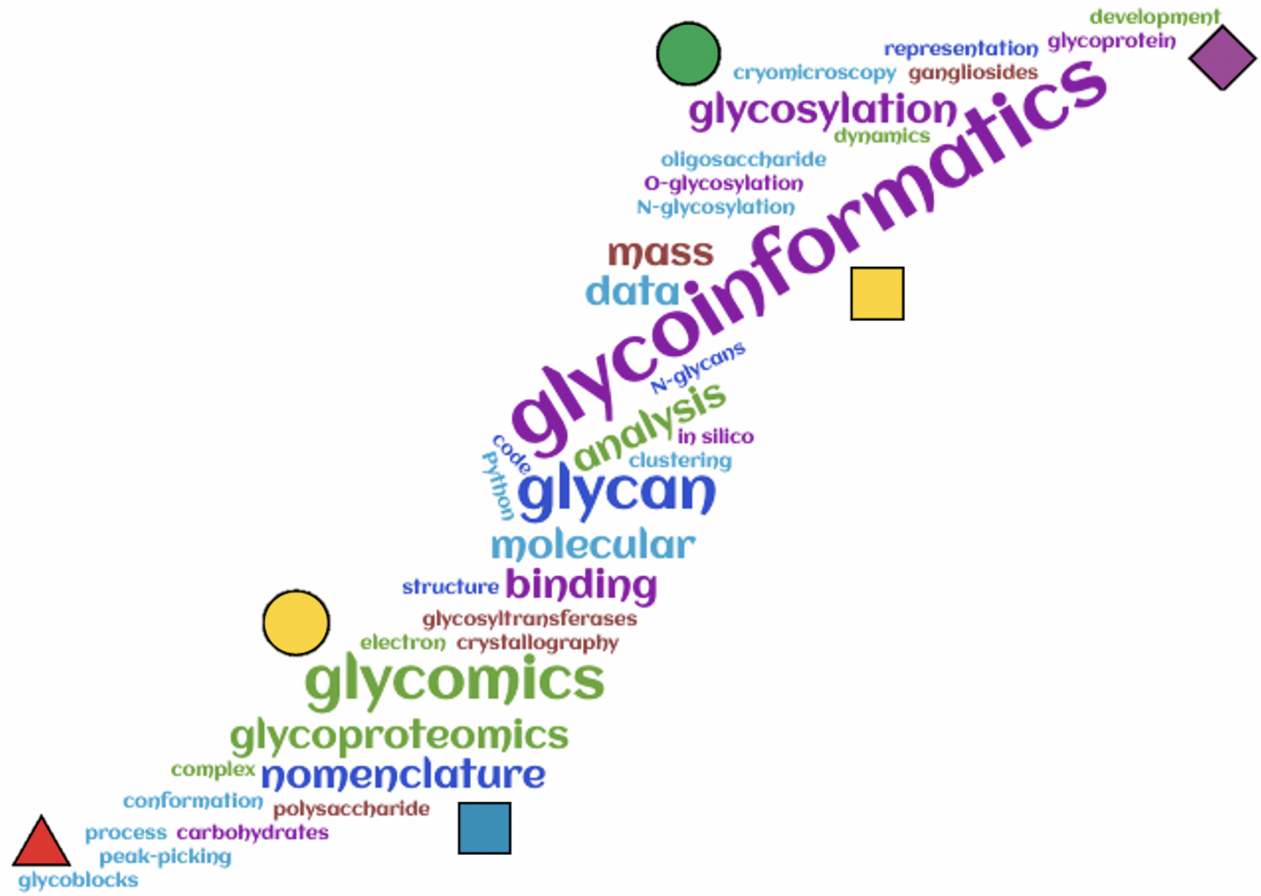
BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

GlycoBioinformatics

Edited by Kiyoko F. Aoki-Kinoshita, Frédérique Lisacek, Niclas Karlsson, Daniel Kolarich and Nicolle Packer



Generated on 08 February 2026, 11:05

Imprint

Beilstein Journal of Organic Chemistry
www.bjoc.org
ISSN 1860-5397
Email: journals-support@beilstein-institut.de

The *Beilstein Journal of Organic Chemistry* is published by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften.

Beilstein-Institut zur Förderung der Chemischen Wissenschaften
Trakehner Straße 7–9
60487 Frankfurt am Main
Germany
www.beilstein-institut.de

The cover image shows a word cloud describing the key concepts presented in this thematic issue. The cloud is in the shape of a monosaccharide, which is outlined by SNFG representations of major human monosaccharides.

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

GlycoBioinformatics

Kiyoko F. Aoki-Kinoshita[1], Frédérique Lisacek[2], Niclas Karlsson[3,4], Daniel Kolarich[5] and Nicolle H. Packer[*6]

Editorial

Open Access

Address:
[1]Faculty of Science and Engineering, Soka University, 1-236 Tangi-machi, Hachioji-shi, Tokyo, Japan, [2]University of Geneva and Swiss Institute of Bioinformatics, CUI - 7, route de Drize, 1211 Geneva, Switzerland, [3]Department of Medical Biochemistry and Cell Biology, University of Gothenburg, Box 440, 40530 Gothenburg, Sweden, [4]Faculty of Health Sciences, Department of Life Sciences and Health, Pharmacy, Oslo Metropolitan University, 0167 Oslo, Norway, [5]Griffith University, Gold Coast Campus, Southport, Queensland 4222, Australia and [6]Department of Molecular Sciences, Macquarie University, Sydney, New South Wales, Australia

Email:
Nicolle H. Packer[*] - nicki.packer@mq.edu.au

* Corresponding author

Keywords:
bioinformatics; glycobioinformatics; glycoinformatics

Beilstein J. Org. Chem. **2021,** *17,* 2726–2728.
https://doi.org/10.3762/bjoc.17.184

Received: 19 October 2021
Accepted: 27 October 2021
Published: 09 November 2021

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editors: K. F. Aoki-Kinoshita, F. Lisacek, N. Karlsson, D. Kolarich and N. H. Packer

In order to introduce this thematic issue “GlycoBioinformatics” [1] in the *Beilstein Journal of Organic Chemistry*, it would be appropriate to define what we actually mean by this term. This is important not only for newcomers to the field but also in order for researchers that have used or developed “glycobioinformatics” to place their work into a wider context of this diverse field. The term “bioinformatics” is described by the National Human Genome Research Institute as “a subdiscipline of biology and computer science concerned with the acquisition, storage, analysis, and dissemination of biological data, most often DNA and amino acid sequences”. Adding the prefix “glyco-“ is about placing genomic and proteomic data into a glycomic context by harvesting information about glyco-related genes and proteins. Glycobioinformatics requires additional information about the expressed glycan, including but not limited to monosaccharide composition, full or partial sequence including linkage and branching structure, type and linkage of glycoconjugate (e.g., N-linked, O-linked glycoprotein, glycolipid, proteoglycan), association with, and regulation of, expression in particular tissues or cell types, and interaction with biological surroundings. With this definition, it is obvious that glycobioinformatics is tightly connected to mainstream bioinformatics. For example, databases and tools from genomics can be used for gaining information about genes encoding for glycosyltransferases, glycosidases, and glycan-binding proteins (lectins), and search engines initially designed for the detection of posttranslational modifications of peptides in proteomics can be adapted to specifically identify glycopeptides. What is also obvious for glycobioinformatics is that it needs an own language that is understood by both computers and researchers to facilitate the exchange of glyco-specific information as well as the development and evolution of dedicated databases that store glyco-related quality information. With glycobioinformatics still being in its infancy, these requirements are continuing to evolve.

The editors of this thematic issue represent both bioinformatics developers as well as users who have the conviction that impor-

tant life science questions more often than not include an element of "glyco". For this thematic issue, we have assembled publications from world-renowned glycoscience researchers who are involved in the current state-of-the-art glycobioinformatics approaches that are needed to find solutions for current global health challenges and to understand just about every biological process.

Molecular dynamic modeling to understand how glycans interact with biomolecules visualizes and allows the development of hypotheses regarding the function of glycans to be tested at a molecular level. The article by Barnett et al. [2] uses molecular dynamics to show that O-linked glycosylation alters peptide conformation, which influences the binding of the peptides to antibodies, despite the fact that glycans are not directly involved in the binding. Another molecular modeling article by Fogarty et al. [3] suggests a new concept of glycoblocks, which are subunits of 3D glycan structures. This concept may become useful in describing specific epitopes and functional units of glycans. With the recent pandemic experience, the need for glycobioinformatics for global health was highlighted, where the laboratory of one of the authors of this article, Fadda, used glyco-adapted molecular dynamics to explain in a separate publication [4] how the COVID-19 spike protein recognition element requires N-linked glycosylation to be exposed. Another approach to understanding glyco-interactions is described in a review paper by Mehta et al. [5], who summarize recent developments and available online resources for glycan array data, a very powerful technique for understanding the structural element(s) of glycans required for different lectin binding. This further emphasizes the role of glycans as mediators of cellular communication.

For newcomers and experienced glycoscience researchers, the review by Lal et al. [6] is a helpful guide to resources currently available for displaying glycan structures in 2D and 3D for scientific publications and presentations. The evolution of the "glyco" language is illustrated by Kellman et al. [7], wherein glycan substrate specificities and glycoenzyme reaction rules are described using an improved linear code that is standardized for use in analytical computational tools. This links with McDonald and Davey's paper [8], which expands on their previously described theoretically derived protein O-linked glycome based on the specificity of mammalian glycoenzymes, in order to generate a theoretical glycolipid glycome.

One of the main tasks of glycobioinformatics is to convert analytical data obtained from biological samples (cell lysates, tissues, isolated proteins) into glycoscience knowledge. Most structural data at this stage is generated by analytical approaches, such as mass spectrometry (MS), high-pressure liquid chromatography (HPLC), and capillary electrophoresis (CE). The articles by Phung et al. [9] and by Lippold et al. [10] suggest ways of combining and customising available MS data analysis tools for glycoproteomic characterization and quantification. The article by Walsh et al. [11], on the other hand, addresses the problems of an irreproducible retention time and peak integration in antibody glycomic analysis using CE, thus allowing small quantitative differences to be detected when comparing similar glycomes by this method.

The articles by Groth et al. [12] and by Bagdonas et al. [13] illustrate how glycoinformation can be harvested and integrated from available -omics databases, with the former paper identifying putative cell signaling molecules and transcription factors using next-generation sequencing expression data of glycoenzymes in cancer cell lines. The latter paper uses knowledge from current open access glycomic databases to curate and validate glycan structures reported on proteins in the Protein Data Bank (PDB) database.

Overall, the wide breadth of glycobioinformatics articles that comprises this special issue only captures a snapshot of the impact that glycosciences and glycobioinformatics is now having across diverse scientific fields. These exciting results indicate the great progress that has been made and illustrates the huge potential for novel developments being made in this rather newly recognized field of life sciences.

Kiyoko F. Aoki-Kinoshita, Frédérique Lisacek, Niclas Karlsson, Daniel Kolarich and Nicolle H. Packer

Tokyo, Geneva, Gothenburg, Southport, Sydney, October 2021

ORCID® iDs

Kiyoko F. Aoki-Kinoshita - https://orcid.org/0000-0002-6662-8015
Frédérique Lisacek - https://orcid.org/0000-0002-0948-4537
Niclas Karlsson - https://orcid.org/0000-0002-3045-2628
Daniel Kolarich - https://orcid.org/0000-0002-8452-1350
Nicolle H. Packer - https://orcid.org/0000-0002-7532-4021

References

1. Thematic issue "GlycoBioinformatics" in the Beilstein Journal of Organic Chemistry. https://www.beilstein-journals.org/bjoc/series/107 (accessed Oct 19, 2021).
2. Barnett, C. B.; Senapathi, T.; Naidoo, K. J. *Beilstein J. Org. Chem.* **2020,** *16,* 2540–2550. doi:10.3762/bjoc.16.206
3. Fogarty, C. A.; Harbison, A. M.; Dugdale, A. R.; Fadda, E. *Beilstein J. Org. Chem.* **2020,** *16,* 2046–2056. doi:10.3762/bjoc.16.171
4. Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. *ACS Cent. Sci.* **2020,** *6,* 1722–1734. doi:10.1021/acscentsci.0c01056

5. Mehta, A. Y.; Heimburg-Molinaro, J.; Cummings, R. D. *Beilstein J. Org. Chem.* **2020,** *16,* 2260–2271. doi:10.3762/bjoc.16.187
6. Lal, K.; Bermeo, R.; Perez, S. *Beilstein J. Org. Chem.* **2020,** *16,* 2448–2468. doi:10.3762/bjoc.16.199
7. Kellman, B. P.; Zhang, Y.; Logomasini, E.; Meinhardt, E.; Godinez-Macias, K. P.; Chiang, A. W. T.; Sorrentino, J. T.; Liang, C.; Bao, B.; Zhou, Y.; Akase, S.; Sogabe, I.; Kouka, T.; Winzeler, E. A.; Wilson, I. B. H.; Campbell, M. P.; Neelamegham, S.; Krambeck, F. J.; Aoki-Kinoshita, K. F.; Lewis, N. E. *Beilstein J. Org. Chem.* **2020,** *16,* 2645–2662. doi:10.3762/bjoc.16.215
8. McDonald, A. G.; Davey, G. P. *Beilstein J. Org. Chem.* **2021,** *17,* 739–748. doi:10.3762/bjoc.17.64
9. Phung, T. K.; Pegg, C. L.; Schulz, B. L. *Beilstein J. Org. Chem.* **2020,** *16,* 2127–2135. doi:10.3762/bjoc.16.180
10. Lippold, S.; de Ru, A. H.; Nouta, J.; van Veelen, P. A.; Palmblad, M.; Wuhrer, M.; de Haan, N. *Beilstein J. Org. Chem.* **2020,** *16,* 3038–3051. doi:10.3762/bjoc.16.253
11. Walsh, I.; Choo, M. S. F.; Chiin, S. L.; Mak, A.; Tay, S. J.; Rudd, P. M.; Yuansheng, Y.; Choo, A.; Swan, H. Y.; Nguyen-Khuong, T. *Beilstein J. Org. Chem.* **2020,** *16,* 2087–2099. doi:10.3762/bjoc.16.176
12. Groth, T.; Gunawan, R.; Neelamegham, S. *Beilstein J. Org. Chem.* **2021,** *17,* 1712–1724. doi:10.3762/bjoc.17.119
13. Bagdonas, H.; Ungar, D.; Agirre, J. *Beilstein J. Org. Chem.* **2020,** *16,* 2523–2533. doi:10.3762/bjoc.16.204

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

How and why plants and human N-glycans are different: Insight from molecular dynamics into the “glycoblocks” architecture of complex carbohydrates

Carl A. Fogarty‡, Aoife M. Harbison‡, Amy R. Dugdale and Elisa Fadda*

Full Research Paper

Open Access

Address:
Department of Chemistry and Hamilton Institute, Maynooth University, Maynooth, Kildare, Ireland

Email:
Elisa Fadda* - elisa.fadda@mu.ie

* Corresponding author ‡ Equal contributors

Keywords:
complex carbohydrates; fucose; glycoblocks; molecular dynamics; molecular recognition; N-glycans; xylose

Beilstein J. Org. Chem. **2020,** *16,* 2046–2056.
https://doi.org/10.3762/bjoc.16.171

Received: 02 June 2020
Accepted: 05 August 2020
Published: 21 August 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: F. Lisacek

Abstract

The N-glycosylation is one of the most abundant and diverse post-translational modifications of proteins, implicated in protein folding and structural stability, and mediating interactions with receptors and with the environment. All N-glycans share a common core from which linear or branched arms stem from, with functionalization specific to different species and to the cells’ health and disease state. This diversity generates a rich collection of structures, all diversely able to trigger molecular cascades and to activate pathways, which also include adverse immunogenic responses. These events are inherently linked to the N-glycans’ 3D architecture and dynamics, which remain for the large part unresolved and undetected because of their intrinsic structural disorder. In this work we use molecular dynamics (MD) simulations to provide insight into N-glycans’ 3D structure by analysing the effects of a set of very specific modifications found in plants and invertebrate N-glycans, which are immunogenic in humans. We also compare these structural motifs and combine them with mammalian N-glycan motifs to devise strategies for the control of the N-glycan 3D structure through sequence. Our results suggest that the N-glycans’ architecture can be described in terms of the local spatial environment of groups of monosaccharides. We define these “glycoblocks” as self-contained 3D units, uniquely identified by the nature of the residues they comprise, their linkages and structural/dynamic features. This alternative description of glycans’ 3D architecture can potentially lead to an easier prediction of sequence-to-structure relationships in complex carbohydrates, with important implications in glycoengineering design.

