
Letter to Glycoforum

GlyTouCan: an accessible glycan structure repository

Michael Tiemeyer², Kazuhiro Aoki², James Paulson³,
Richard D Cummings⁴, William S York², Niclas G Karlsson⁵,
Frederique Lisacek⁶, Nicolle H Packer^{7,8}, Matthew P Campbell⁷,
Nobuyuki P Aoki⁹, Akihiro Fujita⁹, Masaaki Matsubara²,
Daisuke Shinmachi⁹, Shinichiro Tsuchiya⁸, Issaku Yamada¹⁰,
Michael Pierce², René Ranzinger², Hisashi Narimatsu¹¹,
and Kiyoko F Aoki-Kinoshita^{9,1}

²Complex Carbohydrate Research Center, The University of Georgia, 315 Riverbend Rd, Athens, Georgia 30602, USA,

³The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA, ⁴Harvard Medical School, 330 Brookline Ave, Room SL-0408, Boston, MA 02115, USA, ⁵University of Gothenburg, Box 440, 40530 Gothenburg, Sweden, ⁶Swiss Institute of Bioinformatics, CUI - 7, route de Drize, CH-1211 Geneva, Switzerland, ⁷Institute for Glycomics, Gold Coast Campus, Griffith University, Parklands Drive, Gold Coast, QLD 4222, Australia, ⁸Macquarie University, Balacava Road, North Ryde, NSW 2109, Australia, ⁹Soka University, 1-236 Tangi-machi, Hachioji, Tokyo 192-8577, Japan, ¹⁰The Noguchi Institute, 1-9-7, Kaga, Itabashi-ku, Tokyo 173-0003, Japan, and ¹¹National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-0046, Japan

¹To whom correspondence should be addressed: Tel/Fax: +81-42-691-4116; e-mail: kkiyoko@soka.ac.jp

Abstract

Rapid and continued growth in the generation of glycomic data has revealed the need for enhanced development of basic infrastructure for presenting and interpreting these datasets in a manner that engages the broader biomedical research community. Early in their growth, the genomic and proteomic fields implemented mechanisms for assigning unique gene and protein identifiers that were essential for organizing data presentation and for enhancing bioinformatic approaches to extracting knowledge. Similar unique identifiers are currently absent from glycomic data. In order to facilitate continued growth and expanded accessibility of glycomic data, the authors strongly encourage the glycomics community to coordinate the submission of their glycan structures to the GlyTouCan Repository and to make use of GlyTouCan identifiers in their communications and publications. The authors also deeply encourage journals to recommend a submission workflow in which submitted publications utilize GlyTouCan identifiers as a standard reference for explicitly describing glycan structures cited in manuscripts.

Key words: database, glycan identifier, GlyTouCan, repository, structure

A bottleneck in glycomics

The inherent complexity of defining glycan structures and the essential importance of understanding how the glycome influences biological function requires the integration of data from multiple disciplines, including cell biology, genetics and molecular biology, as

well as structural, analytic and synthetic chemistry. The diverse nature of the experimental approaches underlying these disciplines predicts the need for a common language to communicate glycan structures. In addition, technologic advances across a broad range of analytic approaches, including mass spectrometry, liquid

chromatography, capillary electrophoresis and other orthogonal strategies, are generating increasingly expansive glycomic datasets. While this accumulating wealth of diverse datasets presents great opportunities for understanding glycan functions and structural diversity, it also reveals the need for growth in bioinformatic infrastructure to support and enhance data interpretation.

Among “omic” analyses, there is a need for glycomics to catch up with more established fields like genomics and proteomics. Glycomics currently faces many of the same obstacles that genomics and proteomics resolved successfully in the last two decades. Namely, the adoption of standards for data annotation, data interpretation, data presentation, data archiving and database structure have allowed genomics and proteomics to advance rapidly (MIAME and MIAPE); similar standards for glycomics (MIRAGE) are developing robustly along an accelerating trajectory with the development of MS, glycan array and sample preparation guidelines but are still passing through their early growth stages (Brazma et al. 2001; Taylor et al. 2007; Kolarich et al. 2013; York et al. 2014; Struwe et al. 2016; Liu et al. 2017).

