of 6.8 which corresponds to the value used for the microbial production of glycans and monitored during the whole process of fermentation. This pH value is the one maintained during the purification steps of the glycans, which are further submitted to a lyophilization phase and re-dissolved in D2O for the purpose of NMR characterization.

The Avance III 400 MHz Bruker NMR offers an automatic locking procedure in which the solvent used is selected. For each glycan, all the $^1$H experiments were performed at a fixed temperature (which is reported in the database). Depending upon the glycan under investigation, the temperatures lie between 293 and 353 K. The magnitude of the chemical shift of the HOD residual signal is a function of the temperature. Starting from 4.85 ppm at 293 K, there is a regular decrease of 0.1 ppm every 10 K. As a result, the range covered is from 4.85 ppm at 293 K to 4.25 ppm at 353 K. For the $^{13}$C, chemical shifts are expressed in ppm in reference to an external standard (Tetramethylsilane).

$^{13}$C spectra were recorded using 90° pulses, 20,000 Hz spectral width, 65,536 data points, 1.638 s acquisition time, 1 s relaxation delay and between 8192 and 16,834 scans. Proton spectra were recorded with 4006 Hz spectral width, 32,768 data points, 4.089 s acquisition times, 0.1 s relaxation delays and 16 scans. The $^1$H and $^{13}$C-NMR assignments were based on homonuclear $^1$H–$^1$H and heteronuclear $^1$H–$^{13}$C correlation experiments (correlation spectroscopy COSY, heteronuclear multiple bond coherence HMBC, heteronuclear multiple quantum coherence HMQC). They were performed with 4006 Hz spectral width, 2048 data points, 0.255 s acquisition time, 1 s relaxation delay; from 32 up to 512 scans were accumulated.

Fig. 5. An illustration of the results in NMR for Galili antigen pentaose (Gal α1–3 Gal β1–4 GlcNAc β1–3 Gal β1–4 Glc) (A) showing molecule information (B) and display and download. This figure is available in black and white in print and in color at Glycobiology online.
The chemical shifts in ppm are reported in a table with the spectrum picture of the molecule.

Computational methods
All bioactive oligosaccharides contained in the database have been sequentially built using the same protocol. First, the SWEET-II web-based tool on the Glycosciences.de web portal (http://www.dkfz.de/spec/sweet/doc/index.php) was used to generate a 3D model from the oligosaccharide sequence (Lütteke et al. 2006). The resulting 3D model was further optimized using MM3 force field (Allinger et al. 1990) as implemented in the TINKER package (Pappu et al. 1998) and then saved in the Protein Data Bank (PDB) format. Subsequently, the carbohydrate atom and bond typing were manually checked and corrected within the SYBYL X1.3 interface (TRIPOS 2001). The Shape software (Rosen et al. 2009) has been used to perform the high-throughput computational exploration of many di- and all oligosaccharides entries, whose conformations have been reported in the present investigation. Shape uses a genetic algorithm (GA) for searching the conformational space of the glycans. The MM3 force field is used for energy evaluations, which have been performed using a value of 4.0 for the dielectric constant for all calculations. The block diagonal minimization method for geometry optimization was used in MM3 with the default energy-convergence criterion \( E = 0.00008n \) kcal/mol every 5 iterations, where \( n \) is the number of atoms. MM3 allows full relaxation of the glycosidic residues taking into account the exo-anomeric effect (Li and Allinger 1991; Perez et al. 1998) and this force field allows optimization to a nearby transition state (with the full matrix Newton–Raphson method). The genetic algorithm implementation in Shape is a generational parallel population Lamarckian method that follows molecular evolution. The genetic operators
in action are mutation, migration and crossover. A population size of 25 individuals was specified for inclusion in every population of conformations throughout the search, while the total number of parallel populations to be used during the search was set to 20. Each generation produced by the genetic algorithm comprised: Total number of individuals × population size × total number of populations. The energy convergence criterion for the conformers generated was assigned a window size of 20 to search for improvements (i.e. the search was terminated when even after 20 generations, no significant improvement in conformational energy was found). The highest energy difference is the entire window that is accepted as a significant improvement for the search to continue (i.e. the limit) is of −0.5 kcal/mol. This is directly related to the maximum efficiency of the evolution of conformers, since this is the absolute minimum limit to the length of the conformation search. For each run, once the “best” conformer has been found, the search still continues for a number of generations, to the specified window size, till satisfied that the results have converged. To analyze the large amount of conformations generated by the GA, the results were clustered into distinct families of low-energy conformations. The conformations were clustered using atom distances, ignoring hydrogen atoms and a 1 Å tolerance for RMSD from the cluster centroid. After the families of low-energy conformations are clustered, a further filtering is applied based upon possible low-energy regions that could be populated by the conformations of the molecules being investigated. Out of the cluster centroids reported after Shape clustering, the ones that inhabit the low-energy regions are selected and stored as the final results of the conformational sampling in BiOligo. The results were checked with previously reported analysis of conformational clustering using the CICADA heuristic method (Koca et al. 1995). The 3D structures deposited can be viewed on the interface via the Jmol application (Herraez 2006). The provision to download the atomic coordinates for further independent use is provided in the PDB format. All the calculations have been performed using the facilities of the Centre d’Expérimentation et de Calcul Intensif en Chimie (CECIC) on a cluster of computers made up of a 18-node Dell Power Edge C6100 (24 and 48 GB of central memory), and 20-node Bull R424E3 (32 GB of central memory) linked by an Infiniband interconnection network, making a total of 536 cores and with access to a disk storage system offering a global capacity of 1.0 TB. This facility is part of the Grenoble University High Performance Computing Center: CIMENT.

**Supplementary data**

Supplementary data for this article are available online at http://glycob.oxfordjournals.org/.

**Funding**

The research leading to this publication has received funding from the European Commission’s Seventh Framework Programme FP7/2007-2013 under grant agreement n° 215536 and from the Agence Nationale de la Recherche under the “Genomic and Plant Biotechnology” Action throughout the “Wall-Array” project.

**Acknowledgements**

We are grateful to the Marie Curie Initial Training Network as part of the FP7 People Programme for training and funding. Most of the computations presented in this article were performed on the Centre d’Expérimentation et de Calcul Intensif en Chimie (CECIC), part of the CIMENT infrastructure (https://ciment.ujf-grenoble.fr) which benefits from support by Rhône-Alpes Region.

**References**