Introduction

Complex carbohydrates (or glycans) are an essential class of biomolecules, directly implicated in the cell’s interactions with its environment, facilitating communication and infection [1,2]. These processes are often initiated by molecular recognition involving carbohydrate-binding proteins (lectins) or by glycan–glycan interactions [1,3-5], all events that hinge on spe-

cific structural and dynamic features of the glycans. This makes the 3D complementarity of the glycans architecture key towards the success of these processes and an essential piece of information for us to have in order to understand glycan recognition. Because of their chemical nature, glycans are intrinsically flexible and highly dynamic at room temperature, thus their characterization through experimental structural biology methods is hardly straightforward even in cryogenic environments [6]. As an additive layer of difficulty, glycosylation is only indirectly dependent on the genome, which often results in a micro- (or macro-)heterogeneity of glycan sequences at specific sites [7]. These complexities are very difficult to resolve, requiring high levels of expertise and multi-layered orthogonal approaches [7-10]. Within this framework, the contribution of glycoinformatics tools and databases represents an essential resource to advance glycomics [11-15], while molecular simulations fit in very well as complementary and orthogonal techniques to support and advance structural glycobiology research. Indeed, current high performance computing (HPC) technology allows us to study realistic model systems [16,17] and to reach experimental timescales [18], so that computing can now contribute as one of the leading research methods in structural glycobiology.

One of the most interesting and remarkably challenging areas in glycoscience research that HPC simulations can address is the study of the links between glycans' sequence and the 3D structure. This direct relationship is a well-recognized and broadly accepted concept in proteins' structural biology, according to which the amino acid sequence dictates the functional 3D fold and its stability. However, the same notion is not generally invoked when discussing other biopolymers or complex carbohydrates. In the specific case of glycans, the structural complexity, in terms of the diversity of monosaccharides, the linkages' stereochemistry and the branched scaffolds, makes the already difficult case even more intricate. Nevertheless, the fact that glycoforms follow recurrent sequence patterns, clearly suggests that the glycans 3D structure is also non-random and very likely sequence-determined. We use computer modelling to gain insight into these relationships and to define a framework to understand how subtle modifications to the glycans sequence can alter their 3D structure and conformational dynamics, ultimately regulating recognition [19]. In this work we use molecular dynamics (MD) simulations to analyse the effects of the inclusion of motifs typically found in plants and invertebrates N-glycans and immunogenic in mammals [20-23]. More specifically, we investigate how core α(1-3)-linked fucose (Fuc) and β(1-2)-linked xylose (Xyl) affect the structure and dynamics of plants N-glycoforms [23] and of hybrid constructs with mammalian N-glycoforms [24].

At first glance plants protein N-glycosylation [23] is quite similar to the one of higher species [25], carrying the distinctive trimannose core (Man3), which can be further functionalised with β(1-2)-linked GlcNAc residues on the arms. As a trademark feature, shown in Figure 1, plants N-glycans can also have a β(1-2)-Xyl linked to the central mannose and core α(1-3)-Fuc,

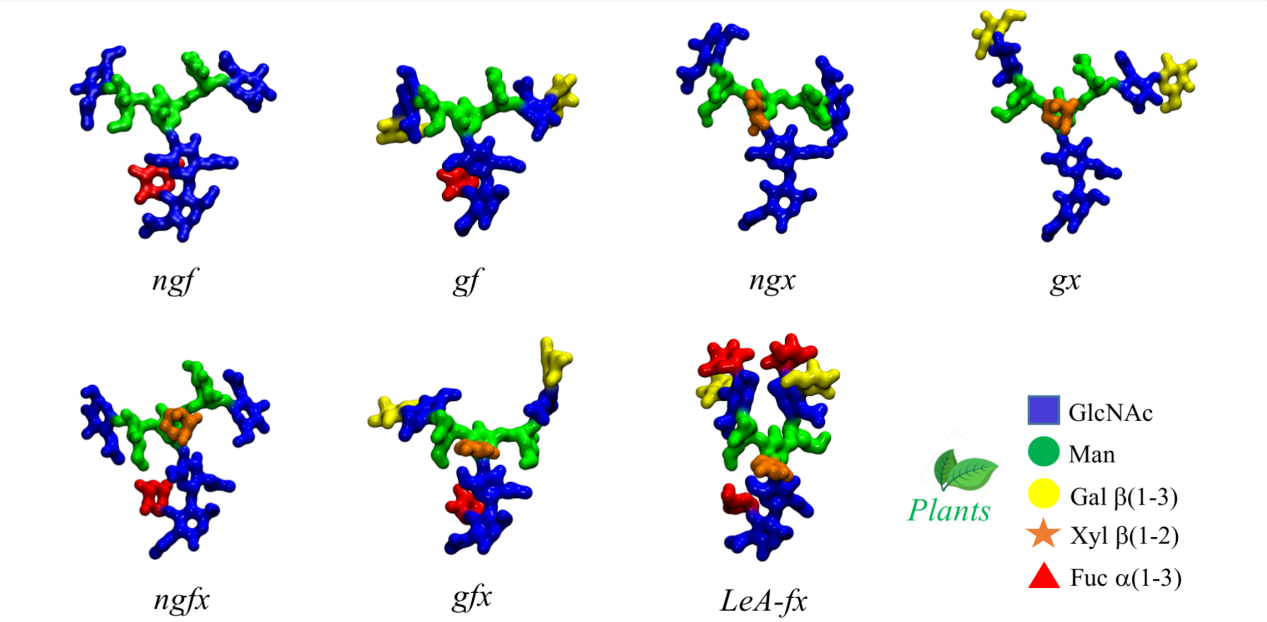


Figure 1: Representative structures of the plant N-glycans studied in this work with corresponding nomenclature. The letters *f*, *x*, and *g* indicate the presence of Fuc, Xyl and β(1-3) Gal, respectively, and *ng* the absence of β(1-3) Gal. *LeA* stands for Lewis A antigen. The N-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The plants N-glycan characteristic linkages are indicated in the legend. Rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/).

instead of the α(1-6)-Fuc commonly found in mammalian complex N-glycans. Additionally, the arms can be further functionalised with terminal galactose (Gal) in β(1-3) instead of β(1-4) [23], commonly found in vertebrates, which forces the addition of fucose in the α(1-4) position of the GlcNAc and results in the occurrence of Lewis A (LeA) instead of Lewis X (LeX) terminal motifs on the arms [23,26]. In a previous study, we characterized through extensive sampling the structure and dynamics of complex biantennary N-glycans commonly found in the human IgGs Fc region [24]. The results of this study indicated a clear sequence-to-structure relationship, especially in the context of the dynamics of the (1-6) arm. More specifically, we found that the outstretched (open) conformation of the (1-6) arm gets progressively less populated as the functionalization of the arm grows, i.e., from 85% in Man3, to 52% in (F)A2, (F)A2[3]G1, and (F)A2[3]G1S1, where the (F) indicates the presence or absence of α(1-6) core fucosylation, to 24% in all structures with (1-6) arm terminating with Gal-β(1-4)-GlcNAc or Sia-α(2-6)-Gal-β(1-4)-GlcNAc, irrespective of the functionalization of the (1-3) arm [24]. As a practical implication of these results, positional isomers, such as (F)A2[3]G1 and (F)A2[6]G1, have different conformational propensities, the latter with a much lower population of outstretched (1-6) arm and therefore quite different 3D average structures, which ultimately explains their differential recognition in glycan arrays [27]. Additionally, the different conformation of the arms explains the known difficulties in sialylating the (1-6) arm by ST6-Gal1, relatively to the (1-3) arm [28]. Also, the different 3D conformational propensity of the arms in function of sequence can have important implications in terms of the N-glycans' biosynthesis and biodegradation [29]. As an additional interesting point, we found that the folding of the (1-6) arm over the chitobiose region is completely independent of core α(1-6) fucosylation [24], with the result that core-fucosylated and non-core fucosylated N-glycans with the same sequence in the (1-6) arm correspond to the same structural ensemble.

In this work we discuss how core α(1-3)-Fuc and β(1-2)-Xyl regulate the conformational propensity of the (1-6) arm to push a predominantly outstretched (open) conformation when the arms are functionalized with terminal β(1-3)-Gal. Within this framework, we explored the possibility of integrating these motifs in the context of mammalian sequences as an exploratory strategy towards the design of N-glycans with the desired 3D structure. For simplicity in the presentation and discussion of the results, we refer to N-glycans as either "plant" or "hybrid" separately. Nevertheless, it is important to underline that some of these motifs, such as β(1-2) xylosylation and difucosylated core are also found in invertebrate N- glycosylation [30]. Finally, we discuss these findings within a framework where the different N-glycoforms can be represented as a combination of spatial self-contained units, named "glycoblocks", rather than in terms of monosaccharides and linkages. We find that this approach helps our understanding of N-glycans architecture in terms of equilibrium structures and relative populations and also of how specific modifications affect molecular recognition.

Computational Methods

All starting structures were generated with the GLYCAM Carbohydrate Builder (http://www.glycam.org). For each sequence we selected the complete set of torsion angle values obtained by variation of the 1-6 dihedrals, namely the three *gg*, *gt* and *tg* conformations for each 1-6 torsion. The topology file for each structure was obtained using *tleap* [31], with parameters from the GLYCAM06-j1 [32] for the carbohydrate atoms and with TIP3P for water molecules [33]. All calculations were run with the AMBER18 software package [31] on NVIDIA Tesla V100 16GB PCIe (Volta architecture) GPUs installed on the HPC infrastructure *kay* at the Irish Centre for High-End Computing (ICHEC). Separate production steps of 500 ns each were run for each rotamer (starting system) and convergence was assessed based on conformational and clustering analysis, see Supporting Information File 1 for all relevant Tables. Simulations were extended, if the sampling was not deemed sufficient, i.e., in case standard deviation values measured were significantly larger than 15° for each cluster in each trajectory. All trajectories were processed using *cpptraj* [31] and visually analysed with the Visual Molecular Dynamics (VMD) software package [34]. Root mean square deviation (RMSD) and torsion angles values were measured using VMD. A density-based clustering method was used to calculate the populations of occupied conformations for each torsion angle in a trajectory and heat maps for each dihedral were generated with a kernel density estimate (KDE) function. Statistical and clustering analysis was done with the R package and data were plotted with RStudio (https://www.rstudio.com). Further details on the simulation set-up and running protocol are included as Supplementary Material.

Results

Core α(1-3) fucose in plant N-glycans: One distinctive feature of plants N-glycans is the occurrence of core fucosylation in α(1-3), rather than α(1-6)-Fuc, normally found in mammalian N-glycans [23,24]. To understand the effects on the 3D structure of this modification, we have considered two biantennary systems, one terminating with β(1-2)-GlcNAc on both arms (*ngf*) and the other with terminal β(1-3)-Gal on both arms (*gf*), shown in Figure 1. In both glycoforms core α(1-3)-Fuc occupies a stable position, with one single conformer populated (100%), see Tables S1 and S2. in Supporting Information File 1. This conformation is supported by a stacking interaction

between the core α(1-3)-Fuc and β(1-4) GlcNAc of the chitobiose in a "closed" conformation, which resembles the stable conformation of LeX [35]. This spatial arrangement imposes a 20° rotation of the GlcNAc-β(1-4)-GlcNAc linkage, see Tables S1 and S2 in Supporting Information File 1, relative to the α(1-6) core fucosylated or non-fucosylated chitobiose [24], where the average psi value is −127.8° (14.8) [24], but doesn't affect the structure of the linkage to the central mannose. As shown by the low standard deviation values and by the lack of multiple minima (clusters), the N-glycan core remains relatively rigid throughout the trajectories. The slight torsion of the GlcNAc-β(1-4)-GlcNAc linkage imposed by the α(1-3)-Fuc has a dramatic effect on the conformational dynamics of the (1-6) arm, which is found predominantly in an outstretched (66%, cluster 1) conformation, rather than folded over (34%, clusters 1 and 2), see Table S1 in Supporting Information File 1. The addition of a terminal β(1-3)-Gal in the *gf* N-glycan pushes the equilibrium towards an outstretched (1-6) arm even further, with the open conformation populated at 72%, see Table S2 in Supporting Information File 1. Interestingly, in the case of α(1-6) core fucosylated N-glycans, and with double fucosylation as discussed later on, the equilibrium of the (1-6) arm was the exact opposite, with a predominance of the folded conformation, especially in the presence of terminal β(1-4) Gal [24]. To note, the folded (1-6) arm conformation can be either a 'front fold', see Figure 2 panel a, where the torsion around the α(1-6) linkage brings the arm towards the reader, or a 'back fold' where the (1-6) arm interacts with the α(1-3)-Fuc, away from the reader. As shown in Tables S1 and S2 in Supporting Information File 1, the equilibrium of the (1-3) arm is not affected by core α(1-3)-Fuc.

β(1-2) xylose in plant N-glycans: Because the β(1-2)-Xyl sits in front of the two arms, it greatly affects their dynamics. Because of steric hindrance, the (1-3) arm is much more rigid relative to non-xylosylated species, see Table S3 in Supporting Information File 1, losing its "two conformer" dynamics characteristic of the biantennary mammalian N-glycans [24], also retained in the plant N-glycans with only α(1-3)-Fuc discussed above, see also Tables S1 and S2 in Supporting Information File 1. In regards to the (1-6) arm, as shown in Figure 2 panel b, the presence of β(1-2)-Xyl has a very similar effect as the

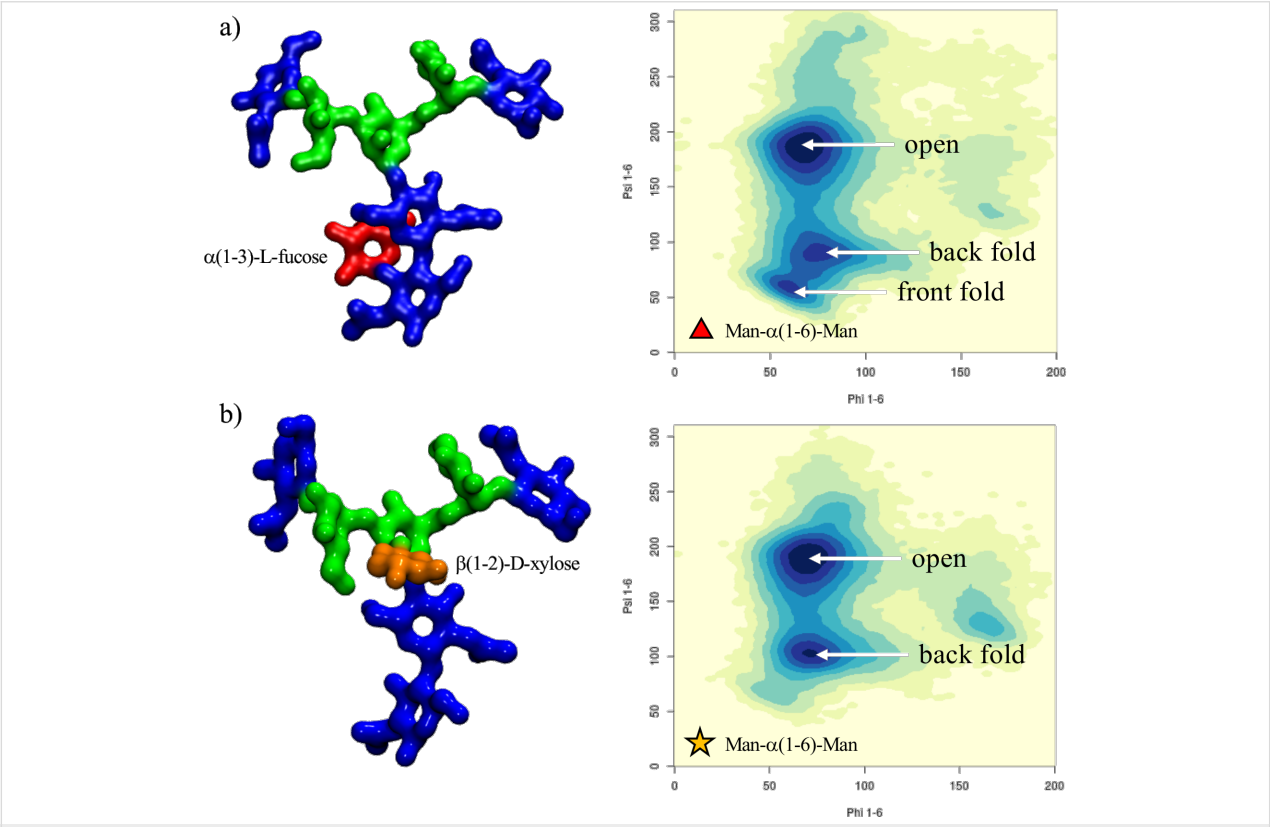


Figure 2: A representative structure of the non-galactosylated N-glycan with α(1-3)-linked core fucose (*ngf*) is shown in panel a), with on the right-hand side the corresponding heat map showing the conformations accessible to the (1-6) arm in terms of the phi/psi torsion angles. A representative structure of the non-galactosylated N-glycan with β(1-2)-linked xylose (*ngx*) is shown in panel b), with on the right-hand side a heat map showing the conformations accessible to the (1-6) arm in terms of the phi/psi torsion angles. The N-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with *RStudio* (https://www.rstudio.com).

