Glyco-Bioinformatics today (August 2011) – Solutions and Problems

René Ranzinger* and William S. York

Complex Carbohydrate Research Center, The University of Georgia,
315 Riverbend Road, Athens GA, 30602, U.S.A.

E-Mail: *rene@ccrc.uga.edu

Received: 24th April 2012/Published: 11th July 2012

Abstract

Glyco-bioinformatics is an emerging sub-field of bioinformatics, aiming to develop tools, databases, web services and workflows to facilitate research in the field of glycomics. Although glyco-bioinformatics is still in its infancy a large set of web applications, stand-alone applications and databases have been developed recently. Most of the programs are available via the Internet and can be used freely by glycoscientists. In the first half of this chapter we give a non-comprehensive overview of the tools that have been developed for the different subfields of glyco-bioinformatics. In the second section we discuss fundamental problems that hinder rapid progress in the field and identify milestones to be achieved in order to overcome these problems.

Introduction

Bioinformatics is a multidisciplinary enterprise that spans the interface of biology and computer science. The power of genomics and proteomics to provide insight into the molecular bases of life stems directly from the application of bioinformatics tools, including algorithms, databases and programs that have been developed over the last decades. Glyco-bioinformatics is a sub-field of bioinformatics, aiming to develop similar tools to facilitate research in the field of glycomics. Glycomics has been overshadowed by its more famous siblings proteomics and genomics, emerging only recently to reveal the structural basis for many of the critical roles that complex carbohydrates play in biological development, cell function and the progression of disease. In contrast to the well accepted, traditional bioinformatics that is widely used in proteomics and genomics, glyco-bioinformatics is still in its infancy. One of the main obstacles to the development of effective glyco-bioinformatics is the fact that the molecular structures of carbohydrates are more complex than those of nucleic acids and proteins. Two amino-acids or two base-pairs are almost always connected by a single type of linkage, while a pair of monosaccharides can be linked by several different linkages. For example, two hexose residues can be linked by 1 – 2, 1 – 3, 1 – 4 or 1 – 6 bonds. Thus, a single monosaccharide can have several other monosaccharides attached to it such that carbohydrates, in contrast to polynucleic acids and proteins, are branched structures rather than linear chains. Furthermore, the number of distinct monosaccharide residues is much larger than the number of amino-acids and base pairs. Although some have argued that that less than 20 monosaccharides (D-Glc, D-GlcNAc, D-Man, L-IdoA, D-GlcA, D-Gal, D-GalNAc, D-Sia, L-Fuc, D-Xyl, each with two different anomeric states [1]), are required to generate the mammalian glycome, the number of natural monosaccharides increases drastically if we include the glycomes of other taxonomic groups such as plants, fungi and bacteria. This complexity and diversity are the main reasons that algorithms developed for genomics and proteomics cannot simply be reused for glycomics. These issues also give rise to problems with the analysis and interpretation of glycomics data, which are reflected in the slow rate at which such tools have been developed.

Although glyco-bioinformatics has not been as widely received as bioinformatics in other fields, several databases and tools for interpretation and analysis of glycomics data have been developed over the last decades. The first half of this chapter provides a non-comprehensive overview of the available tools and databases, which are illustrated by way of example. The second half of the chapter outlines obstacles and problems in the field of glyco-bioinformatics and suggests milestones that should be achieved in order to make the tools developed in this field more generally accessible and applicable to the diverse biological research that will be performed in the coming years.

Glyco-Bioinformatics Today

Although glyco-bioinformatics is still in its infancy, a sizeable collection of applications, databases and tools are available in this area. The next section of this chapter provides a non comprehensive overview of the existing tools.

Sequence representations and sequence formats

In contrast to genes and proteins, carbohydrates are often branched molecules that cannot easily be abstractly represented and displayed as simple character sequences. Several different representational schemes have been developed for the display of glycans in publications and in databases.

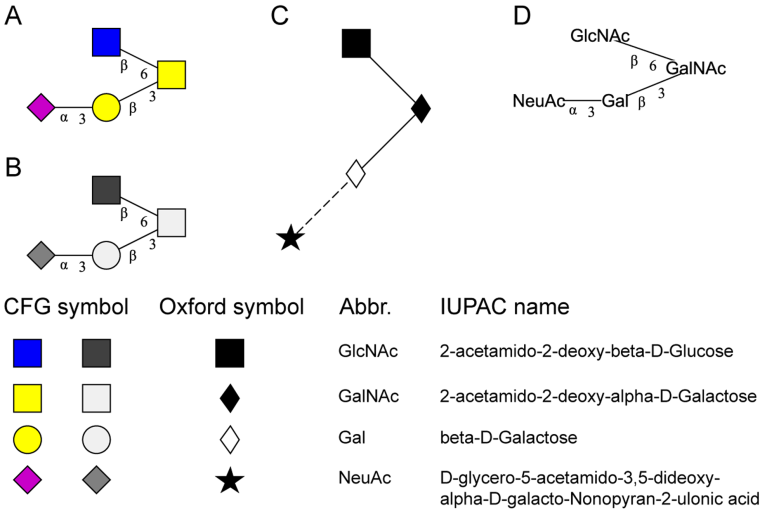

A
β
6
3
α
3
β
B
β
6
3
α
3
β
C
D
GlcNAc
β
6
GalNAc
NeuAc
α
3
Gal
β
3
CFG symbol
Oxford symbol
Abbr.
IUPAC name
GlcNAc
2-acetamido-2-deoxy-beta-D-Glucose
GalNAc
2-acetamido-2-deoxy-alpha-D-Galactose
Gal
beta-D-Galactose
NeuAc
D-glycero-5-acetamido-3,5-dideoxy-alpha-D-galacto-Nonopyran-2-ulonic acid

Figure 1. O-glycan specified by GlycomeDB ID 14282 illustrated using several different representation schemes. **(A)** CFG cartoon representation – color, **(B)** CFG carton representation – gray scale, **(C)** Oxford representation, and **(D)** textual representation using IUPAC names. The legend below the structures shows the different monosaccharide symbols and names used for each scheme together with the full IUPAC name for each monosaccharide.

Figure 1 shows an overview over the most commonly used glycan representation schemes. One of the earliest ways to represent carbohydrates is as an image that specifies each monosaccharide using its IUPAC short name (see Figure 1D) [2]. But these images have proven to be inconvenient when representing a large set of carbohydrates such as those produced in modern high throughput experiments. Therefore, representation schemes that replace the monosaccharide names with colored geometric figures (so called cartoon representations) have been developed. The CFG scheme (Figure 1A and 1B) was developed by the Consortium for Functional Glycomics for use by their web database and can be found in many publications [3]. The representation scheme developed by the Oxford GlycoBiology group (Figure 1C) uses a different set of black and white symbols [4] with the anomer and linkage positions indicated by the line style and angle of the bond, respectively, rather than by annotating it with characters and numbers. Several variations of these cartoons using different colors or shapes to represent monosaccharides can be found in publications and web pages.

