Available online at www.sciencedirect.com
SCIENCE DIRECT
Carbohydrate Research 340 (2005) 2802–2807
Carbohydrate RESEARCH

GLYDE—an expressive XML standard for the representation of glycan structure

Satya S. Sahoo,[a,b] Christopher Thomas,[b] Amit Sheth,[b] Cory Henson[a] and William S. York[a,*]

[a]*Complex Carbohydrate Research Center, University of Georgia, 315 Riverbend Road, Athens, GA 30602-4712, USA*
[b]*Large Scale Distribution Information Systems (LSDIS) Lab, Department of Computer Science, University of Georgia, Athens, GA 30602, USA*

Received 6 June 2005; accepted 1 September 2005
Available online 20 October 2005

Abstract—The amount of glycomics data being generated is rapidly increasing as a result of improvements in analytical and computational methods. Correlation and analysis of this large, distributed data set requires an extensible and flexible representational standard that is also '*understood*' by a wide range of software applications. An XML-based data representation standard that faithfully captures essential structural details of a glycan moiety along with additional information (such as data provenance) to aid the interpretation and usage of glycan data, will facilitate the exchange of glycomics data across the scientific community. To meet this need, we introduce GLYcan Data Exchange (GLYDE) standard as an XML-based representation format to enable interoperability and exchange of glycomics data. An online tool (http://128.192.9.86/stargate/formatIndex.jsp) for the conversion of other representations to GLYDE format has been developed.

Keywords: GLYcan Data Exchange (GLYDE); Glycoinformatics; XML-based glycan representation; Glycan data interoperability

1. Introduction

Glycomics, an emerging member of the '-omics' family, is the study of a collection of all glycan structures that are expressed by a given organism, tissue, or cell type in a particular developmental or physiological state. The current focus of most glycomics studies is the analysis of the glycosylation of proteins and lipids. This is an extremely challenging endeavor that depends on sophisticated experimental and computational methods. Nevertheless, growth of this area of research is accelerating rapidly as high-throughput experimental protocols to generate data on a large scale are developed. The next transformational steps in the classical hierarchy of data > information > knowledge require an integrated approach for the retrieval and analysis of this extensive collection of data. In this context, it is critical to develop an extensible, scalable, and faithful structural representation standard to encapsulate glycomics data.

The comparison of carbohydrate structural information obtained in different laboratories is a fundamental requirement[12] for the most effective realization of glycomics research. Frequently, structural data generated by research labs are stored in relational databases using schemas that are more closely related to a particular experimental data format or biological system than to any standard that would foster comparative analysis and end-user accessibility.[4] This leads to database schema diversity that must be overcome if the potential of glycomics data analysis is to be fully realized. As compared to the databases available in proteomics or genomics, currently available glycomics databases are modest in size. However, the rapid evolution of glycomics will result in very large data collections that will have to be organized, analyzed, and compared, requiring standards for structural representation. The data

* Corresponding author. Tel.: +1 706 542 4628; fax: +1 706 542 4412; e-mail: will@ccrc.uga.edu

0008-6215/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.carres.2005.09.019

representation standard used by SWEET-DB, LInear Notation for Unique description of Carbohydrate Sequences (LINUCS),[1,2] provides a unique representation of carbohydrate structure at the expense of human readability. Thus, LINUCS is more appropriate for machine-mediated data comparison than the standard IUPAC–IUBMB (International Union of Pure and Applied Chemistry–International Union of Biochemistry and Molecular Biology) carbohydrate nomenclature, which allows a single glycan structure to be specified in many different ways. However, as pointed out by von der Lieth et al.,[3] there are no generally accepted ways to exchange glycan related data; hence, the possibility of creating multiple disconnected and incompatible islands of glycomics data is very real. Furthermore, as espoused by the Semantic Web initiative,[6,11] and exemplified by early efforts in applying the principles and tools to bioinformatics including glycomics, future data processing, and analytical tools should be able to '*understand*' the data processes by them.[9,10] These tools, within the framework of the 'Service-Oriented Science',[8] will be available as platform independent applications—*Web Services*,[15] developed by one research group but accessible to the worldwide community of researchers. Hence, an XML-based standard structural representation format for glycans forms the foundation for the development of 'Service-Oriented Glycomics'.