α(1-3)-Fuc, pushing the equilibrium towards an open conformation. To note, in the presence β(1-2)-Xyl, the (1-6) arm cannot fold over the chitobiose core in a 'front fold' either, because of steric hindrance. Also, similarly to the α(1-3) fucosylated glycans, the stability of the open structure is slightly increased when the arm is further functionalized with terminal β(1-3)-Gal, see Table S4 in Supporting Information File 1. As an additional interesting feature, through the cumulative 3 μs MD sampling, the xylose ring repeatedly inverts its conformation from the all equatorial $^4C_1$ chair, to the $^1C_4$ chair, where all hydroxy groups are axial, see Figure 3. This transition may be energetically facilitated by the hydrogen bonding interaction xylose is able to form when in a $^1C_4$ chair with the α(1-6)-Man, which may compensate for the steric compression, making the $^1C_4$ chair the highest populated conformer at 76% within an N-glycan scaffold. Both experimental and ab-initio theoretical studies [36-38] have shown that the $^1C_4$ chair is energetically accessible in isolated β-D-Xyl at room temperature in different dielectric conditions.

Core α(1-3) fucose and β(1-2) xylose in plant N-glycans: The presence of both α(1-3)-Fuc and β(1-2)-Xyl brings in the characteristic features highlighted earlier in the analysis of the structures with either α(1-3)-Fuc or β(1-2)-Xyl. Indeed, we see here again the 20° rotation of the chitobiose GlcNAc-β(1-4)-GlcNAc psi angle caused by the stacking of the α(1-3)-Fuc to the chitobiose β(1-4)-GlcNAc and the conformational restraints imposed by the β(1-2)-Xyl on the (1-3) arm, see Table S5 in Supporting Information File 1. We also observed that both α(1-3)-Fuc and β(1-2)-Xyl push the (1-6) arm equilibrium towards an open conformation, which is also the case when both are present in the *ngfx* N-glycan and to an even higher degree, i.e. 87%, in the *gfx* N-glycan, when both arms are functionalized with terminal β(1-3)-Gal, see Table S6 in Supporting Information File 1. One feature specific to the *ngfx* N-glycan is the higher flexibility of the core Man-β(1-4)-GlcNAc linkage, which allows for the rotation of the trimannose group relative to the chitobiose core. This conformation was accessible, but only populated around 2% when either β(1-2)-Xyl or α(1-3)-Fuc are present, see Tables S1 to S4 in Supporting Information File 1. When both fucose and xylose are present, the population of the rotated trimannose reaches above 20%, see Table S5 in Supporting Information File 1, which can be considered a synergistic effect as this conformation is stabilized by a hydrogen bonding network involving the core fucose, the GlcNAc on the (1-6) arm and the xylose, as shown in Figure S1, Supporting Information File 1. Such folding event has been observed as a stable conformation in two independent simulations. To note, the functionalization of the arms to include terminal β(1-3)-Gal reduces the occurrence of this event down to around 5%, see Table S6 in Supporting Information File 1.

Terminal LeA and LeX motifs in plant N-glycans: To understand how an increased complexity on the arms would affect the dynamics of the α(1-3) fucosylated and β(1-2) xylosylated N-glycans, we considered the functionalization with terminal LeA antigens present in plants N-glycans [26] and with LeX for comparison. As expected [35] the LeA and LeX structures are quite rigid, see Tables S7 and S15 in Supporting Information File 1, and remain in what is known as the "closed" conforma-

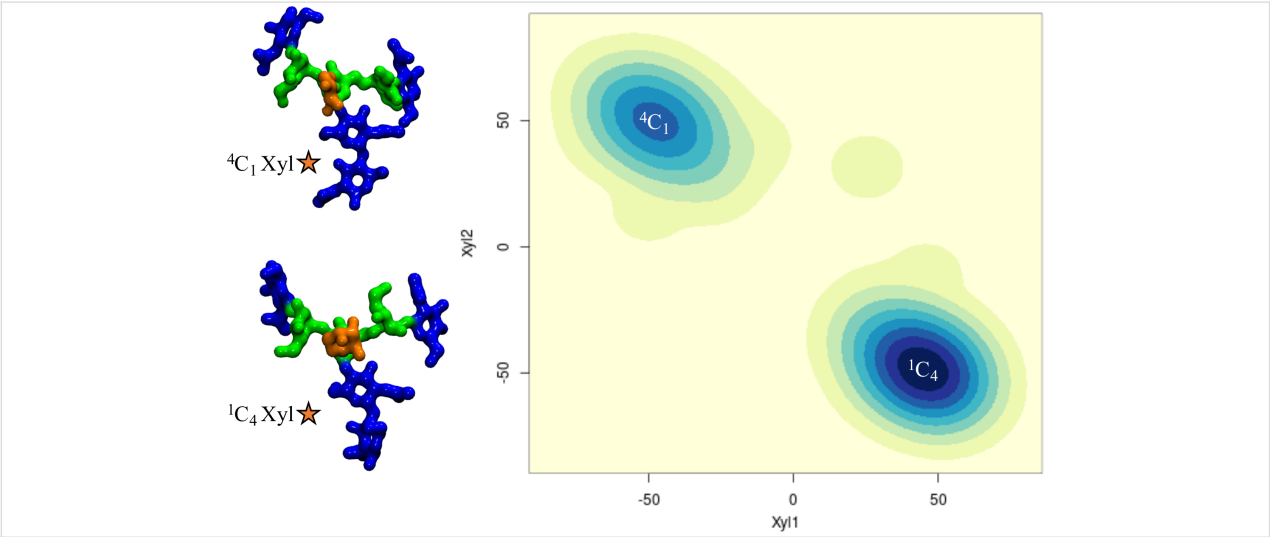


Figure 3: β-D-xylose ring pucker analysis over 3 μs of cumulative MD sampling of the *ngx* N-glycan. The two snapshots on the right-hand side are representative *ngx* conformations corresponding to the two different ring puckers. The Xyl1 and Xyl2 axis labels refer to the torsion angles C1–C2–C3–C4 and C2–C3–C4–C5, respectively. The N-glycan structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with *RStudio* (https://www.rstudio.com).

tion throughout the 1.5 μs cumulative sampling time for each system. One interesting point is that the branching introduced by functionalizing the terminal GlcNAc residues with α(1-4)-Fuc and β(1-3)-Gal, i.e. LeA, promotes the interaction between the two arms, which is not observed when the arms are linear, neither here for plants N-glycans, nor for mammalian IgG-type complex biantennary N-glycans [24]. The interaction between the arms is promoted by the ability to form complex hydrogen bonding networks, which in this specific case, may also involve the central xylose. As outcomes of the complex interaction the branched arms can establish, the equilibrium of the (1-6) arm is restrained in conformations previously not significantly populated, see Figure 4 and Supporting Information File 1, Table S7, and the GlcNAc-β(1-2)-Man linkage in both arms is remarkably flexible, which is also not observed when the arms are not branched. Although not natural in plants, to check the corresponding symmetry, we built a core α(1-3)-Fuc and β(1-2)-Xyl N-glycan with terminal LeX on both arms, a feature actually found in schistosome N-glycosylation [30]. Remarkably, as shown in Figure 4 and Supporting Information File 1, Table S15, within this framework the dynamics of the (1-6) arm is completely different. Contrary to the N-glycan with terminal LeA groups, the two arms with LeX are not interacting and the (1-6) arm is predominantly (90%) in an extended (open) conformation, while the closed conformation, which accounts for the remaining 10% is achieved through a rotation around the core Man-β(1-4)-GlcNAc. The lack of interaction between the arms is due to the inability to establish the same stable hydrogen bonding network due to the non-complementary position of the deoxy-C6 of the fucose in LeX relative to LeA.

Hybrid N-glycans. To understand how characteristic plant N-glycan motifs can affect the structure of mammalian N-glycoforms, we have designed and analysed the dynamics of a set of hybrid systems. In particular, we were interested in the effect of the addition of β(1-2)-Xyl and α(1-3)-Fuc to (F)A2G2 N-glycans scaffolds in terms of potential alteration of the (1-6) arm dynamics.

β(1-2)-xylosylated mammalian N-glycans. Unlike the case of plants N-glycans, the presence of β(1-2)-Xyl hinders but does not completely prevent the (1-6) arm from folding over when the terminal galactose is β(1-4)-linked, as folding over the chitobiose can be stabilized by stacking, see Figure 5 and Supporting Information File 1, Table S8. The folded conformation with a median psi value of 103.5° (± 11.3) is 20° from the average value of 82.9° calculated for the non-xylosylated (mammalian) counterpart [24], so slightly distorted, and its population reduced from 74% to 57%. Nevertheless, the closed conformation is still the predominant form, even with β(1-2)-Xyl. The presence of α(1-6)-linked core fucose to create a β(1-2)-xylosylated FA2G2, which is actually a type of N-glycosylation found in schistosoma [30], brings in yet another change. As shown in Figure 5 and Supporting Information File 1, Table S9, α(1-6)-Fuc and β(1-2)-Xyl are in an optimal conformation to support the closed (folded) (1-6) arm, by stacking of the terminal galactose by fucose and hydrogen bonding by xylose. Within this context the closed (1-6) arm is the highest populated conformer at 70.0% over 4.5 μs of cumulative sampling of this system. To note that the conformation of the α(1-6)-linked core fucose is the same as the one seen in mammalian N-glycans [24], which on its own we have seen is not enough to affect the (1-6) arm equilibrium, see Table S9 in Supporting Information File 1. The interaction of the α(1-6)-Fuc with the terminal β(1-4)-Gal is essential to promote the closed conformation of the (1-6) arm as demonstrated by the results

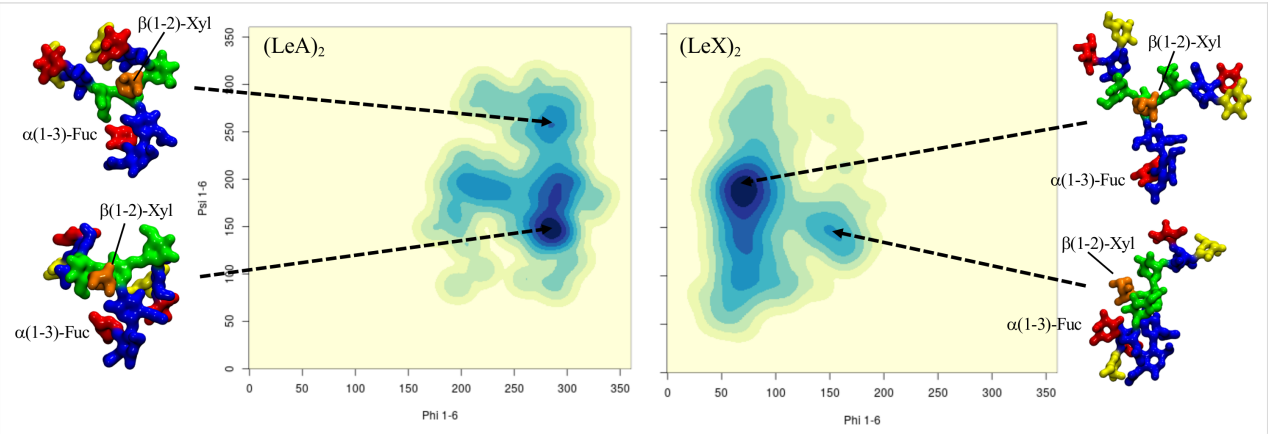


Figure 4: Comparison of the different conformational equilibria of the (1-6) arm in a core α(1-3)-Fuc β(1-2)-Xyl A2 N-glycan with terminal LeA and LeX groups on the left- and right-hand side, respectively. Representative structures from 1.5 μs MD sampling of each system are shown to illustrate the conformations corresponding to the different minima. The N-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with *RStudio* (https://www.rstudio.com).

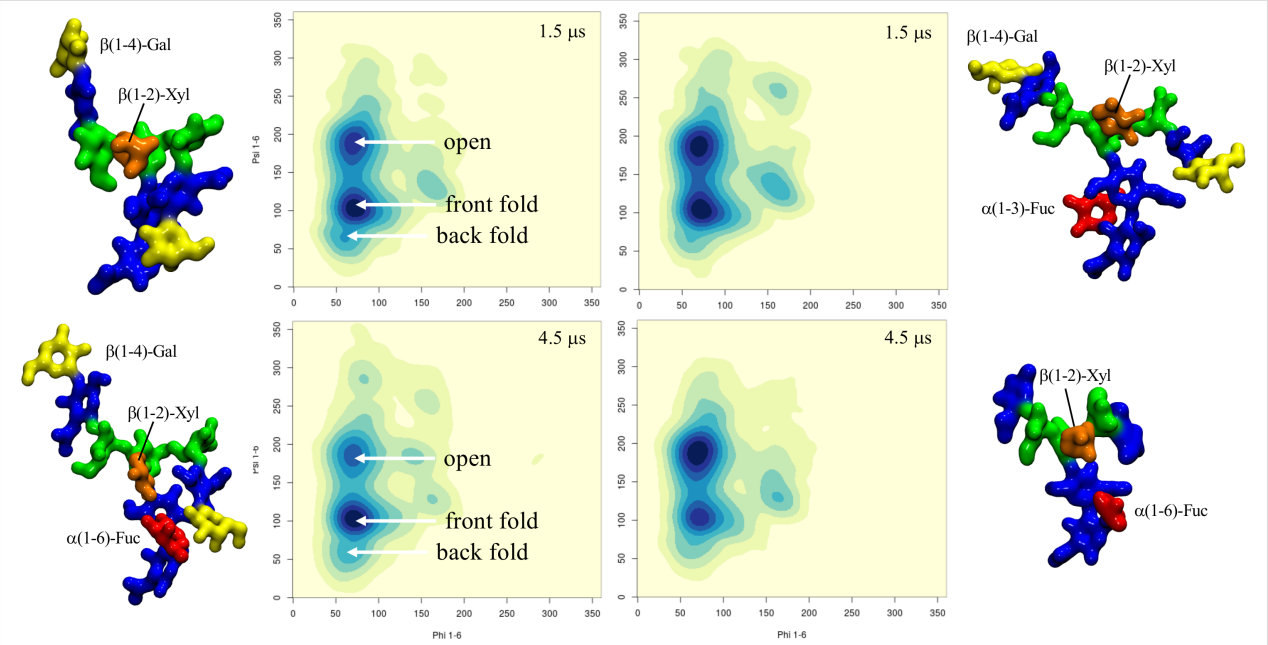


Figure 5: Conformational analysis of the (1-6) arm in four hybrid N-glycoforms, β(1-2)-xylosylated A2G2 (top-left), β(1-2)-xylosylated FA2G2 (bottom-left), β(1-2)-xylosylated α(1-3)-core fucosylated A2G2 (top-right) and β(1-2)-xylosylated FA2 (bottom-right). The predominant conformations are indicated in the top- and bottom-left heat maps for simplicity. The simulation time relative to each system is indicated in the top right corner of each heat map. The N-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with *RStudio* (https://www.rstudio.com).

obtained for the xylosylated FA2 systems, which recovers a conformational propensity similar to the non-fucosylated, xylosylated A2G2, see Figure 5 and Tables S8 and S10 in Supporting Information File 1.