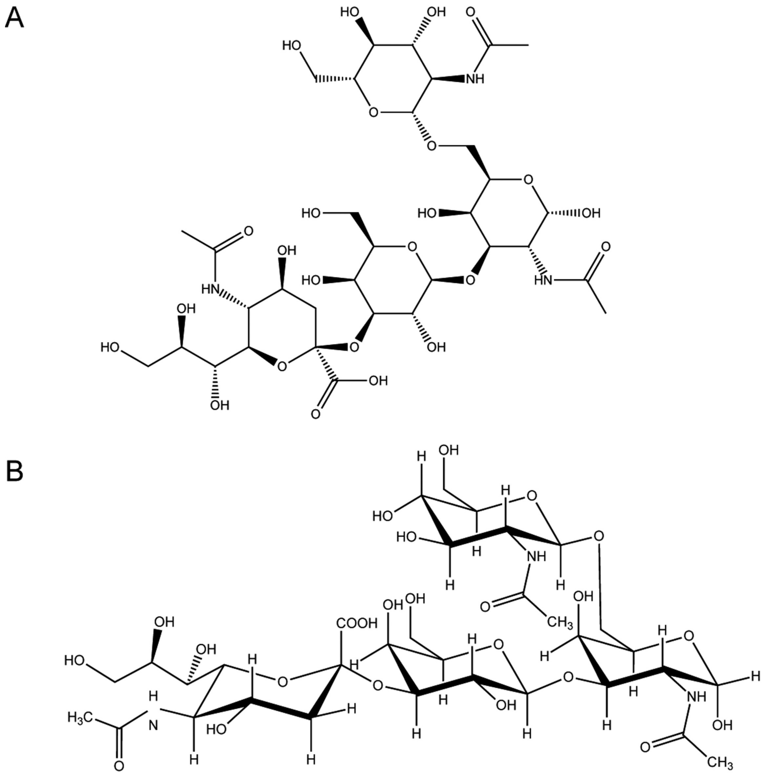


Figure 2. O-glycan specified by GlycomeDB ID 14282 using two explicit chemical representation schemes.

Figure 2 illustrates the same structure shown in Figure 1 but using two different chemical representation schemes that are often used by analytical and synthetic chemists. In contrast to the cartoon representations, structural details (individual atoms and their stereochemistry) are explicitly displayed rather than represented as symbols or names.

However, images of carbohydrates are not suitable as a primary encoding for storing carbohydrate structures in databases or software applications, since it is quite difficult to program a computer to extract monosaccharide and linkage information from such images. In software applications and databases, genes and proteins are usually specified as linear sequences using one character of the English alphabet to represent each residue. In contrast, carbohydrates are usually branched molecules that consist of a larger variety of residues (monosaccharides and their modifications) that cannot be fully represented by the 26 characters in the English alphabet. Software and database developers have therefore been required to create new sequence formats and digital representations for carbohydrates. This unfortunately led to the development of several different sequence formats, many of which are used only in a single application or database.

A

```
[][a-D-GalpNAc]{[(3+1)][b-D-Galp]{[(3+2)][a-D-Neup5Ac]{}}[(6+1)][b-D-GlcpNAc]{}}
```

B

```
Ac(1-5)aXNeup(1-3)bDGalp(1-3)[Ac(1-2)bDGlcpN,Ac(1-2)]bDGalpN
```

C

```
                     b-D-GlcpNAc-(1-6)+
                                      |
                                a-D-GalpNAc
                                      |
a-D-Neup5Ac-(2-3)-b-D-Galp-(1-3)+
```

D

```
TRP type : α-D-GalNAc
Sugars:     [incr. 3,3] α-D-NeuAcp (2-3) (acidic)
            [incr. 3] β-D-Galp (1-3) (neutral)
            [incr. 6] β-D-GlcNAcp (1-6) (neutral)
```

E

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

F

```
ENTRY      G00208   Glycan
NODE        5
        1   Ser/Thr   13    0
        2   GalNAc     3    0
        3   Gal       -5   -4
        4   GlcNAc    -6    4
        5   Neu5Ac   -14   -4
EDGE        4
        1    2:a1    1
        2    3:b1    2:3
        3    4:b1    2:6
        4    5:a2    3:3
///
```

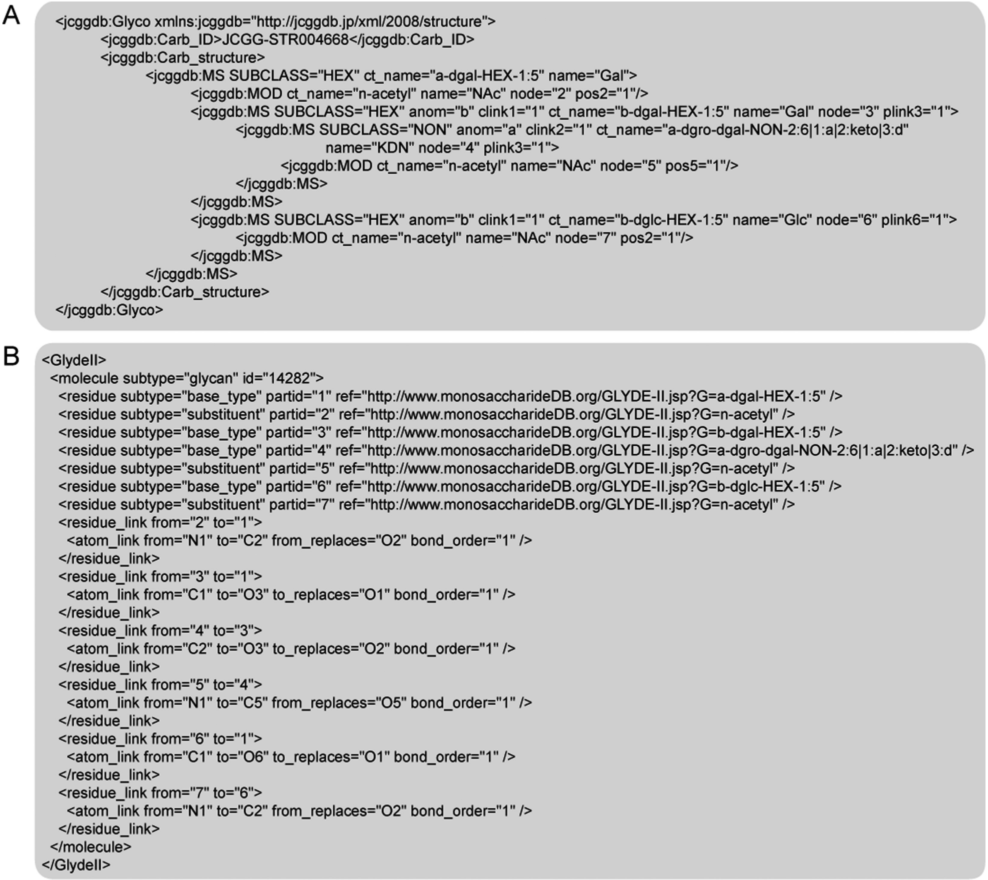
G

```
NNa3Ab3(GNb6)AN
```

Figure 3. O-glycan specified by GlycomeDB ID 14282 using different systematic sequence encodings. **(A)** LINUCS, **(B)** BCSDB encoding, **(C)** CarbBank sequence format, **(D)** sequence format of the GlycoBase(Lille) database, **(E)** GlycoCT format, **(F)** KCF format and **(G)** Linearcode® encoding.

Figure 3 shows several digital sequence formats that are currently in use. The problem of branched carbohydrate structures has been solved using three different approaches, which have been applied in different sequence formats. The first solution is to represent the carbohydrate as several lines of text, mirroring the textual representation often used in scientific journals (see Figure 1D). The sequence format used in the CarbBank database [5, 6] (see Figure 3C) is an example of this. The second solution is to linearize the carbohydrate by putting the branches in parentheses. Examples for this are LINUCS [7] (see Figure 3A), the format used by the BCSDB database [8] (see figure 3B), or the Linearcode® [9] (Figure 3G). In all presented cases, the format definition includes rules for sorting the branches with the aim of generating a unique character sequence for each carbohydrate. The third solution is a connection table approach wherein the residues are listed one by one and the bonds connecting these residues to each other are specified subsequently. Examples are the GlycoCT encoding [10] (Figure 3E) and the KCF format [11] (Figure 3F). The sequence format of the GlycoBase(Lille) database [12] (Figure 3D) is similar but encodes the linkage path to the reducing end before each of the residue names. There are also XML based formats that were created to take advantage of the extensibility and flexibility provided by XML as a data exchange language. Two examples, are shown in Figure 4A (CabosML [13]) and 4B (Glyde-II [14]).