2. Results

To address the issues described in the introduction to this manuscript, we introduce the GLYcan Data Exchange (GLYDE) representation standard for the representation of glycan structures. GLYDE is based on the eXtensible Markup Language (XML), which is a meta language used to define other languages.[16,17] GLYDE uses XML syntax to define a representation schema for glycans, facilitating the seamless exchange and processing of glycomics data. As XML uses Unicode text, it is human-readable as well as machine-processable.[17] Thus, GLYDE was designed to fulfill the requisites of a truly useful representation standard, viz. expressiveness, compatibility with other representation schemes (namely LINUCS,[2] IUPAC–IUBMB, and CabosML[5]), uniqueness, and machine comprehensibility.

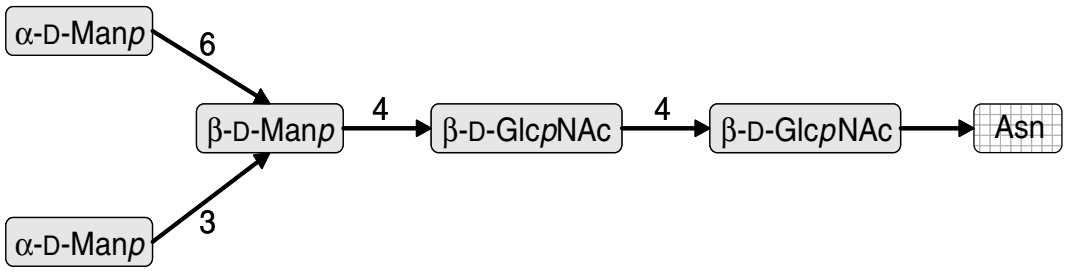
2.1. The structure of GLYDE

GLYDE uses a tree-based approach, closely related to that implemented within LINUCS, thereby bestowing it with a robust topological correspondence to the natural structure of complex carbohydrates. These polymers are composed of glycosyl residues (subunits) that are connected links that have a well-defined directionality. That is, each glycosyl residue is glycosidically linked to exactly one other residue. This may be represented as a directed graph, in which at most one edge points away from each residue node (Fig. 1). However, each glycosyl residue may have up to four other glycosyl residues linked to it, so each node in the graph may have several edges pointing toward it. Thus, a typical glycan structure corresponds to a directed acyclic graph. It is topologically equivalent to a tree in which the set of residues B {b: b is linked to a} are specified as direct children of residue a. This tree is formalized in GLYDE by defining a nested set of *residue* tags (Fig. 1) corresponding to XML elements whose attributes identify each residue and specify its physical connection to other residues. (Note that XML elements and attributes are indicated by *underlined italics* in this manuscript.) Additional tags can be added to the GLYDE schema, making it adaptable and scalable to future modifications (Fig. 2).

The root of the GLYDE tree is an XML element (*glycan*) that subsumes a single glycan structure. The *glycan* element may have several types of elements as its children (described in the next section).

2.1.1. Structural database identifiers. To maximize its interoperability, GLYDE provides element tags that point to specific entries in various databases. The structural database identifier associated with each of these element tags (listed below) is specified as the *ID* attribute of the tag. Each of these database tags may be omitted if the corresponding *ID* has not been assigned to the structure or the *ID* is not known.