Major advances in genomics and proteomics were achieved over the last two decades by community acceptance of unique identifiers for genes and proteins. Unique identifiers allow authors to submit and cite unambiguous references to gene sequences, mRNA sequences, translated protein sequences, and explicit protein structures, enhancing the ability of investigators to interrogate published reports and utilize this data to advance their own research. Gene and protein identifiers also facilitated database development by providing interconnectivity and cross-referencing capabilities. Without the development of a similar, broadly accepted infrastructure for submitting and citing glycan structures, the glycomics community will remain handicapped by the need for each individual investigator or groups of investigators to separately generate their own descriptors for publication and communication. While the standardization of graphical representations of glycan structures has enjoyed broad acceptance within the glycomics community (SNFG, symbol nomenclature for glycans), the adoption of SNFG representations for publication does not ensure that glycomics datasets will be accessible or can be interrogated by current or developing database efforts (Varki et al. 2015).

Glycan identifiers, written in a broadly accepted, machine-readable language, will allow authors and investigators to point to explicit glycan structures in their publications and will facilitate the expansion of current knowledge databases through streamlined incorporation of new glycomic discoveries. Full structural characterization of a new or known glycan generally requires multiple analytic approaches, some of which are not amenable to the amounts of material available from many biologic systems. Therefore, unlike genomics or proteomics, unique glycan structure identifiers must be able to incorporate ambiguity to be broadly useful. A unique glycan identifier should register a structure at the level of structural resolution submitted by the experimentalist, who can then use this identifier to reference their work. And, importantly, consumers of this work will be able to use these identifiers to appreciate the level of resolution of the submitted structures in light of the published experimental approaches.

Development and implementation of a solution

The Complex Carbohydrate Structure Database and CarbBank were early attempts to implement useful identifiers for explicitly describing glycan structures (Doubet et al. 1989; Doubet and Albersheim 1992). By the time funding support for CarbBank ended in the late 1990s,

other efforts (CFG, KEGG, JCGGDB, BCSDb) had been initiated around the world, resulting in a proliferation of database identifiers (Hashimoto et al. 2006; Raman et al. 2006; Toukach et al. 2007). GlycomeDB undertook the interconnection and consolidation of multiple databases, potentially providing an opportunity to assign unique identifiers to database entries (Ranzinger et al. 2008). However, the goals of GlycomeDB and other databases are nobly and ambitiously much broader than simply generating unique identifiers. These goals include the capture of as much metadata and analytic structural validation as possible, inevitably requiring significant database curation and expert intervention. The additional responsibility of such oversight is likely to hinder the rapid assignment, archiving and dissemination of unique identifiers for explicit glycan structures, which may be defined at various levels of ambiguity. However, computer algorithms can perform the core function of assigning identifiers to structures with little human intervention, reducing maintenance costs and fostering continuity over the long-term. The durability of the identifiers provided by such a stable core resource is a key requirement for their use as the semantic foundation for mapping, integrating and correlating the data and metadata compiled in more extensive and diverse databases.

Beginning with discussions at the 4th Warren Workshop in Athens, GA in 2012, and continuing through the 5th ACGG-DB (Asian Consortium for Glycobiology and Glycotechnology) meeting in Dalian, China in 2013, the glycomics and glycobioinformatics communities agreed with renewed vigor that unique glycan identifiers were needed in order to enhance data sharing in publications and across database platforms (Aoki-Kinoshita et al. 2013). A consensus was reached that a stand-alone, internationally recognized glycan structure repository should be developed. The function of the repository would be solely to assign unique identifiers to submitted glycan structures and to store minimal metadata, limited only to submitter and submission date/time for each accession number (Aoki-Kinoshita et al. 2016). The simple functions of such a repository would require minimal human management, allowing immediate assignment of accession numbers.