In addition to the formats shown here, other notations, such as the three notations recommended by IUPAC [2] and several variations of this format have been used in scientific publications.



```
<jcggdb:Glyco xmlns:jcggdb="http://jcggdb.jp/xml/2008/structure">
        <jcggdb:Carb_ID>JCGG-STR004668</jcggdb:Carb_ID>
        <jcggdb:Carb_structure>
                <jcggdb:MS SUBCLASS="HEX" ct_name="a-dgal-HEX-1:5" name="Gal">
                        <jcggdb:MOD ct_name="n-acetyl" name="NAc" node="2" pos2="1"/>
                        <jcggdb:MS SUBCLASS="HEX" anom="b" clink1="1" ct_name="b-dgal-HEX-1:5" name="Gal" node="3" plink3="1">
                                <jcggdb:MS SUBCLASS="NON" anom="a" clink2="1" ct_name="a-dgro-dgal-NON-2:6|1:a|2:keto|3:d"
                                                name="KDN" node="4" plink3="1">
                                        <jcggdb:MOD ct_name="n-acetyl" name="NAc" node="5" pos5="1"/>
                                </jcggdb:MS>
                        </jcggdb:MS>
                        <jcggdb:MS SUBCLASS="HEX" anom="b" clink1="1" ct_name="b-dglc-HEX-1:5" name="Glc" node="6" plink6="1">
                                <jcggdb:MOD ct_name="n-acetyl" name="NAc" node="7" pos2="1"/>
                        </jcggdb:MS>
                </jcggdb:MS>
        </jcggdb:Carb_structure>
</jcggdb:Glyco>
```

```
<GlydeII>
 <molecule subtype="glycan" id="14282">
  <residue subtype="base_type" partid="1" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=a-dgal-HEX-1:5" />
  <residue subtype="substituent" partid="2" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=n-acetyl" />
  <residue subtype="base_type" partid="3" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=b-dgal-HEX-1:5" />
  <residue subtype="base_type" partid="4" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=a-dgro-dgal-NON-2:6|1:a|2:keto|3:d" />
  <residue subtype="substituent" partid="5" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=n-acetyl" />
  <residue subtype="base_type" partid="6" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=b-dglc-HEX-1:5" />
  <residue subtype="substituent" partid="7" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=n-acetyl" />
  <residue_link from="2" to="1">
   <atom_link from="N1" to="C2" from_replaces="O2" bond_order="1" />
  </residue_link>
  <residue_link from="3" to="1">
   <atom_link from="C1" to="O3" to_replaces="O1" bond_order="1" />
  </residue_link>
  <residue_link from="4" to="3">
   <atom_link from="C2" to="O3" to_replaces="O2" bond_order="1" />
  </residue_link>
  <residue_link from="5" to="4">
   <atom_link from="N1" to="C5" from_replaces="O5" bond_order="1" />
  </residue_link>
  <residue_link from="6" to="1">
   <atom_link from="C1" to="O6" to_replaces="O1" bond_order="1" />
  </residue_link>
  <residue_link from="7" to="6">
   <atom_link from="N1" to="C2" from_replaces="O2" bond_order="1" />
  </residue_link>
 </molecule>
</GlydeII>
```

Figure 4. O-glycan specified by GlycomeDB ID 14282 in the XML based sequence format CabosML **(A)** and Glyde-II **(B).**

Carbohydrate structure databases

The first large publicly available database for carbohydrate structures was the CarbBank database developed and maintained in the 1980 s and 1990 s. This resource contains approximately 50,000 records with more than 23,000 structures from more than 13,000 publications, although no new structures have been added since funding was discontinued in the mid 1990s. Several subsequent database initiatives, such as GLYCOSCIENCES.de [15], CFG [16] and BCSDB, began by using some or all the structures from CarbBank as the initial information content. Currently active carbohydrate structure database projects are shown in Table 1 along with the URL of the database webpage, the sequence format used to encode carbohydrate structures and the number of structures in each database. Each of these resources also stores a set of meta-information that is associated with each structure. Table 2 provides an overview of this information.

Table 1. List of carbohydrate structure database projects. For each database the name of the project/database, the web page (URL), the used carbohydrate sequence format with the reference to Figure 3 in parentheses and the number of carbohydrate sequences in the database in the used sequence format (as of February 2012) is shown.

Database name	URL	Sequence Format (Illustrated in Figure)	Number of Sequences
CarbBank	http://www.genome.jp/dbget-bin/www_bfind? carbbank	CarbBank format (3C)	23402
CFG database	http://www.functionalglycomics.org/	Linearcode® (3G)	9201
GLYCOSCIENCES.de	http://www.glycosciences.de/	LINUCS (3A)	23367
KEGG	http://www.genome.jp/kegg/glycan/	KCF (3F)	10978
EUROCarbDB	http://www.ebi.ac.uk/eurocarb/	GlycoCT (3E)	13471
GlycoBase(NIBRT)	http://glycobase.nibrt.ie/glycobase/show_nibrt.action	GlycoCT (3E)	7365
GlycoBase(Lille)	http://glycobase.univ-lille1.fr/base/	GlycoBase (Lille) format (3D)	248
BCSDB	http://csdb.glycoscience.ru/bacterial/	BCSDB format (3B)	11565

Table 2. Meta-information archived in carbohydrate structure databases.

Database	Biological source information	Structural Provenance Information	Other information
CarbBank	Species, organ, tissue, disease, aglycon		Literature References
CFG database	Species, organ, tissue, cell type, disease	MS, Glycan array	Literature References
GLYCOSCIENCES.de	Species	NMR	Literature References
KEGG			Pathway
EUROCarbDB	Species, organ, tissue, disease	NMR, HPLC, MS	Literature References
GlycoBase(NIBRT)	Species, organ, tissue, disease	HPLC	Literature References
GlycoBase(Lille)	Species	NMR	Literature References
BCSDB	Species, strain, sero group	NMR	Literature References

The available databases differ not only in the numbers of structures each contains, but also in the amount of additional information that is stored. All of the databases listed in Table 2 (with the exception of KEGG) provide the species information. Several of the databases provide a more detailed specification of the biological source, allowing the organ, the tissue or the cell type of the source sample to be described along with the disease-state of the source tissue. The BCSDB allows the strain and sero group of the bacterial source of the carbohydrate to be recorded. Most of the databases identify the experimental data that was used to characterize the carbohydrate structure and its interactions with other molecules. This data is either extracted from the literature (GLYCOSCEICNES.de, BCSDB) or is provided by the scientist that performed the experiment (CFG, GlycoBase [NIBRT], GlycoBase [Lille]). Finally almost all of the databases provide references to publications in which the carbohydrates were described or the experimental data was published.

Carbohydrate related databases

In addition to the databases that store carbohydrate structures, there is a large set of databases that store carbohydrate related information of interest to glycobiologists.

Table 3. Databases with information related to glycobiology.