α(1-3)-fucosylated mammalian N-glycans. Because of its orientation tucked "behind" the chitobiose core defined in the context of plants N-glycans earlier, the effect of core α(1-3)-Fuc on the (1-6) arm equilibrium within an A2G2-xylosylated scaffold is not as significant as α(1-6)-Fuc. As shown in Figure 5 and in Supporting Information File 1, Table S11, this lack of direct effect is demonstrated by the recovery of the same equilibrium as the non-fucosylated A2G2-xylosylated system. The dynamics of the chitobiose core is very similar to the one determined for the corresponding plant N-glycan. To analyse the effect of core α(1-3) fucosylation without β(1-2)-Xyl, we have looked at two A2G2 hybrid systems, one with only α(1-3)-linked fucose and one with both core α(1-3)- and α(1-6)-linked fucose, a characteristic "double-fucose" glycosylation found in worm and fly cells [30]. As shown in Supporting Information File 1, Table S12 unlike in plants N-glycans, the α(1-3)-Fuc alone does not affect the A2G2 (1-6) arm equilibrium [24], as the folding of the (1-6) arm with terminal β(1-4)-Gal is not obstructed by the rotation of the chitobiose core imposed by the α(1-3)-Fuc position. When both α(1-3)- and α(1-6)-linked fucoses are present the (1-6) arm with terminal β(1-4)-Gal is predominantly folded (closed) at 85%, see Figure 6 and Supporting Information File 1, Table S13, which is higher than in the absence of α(1-3)-Fuc [24]. Indeed, the latter can actively contribute in stabilizing the interaction with the terminal β(1-4)-Gal of the folded (1-6) arm. We also observed interesting events, one representing 10% of 2 µs as indicated by the values of the GlcNAc-β(1-4)-GlcNAc torsion, where the GlcNAc is stacked in between the two fucose residues and another one, contributing to 18% of the simulation time, 14% when the system is also xylosylated, in which the GlcNAc ring transitions from $^4C_1$ to $^1C_4$ allowing the two fucose to stack, see Tables S13 and S14 and Figure S2 in Supporting Information File 1.

Discussion

Differences and similarities in N-glycan sequences are highly cell-specific as well as important indicators of health and disease states [1,39]. Exogenous N-glycans motifs can be quite subtle, yet trigger profound differences in terms of molecular recognition [19,27] and dangerous immunogenic responses [20-22]. In this work we have analysed the effects on the N-glycans structure and dynamics of two motifs in particular, namely β(1-2)-Xyl and core α(1-3)-Fuc, common in plants [23] and invertebrates [30], but completely absent in mammalian N-glycans. Within the context of plant-type N-glycans, which have a terminal β(1-3)-Gal, rather than β(1-4)-Gal, both

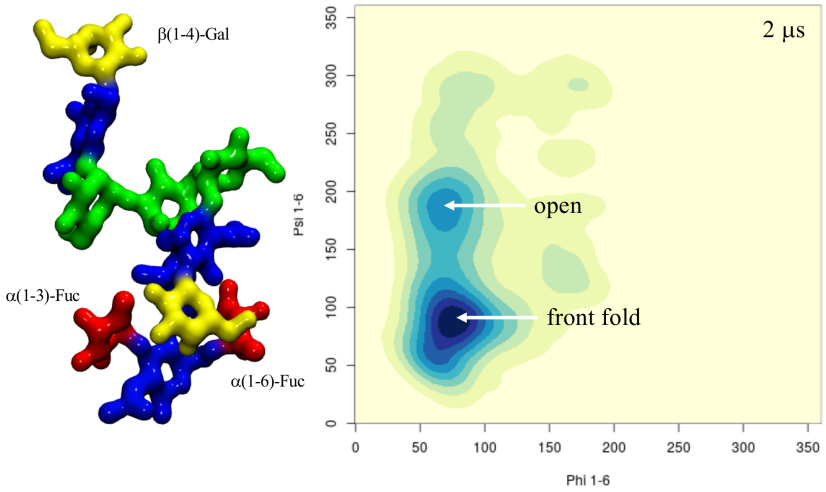


Figure 6: Conformational equilibrium of the (1-6) arm in terms of phi/psi torsion angle values for the α(1-3)-fucosylated FA2G2 N-glycoform. The structure with the folded (1-6) arm where the terminal β(1-4)-Gal interacts with both fucose residues is shown on the left-hand side. The N-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with *RStudio* (https://www.rstudio.com).

β(1-2)-Xyl and α(1-3)-Fuc contribute independently in promoting an outstretched (open) conformation of the (1-6) arm because of steric hindrance of the xylose and of the rotation forced upon the chitobiose core by the α(1-3)-Fuc. The latter is not an obstruction for the folding of a β(1-4)-Gal terminated (1-6) arm, as we have seen in the hybrid N-glycans constructs. Therefore, in β(1-2) xylosylated N-glycans terminating with β(1-3)-Gal, both arms should be more available for recognition, binding and further functionalization [30], unlike in mammalian N-glycans where the β(1-4)-Gal determines a prevalently closed and inaccessible (1-6) arm [24,27]. Also, the analysis of the structure and dynamics of the LeA terminating plant N-glycans showed that the specific branching and spatial orientation of the motif allowed for a stable interaction between the arms, which is not observed in complex N-glycans with a linear functionalization of the arms [24]. Notably, the same hydrogen bonding network between the arms cannot be established when the same N-glycan terminates with LeX, because of the non-complementary position of the α(1-3)-Fuc deoxy-C6.

The analysis of all these different complex N-glycoforms clearly shows that every modification, addition or removal of a specific motif, can greatly affect the 3D architecture of the N-glycan, thus its accessibility and complementarity to a receptor. However, these effects are rather complex to understand or to predict, if we think of the N-glycans 3D structure in terms of sequence of monosaccharides, a view that stems from the way we think about proteins. Our results show that the main effect of all functionalizations is actually local. For example, the core α(1-3)-Fuc forces a rotation of the chitobiose, a degree of freedom very lowly populated otherwise; meanwhile, β(1-2)-Xyl restricts the flexibility of the trimannose core and occupies its centre. Within this framework, the 3D structural and dynamics features of the N-glycoforms can be rationalized by discretizing their architecture in terms 3D units, or "glycoblocks", that group monosaccharides and their linkages within their immediate spatial vicinity, e.g., the core α(1-3)-Fuc and the chitobiose which structure it has modified. A list of the glycoblocks that we have identified with the corresponding descriptors of their 3D features are listed in Figure 7. The whole N-glycan 3D architecture, in terms of the structures accessible and their conformational propensity, can be then described through the combination of these glycoblocks, together with the knowledge of their dynamic properties and flexibility. Also, consideration of these glycoblocks as spatial units can be useful to understand recognition by lectins and antibodies, which is often affected primarily by the targeted monosaccharide's immediate vicinity and by its accessibility within a specific glycoform. For example, if we consider the 3D structure of the β(1-2)-Xyl Man3 glycoblock vs the Man3 without Xyl, we can understand how the β(1-2)-Xyl position within that unit negates binding to DC-SIGN lectins [19], see Supporting Information File 1, Figure S3 panels a and b. Additionally, we can see that the slight rotation on the chitobiose imposed by the core α(1-3)-Fuc does not prevent recognition and binding, see Supporting Information File 1, Figure S3 panel c.

Conclusion

In this work we used extensive sampling through MD simulations to study the effects on the N-glycan architecture of subtle, yet highly consequential modifications, namely core α(1-3)-Fuc and β(1-2)-Xyl [19]. These are part of standard N-glycoforms

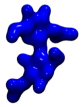
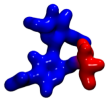
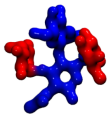
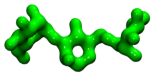
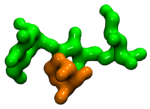
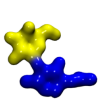
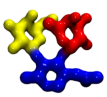
Glycoblock	Unit 3D structure	Dominant motif descriptor	Characteristic glycosidic linkage
		Very stable conformation, with a low degree of flexibility around the equilibrium structure	**GlcNAc β(1-4) GlcNAc** phi = -78.7 (11.1)/100 psi = -130.8 (15.7)/99
		The α(1-3)-Fuc stacking interaction imposes a 20° rotation around the chitobiose glycosidic linkage and further stabilizes the chitobiose structure	**GlcNAc β(1-4) GlcNAc** phi = -72.1 (8.3)/100 psi = -107.1 (7.6)/100
		The longer glycosydic linkage provides the α(1-6)-Fuc with more degrees of freedom and does not alter the chitobiose structure	**GlcNAc β(1-4) GlcNAc** phi = -77.8 (11.2)/100 psi = -127.8 (14.8)/100
		This motif is structurally similar to the one with the α(1-3)-Fuc, which acts as the dominant restraint. The presence of the α(1-6)-Fuc increases flexibility	**GlcNAc β(1-4) GlcNAc** phi = -73.9 (9.5)/100 psi = -106.4 (14.1)/90
		Joint with differently flexible arms. The α(1-6) torsion has two dominant, different conformations accessible that can open and close the (1-6) arm. Populations depend on the attached glycoblock units	**Man α(1-6) Man** (with terminal β(1-4)-Gal) open phi = 76.3 (15.0)/100 psi = -185.1 (22.1)/23 closed phi = 76.3 (15.0)/100 psi = 85.3 (15.8)/76
		The addition of the β(1-2)-Xyl further restricts the (1-3) arm and hinders the closing forward of the (1-6) arm, which is only favored in (F)A2G2 hybrids	**Man α(1-6) Man** (with terminal β(1-3)-Gal) open phi = 70.5 (10.4)/100 psi = -173.6 (16.3)/70 closed phi = 70.5 (10.4)/100 psi = 103.8 (12.7)/26
		This unit is rather rigid, while the orientation of the terminal Gal in β(1-3/4) is determinant for the interaction with the chitobiose in the closed (1-6) arm. These units in the two arms do not interact	**Gal β(1-3) GlcNAc** phi = -72.9 (11.2)/100 psi = -124.0 (18.2)/96
		The Lewis A epitope is stable in its closed conformation and its presence greatly affects the (1-6) arm conformation because branching promotes association of the arms	**Man α(1-6) Man** (with terminal LeA and α(1-3)-Fuc) open phi = -70.1 (5.7)/100 psi = -177.1 (16.3)/28 closed phi = -70.1 (5.7)/100 psi = 148.2 (7.4)/54
		This unit is rather rigid, while the orientation of the terminal Gal in β(1-3/4) is determinant for the interaction with the chitobiose in the closed (1-6) arm. These units in the two arms do not interact	**Gal β(1-4) GlcNAc** phi = -76.2 (15.1)/100 psi = -125.7 (15.4)/97
		The Lewis X epitope is stable in its closed conformation. Contrary to Lewis A its branching does not favor association of the arms because of the relative positions of the deoxy-C_6.	**Man α(1-6) Man** (with terminal LeX, core α(1-3)-Fuc and β(1-2)-Xyl) open phi = 70.7 (9.2)/90 psi = -173.6 (13.4)/90 closed phi = 152.4 (12.9)/10 psi = 145.6 (12.3)/10

Figure 7: List of 3D structural units of monosaccharides (glycoblocks) that regulate the 3D architecture and dynamics of complex biantennary N-glycans from plants and invertebrate sources and hybrid mammalian constructs. The SFNG representation of each glycoblock is indicated in the first column from the left, 3D structures from the highest populated conformers are shown in the second column, rendered with VMD (https://www.ks.uiuc.edu/Research/vmd/). A brief summary of the conformational features of each glycoblock and the characteristic linkage or its effect on the (1-6) arm conformation are indicates in the last two columns, respectively.

found in plants [23] and invertebrates [30], but immunogenic in humans [21,22,26]. Our results show that these modifications can greatly affect the 3D structure of the N-glycan and its structural dynamics, therefore its selective recognition by lectin receptors and antibodies. The atomistic-level of detail information that the MD simulations provide us with, highlights that the effects of different functionalizations, in terms of monosaccharide types and linkages, are primarily local, affecting the immediate spatial vicinity of the monosaccharide within the N-glycan structure. Within this framework, we propose an alternative approach that can help to describe and predict the architecture of N-glycans based on the combination of structural 3D units, or glycoblocks. Unlike a description based on the monosaccharide sequence and linkages as two separate features, the transition to well-defined and self-contained units, integrating information on both monosaccharides and linkages, can help us rationalize and deconvolute the glycans structural disorder and ultimately understand more clearly the relationships between sequence and structure in complex carbohydrates.

Supporting Information

Supporting Information File 1

Computational methods and supplementary figures and tables.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-171-S1.pdf]

Acknowledgements

EF would like to thank Prof. Iain B.H. Wilson for insightful feedback on an earlier version of the manuscript.

Funding

The Irish Centre for High-End Computing (ICHEC) is gratefully acknowledged for generous allocation of computational resources. EF and CAF acknowledge the Irish Research Council (IRC) for funding through the Government of Ireland Postgraduate Scholarship Programme. EF and AMH acknowledge the John and Pat Hume Doctoral Scholarship Programme at Maynooth University for funding.

ORCID® iDs

Aoife M. Harbison - https://orcid.org/0000-0001-7471-8064
Elisa Fadda - https://orcid.org/0000-0002-2898-7770