- *CBANK* specifies a unique accession number in the Complex Carbohydrate Structural Database (CCSD), also known as CARBBANK. Although CARBBANK is no longer actively supported, it comprises the basis for many current carbohydrate structural databases, and CARBBANK records can still be accessed indirectly, for example, via the Kyoto Encyclopedia of Genes and Genomes (KEGG—see HYPERLINK 'http://www.genome.jp/dbget-bin/www_bfind?glycan', http://www.genome.jp/dbget-bin/www_bfind?glycan').
- *KEGG* specifies a unique identifier in the 'KEGG gl' structural database. (See, for example HYPERLINK: 'http://www.genome.jp/dbget-bin/www_bget?gl:G00009', 'http://www.genome.jp/dbget-bin/www_bget?gl:G00009'.)
- *LINUCS* specifies a unique identifier corresponding to a specific structure in the sweetdb system. (See HYPERLINK 'http://www.dkfz-heidelberg.de/spec2/sweetdb/', 'http://www.dkfz-heidelberg.de/spec2/sweetdb/'.)
- *GlycO* is a unique identifier corresponding to a specific structure in the GlycO ontology (http://lsdis.cs.uga.edu/projects/glycomics/glyco/).




```
a-D-Manp-(1-6)+
               |
        b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc-(1-4)-Asn
               |
a-D-Manp-(1-3)+
```

```
<Glycan>
  <residue >
    <residue link="4" anomeric_carbon="1" anomer="b" chirality="D" monosaccharide="GlcNAc" ring_form="p" >
      <residue link="4" anomeric_carbon="1" anomer="b" chirality="D" monosaccharide="GlcNAc" ring_form="p" >
        <residue link="4" anomeric_carbon="1" anomer="b" chirality="D" monosaccharide="Man" ring_form="p" >
          <residue link="3" anomeric_carbon="1" anomer="a" chirality="D" monosaccharide="Man" ring_form="p" >
          </residue>
          <residue link="6" anomeric_carbon="1" anomer="a" chirality="D" monosaccharide="Man" ring_form="p" >
          </residue>
        </residue>
      </residue>
    </residue>
  </residue>
</Glycan>
```

Figure 1. The tree structure of the GLYDE formalism corresponds to the natural topology of complex glycans. (Top) graphical representation of an *N*-glycan. (Middle) IUPAC representation of the same structure. (Bottom) Minimal GLYDE representation.

```
<!-- ********************************************************************************************** -->
<!-- 2. Defining the 'chirality' orientation of the residue: enumerated attribute types
        If the chirality of the molecule is not known/discovered, the default value for the
        attribute is 'null'. Also, some residues are ' meso' i.e. they do not have chirality.
        These are designated to have 'm' chirality.                                          -->
<!-- ********************************************************************************************** -->
  <!ATTLIST residue chirality (D|L|m|null) "null">
```

Figure 2. A snapshot of the GLYDE DTD document, detailing the attribute 'chirality' of the XML element 'residue' with the relevant documentation.

2.1.2. Chemical properties. The list of element tags that specify the overall chemical properties of the glycan are listed below:

- *MW* (the molecular weight) of the glycan can be expressed in two different ways, as a monoisotopic mass or as an average mass. These two values (*monoisotopic_mass* and *average_mass*) are attributes of the *MW* element.
- The *composition* tag represents the number of atoms of each chemical element in the glycan, each of which is specified by an attribute (H, Li, B, C, N, O, etc.). For example, the elemental composition $C_{50}H_{98}N_2O_{48}$ is specified as ⟨composition C = '50' H = '98' N = '2' O = '48'⟩.

2.1.3. Structural elements. The list of structural element tags that specify the chemical constituents of the glycan are listed next:

- As introduced above, *residue* is an element tag that represents the basic building blocks of the *glycan*. Each *residue* can have children that are themselves *residues*, and each child *residue* is glycosidically linked to its parent *residue*. The site on the parent *residue* to which the child *residue* is attached is specified by the *link* attribute. Each *glycan* has a single root *residue*, consistent with the tree formalism. The *monosaccharide* attribute of the residue defines the name of the monosaccharide that will be generated from the residue by hydrolysis of glycan, and therefore *monosaccharide* defines the name of the residue. The ring form of the residue is defined by using the *ring form* attribute tag. The value of the *ring form* specifies whether the residue adopts a five-membered (i.e., furanose) or six-membered (i.e., pyranose) ring form or no ring at all (i.e., an acyclic alditol form). The attribute *anomer* describes the chemical configuration of the glycosidic carbon. It is either a (alpha) or b (beta), or null (e.g., for alditol residues). The *cyclic* attribute indicates that the glycan has a macrocyclic structure: setting the *cyclic* attribute to TRUE for the current residue designates that the root residue is glycosidically linked to the current residue, making the graph (and the glycan itself) cyclic. The *link* attribute of the root *residue* defines the point of attachment (on the current *residue*) for this glycosidic linkage. That is, cyclization makes the current *residue* a pseudo-parent of the root *residue*. The *repeat* attribute indicates that part of the glycan sequence is repeated: setting the *repeat* attribute of the current *residue* to a number ($n > 1$) indicates that all descendent *residues* of the current *residue* are tandemly repeated n times. The *end repeat* attribute negates the effect of the repeat attribute for the current *residue* and all of its descendents, facilitating the specification of glycans in which internal sequences are repeated.
- The *substituent* is an element tag that, as a child of the *residue* tag, specifies that the *residue* has a specific chemical substituent. The identity of the *substituent* (e.g., an *O*-acetyl group) is specified by the *name* attribute of the *substituent* and the site on the parent *residue* to which the *substituent* is attached is specified by the *link* attribute of the *substituent*. Each *residue* can have zero or more *substituents*. Furthermore, each *substituent* may have a residue or another substituent as its child. This allows the description of structures such as teichoic acids in which carbohydrate residues are connected via non-carbohydrate substituents, such as phosphate groups.
- The *aglycon* is an element tag that identifies the chemical moiety to which the root *residue* is glycosidically attached (under the restriction that the glycan is not cyclic—see above). The aglycon is not intended to be another glycosyl residue, but rather a distinct chemical entity, such as a protein, peptide, lipid, or other non-carbohydrate moiety. The *link* attribute of the root residue specifies the molecular attachment site of this linkage on the *aglycon* (e.g., '4' for N4 of an Asn residue to which an *N*-glycan is linked). Considerable flexibility is allowed when specifying the *aglycon*: the aglycon can be a simple string (e.g., 'Asn') or URI that points to a more complex specification. For example, the URI could point to a table that describes a collection of peptide sequences, their N-glycosylation sites, and the fractional residence of the glycan at each site. Due to the flexible nature of XML, the URI could point to another section of the same document, providing a succinct way to completely describe the distribution of the glycan in a glycomics sample.

As described above for the *aglycon* element, GLYDE is designed to use Universal Resource Identifiers (URI) as references that link concepts and terms defined in the schema to external resources. This allows structural, physical, and functional features of the glycan to be specified, even if they are not explicitly defined within the GLYDE schema. Another example of this is the *GlycO ID*, which allows glycan structures defined using GLYDE to be mapped to GlycO, a focused and deep domain ontology that is designed to contain glycomics knowledge and provenance information. GlycO is designed to facilitate the correlation of specific structural features of complex glycans to their biosynthesis, metabolism, and biological relevance. GlycO embodies semantically rich descriptions of carbohydrate structures by modeling them as sets of simpler structures (i.e., individual glycosyl residues in distinct molecular environments). Ultimately, GlycO will provide a semantic description of glycan-binding relationships, glycan biosynthetic pathways, and functional relationships of glycan structure to developmental biology, although full realization of GlycO depends on the incorporation of additional instances of glycan structures and proteins that interact with them.