Database	URL	Information Content
CAZy	http://www.cazy.org/	Enzymes and Carbohydrate Binding Modules
BRENDA	http://www.brenda-enzymes.org/	Enzymes
CFG Glycosyltransferases database	http://www.functionalglycomics.org/glycomics/molecule/jsp/glycoEnzyme/geMolecule.jsp	Enzymes
CFG Glycan Binding Proteins	http://www.functionalglycomics.org/glycomics/molecule/jsp/gbpMolecule-home.jsp	Lectins
CancerLectinDB	http://proline.physics.iisc.ernet.in/cancerdb/	Lectins
LectinDB	http://proline.physics.iisc.ernet.in/lectindb/	Lectins
UniProt	http://www.uniprot.org/	Proteins and Glycosylation position
O-GlycBase	http://www.cbs.dtu.dk/databases/OGLYCBASE/	Proteins and Glycosylation position
KEGG Pathway	http://www.genome.jp/kegg/pathway.html	Pathways
GlycoEpitope	http://www.glyco.is.ritsumei.ac.jp/epitope/	GlycoEpitopes

Table 3 shows some examples for databases with information related to glycobiology, such as descriptions of carbohydrate active enzymes (e.g., CAZy [17] and the CFG Glycosyltransferases database) or enzymes in general (e.g. BRENDA [18]). The KEGG pathway collection contains visual renderings of complex metabolic pathways that include references to carbohydrate structures and the enzymes that are involved in their metabolism. Several available databases store information about carbohydrate binding proteins, including both lectins and carbohydrate binding modules. These include the CFG Glycan Binding Proteins database, the LectinDB [19], the CancerLectinDB [20] and CAZy. A few databases (such as the GlycoEpitope database) contain information about carbohydrate motifs that are bound by proteins and antibodies. Some of the established proteomics databases such as UniProt [21] and O-GlycBase [22] contain other information that is relevant to glycobiology, such as predicted or experimentally verified glycosylation sites in glycoproteins.

Integration and standardization efforts

Although a considerable amount of structural and biological information for complex carbohydrates is available in the databases described above, most of these resources are isolated islands of data with little or no connection to other data sources. Nevertheless, carbohydrate structure databases have a significant amount of overlap due to the fact that many of the structures they contain were initially derived from the CarbBank database. In addition, cross-referencing of the CFG database and GLYCOSCIENCES.de was implemented to provide links from the CFG to GLYCOSCIENCES.de structure pages. Cross-database search methods were also established [23] to allow the BCSDB and GLYCOSCIENCES.de to be simultaneously searched using either's Web interface. However, beyond these few established interconnections, the databases are not connected to each other or to databases in other domains.

Scientists interested in finding all of the information available for a particular glycan structure have thus been forced to search for that structure several times using the web interface of each database separately. This is not an easy task, since the Web interfaces and the notations used by each database to describe structures and their searchable properties are quite different. Therefore, GlycomeDB [24, 25] was developed to provide a single access point for all of the structures in the carbohydrate structure databases described above. GlycomeDB integrates these databases and other information resources by specifying a single identifier (index) for each unique structure that they contain. GlycomeDB uses this index to maintain and map references for each structure to the data resources in which it is described, providing an unparalleled ability to identify an extensive set of specific database entries for each structure. Creation of a single index for each unique structure requires a single, consistent format to represent structures described in the different resources. Therefore, all structures and their monosaccharides are translated to GlycoCT format [10] and archived using customized data importers written for GlycomeDB along with tools, such as MonosaccharideDB (http://www.monosaccharidedb.org/), that allow monosaccharide representations to be translated from diverse namespaces to the GlycoCT namespace.

Another integration approach is implemented by the JCGGDB (http://jcggdb.jp/index_en.html), which provides a search interface comprising almost all glycosciences databases in Japan. This resource facilitates searching not only for carbohydrate structures in different databases but also for terms and keywords in various carbohydrate-related databases, including those holding information about lectins, antibody binding, glycolipids, glycoproteins and other aspects of glycoscience. The result of a JCGGDB search is a list of hyperlinks to resources containing information related to the user query.

Analytical tools

In addition to the databases described above, a large set of web applications and stand-alone applications has been developed to assist in the interpretation of experimental data.

Table 4. List of analytical tools. For each tool the name, the project URL, the type of processed data and a short description is given.

Name	URL	Data Type	Description
CcpnNMR Analysis	http://www.ccpn.ac.uk/software/analysis	NMR	Annotation assistance for NMR data
Caspar	http://www.casper.organ.su.se/casper/	NMR	Annotation of chemical shifts, spectra simulation
GlyNest	http://www.glycosciences.de/database/nmr/	NMR	Matching of shifts with stored NMR data, spectra simulation
GlycoMod	http://web.expasy.org/glycomod/	MS	Composition analysis
GlycoPeakfinder	http://www.glyco-peakfinder.org/	MS	Composition analysis
GlycoWorkbench	http://www.glycoworkbench.org/	MS	Spectrum annotation with user defined structures or structures from a database

Most of the available tools have been developed to assist in the interpretation of NMR data and mass spectrometry data. For NMR tools there are two different approaches: 1) annotation of chemical shifts based on previously recorded experimental data or 2) simulation of a theoretical spectrum of a structure that can be compared with experimental data. These approaches are followed by the web based applications Caspar [26] and GlyNest [27]. A similar tool is available as part of the BCSDB database. The CcpnNMR Analysis application is a large stand-alone software suite for displaying, analyzing and interpreting NMR data of diverse types of molecules, including carbohydrates.

Another set of tools has been developed to assist in the interpretation of mass spectrometric data. Three main approaches are used: 1) *de novo* composition analysis, where each peak is annotated with a carbohydrate composition or the composition of a fragment. This approach is implemented in the web applications GlycoMod [28] and GlycoPeakfinder [29]. Both tools allow possible structures corresponding to an identified composition to be located in a structural database (GlycoSuiteDB [30] or GLYCOSCIENCES.de, respectively). 2) Annotation of the spectra with user defined structures or structures retrieved from a database. This strategy is used in the stand alone application GlycoWorkbench [31], which has access to several of the carbohydrate structure databases shown above. Each of the structures is matched against the spectrum and scored. There are also commercial tools available following the same strategy (e. g. SimGlycan® [32] or ProteinScape [33]). 3) Annotation of spectra with theoretical fuzzy structures (cartoons) generated for each observed monosaccharide composition using implicit biosynthetic rules rather than selection from a database. This strategy is implemented in the Cartoonist applications [34, 35] developed by the CFG.

3D structure tools

Due to the importance of molecular geometry in mediating the biological functions of glycans and glycoconjugates, 3D structural analysis of these molecules has emerged as a key research area in glycobiology. Nevertheless, carbohydrate moieties of glycoconjugates are still often neglected in crystallographic or spectroscopic analyses of the 3D structures of these complex molecules, having either been removed by chemical or enzymatic treatment before the experimental data is recorded or simply ignored when processing the data. Failure to fully elucidate the 3D structure of the glycan moieties often stems from their molecular flexibility, which confounds the interpretation of the experimental data. Thus, primary 3D structure databases such as the PDB [36] often contain truncated carbohydrate structures, which are reflected in carbohydrate structure databases whose contents are extracted from the PDB. These include GLYCOSCIENCES.de, which uses pdb2linucs [37] to extract this information and the GlycoConjugate Data Bank [38]. However, complete 3D structures are required as starting points for molecular dynamics simulations of glycoconjugates. Therefore programs have been developed to generate energetically feasible 3D structures for carbohydrates *in silico*. These tools include Sweet [39] and the Glycam molecular structure builder (http://glycam.ccrc.uga.edu/ccrc/biombuilder/biomb_index.jsp). The hypothetical 3D structure can then be attached to protein structure models, either manually or by using other programs such as GlyProt [40], for use as starting points for molecular dynamics simulations or docking studies. Other resources such as GlycoMapsDB [41] and GlyTorsion [42] provide conformational maps and/or statistical analyses of glycan geometry, which can be used to evaluate and compare the 3D models.