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.1101/2020.05.22.110528

References

1. Varki, A. *Glycobiology* **2017,** *27,* 3–49. doi:10.1093/glycob/cww086
2. Stanley, P.; Taniguchi, N.; Aebi, M. N-Glycans. *Essentials of Glycobiology,* 3rd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2017; pp 99–112.
3. Cobb, B. A. *Glycobiology* **2020,** *30,* 202–213. doi:10.1093/glycob/cwz065
4. Ferrara, C.; Grau, S.; Jäger, C.; Sondermann, P.; Brünker, P.; Waldhauer, I.; Hennig, M.; Ruf, A.; Rufer, A. C.; Stihle, M.; Umaña, P.; Benz, J. *Proc. Natl. Acad. Sci. U. S. A.* **2011,** *108,* 12669–12674. doi:10.1073/pnas.1108455108
5. Day, C. J.; Tran, E. N.; Semchenko, E. A.; Tram, G.; Hartley-Tassell, L. E.; Ng, P. S. K.; King, R. M.; Ulanovsky, R.; McAtamney, S.; Apicella, M. A.; Tiralongo, J.; Morona, R.; Korolik, V.; Jennings, M. P. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112,* E7266–E7275. doi:10.1073/pnas.1421082112
6. Nagae, M.; Yamaguchi, Y. *Int. J. Mol. Sci.* **2012,** *13,* 8398–8429. doi:10.3390/ijms13078398
7. Zacchi, L. F.; Schulz, B. L. *Glycoconjugate J.* **2016,** *33,* 359–376. doi:10.1007/s10719-015-9641-3
8. De Leoz, M. L. A.; Duewer, D. L.; Fung, A.; Liu, L.; Yau, H. K.; Potter, O.; Staples, G. O.; Furuki, K.; Frenkel, R.; Hu, Y.; Sosic, Z.; Zhang, P.; Altmann, F.; Gru«nwald-Grube, C.; Shao, C.; Zaia, J.; Evers, W.; Pengelley, S.; Suckau, D.; Wiechmann, A.; Resemann, A.; Jabs, W.; Beck, A.; Froehlich, J. W.; Huang, C.; Li, Y.; Liu, Y.; Sun, S.; Wang, Y.; Seo, Y.; An, H. J.; Reichardt, N.-C.; Ruiz, J. E.; Archer-Hartmann, S.; Azadi, P.; Bell, L.; Lakos, Z.; An, Y.; Cipollo, J. F.; Pucic-Bakovic, M.; Štambuk, J.; Lauc, G.; Li, X.; Wang, P. G.; Bock, A.; Hennig, R.; Rapp, E.; Creskey, M.; Cyr, T. D.; Nakano, M.; Sugiyama, T.; Leung, P.-K. A.; Link-Lenczowski, P.; Jaworek, J.; Yang, S.; Zhang, H.; Kelly, T.; Klapoetke, S.; Cao, R.; Kim, J. Y.; Lee, H. K.; Lee, J. Y.; Yoo, J. S.; Kim, S.-R.; Suh, S.-K.; de Haan, N.; Falck, D.; Lageveen-Kammeijer, G. S. M.; Wuhrer, M.; Emery, R. J.; Kozak, R. P.; Liew, L. P.; Royle, L.; Urbanowicz, P. A.; Packer, N. H.; Song, X.; Everest-Dass, A.; Lattová, E.; Cajic, S.; Alagesan, K.; Kolarich, D.; Kasali, T.; Lindo, V.; Chen, Y.; Goswami, K.; Gau, B.; Amunugama, R.; Jones, R.; Stroop, C. J. M.; Kato, K.; Yagi, H.; Kondo, S.; Yuen, C. T.; Harazono, A.; Shi, X.; Magnelli, P. E.; Kasper, B. T.; Mahal, L.; Harvey, D. J.; O'Flaherty, R.; Rudd, P. M.; Saldova, R.; Hecht, E. S.; Muddiman, D. C.; Kang, J.; Bhoskar, P.; Menard, D.; Saati, A.; Merle, C.; Mast, S.; Tep, S.; Truong, J.; Nishikaze, T.; Sekiya, S.; Shafer, A.; Funaoka, S.; Toyoda, M.; de Vreugd, P.; Caron, C.; Pradhan, P.; Tan, N. C.; Mechref, Y.; Patil, S.; Rohrer, J. S.; Chakrabarti, R.; Dadke, D.; Lahori, M.; Zou, C.; Cairo, C.; Reiz, B.; Whittal, R. M.; Lebrilla, C. B.; Wu, L.; Guttman, A.; Szigeti, M.; Kremkow, B. G.; Lee, K. H.; Sihlbom, C.; Adamczyk, B.; Jin, C.; Karlsson, N. G.; Örnros, J.; Larson, G.; Nilsson, J.; Meyer, B.; Wiegandt, A.; Komatsu, E.; Perreault, H.; Bodnar, E. D.; Said, N.; Francois, Y.-N.; Leize-Wagner, E.; Maier, S.; Zeck, A.; Heck, A. J. R.; Yang, Y.; Haselberg, R.; Yu, Y. Q.; Alley, W.; Leone, J. W.; Yuan, H.; Stein, S. E. *Mol. Cell. Proteomics* **2020,** *19,* 11–30. doi:10.1074/mcp.ra119.001677
9. Mimura, Y.; Katoh, T.; Saldova, R.; O'Flaherty, R.; Izumi, T.; Mimura-Kimura, Y.; Utsunomiya, T.; Mizukami, Y.; Yamamoto, K.; Matsumoto, T.; Rudd, P. M. *Protein Cell* **2018,** *9,* 47–62. doi:10.1007/s13238-017-0433-3
10. Thaysen-Andersen, M.; Packer, N. H. *Biochim. Biophys. Acta, Proteins Proteomics* **2014,** *1844,* 1437–1452. doi:10.1016/j.bbapap.2014.05.002

11. Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.; Packer, N. H.; Lisacek, F. *J. Proteome Res.* **2019,** *18,* 664–677. doi:10.1021/acs.jproteome.8b00766
12. Lisacek, F.; Mariethoz, J.; Alocci, D.; Rudd, P. M.; Abrahams, J. L.; Campbell, M. P.; Packer, N. H.; Ståhle, J.; Widmalm, G.; Mullen, E.; Adamczyk, B.; Rojas-Macias, M. A.; Jin, C.; Karlsson, N. G. Databases and Associated Tools for Glycomics and Glycoproteomics. In *High-Throughput Glycomics and Glycoproteomics;* Lauc, G.; Wuhrer, M., Eds.; Methods in Molecular Biology, Vol. 1503; Humana Press: New York, NY, USA, 2017; pp 235–264. doi:10.1007/978-1-4939-6493-2_18
13. Mariethoz, J.; Alocci, D.; Gastaldello, A.; Horlacher, O.; Gasteiger, E.; Rojas-Macias, M.; Karlsson, N. G.; Packer, N. H.; Lisacek, F. *Mol. Cell. Proteomics* **2018,** *17,* 2164–2176. doi:10.1074/mcp.ra118.000799
14. Aoki-Kinoshita, K.; Agravat, S.; Aoki, N. P.; Arpinar, S.; Cummings, R. D.; Fujita, A.; Fujita, N.; Hart, G. M.; Haslam, S. M.; Kawasaki, T.; Matsubara, M.; Moreman, K. W.; Okuda, S.; Pierce, M.; Ranzinger, R.; Shikanai, T.; Shinmachi, D.; Solovieva, E.; Suzuki, Y.; Tsuchiya, S.; Yamada, I.; York, W. S.; Zaia, J.; Narimatsu, H. *Nucleic Acids Res.* **2016,** *44,* D1237–D1242. doi:10.1093/nar/gkv1041
15. Rojas-Macias, M. A.; Mariethoz, J.; Andersson, P.; Jin, C.; Venkatakrishnan, V.; Aoki, N. P.; Shinmachi, D.; Ashwood, C.; Madunic, K.; Zhang, T.; Miller, R. L.; Horlacher, O.; Struwe, W. B.; Watanabe, Y.; Okuda, S.; Levander, F.; Kolarich, D.; Rudd, P. M.; Wuhrer, M.; Kettner, C.; Packer, N. H.; Aoki-Kinoshita, K. F.; Lisacek, F.; Karlsson, N. G. *Nat. Commun.* **2019,** *10,* 3275. doi:10.1038/s41467-019-11131-x
16. Durrant, J. D.; Kochanek, S. E.; Casalino, L.; Ieong, P. U.; Dommer, A. C.; Amaro, R. E. *ACS Cent. Sci.* **2020,** *6,* 189–196. doi:10.1021/acscentsci.9b01071
17. Yu, I. M.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. *eLife* **2016,** *5,* e19274. doi:10.7554/elife.19274
18. Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Shaw, D. E. *J. Phys. Chem. B* **2016,** *120,* 8313–8320. doi:10.1021/acs.jpcb.6b02024
19. Brzezicka, K.; Echeverria, B.; Serna, S.; van Diepen, A.; Hokke, C. H.; Reichardt, N.-C. *ACS Chem. Biol.* **2015,** *10,* 1290–1302. doi:10.1021/cb501023u
20. Wilson, I. B. H.; Harthill, J. E.; Mullin, N. P.; Ashford, D. A.; Altmann, F. *Glycobiology* **1998,** *8,* 651–661. doi:10.1093/glycob/8.7.651
21. van Ree, R.; Cabanes-Macheteau, M.; Akkerdaas, J.; Milazzo, J.-P.; Loutelier-Bourhis, C.; Rayon, C.; Villalba, M.; Koppelman, S.; Aalberse, R.; Rodriguez, R.; Faye, L.; Lerouge, P. *J. Biol. Chem.* **2000,** *275,* 11451–11458. doi:10.1074/jbc.275.15.11451
22. Bardor, M.; Faveeuw, C.; Fitchette, A.-C.; Gilbert, D.; Galas, L.; Trottein, F.; Faye, L.; Lerouge, P. *Glycobiology* **2003,** *13,* 427–434. doi:10.1093/glycob/cwg024
23. Strasser, R. *Glycobiology* **2016,** *26,* 926–939. doi:10.1093/glycob/cww023
24. Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. *Glycobiology* **2019,** *29,* 94–103. doi:10.1093/glycob/cwy097
25. Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H. *Essentials of Glycobiology,* 3rd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2017.
26. Fitchette-Lainé, A.-C.; Gomord, V.; Cabanes, M.; Michalski, J.-C.; Saint Macary, M.; Foucher, B.; Cavelier, B.; Hawes, C.; Lerouge, P.; Faye, L. *Plant J.* **1997,** *12,* 1411–1417. doi:10.1046/j.1365-313x.1997.12061411.x
27. Echeverria, B.; Serna, S.; Achilli, S.; Vivès, C.; Pham, J.; Thépaut, M.; Hokke, C. H.; Fieschi, F.; Reichardt, N.-C. *ACS Chem. Biol.* **2018,** *13,* 2269–2279. doi:10.1021/acschembio.8b00431
28. Barb, A. W.; Brady, E. K.; Prestegard, J. H. *Biochemistry* **2009,** *48,* 9705–9707. doi:10.1021/bi901430h
29. Möginger, U.; Grunewald, S.; Hennig, R.; Kuo, C.-W.; Schirmeister, F.; Voth, H.; Rapp, E.; Khoo, K.-H.; Seeberger, P. H.; Simon, J. C.; Kolarich, D. *Front. Oncol.* **2018,** *8,* 70. doi:10.3389/fonc.2018.00070
30. Paschinger, K.; Wilson, I. B. H. *Parasitology* **2019,** *146,* 1733–1742. doi:10.1017/s0031182019000398
31. *AMBER 2018;* University of California: San Francisco, CA, USA, 2018.
32. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008,** *29,* 622–655. doi:10.1002/jcc.20820
33. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983,** *79,* 926–935. doi:10.1063/1.445869
34. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996,** *14,* 33–38. doi:10.1016/0263-7855(96)00018-5
35. Alibay, I.; Burusco, K. K.; Bruce, N. J.; Bryce, R. A. *J. Phys. Chem. B* **2018,** *122,* 2462–2474. doi:10.1021/acs.jpcb.7b09841
36. Mayes, H. B.; Broadbelt, L. J.; Beckham, G. T. *J. Am. Chem. Soc.* **2014,** *136,* 1008–1022. doi:10.1021/ja410264d
37. Rönnols, J.; Manner, S.; Siegbahn, A.; Ellervik, U.; Widmalm, G. *Org. Biomol. Chem.* **2013,** *11,* 5465–5472. doi:10.1039/c3ob40991k
38. Iglesias-Fernández, J.; Raich, L.; Ardèvol, A.; Rovira, C. *Chem. Sci.* **2015,** *6,* 1167–1177. doi:10.1039/c4sc02240h
39. Reily, C.; Stewart, T. J.; Renfrow, M. B.; Novak, J. *Nat. Rev. Nephrol.* **2019,** *15,* 346–366. doi:10.1038/s41581-019-0129-4

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Clustering and curation of electropherograms: an efficient method for analyzing large cohorts of capillary electrophoresis glycomic profiles for bioprocessing operations

Ian Walsh[*1], Matthew S. F. Choo[1], Sim Lyn Chiin[1], Amelia Mak[1], Shi Jie Tay[1], Pauline M. Rudd[1,2], Yang Yuansheng[3], Andre Choo[4,5], Ho Ying Swan[5] and Terry Nguyen-Khuong[*1]

Full Research Paper

Open Access

Address:
[1]Analytics Group, Bioprocessing Technology Institute - Agency for Science Technology and Research. Singapore 138668, [2]University College Dublin, Belfield, Dublin, Ireland, [3]Animal Cell Technology Group, Bioprocessing Technology Institute, Agency for Science Technology and Research, Singapore 138668, [4]Stem Cells 1 Group, Bioprocessing Technology Institute - Agency for Science Technology and Research, Singapore 138668 and [5]Department of Biomedical Engineering, Faculty of Engineering, National University of Singapore (NUS), Singapore 117575

Email:
Ian Walsh[*] - ian_walsh@bti.a-star.edu.sg; Terry Nguyen-Khuong[*] - terry_nguyen_khuong@bti.a-star.edu.sg

* Corresponding author

Keywords:
capillary electrophoresis; clustering; data analysis; electropherogram; glycosylation; monoclonal antibodies; peak picking; process development

Beilstein J. Org. Chem. **2020,** *16,* 2087–2099.
https://doi.org/10.3762/bjoc.16.176

Received: 30 May 2020
Accepted: 13 August 2020
Published: 27 August 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editors: N. H. Packer and F. Lisacek

Abstract

The accurate assessment of antibody glycosylation during bioprocessing requires the high-throughput generation of large amounts of glycomics data. This allows bioprocess engineers to identify critical process parameters that control the glycosylation critical quality attributes. The advances made in protocols for capillary electrophoresis-laser-induced fluorescence (CE-LIF) measurements of antibody N-glycans have increased the potential for generating large datasets of N-glycosylation values for assessment. With large cohorts of CE-LIF data, peak picking and peak area calculations still remain a problem for fast and accurate quantitation, despite the presence of internal and external standards to reduce misalignment for the qualitative analysis. The peak picking and area calculation problems are often due to fluctuations introduced by varying process conditions resulting in heterogeneous peak shapes. Additionally, peaks with co-eluting glycans can produce peaks of a non-Gaussian nature in some process conditions and not in others. Here, we describe an approach to quantitatively and qualitatively curate large cohort CE-LIF glycomics data. For glycan

identification, a previously reported method based on internal triple standards is used. For determining the glycan relative quantities our method uses a clustering algorithm to 'divide and conquer' highly heterogeneous electropherograms into similar groups, making it easier to define peaks manually. Open-source software is then used to determine peak areas of the manually defined peaks. We successfully applied this semi-automated method to a dataset (containing 391 glycoprofiles) of monoclonal antibody biosimilars from a bioreactor optimization study. The key advantage of this computational approach is that all runs can be analyzed simultaneously with high accuracy in glycan identification and quantitation and there is no theoretical limit to the scale of this method.

Introduction

Glycosylation is important for the efficacy and function of a majority of the most dominant biologic drugs currently on the global market. In the case of antibody-based biotherapeutics, the absence of fucosylation or increase in galactosylation is needed for either antibody dependent cell cytotoxicity [1,2] or complement-dependent cytotoxicity [3,4], respectively, whilst additionally, antibody mannosylation is important for clearance [5]. For these reasons, glycosylation is a critical quality attribute (CQA) of most biologics. This necessitates control of glycosylation processing during a drug process development stage to ultimately relay a consistent glycosylation of the biologic product during manufacturing [6]. This is difficult because glycosylation during fermentation occurs with a high degree of heterogeneity and is influenced by several factors including the host expression system and process parameters such as temperature shifts, pH, and the type of basal/feed media [7]. To understand how these environmental factors impact the glycosylation of a biologic, analytical methods are needed to assess how glycans behave under these diverse conditions. During this process development of antibody-based drugs, the N-glycosylation of an antibody can deviate from their expected glycomic profiles as a result of fluctuations in culture conditions and operating parameters. Therefore, to assess antibody glycosylation accurately, high-throughput analysis of hundreds to thousands of profiles is required for the identification of critical process parameters that control the glycosylation CQAs [8].

For complete bioprocessing analysis, favorable glyco-analytical methods need to convey a qualitative description of the glycans, their relative abundance, and most importantly be high-throughput in terms of quantity, comprehensiveness, and speed of data generation. Capillary electrophoresis-laser-induced fluorescence (CE-LIF) is a glycomic analytical technology that has been adapted for automated and high-throughput analysis [9]. In CE-LIF, released and fluorescently labelled glycans migrate over a capillary and are identified by comparison to the standardized migration time with external or internal oligosaccharide standards. In order to achieve standardized migration time in a high-throughput manner, migration time is generally calculated by correlation with internal standards that bracket the time of elution of the glycans of interest [10]. This process is used to calculate a glucose unit (GU) which helps to align the datasets so that the GU of each glycan can be used to identify the glycan through available GU-based glycan databases [11-13]. The technique is suitable for the assessment of glycosimilarity of biologics [14] and most importantly has potential for analyzing large cohort studies to assess the aforementioned process parameters and their correlations with antibody glycosylation [7]. GU databases and software (among others) are discussed in a recent review [15].

A long standing problem associated with the analysis of large sets of electrophoretic data generated during bioprocessing is inevitably the drift of the peak migration time and area under the curve pertaining to glycan structures. This can be caused by a combination of sample complexity, temperature, pH, day of analysis, and other physicochemical fluctuations during the operation of the analysis. Although GU calculation can help solve this for the qualitative analysis, there is still difficulty automating peak picking due to small peaks and peaks that can lose their "Gaussian-ness" when multiple peaks migrate close together. This is especially true for large sets of diverse CE electropherograms collected over days or months under varied conditions. Consequently, they are often processed with automated software using different parameter settings for each electropherogram (or groups of similar electropherograms) requiring substantial human intervention to check correctness of the automated picked peaks and tuning parameters. This level of human manual data analysis is impractical when dealing with thousands of samples.