One advantage of the LINUCS representation of glycan structure is its capacity to define a unique representation of each distinct glycan structure. This cannot be said for the IUPAC representation. LINUCS representations are text strings that can be lexically compared and checked for equivalence, facilitating database searching and other computational tasks. In principle, GLYDE can provide unique representations of glycan structure if sibling residues are sorted using the same criteria (linkage position) employed in the LINUCS scheme. This is true for textual representation of a GLYDE structure if tags that are not explicitly used to define structural features are stripped out, but more importantly, it is true for (tree-based) representations of glycan structure that are implicit in the GLYDE formalism and readily translated to data structures generated by various XML parsers.

CabosML[5] is another XML-based format that has been proposed for the representation of carbohydrate structural data. CabosML formalizes the glycan as a tree structure in a manner that is very similar to GLYDE format. However, CabosML representations are considerably simpler than GLYDE representations, as they do not explicitly describe all of the structural features (such as absolute configuration and ring form) of each monosaccharide residue in the glycan. Rather, this information is implied from the name of the monosaccharide residue. For example, a D-Man*p* residue is implied when the CabosML monosaccharide residue attribute *name* is equal to 'Man'. Similar to GLYDE, CabosML describes the anomeric configuration and linkage of the monosaccharide residue as attributes of the monosaccharide. But, CabosML does not explicitly account for the aglycon that may be attached to the glycan; neither does it provide fields to specify identifiers in the various structural databases nor does it account for macrocyclic or repeating structures. Interconversion of CabosML to GLYDE format is straightforward using standard XML parser methodology. Conversion of CabosML to GLYDE requires a predefined mapping of each CabosML monosaccharide *name* to its full structural representation (e.g., 'Man' is mapped to 'D-Man*p*'), but this still produces an incomplete GLYDE representation. Conversely, conversion of GLYDE to CabosML is essentially 'lossy', as it requires removal of some of the explicit structural information that is not available in CabosML.

2.1.4. Interoperability of GLYDE with LINUCS and IUPAC–IUBMB. We have developed and deployed a suite of web-based tools to enable conversion of glycan structures described using the two popular representations into GLYDE, using the following chain of conversion:

IUPAC–IUBMB > LINUCS > GLYDE

The resource is available at URL http://128.192.9.86/stargate/formatIndex.jsp.

The tools allow users to upload an existing file containing glycan structure represented using either IUPAC–IUBMB or LINUCS. Also, the user may manually input relevant glycan structures in the data input section of the tool to be converted into the next representation format (according to the conversion chain described above). We detail the two subcomponents of the conversion chain suite:

(a) IUPAC to LINUCS: We have implemented a wrapper application that utilizes the IUPAC–IUBMB to LINUCS tool available on the SWEETDB site (http://www.glycosciences.de/linucs/).
(b) LINUCS to GLYDE: We use java servlet technology to convert a LINUCS-based glycan structure representation to an XML format, conforming to the GLYDE XML-schema.

As future enhancements, we will include utilities to interconvert GLYDE and CabosML representations and create a Web process constituted of multiple Web services to implement the interconversion of GLYDE, LINUCS, and CabosML.

3. Discussion

XML-based meta languages have become the touchstone in the biosciences domain for data structure representation. The Systems Biology Markup Language (SBML),[7] Genome Annotation Markup Elements (GAME),[14] and Human Proteome Markup Language for Proteomics Database (HUP-ML)[13] are some significant examples. We believe GLYDE similarly fulfills an important requirement for rapid evolution of glycomics by providing a common standard for glycomics data representation. We developed GLYDE as a flexible representation, keeping in mind the requirements of end users—glycobiologists, who may use the GLYDE framework intuitively to use and understand glycans represented in GLYDE.