Glycosylation prediction and analysis tools

As mentioned above, several databases contain information about the glycosylation sites of proteins, either directly stored as meta-information for the protein (e. g. Uniprot, O-GlycBase) or as part of the primary data/structure (e. g. PDB). This information has been analyzed and used to develop tools for predicting the glycosylation sites of other proteins. These tools include NetNGlyc and NetOGlyc [43], which predict glycosylation sites for N- and O-glycans, respectively, using artificial neural networks (ANNs) trained with glycosylation-site information gleaned from databases. Similar tools are available for other types of glycosylation, such as C-mannosylation [44], GPI-anchors [45] or O-GlcNAc substitution [46]. Several other data processing and mining strategies have been used in addition to ANNs for this purpose. An example is EnsembleGly [47], which implements support vector machines for glycosylation site prediction. However, users should be aware that techniques such as ANNs and support vector machines do not provide predictions that are 100% correct, often generating false positive or false negative results. The predictive accuracy of these tools depends on the specific algorithm that is implemented and the range of structural diversity for the molecules included in the training set.

Table 5. Tools for the prediction of glycosylation sites and analysis of glycosylation.

Name	URL	Description
NetNGlyc	http://www.cbs.dtu.dk/services/NetNGlyc/	Prediction of N-glycosylation sites using neuronal networks.
NetOGlyc	http://www.cbs.dtu.dk/services/NetOGlyc/	Prediction of O-glycosylation sites using neuronal networks.
Big-PI Predictor	http://mendel.imp.ac.at/gpi/gpi_server.html	Prediction of GPI modification sites.
EnsembleGly	http://turing.cs.iastate.edu/EnsembleGly/	Prediction of glycosylation sites using support vector machines.
GlySeq	http://www.glycosciences.de/tools/glyseq/	Statistical analysis of amino acid sequences that are glycosylated, based on information from PDB.
GlyVicinity	http://www.glycosciences.de/tools/glyvicinity/	Statistical analysis of amino acids that are physically close to carbohydrates, based on information from the PDB.

Other web applications such as GlySeq and GlyVicinity [42] provide statistical analysis of the amino acid sequences in the neighborhood of the glycosylation site or which are close in 3D space to the sugars. This information, which was extracted from glycosylated proteins in the PDB, can be used as the basis for other glycosylation-site prediction tools.

Obstacles and Challenges in Glyco-Bioinformatics

As the previous chapter section has shown, a basic set of databases and applications is now available to assist in glycomics research. The number of such tools will likely grow as analysts develop new high-throughput analytical methods, which require considerable software support. In addition, the mining and integration of data from several different resources will become more important for discovering relationships between diverse aspects of glycobiology, including molecular structure, cell compartmentalization, metabolism, biological development and disease. However, several obstacles currently slow the development of software and other glyco-bioinformatics tools for these purposes. This chapter section attempts to identify some of these obstacles from the point of view of the authors.

Sequence representations and sequence formats

The existence of competing graphical representation schemes for carbohydrate structure and the non-uniform usage of these schemes has long been considered a problem by many glycobiologists. Adoption of a single graphical representation scheme (or at least one scheme for each scientific journal) would be preferable. However, it is difficult to get glycoscientists to agree on a single scheme, because each of the competing graphical representations has its own advantages. Due to their complexity, explicit chemical representation schemes (Figure 2) are not appropriate to describe a large number of carbohydrates identified in a high-throughput experiment. Generating images of spectra annotated with structural assignments rendered using this kind of representation is even more problematic, so cartoon structures are often used for this purpose. On the other hand, cartoon representations are typically limited to carbohydrate structures composed of small, well-defined sets of biologically relevant monosaccharide residues and rarely provide the level of structural detail required to fully describe intermediates in the chemical synthesis of a complex carbohydrate. A reasonable compromise may be the selection of one explicit representation scheme for depicting chemically diverse, detailed structures and one cartoon representation for biological structures composed of common monosaccharide residues. In any case, this decision is up to the journal editors and the research community; the duty of glyco-bioinformaticians is to provide the tools required to generate and process structural representations using the schemes that are selected by the community. It should be noted that, so far, no tools that translate carbohydrate structures from one of the digital sequence formats directly to a chemical representation (Figure 2) are available.

From the view point of bioinformatics, the existence of the different graphical representation schemes is not a major problem. As pointed out above, images are not suitable as a primary digital format for storing and exchanging carbohydrate structural information. However, images that conform to an individual user's requirements can be generated automatically from digital sequence information, which is usually encoded as highly formatted text (e.g., GlycoCT and XML). Several tools and databases, including GlycoWorkbench [31], GlycomeDB [24] and EUROCarbDB [48], already allow users to choose among several different graphical representation schemes to generate images from such digitally encoded sequences. The major problem is that each of the various databases and tools uses its own textual description of carbohydrate structure rather than using a common standard, even though there is an agreement within the glycobiology community that such a standard is needed. Although an XML-based format for GLYcan Data Exchange (GLYDE-II) has been selected as the standard [49], this format is so far supported only by a few databases and applications.

Two impediments to the implementation of GLYDE-II exist.

1. Algorithms designed to search or manipulate structures (e.g. substructure searches and composition searches) are almost always tied closely to the sequence format used by the program or database. Thus, adoption of GLYDE-II as the internal format for structural representation would require a large portion of the internal logic to be rewritten.
2. No currently available software application or source code supports the translation of all of the various sequence formats (see above) to GLYDE-II and back. Only a

part of the translation functionality is available in the existing program code in GlycomeDB and MonosaccharideDB.

Thus, there is a critical need to rapidly develop freely available software that translates the different sequence formats into each other. Although it is not practical to replace the internal encoding of most databases and data-processing tools with GLYDE-II, translation software would allow all the tools to read and write into the standard format for data export, import and exchange. Developers of new projects would probably benefit by considering GLYDE-II (or one of the other existing formats) for structural representations rather than reinventing the wheel by creating yet another sequence format.

Meta-information and standards

Recently, interest in maintaining and organizing meta-information for the annotation of glycan sequences has grown since this type of information provides the basis for knowledge discovery when data mining strategies are implemented. Meta data includes information about the biological source of a glycan, pointers to relevant database resources and literature references describing the glycan, its characterization, biosynthesis and biological properties. Each carbohydrate structure database stores a different set of meta-information related to the carbohydrate structures that it contains. The first large carbohydrate structure database initiative, CarbBank, attempted to collect several types of meta data for each glycan structure, including bibliographic information, biological origin (species, tissue, organ type, cell line, disease) and an identifier (such as a protein ID) for any aglycon attached to the reducing end. A large subset of CarbBank records contains most of this information, although it is not present in every entry. Most of the database projects that succeeded CarbBank took a step back and reduced the amount of meta-information that they stored. For example, description of the biological origin might be limited to species only or bibliographic information might not be included. Unfortunately, many of these databases lack a common vocabulary or syntax to describe the information (such as the identity of the source species) that has been retained. Sometimes, identifiers that are specified in publicly available ontologies or dictionaries (e. g. NCBI taxonomy) are used but in other cases local dictionaries have been created or free text fields are utilized. This makes it very difficult to integrate or compare data that is retrieved from more than one database.

Solving these problems requires general agreement regarding a minimal set of meta-information that should be stored in a glycan structure database to provide a universally informative data set. Development of such guidelines will require input from members of the glycobiology community with diverse expertise in order to ensure that they foster datasets that address current and future research needs of the community. On the other hand, there is also a critical need for input from computer scientists who should agree, for example, on digital formats for the storage and exchange of this information. As described above in the context of carbohydrate structures, efficient data integration across different databases requires selection of a single standard representation for exchange of each type of data that has to be stored. Adaptation of the agreed-upon standards by databases and software applications will create many advantages, including facile exchange of readily parsable information among these resources, which is required for the development of effective approaches for data retrieval, browsing and mining.