To this end, GlyTouCan was developed and deployed as a website (<http://glytoucan.org>) in 2015. At the 6th Warren Workshop in Sapporo, Japan in 2016, the assembled glycomics community was presented with the functionalities of the current GlyTouCan release (version 2). Ensuing discussion led to a consensus of support for the broadest possible acceptance of GlyTouCan as the essential resource for obtaining unique identifiers for glycan structures and as the international glycan structure repository of choice.

Based on the pressing need for generating unique glycan structure identifiers that possess appropriate utility for informatics platforms and sufficient stability for journal publications, the authors and the undersigned concurring colleagues strongly endorse GlyTouCan as an accepted, international repository for glycan structures. In order to fully realize the potential of GlyTouCan, the authors encourage the community to submit glycan structures to the repository and use the assigned identifiers in their submitted manuscripts. The authors further recommend that journals endorse the same goal, with the objective of eventually incorporating structural submission to GlyTouCan as an expected part of the manuscript submission workflow for manuscripts that describe the structure or function of glycans.

Enhanced GlyTouCan functionalities are already in place or under development that do or will interconnect structure repository identifiers with glycomic databases such as GlycomeDB (Ranzinger et al. 2008), Carbohydrate Structure Database CSDB (Toukach and Egorova 2016), GlycoEpitope (Okuda et al. 2017), GlycoNAVI (<http://glyconavi.org>),

UniCarb-DB (Hayes et al. 2011), SugarBindDB (Mariethoz et al. 2016) and UniCarbKB (Campbell and Packer 2016); these linked databases will continue to be the appropriate repository for expansive metadata and analytic data that validates structural assignments. The acceptance of GlyTouCan within the glycomunity allows these databases to link their metadata to specific glycan identifiers (Aoki-Kinoshita et al. 2016). These databases, known as GlyTouCan Partners, can directly register structures into GlyTouCan. Therefore, any associated metadata for a “new” glycan can be submitted to a GlyTouCan Partner, which will subsequently register structures to provide a GlyTouCan link to the associated metadata. After GlyTouCan registration, a submitter will be able to link deposited structures to accepted publications. Registration directly to GlyTouCan can be performed either as a single structure or using batch downloads. Glycan drawing tools derived from familiar resources such as GlycanBuilder facilitate the submission.

Using GlyTouCan

GlyTouCan (Figure 1) has been developed to be as user-friendly as possible. It provides an intuitive portal for searching and depositing structures. Note that glycans with unknown linkages, glycans known only as monosaccharide compositions (e.g. Gal2GlcNAc4Man2), and even glycan compositions with undefined monosaccharides (e.g. Hex4HexNAc4) can all be registered and retrieved.

Searching glycan structures in GlyTouCan

Glycans can be searched by either (1) browsing through the list of registered glycans or (2) specifying a particular glycan (sub)-structure and querying for similar registered structures. The “Glycan List” option under “View All” provides functionality to allow the user to filter down the list of glycans to search. Figure 2 is a snapshot of the full list that is shown initially after choosing the Glycan List option. Here, the list can be filtered by selecting structural components of the



Fig. 1. GlyTouCan Logo. GlyTouCan is named by a combination of “Glycan” and “Tou (糖)”, which means “sugar” in Japanese. The word “Can (Jpn. 缶)” of GlyTouCan signifies that it is a container where glycans are accumulated. Therefore, a can is used in the logo to indicate that it is a repository with accumulated glycan data. In addition, a toucan is used as a mascot character since the word “TouCan” is included in GlyTouCan. The GlyTouCan toucan sits, appropriately, on the branch of a Lewis type glycan structure. This figure is available in black and white in print and in color at *Glycobiology* online.

glycans that are being searched for, such as Motif (e.g. “Sialyl-Lewis” or “Lactosamine”) or Monosaccharide component. A list of Databases is also available if the target glycan is known to be stored in a particular database. Moreover, a mass range can be specified to filter the Glycan List by mass.