Here, we describe a computational solution for the identification and quantitation of glycans in a large glycomics CE dataset generated during process development of an anti-HER-2 antibody. The method is a semi-automated approach and improved accurate glycan assignments and quantitation compared to other tested fully automated software. Briefly, the method performs clustering analysis of glycomic electropherograms to group them into manageable clusters, followed by subsequent quantitation after semi-automated curation using the open source software HappyTools [16]. The clustering and migration time calibration in HappyTools allows for easy manual peak picking (spending 1 to 3 hours) before quantitation begins. After peaks are defined, large sets of electropherograms can be processed

expediently and efficiently without any further need for human intervention either pre or post-quantitation. To the best of our knowledge, we are the first to apply this computational approach to a large set of CE-LIF glycomic data. The result of this new method is that large cohorts (thousands) of bioreactor runs can be analyzed at once with high accuracy in quantitation and glycan identification. We demonstrate this approach through the high-throughput qualitative and quantitation of CE-LIF glycomic data, displaying glycan trends that exist in eleven in-house bioreactor culture conditions. Most importantly we show that the quantitation is consistent with respect to other software. The key advantage of this computational approach is that all runs can be analyzed simultaneously with high accuracy in glycan identification and quantitation and there is no theoretical limit to the scale of samples that can be processed using this method.

Results and Discussion

Anti-HER-2 cultures

A comparison of a large set of glycosylation profiles derived from the bioprocessing and harvesting of Anti-HER-2 antibodies every day across 11 different culturing conditions. Specifically, Anti-HER-2 antibodies were harvested from 3 technical replicates for biological replicate across 12 days and 11 different culturing conditions (Supporting Information File 1, Table S1). Five of the replicates failed due to sampling errors, leaving a total of 391 electropherograms to identify and quantitate glycans. The N-glycans were enzymatically removed, fluorescently labelled with aminopyrene trisulfonate (APTS), and analyzed by capillary electrophoresis. The N-glycans were separated using a 5 minute separation across a 30 cm capillary. N-Glycan peaks in the electropherograms were annotated for all 391 electropherograms separately demonstrating that varying culture conditions resulted in significant differences in the electropherograms, i.e., certain glycan peaks became absent or present depending on the conditions and day of culture.

Problems with automated identification and quantitation of glycans using Gaussian approximations

Several approaches were examined to compare glycan identification and relative quantities between electropherograms for one set of results. This single set consisted of one bioreactor condition containing 12 days of CE-LIF electropherograms with 3 technical replicates (12 × 3 = 36 electropherograms). Figure 1 shows the approach we found to be optimal for this batch. The

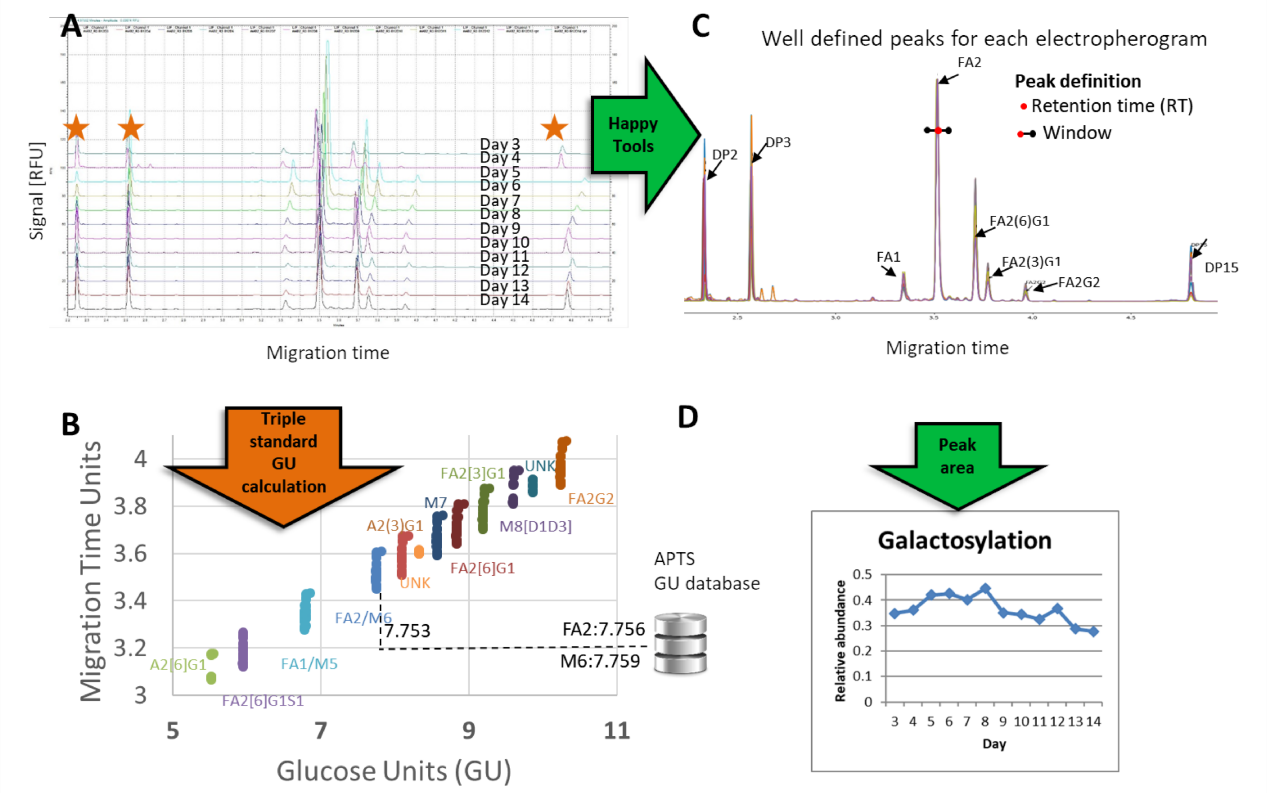


Figure 1: A single bioreactor run with defined culture conditions for twelve days. (A to B) batch GU calculation using the triple standard approach. The orange star marks the three bracketing standards. (B) Dotplot of the GU value vs. migration time. (A to C to D) HappyTools software allowed easier quantitation since all peaks can be aligned/calibrated and all peaks start and end migration times can be defined before quantitation begins using HappyTools.

approach used to identify glycans was based on a triple standard GU calculation [10,11] and database matching whilst the quantitation used HappyTools calibration and area calculation [14]. The GU calculation involved standardizing the migration time of the peaks by generating a 'virtual' glucose unit (GU) ladder calculated using the migration time of the 3 oligosaccharide standards that were separated with each sample. The migration times of glycan peaks were then translated to a calculated GU that made it easy to compare peaks and identify glycans across electropherograms (Figure 1B). Despite major misalignment of migration time and bracketing standards in the electropherograms, the variation of GU values for glycan peaks generally were within a very small range (Figure 1A and 1B) allowing for consistent database matching against a GU CE database (APTS fluorescent labelled) [10].

HappyTools was used to calibrate the migration times of all the electrophoretic peaks (Figure 1C), define peak boundaries, and quantitate the glycans. Migration time calibration involved aligning peaks so that each glycan peak fell under the same migration time (Figure 1C). After alignment/calibration the peaks were easy to define manually and thus quantitation could be achieved using defined peak windows, migration time positions, and HappyTools peak area calculations (Figure 1C and 1D). Quantitation could be achieved with a simple user interface and HappyTools returned the glycan profile and quantitation results efficiently.

Unfortunately, the automated quantitation with HappyTools (Gaussian mode) and other software were hampered by complications in peak picking and peak area calculations of non-Gaussian peaks (see next section). This required significant time (2–3 days) to manually inspect and correct the quantitative values of the peaks in the subset of samples. The simplified approach shown in Figure 1 although useful for electropherograms with homogenous peaks would not be practical given the scale of sample numbers and heterogeneous nature of the samples we needed to investigate. Further investigation and alternative approaches were needed to facilitate better peak picking and quantitation with electropherograms that were composed of heterogeneous peak shapes in our glycan analysis.

Peak detection and quantitation of non-Gaussian peaks using Riemann approximation

Quantitation of sets of electropherograms that were similar (i.e., technical replicates or biological replicates with the similar operating conditions) was feasible by manually tuning the parameters in software such as 32 Karat (Sciex) [17] (results not shown). However, when there was large heterogeneity in the CE-LIF electropherograms, as would be the case in a bioprocessing operation, a single set of tuned parameters failed to detect peaks and therefore quantitate them. On our dataset, the 32 Karat software needed tuning of parameters for multiple clusters of similar electropherograms; this job was laborious and thus motivated the implementation of our computational approach. The main reason quantitation was complicated by automated data analysis methods, whether using 32 Karat software [17] or HappyTools quantitation functions, was because of Gaussian peak approximation [16]. Using 32 Karat with a single set of default parameters there were difficulties with consistently peak picking closely eluting peaks and this led to inconsistent peak quantitation. Figure 2 shows two peaks that had similar migration time containing glycans FA1/FA2G2S1/A2 and M5 (identification results shown later). Using default settings, the 32 Karat software gave a peak area in one of three ways: for both (Figure 2A), only FA1/FA2G2S1/A2 (Figure 2B), or only M5 (Figure 2C). Similar anomalies in the FA1/FA2G2S1/A2 and M5 peak quantitation were also found when we attempted to fit the peaks using the Gaussian functions in HappyTools (results not shown). Peak quantitation was improved once we switched to non-Gaussian area calculation in HappyTools that used a Riemann sum between manually determined start and end migration times. The Riemann sum setting was recommended previously [16] for the quantitation of asymmetric, non-Gaussian peaks. In Figure 2, the improvement is shown for the FA1/FA2G2S1/A2 and M5 peaks where their relative sizes on the electropherogram compared well with the Riemann sum peak area calculations, i.e, almost equal areas (Figure 2A vs 2D), FA1/FA2G2S1/A2>M5 (Figure 2B vs 2E), and FA1/FA2G2S1/A2<M5 (Figure 2C vs 2F). Further, peak area consistency was achieved between 32 Karat and the HappyTools/Riemann sum approach when FA1/FA2G2S1/A2 and M5 peaks were combined (Figure 2G and 2H vs 2I). Thus, both approaches had consistent peak area calculations when considering the two peaks as one. However, the HappyTools/Riemann sum approach was advantageous because it allowed for a separation of the two distinct peaks thus giving a finer level of glycan detail.

Problems aligning and comparing the large cohort data with Gaussian modelling of electrophoretic data

HappyTools calibration and quantitation worked well with the single bioreactor condition shown in Figure 1A–C because the electropherograms were similar. Upon expanding the same analysis workflow across all 391 electropherograms, inconsistent calibration was observed for the different electropherograms, resulting in peaks that were hard to define (Figure 3A). The reason for this difficulty in defining peaks was because of peak misalignment caused by the heterogeneous nature that resulted from fluctuations in day-to-day electrophoretic oper-

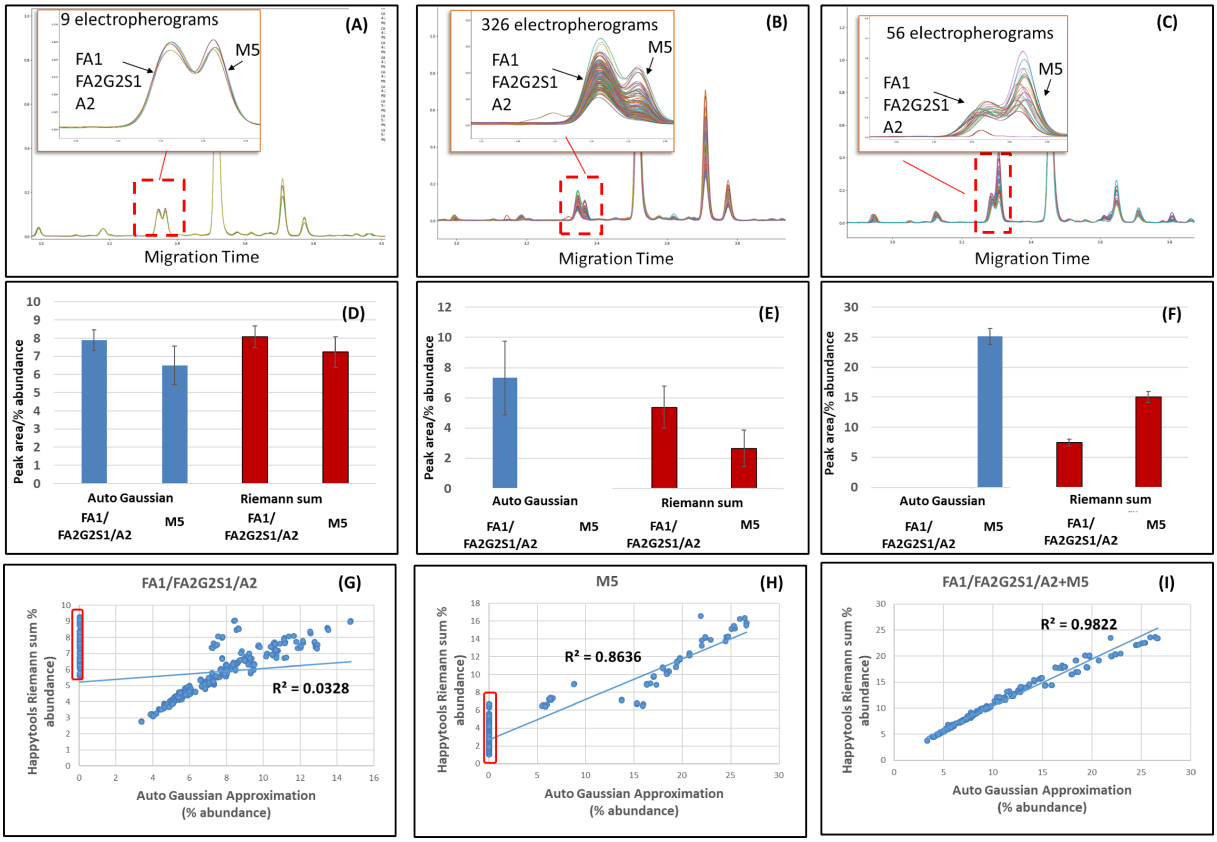


Figure 2: Problems when integrating poorly resolved peaks using FA1/FA2G2S1/A2 and M5 peaks as an example. (A) FA1/FA2G2S1/A2 and M5 had similar peak areas. (B) FA1/FA2G2S1/A2 had a greater peak area than M5. (C) FA1/FA2G2S1/A2 had less peak area than M5. (D) Average peak area and standard deviation (error bars) for 9 electropherograms in A. (E) Average peak area and standard deviation (error bars) for 326 electropherograms in B. (F) Average peak area and standard deviation (error bars) for 56 electropherograms in C. (G). Correlations between 32 Karat and HappyTools/Riemann sums for FA1/FA2G2S1/A2 only (red box peak not picked by 32 Karat), (H) M5 peaks only (red box peak not picked by 32 Karat, and (I) when FA1/FA2G2S1/A2 and M5 peak areas were combined there was excellent correlation between both approaches suggesting 32 Karat integrates the two peaks as a whole.

ating conditions such as temperature, voltage changes, etc. For some electropherograms, differences were the result of new glycan peaks attributed to the biological variations introduced via multivariate culturing conditions. We therefore implemented a clustering algorithm that allowed us to group electropherograms before applying HappyTools calibration (Figure 3B).

Clustering and manual peak picking efficiently and comprehensively quantitate large cohorts of glyco-profiles

The electropherograms were grouped using unsupervised clustering and the peak intensity as input variables. From our analysis, the 391 electropherograms were clustered into three distinct groups (Figure 3B) of electropherograms. Visualization by overlaying the electropherograms in each cluster (Figure 3B) showed that it was easier to define user-generated peak migration times and delta-windows (Supporting Information File 1, Table S2). The clustering simply facilitated manual peak picking thus avoiding the pitfalls associated with automated peak picking. The manually determined data in Supporting Information File 1, Table S3 was transferred to HappyTools quantitation Riemann peak area functions via its analysis file. Therefore, once we defined the peaks manually our clustering + HappyTools computational approach could quantitate similar groups of electropherograms separately on a large scale. In total, Supporting Information File 1, Table S3 shows there were 17 peaks manually identified that required glycan annotations and quantitation.