Experimental data and procedures

The glycobiology community has come to recognize the considerable advantages of archiving primary experimental data along with meta data that is associated with each glycan structure. Such primary data, which include NMR and mass spectra that were used to elucidate molecular structures, can be reused as fingerprints that allow those structures to be identified in biological samples. Primary data of this type is rarely stored in carbohydrate databases, and most of the available spectral information is in the form of highly processed lists of parameter values, such as chemical shifts and lists of ion abundances. This data reduction is not always desirable, as the raw data is usually a richer source of information for feature extraction and other data mining approaches or for re-evaluation of the experimental results. A major impediment to the collection and organization of primary experimental data is that they are often represented using proprietary, device-specific formats that can only be processed by customized software applications. As is the case for the meta-information, community agreement is needed to identify the most valuable types of experimental data that should be archived and the standard digital formats that should be utilized to store and export the data.

In addition to the data itself, it is important to record information such as sample preparation protocols and device setups that were used to generate the data. This information is required to generate transparent results that can be easily understood and reproduced by other researchers. In some of the existing resources these parameters are documented as short text blocks or provided indirectly by reference to the Methods section of a published manuscript. However, these approaches by themselves rarely provide information that is sufficiently detailed to allow complete reproduction of the data set, especially by novices in glycoanalytics.

Two aspects should be considered when storing such information:

1. *sufficiency* – the information should be sufficiently detailed to allow the experiment to be understood and reproduced
2. *practicability* – the minimum information required to achieve consideration 1 should be included, since the collection of large amounts of such information is often impractical, inefficient and time consuming.

For example, the level of detail specified by the EXACT ontology [50] may lead to the storage of too much information, which is difficult to collect and support. Minimum information checklists based on these considerations have been implemented in other research areas (e. g. MIAPE [51] for proteomics and MIAME [52] for microarray analysis). Once the checklists specifying the minimum information describing an experiment are defined, it is important to standardize the digital format for this information. This will not only facilitate automatic checking for compliance with the checklists but can also be used to exchange this information among different systems.

It should be noted that currently accepted checklists are not designed to store information at the level of detail required by non-experts to reproduce the experiment. Therefore, it would be very useful to supplement these checklists with references to freely available textual descriptions of experimental protocols that can be used to perform identical experiments in a different laboratory.

Data integration

The data integration and cross referencing efforts that have started with GlycomeDB and JCGGDB bring considerable benefit to the community, bringing several different data sources together and allowing information to be explored simply by following hyperlinks to different resources and web pages. We consider continuation of this work and integration of all carbohydrate databases into a network of interconnected resources to be a high priority. The next obvious step to this end is to link the different data resources not only to GlycomeDB but also to each other, developing an easily accessible and comprehensive glyco-bioinformatics infrastructure. As envisioned by the Japanese JCGGDB initiative, data integration efforts should extend beyond glycomics resources, providing access to tools and databases (such as Lipidbank or UniProt) that support genomics, proteomics and lipidomics research. Together, this infrastructure should establish methods to identify glycoconjugates that consist of specific carbohydrate structures along with the specific proteins and/or lipids to which they are attached. Implementation of this functionality should facilitate searches to find glycoconjugates that contain a moiety (selected from a glycan, protein or lipid database) connected to moieties that may be fully described in orthogonal databases. Success in this area clearly demands collaboration and cooperation between informaticians who develop databases with diverse information content. Again, this will require the adoption of standard descriptors and exchange formats that allow communication between these data resources.

What is out there?

Not all of the existing glyco-bioinformatics programs and databases have been described in the scientific literature. This makes it difficult for bench scientists and software developers to determine whether a software application or database that addresses a specific informatics need already exists. This lack of information has led to the development of many programs with redundant functionality, essentially reinventing the wheel over and over again. To improve the visibility of the available software applications, several review articles (e. g. [53 – 55]) have been published. In addition, the Web pages of large research initiatives such as the CFG (http://www.functionalglycomics.org/static/consortium/links.shtml) and EUROCarbDB (http://www.eurocarbdb.org/links) often provide links to computational tools that support their scientific domain of interest. Often, descriptions of the tools in review papers or on Web pages are rather short, making it difficult to evaluate the usefulness or relevance of the tools that are listed there. To make matters worse, the URL of the Web page providing access to a tool may change, leading to dead links that give the mistaken impression the software project no longer exists. Since the review papers and on Web pages are often written by a "power user" who may not have been directly involved in the development of the tool, the technical description may be incomplete, inaccurate or simply irrelevant. Finally, review papers and Web links tend to focus on databases, web applications or stand alone applications. Other classes of informatics tools such as Web services and workflows, which may constitute more efficient ways of solving various problems, are usually not described since they are more difficult to identify, understand and invoke.

For the reasons described above, we are developing a Web-based, open access system that allows individuals to register software applications, Web services, workflows, databases and programming libraries that they have developed to support research in the glycobiology domain. The information provided by this GlycomicsPortal is entered directly by registered software developers, making it more likely that it is accurate and up to date. Methods that facilitate revision of this information ensure that URL changes do not result in dead links. A Web based search implemented by the Portal can be used by bench scientists and theoreticians to find tools and databases that provide or process data critical for the success of their research. The system is also of interest to software developers, who can use it to find existing tools, programming libraries and Web services that they can integrate into their own informatics systems.

General considerations for glyco-bioinformatics projects

The success of glyco-bioinformatics in a rapidly changing research environment will benefit immensely by the unconstrained, open exchange of programs, functionality and source code. Many software applications with nearly identical functionality have been developed from scratch because individuals did not have access to previously written source code or fully developed tools that address their current needs. This reinvention of the wheel has resulted in a considerable waste of human talent and time. There are several approaches that a software developer can use to make tools available to the community, allowing others to access their work and benefit from it. One often-used strategy is the development of freely available Web based applications and Web databases, or to provide programs for free download. Alternatively, Web services allow other software developers to remotely invoke computational resources using standardized interfaces. Perhaps, the most effective way of sharing the results of software development is to make the source code of a program freely available. This not only facilitates the reuse of functionality that required extensive effort to implement and test, but also provides an opportunity to extend and optimize the programs providing this functionality. Although this may seem like a no-win situation for a software developer who provides the source code for the benefit of others – nurturing of a culture that encourages sharing of these resources will eventually benefit everyone. Progress in this area will require the leadership by funding agencies, who increasingly demand the free and open dissemination of software developed as a result of projects that they sponsor.

Acknowledgments

The authors wish to thank the NIH National Institute for General Medical Sciences for support of bioinformatics infrastructure development through the Integrated Technology Resource for Biomedical Glycomics (8P41GM103490).

References

[1] Werz, D.B., Ranzinger, R., Herget, S. Adibekian, A., von der Lieth, C.-W., and Seeberger, P.H. (2007) Exploring the Structural Diversity of Mammalian Carbohydrates ("Glycospace") by Statistical Databank Analysis. *ACS Chemical Biology* **2:**685 – 691.
doi: http://dx.doi.org/10.1021/cb700178s.

[2] McNaught, A.D. (1997) Nomenclature of Carbohydrates (Recommendations 1996). *Adv. Carbohydr. Chem. Biochem.* **52:**44 – 177.
doi: http://dx.doi.org/10.1016/S0065-2318(08)60090-6.

[3] Raman, R., Raguram, S., Venkataraman, G., Paulson, J.C., and Sasisekharan, R. (2005) Glycomics: An Integrated Systems Approach to Structure-Function Relationships of Glycans. *Nature Methods* **2:**817 – 824.
doi: http://dx.doi.org/10.1038/nmeth807.

[4] Campbell, M.P., Royle, L., Radcliffe, C.M., Dwek, R.A., and Rudd, P.M. (2008) GlycoBase and Auto GU: Tools for HPLC-based Glycan Analysis. *Bioinformatics* **24:**1214 – 1216.
doi: http://dx.doi.org/10.1093/bioinformatics/btn090.

[5] Doubet, S., and Albersheim, P. (1992) CarbBank. *Glycobiology* **2:**505.
doi: http://dx.doi.org/10.1093/glycob/2.6.505.