To ensure adaptability of GLYDE to modifications and extensions, we have made the XML Data Type Definition (DTD) and XML Schema web accessible (URL: http://lsdis.cs.uga.edu/projects/glycomics/index.php?page=4). The DTD is richly annotated with extensive and comprehensive documentation so that users can understand the structure of GLYDE and when needed, make changes to it. Such changes will clearly be required for GLYDE to fulfill the diverse requirements of glycan representation that will arise in the rapidly evolving field of glycoinformatics. We are actively seeking input from scientists who are interested the development of an XML-based standard for the representation of glycan structures and are currently collaborating in this context with scientists associated with the EuroCarbDB design study (http://www.eurocarbdb.org/about). The current focus of this collaborative effort is the establishment of standards for naming carbohydrate residues, and development of protocols for describing glycans whose structures are incompletely or ambiguously defined.

Acknowledgments

GLYDE was developed as part of the Integrated Technology Resource for Biomedical Glycomics (5 P41 RR18502-02) funded by the National Institutes of Health National Center for Research Resources.

References

1. Loß, A.; Bunsmann, P.; Bohne, A.; Loß, A.; Schwarzer, E.; Lang, E.; von der Lieth, C. W. *Nucleic Acids Res.* **2002**, *30*, 405–408.
2. Bohne-Lang, A.; Lang, E.; Forster, T.; von der Lieth, C. W. *Carbohydr Res.* **2001**, *336*, 1–11.
3. von der Lieth, C. W.; Bohne-Lang, A.; Lohmann, K. K.; Frank, M. *Brief Bioinform.* **2004**, *5*, 164–178.
4. Karp, P. D. *J. Comput. Biol.* **1995**, *2*, 573–586.
5. Kikuchi, N.; Kameyama, A.; Nakaya, S.; Ito, H.; Sato, T.; Shikanai, T.; Takahashi, Y.; Narimatsu, H. *Bioinformatics* **2005**, *21*, 1717–1718.
6. W3C Semantic Web Activity, http://www.w3.org/2001/sw/.
7. Hucka, M.; Finney, A.; Bornstein, B. J.; Keating, S. M.; Shapiro, B. E.; Matthews, J.; Kovitz, B. L.; Schilstra, M. J.; Funahashi, A.; Doyle, J. C.; Kitano, H. *Syst. Biol.* **2004**, *1*, 41–53.
8. Foster, I. *Science* **2005**, *308*, 814–817.
9. Fenyo, D.; Beavis, R. C. *Trends Biotechnol.* **2002**, *20*, S35–S38.
10. Crampin, E. J.; Halstead, M.; Hunter, P.; Nielsen, P.; Noble, D.; Smith, N.; Tawhai, M. *Exp. Physiol.* **2003**, *89*, 1–26.
11. Sahoo, S. S.; Sheth, A. P.; York, W. S.; Miller, J. A. Semantic Web Services for N-Glycosylation Process, International Symposium on Web Services for Computational Biology and Bioinformatics, VBI, Blacksburg, VA, May 26–27, 2005.
12. Taylor, C. F.; Paton, N. W.; Garwood, K. L.; Kirby, P. D.; Stead, D. A.; Yin, Z.; Deutsch, E. W.; Selway, L.; Walker, J.; Riba-Garcia, I.; Mohammed, S.; Deery, M. J.; Howard, J. A.; Dunkley, T.; Aebersold, R.; Kell, D. B.; Lilley, K. S.; Roepstorff, P.; Yates, J. R., III; Brass, A.; Brown, A. J.; Cash, P.; Gaskell, S. J.; Hubbard, S. J.; Oliver, S. G. *Nature Biotechnol.* **2003**, *21*, 247–254.
13. Kamijo, K.; Mizuguchi, H.; Kenmochi, A.; Sato, M.; Takaki, Y.; Tsugita, A. *J. Mass Spectrom. Soc. Jpn.* **2003**, *51*, 542–549.
14. Genome Annotation Markup Elements (GAME), http://www.bioxml.org/Projects.
15. W3C Web Services Activity http://www.w3.org/2002/ws/.
16. Extensible Markup Language (XML), http://www.w3.org/XML/.
17. Monson-Haefel, R. In *The Ultimate Guide: J2EE Web Services*; Addison-Wesley: Boston, 2004; pp 17–25.