The second popular search option is the “Graphic Input” option under the “Search” menu, which allows the user to draw their target glycan, using GlycanBuilder (Tsuchiya et al. 2017), and subsequently use it as a query. The resulting list of matching similar glycans will be shown, and if the query is already registered, its accession number will be displayed.

Depositing glycan structures to GlyTouCan

It is possible to deposit a glycan structure directly to GlyTouCan by signing in to a Google account. There is no need to enter or remember a new password as long as the user has a Google account (other types of accounts will be supported in the future). After signing in, a “Registration” menu will be displayed, via which glycan structure(s) can be added using Graphic Input (similar to Graphic Search via GlycanBuilder), Text Input using GlycoCT{condensed} or WURCS format, or File Upload. Every submitted structure will first be compared with the existing GlyTouCan registrations to ensure that duplicate deposits are not generated. A confirmation screen is shown (1) to list existing GlyTouCan IDs for those that are already registered and (2) to display images of the new structure(s) that will be registered.

Future perspectives

As has been true for all database efforts over the past 30 years, those that are heavily used and that prove to be most useful (e.g. UniProt, GenBank and PDB) are likely to achieve stable funding and long-term support. It will ultimately be in the hands of the glycomics community to demonstrate that GlyTouCan is essential infrastructure worth the continued investment of financial resources.

Existing proteomic databases (e.g. Uniprot and PDB) offer minimal characterization of glycoproteins and currently indicate only the positions of predicted or experimentally validated N- and O-linked glycosylation sites or GPI-anchor attachment sites. As unique GlyTouCan identifiers are integrated with existing proteomic resources, future queries will allow elucidation of the functional significance of specific glycan features across or within protein families to permit advances in the biomedical application of glycomics.

Moreover, with the inclusion of glycan identifiers in other omics databases, the authors and Supporting Investigators anticipate that glycomics data will achieve ever higher visibility and enhanced appreciation within the life science research community.

Supporting Investigators as of May, 2017 (in alphabetical order of last name)

1. Friedrich Altmann, University of Natural Resources and Life Sciences, Vienna, Austria
2. Antony Bacic, University of Melbourne, Australia
3. Christopher B. Barnett, University of Cape Town, South Africa
4. Júlia Costa, Laboratory of Glycobiology, ITQB NOVA, Portal
5. Vivien J. Coulson-Thomas, University of Houston, USA
6. Tamara L. Doering, Washington University School of Medicine, USA
7. Nathan Edwards, Georgetown University, USA
8. Michiko Ehara, Asahi University, Japan

Fig. 2. Glycan List for browsing/filtering glycans registered in GlyYouCan. This interface is accessible through the pull-down menu under View All on the GlyYouCan home page and provides a rapid, intuitive way to interrogate existing entries. This figure is available in black and white in print and in color at *Glycobiology* online.

9. Tamao Endo, Tokyo Metropolitan Institute of Gerontology, Tokyo, Japan
10. Ten Feizi, Imperial College London, UK
11. Martin Frank, Biognos AB, Sweden
12. Morihisa Fujita, Jiangnan University, China
13. Koichi Fukase, Osaka University, Japan
14. Yuzuru Ikehara, AIST and Chiba University, Japan
15. Makoto Ito, Kyushu University, Japan
16. Yukishige Ito, RIKEN, Japan
17. Kenji Kadomatsu, Nagoya University Graduate School of Medicine, Japan
18. Osamu Kanie, Tokai University, Japan
19. Takane Katayama, Kyoto University, Japan
20. Toshiyuki Kawasaki, Ritsumeikan University, Japan
21. Hiroto Kawashima, Chiba University, Japan
22. Carsten Kettner, Beilstein Institut, Germany
23. Kshitij Khatri, Boston University, USA
24. Yoshinobu Kimura, Okayama University, Japan
25. Hiroshi Kitagawa, Kobe Pharmaceutical University, Japan
26. Shinobu Kitazume, RIKEN, Japan
27. Yuriy A. Knirel, N.D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Moscow, Russia
28. Kyoko Kojima-Aikawa, Ochanomizu University, Japan
29. Daniel Kolarich, Griffith University, Australia
30. Matthew R. Kudelka, Emory University, USA
31. Todd L. Lowary, Canadian Glycomics Network Scientific Director and University of Alberta, Canada
32. Thomas Lueteteke, ITech Progress GmbH, Germany
33. Shino Manabe, RIKEN, Japan
34. David Matten, University of Cape Town, South Africa
35. Raja Mazumder, George Washington University, USA
36. Eiji Miyoshi, Osaka University, Japan
37. Antonio Molinaro, University of Napoli Federico II, Italy
38. Yasu S. Morita, University of Massachusetts Amherst, USA
39. Toni M. Mueller, University of Alabama at Birmingham, USA
40. Shunji Natsuka, Niigata University, Japan
41. Shoko Nishihara, Soka University, Japan
42. Sriram Neelamegham, State University of New York, USA
43. Tetsuya Okajima, Nagoya University School of Medicine, Japan
44. Shujiro Okuda, Niigata University, Japan