[6] Doubet, S., Bock, K., Smith, D., Darvill, A., and Albersheim, P. (1989) The Complex Carbohydrate Structure Database. *Trends Biochem. Sci.* **14:**475 – 477.
doi: http://dx.doi.org/10.1016/0968-0004(89)90175-8.

[7] Bohne-Lang, A., Lang, E., Förster, T., and von der Lieth, C.-W. (2001) LINUCS: Linear Notation for Unique Description of Carbohydrate Sequences. *Carbohydr. Res.* **336:**1 – 11.
doi: http://dx.doi.org/10.1016/S0008-6215(01)00230-0.

[8] Toukach, P.V. (2011) Bacterial Carbohydrate Structure Database 3: Principles and Realization. *J. Chem. Info. Model.* **51:**159 – 170.
doi: http://dx.doi.org/10.1021/ci100150d.

[9] Banin, E., Neuberger, Y., Altshuler, Y. Halevi, A., Inbar, O., Dotan, N., and Dukler, A. (2002) A Novel LinearCode® Nomenclature for Complex Carbohydrates. *Trends Glycosci. Glycotechnol.* **14:**127 – 137.
doi: http://dx.doi.org/10.4052/tigg.14.127.

[10] 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.
doi: http://dx.doi.org/10.1016/j.carres.2008.03.011.

[11] Aoki, K.F., Yamaguchi, A. Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S., and Kanehisa, M. (2004) KCaM (KEGG Carbohydrate Matcher): A Software Tool for Analyzing the Structures of Carbohydrate Sugar Chains. *Nucleic Acids Res.* **32:**W267-W272 (web server issue).
doi: http://dx.doi.org/10.1093/nar/gkh473.

[12] Maes, E., Bonachera, F., Strecker, G. and Guerardel, Y. (2009) SOACS Index: An Easy NMR-based Query for Glycan Retrieval. *Carbohydr. Res.* **344:**322 – 330.
doi: http://dx.doi.org/10.1016/j.carres.2008.11.001.

[13] Kikuchi, N., Kameyama, A., Nakaya, S., Ito, H., Sato, T., Shikanai, T., Takahashi, Y., and Narimatsu, H. (2005) The Carbohydrate Sequence Markup Language (CabosML): An XML Description of Carbohydrate Structures. *Bioinformatics* **21:**1717 – 1718.
doi: http://dx.doi.org/10.1093/bioinformatics/bti152.

[14] Sahoo, S.S., Thomas, C., Sheth, A. Henson, C. and York, W.S. (2005) GLYDE – An Expressive XML Standard for the Representation of Glycan Structure. *Carbohydr. Res.* **340:**2802 – 2807.
doi: http://dx.doi.org/10.1016/j.carres.2005.09.019.

[15] Lutteke, T., Bohne-Lang, A., Loss, A. Goetz, T. Frank, M., and von der Lieth, C.-W. (2006) GLYCOSCIENCES.de: An Internet Portal to Support Glycomics and Glycobiology Research. *Glycobiology* **16:**71R-81R.
doi: http://dx.doi.org/10.1093/glycob/cwj049.

[16] Raman, R., Venkataraman, M., Ramakrishnan, S., Lang, W., Raguram, S., and Sasisekharan, R. (2006) Advancing Glycomics: Implementation Strategies at the Consortium for Functional Glycomics. *Glycobiology* **16(5):**82R-90R.
doi: http://dx.doi.org/10.1093/glycob/cwj080.

[17] Cantarel, B.L., Coutinho, P.M., Rancurel, C. Bernard, T., Lombard, V., and Henrissat, B. (2009) The Carbohydrate-Active EnZymes Database (CAZy): An Expert Resource for Glycogenomics. *Nucleic Acids Res.* **37:**D233-D238 (Database issue).
doi: http://dx.doi.org/10.1093/nar/gkn663.

[18] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G. and Schomburg, D. (2004) BRENDA, the Enzyme Database: Updates and Major New Developments. *Nucleic Acids Res.* **32:**D431-D433 (Database issue).
doi: http://dx.doi.org/10.1093/nar/gkh081.

[19] Chandra, N.R., Kumar, N., Jeyakani, J., Singh, D.D., Gowda, S.B., and Prathima, M.N. (2006) Lectindb: A Plant Lectin Database. *Glycobiology* **16:**938 – 946.
doi: http://dx.doi.org/10.1093/glycob/cwl012

[20] Damodaran, D., Jeyakani, J., Chauhan, A., Kumar, N., Chandra, N.R., and Surolia, A. (2008) CancerLectinDB: A Database of Lectins Relevant to Cancer. *Glycoconj. J.* **25:**1919 – 198.
doi: http://dx.doi.org/10.1007/s10719-007-9085-5.

[21] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* (2004) UniProt: The Universal Protein Knowedgebase. *Nucleic Acids Res.* **32:**D115-D119 (Database issue).
doi: http://dx.doi.org/10.1093/nar/gkh131.

[22] Gupta, R., Birch, H., Rapacki, K., Brunak, S., and Hansen, J.E. (1999) O-GLYCBASE Version 4.0: A Revised Database of O-glycosylated Proteins. *Nucleic Acids Res.* 1**27:**370 – 372.
doi: http://dx.doi.org/10.1093/nar/27.1.370.

[23] Toukach, P., Joshi, H.J., Ranzinger, R., Knirel, Y., and von der Lieth, C.-W. (2007) Sharing of Worldwide Distributed Carbohydrate-related Digital Resources: Online Connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de. *Nucleic Acids Res.* **35:**D280-D286 (Database issue).
doi: http://dx.doi.org/10.1093/nar/gkl883.

[24] Ranzinger, R., Frank, M., von der Lieth, C.-W., and Herget, S. (2009) GlycomeDB.org: A Portal for Querying Across the Digital World of Carbohydrate Sequences. *Glycobiology* **19:**1563 – 1567.
doi: http://dx.doi.org/10.1093/glycob/cwp137.

[25] Ranzinger, R., Herget, S., Wetter, T., and von der Lieth, C.-W. (2008) GlycomeDB – Integration of Open-access Carbohydrate Structure Databases. *BMC Bioinformatics* **9:**384.
doi: http://dx.doi.org/10.1186/1471-2105-9-384.

[26] Lundborg, M., and Widmalm, G. (2011) Structural Analysis of Glycans by NMR Chemical Shift Prediction. *Anal. Chem.* **83:**1514 – 1517.
doi: http://dx.doi.org/10.1021/ac1032534.

[27] Loss, A., Stenutz, R., Schwarzer, E. and von der Lieth, C.-W. (2006) GlyNest and CASPER: Two Independent Approaches to Estimate 1 H and 13C NMR Shifts of Glycans Available Through a Common Web-interface. *Nucleic Acids Res.* **34:**W733-W737 (Web server issue).
doi: http://dx.doi.org/10.1093/nar/gkl265.

[28] Cooper, C.A., Gasteiger, E., and N.H. Packer (2001) GlycoMod – A Software Tool for Determining Glycosylation Compositions from Mass Spectrometric Data. *Proteomics* **1:**340 – 349.
doi: http://dx.doi.org/10.1002/1615-9861(200102)1:2<340::AID-PROT340>3.0.CO;2-B.

[29] Maass, K., Ranzinger, R., Geyer, H., von der Lieth, C.W., and Geyer, R. (2007) Glyco-peakfinder – *de novo* Composition Analysis of Glycoconjugates. Proteomics **7:**4435 – 4444.
doi: http://dx.doi.org/10.1002/pmic.200700253.

[30] Cooper, C.A., Joshi, H.J., Harrison, M.J., Wilkins, M.R., and Packer, N.H. (2003) GlycoSuiteDB: A Curated Relational Database of Glycoprotein Glycan Structures and their Biological Sources. *Nucleic Acids Res.* **31:**511 – 513.
doi: http://dx.doi.org/10.1093/nar/gkg099.