The semi-automated approach of clustering electropherograms combined with manual peak curation and HappyTools consistently outperformed automated approaches: 32 Karat (Sciex) and HappyTools automated functionality. On close inspection, it was noted that peak integration under a Gaussian approximation would yield a high variation in the number of picked peaks per electropherogram (Figure 4A). The number of peaks picked using the automated approaches was on average 7 peaks lower

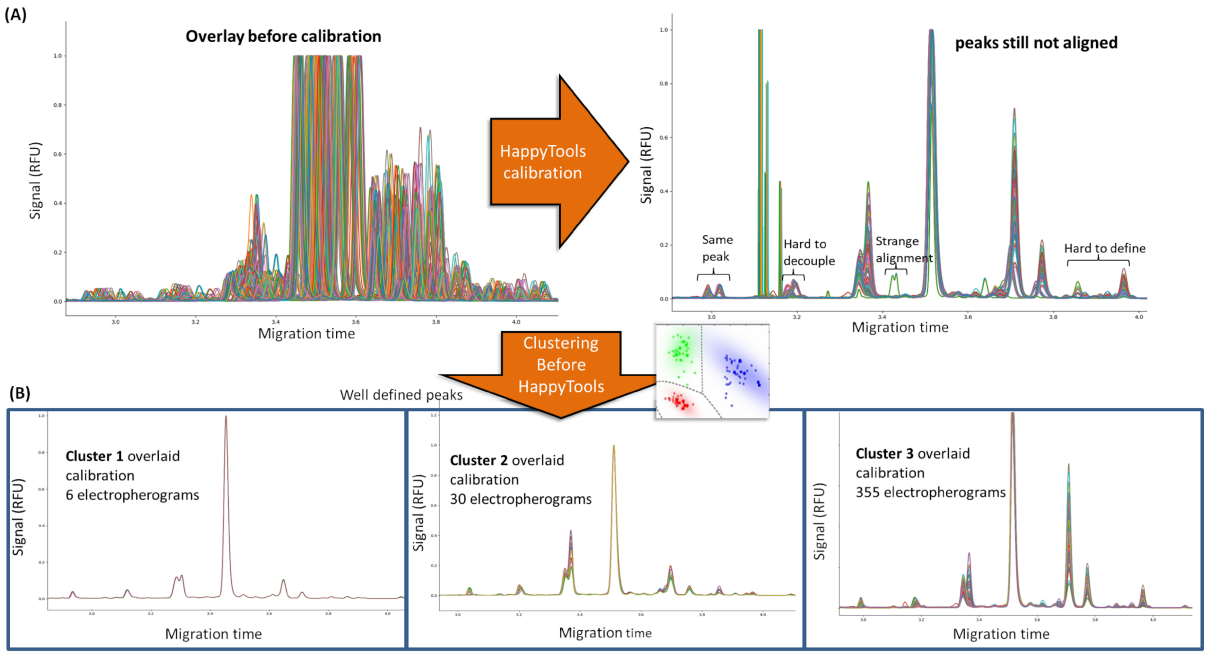


Figure 3: The clustering function allowed grouping of similar electropherograms and therefore clean the HappyTools calibration output. (A) HappyTools calibration of all 391 electropherograms was largely misaligned. This made peaks not well aligned and made it difficult to define the peak positions for quantitation. (B) The clustering grouped the HappyTools calibration into three clusters of similar electropherograms. Each cluster had well defined peaks that were manually determined.

than our clustering, manual peak curation, and quantitation using HappyTools (Figure 4B). Furthermore, for low abundant peaks and peaks with close migration times, automated peak quantitation randomly misses out peaks which were discovered using clustering and manual curation (Figure 4B). These low abundant peaks accounted for a significant share of the glycan peak abundance. Peaks that constituted <1% abundance account for 3.8% of the total average abundance. Peaks that constituted 1–2% abundance accounted for 4.0% of the total average abundance and peaks >2% accounted for 92.2% of the average total abundance. If peaks <1% are not considered it might seem insignificant but it affects the relative abundance of other peaks to an extent of 3.8% in total.

Supporting Information File 1, Figure S1 shows the correlation between clustering + HappyTools and the automated software quantitative strategies. The high correlations in Supporting Information File 1, Figure S1 for the major peaks show that the quantitative calculations of both approaches were similar. However, as mentioned previously we found that the quantitation algorithms provided by the automated software required a substantial amount of human intervention (2 to 3 days approximately). The manual checking was needed because of missing peaks (Supporting Information File 1, Figure S1; red rectangles, Figure 4) and erroneous peak area calculation for close peaks (Figure 2). The user would have to identify exactly which samples out of hundreds or thousands of samples were incorrectly determined, introducing human error and slowness back into the automated process. On the other hand, our clustering + HappyTools the quantitative approach was quick, taking only 2 minutes on a standard personal computer for all 391 electropherograms. However, there was the need to manually define the peaks in Supporting Information File 1, Table S3 which required 1 to 2 hours of examining the electropherogram overlays in Figure 3B. At this stage we had optimized our quantitation protocol but prior to applying the quantitation algorithm, we needed to identify what glycans were eluted at each peak.

Glycan identification for all 11 bioreactor runs

Glycan identifications to all peaks in all 391 electropherograms were confirmed using evidence from two orthogonal approaches, UPLC-MS (RFMS fluorescent label) analysis and a CE-LIF analysis (APTS label). The UPLC-MS approach was used to characterize a commercial Anti-HER-2 reference standard (Section 4.3.4 from mentors) using the UNIFI software and GU/mass database (RFMS labelled) provided by Waters Corp [18]. Figure 5A shows that 14 glycans were identified in 13 UPLC peaks using UPLC-MS glycomics analysis. Figure 5B shows that 17 glycans were identified for 14 peaks using CE. Out of the 17 glycans, there were 11 also confirmed by UPLC-

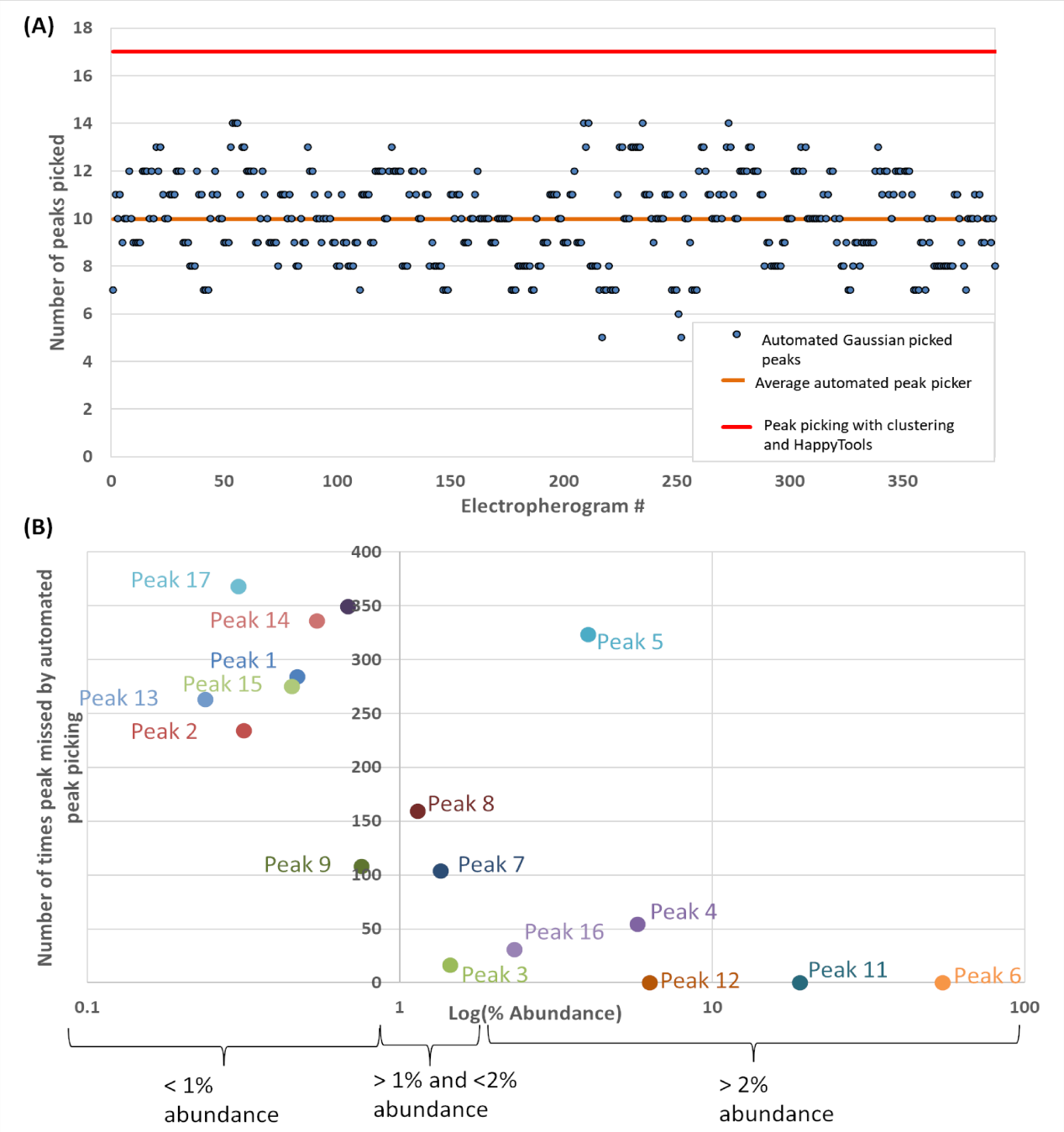


Figure 4: Comparison of the performance of the automated peak picking and semi-automated clustering and HappyTools quantitation for the 391 electropherograms. (A) Automated quantitation using Gaussian approximation approach picked on average 10 peaks (min = 5, max = 14) while our HappyTools + Clustering approach constantly picked 17 peaks. (B) Number of times automated Gaussian peak picking missed one of the 17 peaks listed in Supporting Information File 1, Table S3 as a function of % abundance.

MS (Figure 5A green boxes; Supporting Information File 1, Table S4 green column headers) thus increasing assignment confidence for those. Out of the 17 peaks in Supporting Information File 1, Table S3, 14 could be assigned glycan structures (Figure 5B) using GU database matching. Figure 5A shows that 3 UPLC-MS identified glycans, A1, FA1[3]G1, or A2[6]G1, were not found in the CE-LIF analysis. The reason for lack of A1 and FA1[3]G1 annotation in the CE-LIF was likely because they were not in the APTS GU database. Therefore, glycans A1 and FA1[3]G1 could be any one of the 3 unidentified glycans in the CE-LIF (Figure 5B marked UNK1, UNK2 and UNK3) and further investigation is needed. Glycan peak A2[6]G1 was in the APTS database with a GU of 8.153 and it did not match any of our peaks in Figure 5B. The fact that there were 17 glycans

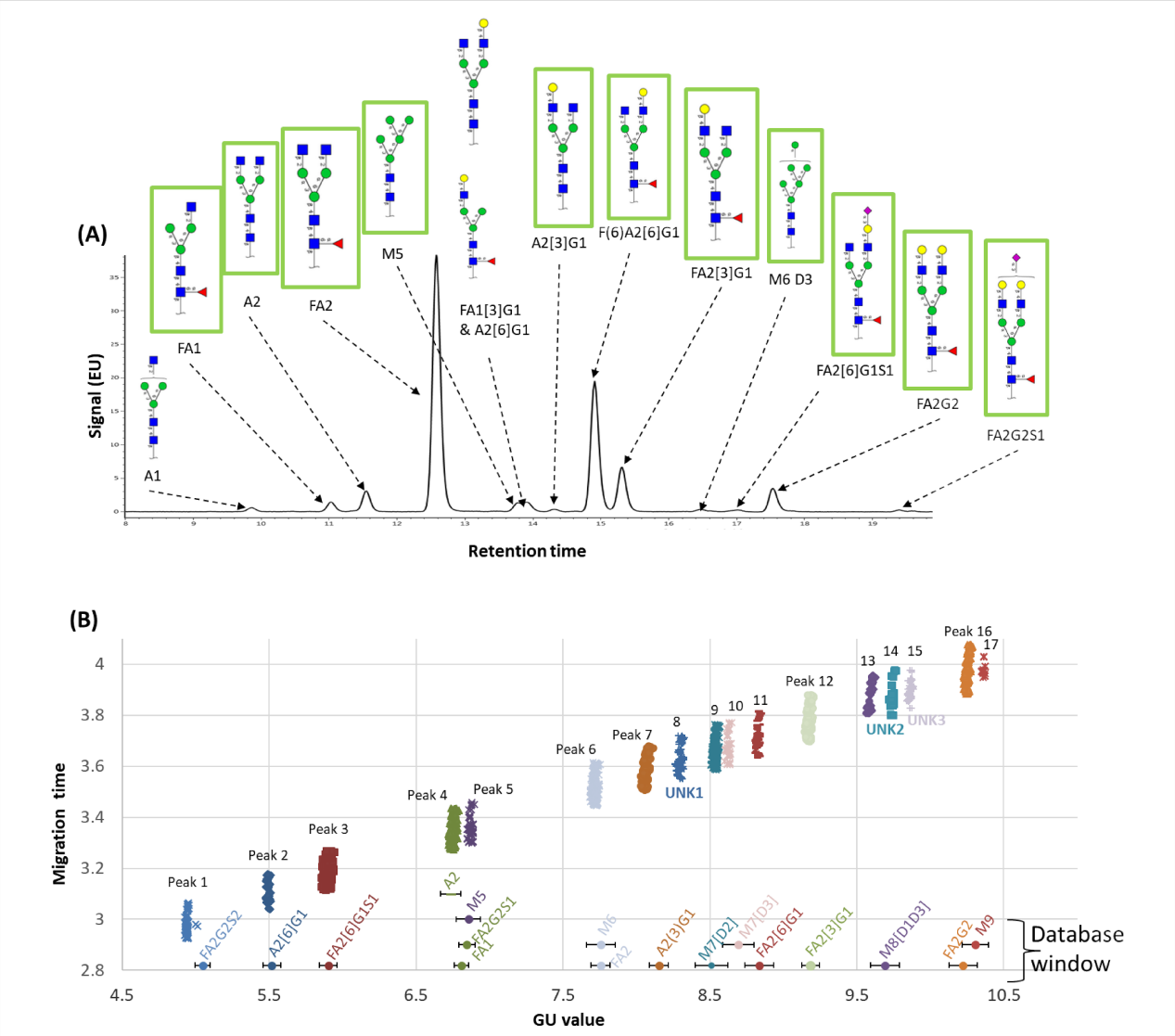


Figure 5: Glycans identified in anti-HER-2 samples using UPLC-MS and CE. (A) the UPLC chromatogram confirmed the 14 glycans using GU and mass. Green boxed glycans were also identified in CE. (B) Glucose units vs. migration time for all 391 CE electropherograms. Database matched glycans are shown in Oxford linear notation [19]. The CE APTS database hits are marked with a circle and a corresponding error bar showing the GU tolerance. All glycans with core fucose were α-1→6 linkage, galactose were β-1→4 linkage and all sialic acid linkages were α-2→3 linkage. All glycans are drawn in SNFG notation [20].

identified in the CE-LIF and 14 in UPLC-MS seems counter-intuitive but it can be explained by the vast degree of variation in our bioreactor conditions while the anti-HER-2 innovator was produced from a single harvested condition.

Using the UPLC-FLR-MS quantitation we could estimate the area under the curve contribution for each glycan in the co-eluting peaks 4 and 6. In UPLC-FLR-MS, FA2G2S1 was very minor (0.21%), while FA1 and A2 had an abundance of 1.8% and 3.9%, respectively. This suggests that A2 is the major component of peak 4 followed by FA1. Similarly, for FA2/M6 UPLC-MS quantitation showed FA2 with 49.3% and M6 with 0.32% abundance, respectively, suggesting the M6 component was relatively smaller for peak 6 quantitation. The ability to estimate the relative contributions of glycans in a co-eluting peak was another advantage of combining our CE analysis with UPLC-MS characterization.