45. Noorjahan Panjwani, Tufts University School of Medicine, USA
46. Dayoung Park, University of California, Davis, USA
47. Serge Perez, France
48. Salomé S. Pinho, University of Porto and Institute for Research and Innovation in Health, Portugal
49. Melody Porterfield, University of Georgia, USA
50. Alka Rao, CSIR-Institute of Microbial Technology, Chandigarh, India
51. Celso A. Reis, University of Porto, Portugal
52. Sylvie Ricard-Blum, University of Lyon 1, France
53. Rafael Ricci de Azevedo, University of Sao Paulo, Brazil
54. Nancy Schwartz, University of Chicago, USA
55. Siro Simizu, Keio University, Japan
56. Avadhesh Surolia, Indian Institute of Science, Bangalore, India
57. Naoyuki Taniguchi, RIKEN, Japan
58. Carlo Unverzagt, University of Bayreuth, Germany
59. Ajit Varki, University of California, San Diego, USA
60. Masahiro Wakao, Kagoshima University, Japan
61. Christopher M. West, University of Georgia, USA
62. Robert J. Woods, University of Georgia, USA
63. Ajit Varki, University of California, San Diego, USA
64. Yoshiki Yamaguchi, RIKEN, Japan
65. Kazuo Yamamoto, The University of Tokyo, Japan
66. Heng Yin, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, China
67. Joseph Zaia, Boston University, USA

Funding

Integrated Database Project sponsored by the Japan Science and Technology Agency (JST) and the National Bioscience Database Center of Japan to GlyTouCan. Contributions of The National Center for Biomedical Glycomics (Grant P41GM103490) and the National Center for Functional Glycomics (Grant P41GM103694) were supported by The National Institute of General Medical Sciences, a part of the United States National Institutes of Health.

Conflict of interest statement

None declared.

Abbreviation

SNFG, Symbol Nomenclature for Glycans.