[31] Ceroni, A., K. Maass, H. Geyer, R. Geyer, Dell, A., and Haslam, S.M. (2008) GlycoWorkbench: A Tool for the Computer-assisted Annotation of Mass Spectra of Glycans. *J. Proteome Res.* **7:**1650 – 1659.
doi: http://dx.doi.org/10.1021/pr7008252.

[32] Apte, A., and Meitei, N.S. (2010) Bioinformatics in Glycomics: Glycan Characterization with Mass Spectrometric Data Using SimGlycan. *Methods Mol. Biol.* **600:**269 – 281.
doi: http://dx.doi.org/10.1007/978-1-60761-454-8_19.

[33] Thiele, H., Glandorf, J., and Hufnagel, P. (2010) Bioinformatics Strategies in Life Sciences: From Data Processing and Data Warehousing to Biological Knowledge Extraction. *J. Integrative Bioinformatics* **7:**141.
doi: http://dx.doi.org/10.2390/biecoll-jib-2010-141.

[34] Goldberg, D., Bern, M., Li, B., and Lebrilla, C.B. (2006) Automatic Determination of O-glycan Structure from Fragmentation Spectra. *J. Proteome Res.* **5:**1429 – 1434.
doi: http://dx.doi.org/10.1021/pr060035j.

[35] Goldberg, D., Sutton-Smith, M., Paulson, J., and Dell, A. (2005) Automatic Annotation of Maatrix-assisted Laser Desorption/Ionization N-glycan Spectra. *Proteomics* **5:**865 – 875.
doi: http://dx.doi.org/10.1002/pmic.200401071.

[36] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28:**235 – 242.
doi: http://dx.doi.org/10.1093/nar/28.1.235.

[37] Lutteke, T., Frank, M., and von der Lieth, C.-W. (2004) Data Mining the Protein Data Bank: Automatic Detection and Assignment of Carbohydrate Structures. *Carbohydr. Res.* **339:**1015 – 1020.
doi: http://dx.doi.org/10.1016/j.carres.2003.09.038.

[38] Nakahara, T., Hashimoto, R., Nakagawa, H., Monde, K., Miura, N. and Nishimura, S. (2008) Glycoconjugate Data Bank: Structures – an Annotated Glycan Structure Database and N-glycan Primary Structure Verification Service. *Nucleic Acids Res.* **36:**D368-D371 (Database issue).
doi: http://dx.doi.org/10.1093/nar/gkm833.

[39] Bohne, A., Lang, E., and von der Lieth, C.-W. (1999) SWEET – WWW-based Rapid 3D Construction of Oligo- and Polysaccharides. *Bioinformatics* **15:**767 – 768.
doi: http://dx.doi.org/10.1093/bioinformatics/15.9.767.

[40] Bohne-Lang, A., and von der Lieth, C.-W. (2005) GlyProt: *in silico* Glycosylation of Proteins. *Nucleic Acids Res.* **33:**W214-W219 (Web server issue).
doi: http://dx.doi.org/10.1093/nar/gki385.

[41] Frank, M., Lutteke, T., and von der Lieth, C.-W. (2007) GlycoMapsDB: A Database of the Accessible Conformational Space of Glycosidic Linkages. *Nucleic Acids Res.* **35:**287 – 290 (Database issue).
doi: http://dx.doi.org/10.1093/nar/gkl907.

[42] Lutteke, T., Frank, M., and von der Lieth, C.-W. (2005) Carbohydrate Structure Suite (CSS): Analysis of Carbohydrate 3D Structures Derived from the PDB. *Nucleic Acids Res.* **33:**D242-D246 (Database issue).
doi: http://dx.doi.org/10.1093/nar/gki013.

[43] Hansen, J.E., Lund, O., Tolstrup, N., Gooley, A.A., Williams, K.L., and Brunak, S. (1998) NetOglyc: Prediction of Mucin Type O-glycosylation Sites Based on Sequence Context and Surface Accessibility. *Glycoconj. J.* **15:**115 – 130.
doi: http://dx.doi.org/10.1023/A:1006960004440.

[44] Julenius, K. (2007) NetCGlyc 1.0: Prediction of Mammalian C-mannosylation Sites. *Glycobiology* **17:**868 – 876.
doi: http://dx.doi.org/10.1093/glycob/cwm050.

[45] Eisenhaber, B., Bork, P., and Eisenhaber, F. (1999) Prediction of Potential GPI-modification Sites in Proprotein Sequences. *J. Mol. Biol.* **292:**741 – 758.
doi: http://dx.doi.org/10.1006/jmbi.1999.3069.

[46] Gupta, R. and Brunak, S. (2002) Prediction of Glycosylation Across the Human Proteome and the Correlation to Protein Function. *Pac. Symp. Biocomput.* **2002:**310 – 322.

[47] Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., Honavar, V. (2007) Glycosylation Site Prediction Using Ensembles of Support Vector Machine Classifiers. *BMC Bioinformatics* **8:**438.
doi: http://dx.doi.org/10.1186/1471-2105-8-438.

[48] von der Lieth, C.-W., Freire, A.A., Blank, D., Campbell, M.P., Ceroni, A., Damerell, D.R., Dell, A., Dwek, R.A., Ernst, B., Fogh, R. *et al.* (2011) EUROCarbDB: An Open-access Platform for Glycoinformatics. *Glycobiology* **21:**493 – 502.
doi: http://dx.doi.org/10.1093/glycob/cwq188.

[49] Packer, N.H., von der Lieth, C.-W., Aoki-Kinoshita, K.F., Lebrilla, C.B., Paulson, J.C., Raman, R., Rudd, P., Sasisekharan, R., Taniguchi, N., and York, W.S. (2008) Frontiers in Glycomics: Bioinformatics and Biomarkers in Disease. An NIH White Paper Prepared from Discussions by the Focus Groups at a Workshop on the NIH Campus, Bethesda MD (September 11 – 13, 2006). Proteomics **8:**8 – 20.
doi: http://dx.doi.org/10.1002/pmic.200700917.

[50] Soldatova, L.N., Aubrey, W., King, R.D., and Clare, A. (2008) The EXACT Description of Biomedical Protocols. *Bioinformatics* **24:**i295-i303.
doi: http://dx.doi.org/10.1093/bioinformatics/btn156.

[51] Taylor, C.F., Paton, N.W., Lilley, K.S., Binz, P.A., Julian, R.K. Jr., Jones, A.R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E.W., *et al.* (2007) The Minimum Information About a Proteomics Experiment (MIAPE). *Nature Biotechnol.* **25:**887 – 893.
doi: http://dx.doi.org/10.1038/nbt1329.

[52] Brazma, A. (2009) Minimum Information About a Microarray Experiment (MIAME) – Successes, Failures, Challenges. *Sci. World J.* **9:**420 – 423.
doi: http://dx.doi.org/10.1100/tsw.2009.57.

[53] von der Lieth, C.-W., Bohne-Lang, A., Lohmann, K.K., and Frank, M. (2004) Bioinformatics for Glycomics: Status, Methods, Requirements and Perspectives. *Brief Bioinform.* **5:**164 – 178.
doi: http://dx.doi.org/10.1093/bib/5.2.164.

[54] Aoki-Kinoshita, K.F., and Kanehisa, M. (2006) Bioinformatics Approaches in Glycomics and Drug Discovery. *Curr. Opin. Mol. Ther.* **8:**514 – 520.

[55] Mamitsuka, H. (2008) Informatic Innovations in Glycobiology: Relevance to Drug Discovery. *Drug Discov. Today* **13:**118 – 123.
doi: http://dx.doi.org/10.1016/j.drudis.2007.10.013.