References

- Aoki-Kinoshita K, Agravat S, Aoki NP, Arpinar S, Cummings RD, Fujita A, Fujita N, Hart GM, Haslam SM, Kawasaki T et al. 2016. GlyTouCan 1.0—The international glycan structure repository. *Nucleic Acids Res.* 44:D1237–D1242.
- Aoki-Kinoshita KF, Sawaki H, An HJ, Campbell MP, Cao Q, Cummings R, Hsu DK, Kato M, Kawasaki T, Khoo K-H et al. 2013. The fifth ACGG-DB meeting report: Towards an International Glycan Structure Repository. *Glycobiology.* 23:1422–1424.
- Brazna A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet.* 29:365–371.
- Campbell MP, Packer NH. 2016. UniCarbKB: New database features for integrating glycan structure abundance, compositional glycoproteomics data, and disease associations. *Biochim Biophys Acta.* 1860(8):1669–1675.
- Doubet S, Albersheim P. 1992. CarBANK. *Glycobiology.* 2:505.
- Doubet S, Bock K, Smith D, Darvill A, Albersheim P. 1989. The complex carbohydrate structure database. *Trends Biochem Sci.* 14:475–477.
- Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita K, Ueda N, Hamajima M, Kawasaki T, Kanehisa M. 2006. KEGG as a glycome informatics resource. *Glycobiology.* 16(5):63R–70R.
- Hayes CA, Karlsson NG, Struwe WB, Lisacek F, Rudd PM, Packer NH, Campbell MP. 2011. UniCarb-DB: A database resource for glycomic discovery. *Bioinformatics.* 27(9):1343–1344.
- Kolarich D, Rapp E, Struwe WB, Haslam SM, Zaia J, McBride R, Agravat S, Campbell MP, Kato M, Ranzinger R et al. 2013. The minimum information required for a glycomics experiment (MIRAGE) project: Improving the standards for reporting mass-spectrometry-based glycoanalytic data. *Mol Cell Proteomics.* 12(4):991–995.
- Liu Y, McBride R, Stoll M, Palma AS, Silva L, Agravat S, Aoki-Kinoshita KF, Campbell MP, Costello CE, Dell A et al. 2017. The minimum information required for a glycomics experiment (MIRAGE) project: Improving the standards for reporting glycan microarray-based data. *Glycobiology.* 27(4):280–284.
- Mariethoz J, Khatib K, Alocci D, Campbell MP, Karlsson NG, Packer NH, Mullen EH, Lisacek F. 2016. SugarBindDB, a resource of glycan-mediated host–pathogen interactions. *Nucleic Acids Res.* 44(D1):D1243–D1250.
- Okuda S, Nakao H, Kawasaki T. 2017. GlycoEpitope. In: Aoki-Kinoshita KF, editor. *A Practical Guide to Using Glycomics Databases.* Tokyo, Japan: Springer Japan. p. 227–245.
- Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R. 2006. Advancing glycomics: Implementation strategies at the Consortium for Functional Glycomics. *Glycobiology.* 16(5):82R–90R.
- Ranzinger R, Herget S, Wetter T, von der Lieth CW. 2008. GlycomeDB—Integration of open-access carbohydrate structure databases. *BMC Bioinformatics.* 9:384.
- Struwe WB, Agravat S, Aoki-Kinoshita KF, Campbell MP, Costello CE, Dell A, Feizi T, Haslam SM, Karlsson NG, Khoo KH et al. 2016. The minimum information required for a glycomics experiment (MIRAGE) project: Sample preparation guidelines for reliable reporting of glycomics datasets. *Glycobiology.* 26(9):907–910.
- Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RKJr., Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW et al. 2007. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol.* 25:887–893.
- Toukach P, Egorova K. 2016. Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. *Nucleic Acids Res.* 44(D1):D1229–D1236.
- Toukach P, Joshi H, Ranzinger R, Knirel Y, von der Lieth CW. 2007. Sharing of worldwide distributed carbohydrate-related digital resources: Online connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de. *Nucleic Acids Res.* 35(Database issue):D280–D286.
- Tsuchiya S, Aoki NP, Shinmachi D, Matsubara M, Yamada I, Aoki-Kinoshita KF, Narimatsu H. 2017. Implementation of GlycanBuilder to draw a wide variety of ambiguous glycans. *Carbohydr Res.* 445:104–116.
- Varki A, Cummings RD, Aebi M, Packer NH, Seeberger PH, Esko JD, Stanley P, Hart G, Darvill A, Kinoshita T et al. 2015. Symbol nomenclature for graphical representations of glycans. *Glycobiology.* 25:1323–1324.
- York WS, Agravat S, Aoki-Kinoshita KF, McBride R, Campbell MP, Costello CE, Dell A, Feizi T, Haslam SM, Karlsson N et al. 2014. MIRAGE: The minimum information required for a glycomics experiment. *Glycobiology.* 24:402–406.