Peak quantitation for all 11 bioreactor runs and all 17 peaks

Upon successful calibration and annotation of the 391 samples separated by CE-LIF, and analysis using clustering + HappyTools, we were able to compare the overall glycosylation profiles that resulted from the different bioprocessing operating conditions. Figure 6 shows the final quantitation for all 17 peaks in Supporting Information File 1, Table S3 and all

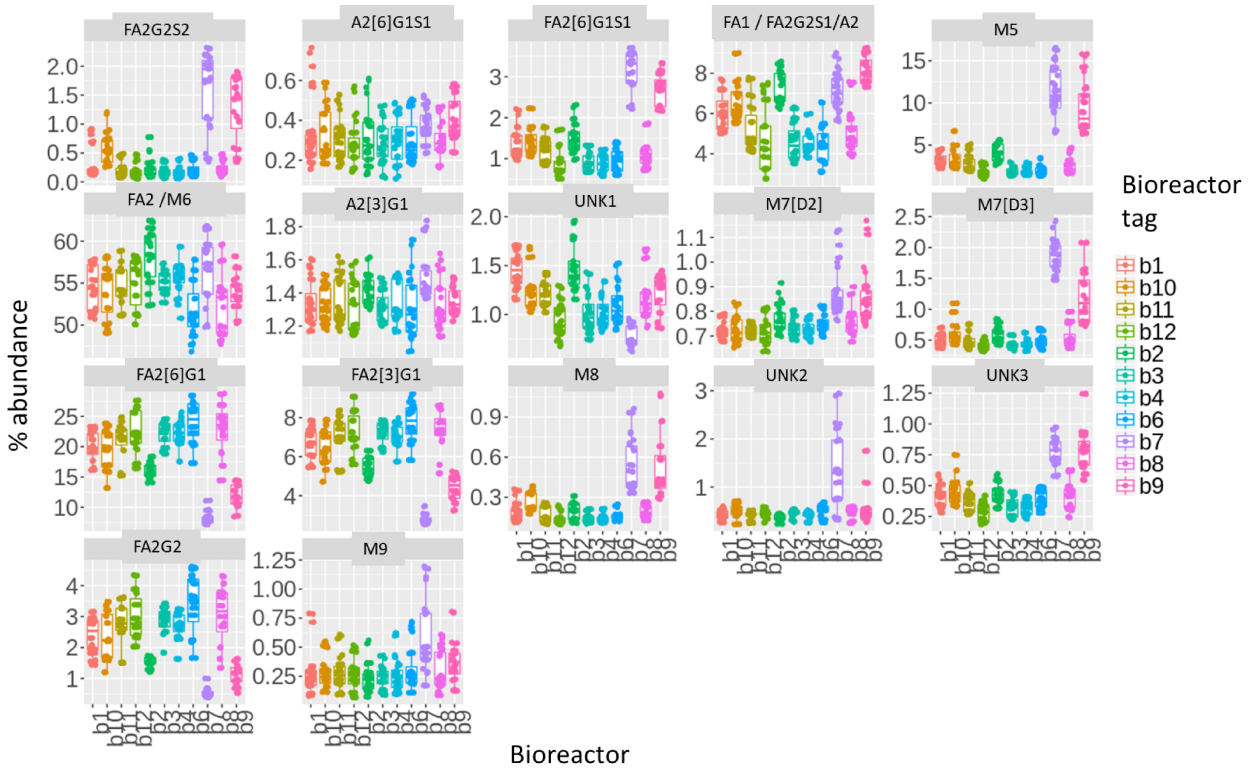


Figure 6: Boxplots showing the quantitation of the 11 different bioreactor conditions. The boxplots show the peak area distribution (expressed as % relative abundance) for each of the 11 bioreactor conditions. Points are averaged relative abundances for three replicates.

glycans using our proposed computational approach. The final relative abundances were an averaged percentage value from three technical replicates resulting in 132 observed glycan profiles. The variability between technical replicates was low as shown by their standard deviations in Supporting Information File 2. Conversely, the variability between bioreactor conditions was high and a number of other interesting observations can be concluded from the quantitation including: a) core fucose sialic acid based glycans had increased abundances in bioreactor condition b7, b8, and b9, b) mannosylation was increased in b7 and b9, c) neutral fucosylation decreased for b7 and b9. However, FA1 did not follow this trend perhaps because it was co-eluting with A2 and FA2G2S1, d) neutral galactosylation decreased for b7 and b9, e) fucosylated and galactosylated glycans FA2G2, FA2[6]G1, and FA2[3]G1 decreased for b7, b9 and b2. Conversely, the fucose agalactosylation glycan FA2 was increased for the b2 condition suggesting a change in the galactose transferase activity in the b2 culture condition. Additionally, the unidentified peaks in the qualitative analysis (UNK1, UNK2, and UNK3) had very small impact.

Supporting Information File 1, Figure S2 shows the intercluster quantitation variation for clusters 1, 2, and 3. It shows different quantitation patterns for 7 abundant glycans (average > 1% abundance) in each cluster. Interestingly, cluster 1 and 2 (Figure 3) were mainly composed of two particular bioreactor conditions (b7 and b9; Figure 6) and contained higher levels of FA2, FA1, and M5 and lower levels of FA2[6]G1 and FA2[3]G1 (Supporting Information File 1, Figure S2). This suggested that the clustering was also useful to group electropherograms based on culture conditions and glycan abundances and could be an important characteristic of the approach as it could allow quick identification of bioreactor conditions important for some glycan types.

Future work: process development of the glycans against various physicochemical parameters of the bioprocess

This work reports a computational pipeline we utilized for batch identification and quantitation of glycans in CE electropherograms from a large sample cohort. This approach will be subsequently utilized to evaluate the time-based glycosylation profiles of an antibody product (and therefore biologic quality) derived from bioreactors operating under varying culture conditions. The number of glycosylation profiles that can be processed has no theoretical upper limit. In this work, the computational approach was optimized on one bioreactor sam-

ple set and proved to be efficient and accurate. In future work we will apply this approach to substantially more bioreactor culture conditions thus producing massive amounts of data that could be used to optimize bioprocessing and biologic quality.

Conclusion

Fluctuations of glycomic profiles are a result of environmental and biochemical pathways of the glycoprotein. Understanding these effects requires a high-throughput analytical technique whereby diverse glycomic profiles of the glycoprotein/s can be compared qualitatively and quantitatively. For capillary electrophoresis-based glycomics data processing, this is complicated by heterogenous glycan peaks which may not fit under a Gaussian approximation. As such, automated Gaussian fitting of these peaks will yield inaccurate representation of the glycans expressed in the complex system. This is particularly true for peaks with very close migration times.

We observed that for large sets of heterogeneous electropherograms current software for quantitation was inadequate (although software such as 32 Karat proved excellent for glycan identification). We therefore describe a method whereby we perform clustering analysis of large cohorts of glycomic CE-LIF electropherograms, breaking them down into smaller, more manageable groups of similar electropherograms, followed by using open-source software for quantitation. The computational approach can be adapted to any analytical technique that produces large amounts of heterogeneous profiles as it allows for easy manual peak picking before quantitation begins. After peaks are defined, large cohorts of profiles can be processed expediently and accurately without any further need for human intervention pre or post-quantitation. Thus, the approach is semi-automated, achieving the scale of automation while still maintaining the accuracy of manual assignment.

We used this technique to comprehensively and accurately characterize the effect of multivariate bioprocessing conditions upon the glycosylation profile of an anti-HER2 antibody product. We found that our clustering + Happytools method reduces 2–3 days of human intervention needed for the automated software down to 1–2 hours for first-time analyses, but down to minutes for repeated analyses. We envision that this approach may be widely applicable to large cohort glycomic studies, where the comparison of glyco-profiles is important to clinical studies, cellular biology, and glycobiology in general.

Experimental

Materials

Sodium phosphate (Merck) and glycine·HCl (Merck) were purchased from Merck. Tris-HCl and EX-CELL Advanced CHO Fed-batch medium were all purchased from Sigma Aldrich. ACN (Part no: A955-4, Fisher Scientific), PVDF syringe filter (Part no: SLGV013SL, Millex 0.22um PVDF 13 mm sterile syringe filter) were obtained from Merck Millipore (Ireland), whilst centrifugal filters, (Part no: UFC3096, Amicon Ultra Centrifugal Filters) was purchased from Merck Millipore, (USA). Protein A HP Spintrap (28-9031-32) was purchased from GE Healthcare, USA. FAST Glycan Kit (Part no: B94499PTO, SCIEX, USA). Ammonium formate (Part No. 186007081), RapiGest SF (Part No. 186001860), RapiFluor-MS Reagent Solution (Part No. 186008091), and ACQUITY UPLC® Glycan BEH Amide Column were purchased from Waters Corp.

Cell culturing of samples

CHO-K1 cells producing Adalimumab biosimilar were cultured in 14-day fed-batch cultures using Ambr250 bioreactors. The cells were thawed and passaged three times in 30 mL of EX-CELL Advanced CHO Fed-batch medium supplemented with 6 mM of glutamine and 250 nM of MTX in 50 mL tube-spin cultures prior to bioreactor inoculation. Cells were inoculated into 200 mL of EX-CELL Advanced CHO Fed-batch medium supplemented with 6 mM of glutamine but without MTX at a viable cell density of 3×10^5 cells/mL. The cultures were mixed using dual pitch blade impellers stirring at 300 rpm. Different bioreactor operating conditions were evaluated. Triplicate experiments were performed for all operating conditions under study. In terms of the feeding strategy, 10% of EX-CELL Advanced CHO Feed 1 (with glucose) were added to all cultures on days 3, 5, 7, 9, and 11. When the concentration of glucose dropped to below 2 g/L, a specified volume of 45% glucose stock was added into fed-batch cultures in order to achieve a final glucose concentration of 6 g/L. Glycosylation analysis was performed for samples obtained daily from days 3 to 14.

Sample preparation of antibody N-glycans

Protein A purification and buffer exchange: The collected cell supernatant was filtered through a 0.22 μM PVDF syringe filter (Sterile Millex Filter, Merck Millipore, Ireland). Antibodies were then purified using protein A spin trap columns (Protein A HP Spintrap, GE Healthcare, USA). The columns were equilibrated and washed via centrifugation (100*g*, 30 s) with 20 mM sodium phosphate (pH 7.0). The sample was loaded to each column (maximum volume of 600 μL) and incubated end-over-end for 10 min. The column was washed with 20 mM sodium phosphate (pH 7.0) via centrifugation (100*g*, 30 s). The antibody samples were eluted from these columns with 0.1 M glycine·HCl (pH 2.7) and neutralized with 1 M Tris-HCl (pH 8.0). Samples were then buffer exchanged using 30 kDa Centrifugal filters, (Amicon Ultra Centrifugal Filters,

Merck Millipore, USA) into dH_2O and dried into 100 µg aliquots using a CentriVap benchtop vacuum concentrator (Labconco, USA).

Free N-glycan labelling with APTS: Free-N-glycans from purified antibodies were labelled with 8-aminopyrene-1,3,6-trisulfonic acid (APTS) using the FAST Glycan Kit (SCIEX, USA). Digestion, denaturing and labelling solutions were made according to the manufacturer's instructions. This protocol was adapted to a 96-well PCR plate. Two hundred (200 µL) of magnetic beads were used per 100 µg of glycoprotein. The magnetic bead storage solution was removed using a plate magnet. Antibodies of 100 µg in 10 µL (10 µg/µL) aliquots were added to the beads. The samples were incubated for 8 min at 60 °C with the denaturing solution. Then, digestion solution was added, and the sample was incubated for 20 min at 60 °C. Acetonitrile was then added to the sample, and then placed on a magnetic plate to separate the beads from the supernatant. The supernatant was removed, labelling solution containing an internal standard (DP3) was added, and samples incubated at 60 °C for 20 min in the dark. After incubation, a cleanup solution and acetonitrile were added, followed by separation on a magnetic plate, and removal of supernatant. This cleaning step was repeated a further 2 times. The labelled glycans were eluted from the beads using 100 µL deionized water and placed on the plate magnet. Eluted labelled glycans were then stored at −21 °C in the dark until further analysis.

Analysis of the released glycans using capillary electrophoresis

Capillary electrophoresis of the released and APTS-labelled antibody N-glycans was performed on a CESI8000 CE instrument (Sciex) equipped with a solid state laser-induced fluorescent detector (excitation 488 nm, emission 520 nm). Separations were made across a 20 cm effective length (30 cm total length), 50 µm i.d. uncoated bare fused capillary, HR-NCHO separation gel buffer (Sciex). The applied electric field strength was 1500 V/cm with the cathode at the injection side and the anode at the detection side (reversed polarity). Samples were electrokinetically injected using 1kV for 5 s. For migration time correction, a bracketing standard (BST) was co-injected with each sample. Samples were run in triplicate and a blank water injection without BST was run periodically throughout the analysis. 32Karat version 10.1 was used to control the instrument

Analysis of Anti-HER-2 innovator released glycans using UPLC-MS

The supernatant was purified using Protein A HP SpinTrap (GE Healthcare). The purified glycoprotein obtained was buffer exchanged into water using a 10 kDa molecular weight cut-off filter (Merck Millipore) to eliminate any salts and nucleophiles that could interfere with the subsequent steps.

N-glycans were analyzed from the anti-HER-2 innovator monoclonal antibody using the RapiFluor-MS (RFMS) N-glycan kit (Waters Corp). Fifteen micrograms (15 µg) of glycoprotein was dried down and reconstituted in a digestion buffer (final concentration 0.01% RapiGest) and heated to 95 °C for 5 min to denature the glycoprotein. After the mixture was cooled to room temperature, 600 U of recombinant PNGase F (New England Biolabs) were added. The mixture was incubated at 55 °C for 10 min to enzymatically cleave the N-glycans from the protein and then cooled to room temperature. RapiFluor-MS Reagent Solution (0.07 mg/µL in anhydrous dimethylformamide (DMF, Waters Corp.) was added to the released glycans, and the labelling proceeded at room temperature for 5 min. The reaction mixture was diluted by adding acetonitrile (ACN; final concentration 89.5%) in preparation for HILIC SPE. Purification of the RFMS-labelled glycans was performed using a 96-well GlycoWorks HILIC µElution Plate (Waters Corp). The plate was initially equilibrated with sequential washes with dH_2O and 85% ACN. The samples were loaded onto the wells and washed with 90% ACN/1% formic acid (FA). The glycans were eluted with GlycoWorks SPE Elution Buffer (Waters Corp). The eluted glycans were dried and reconstituted in a 22.5% (v/v) DMF, 25% (v/v) ACN. The glycans were analyzed via UPLC-MS using a H-Class UPLC equipped ACQUITY UPLC® Glycan BEH Amide Column, (130 Å, 1.7 µm, 2.1 mm × 150 mm) (Waters Corp) which was coupled to a Xevo G2S QToF (Waters Corp). The flow rate was set at 0.4 mL/min and a linear gradient was used: 25–49% of buffer A (50 mM ammonium formate solution, pH 4.4) and buffer B (100% ACN) was run across 40 minutes, followed by a 3 min wash step using buffer A. The column was then equilibrated back to 25% buffer A. The labelled glycans were detected with an FLR detector (Ex 265 / Em 425 nm). The sample manager was set at 10 °C and the temperature of the column was kept at 60 °C throughout the analysis. Glycan masses were measured on the Xevo G2S QToF using sensitivity mode in positive mode. A mass range of *m*/*z* 400–2000 was used, with an acquisition speed of 1 Hz, and the mass spectrometry was set at the following conditions: 2.75 kV electrospray ionization capillary voltage, 15 V cone voltage, 120 °C ion source temperature, 300 °C desolvation temperature, 800 L/h desolvation gas flow. A lockspray [Glu1]-Fibrinopeptide B Standard (Waters Corp.) was also used throughout the run to maintain mass accuracy. Dextran ladder (Waters Corp.) was run to obtain a calibration curve with a cubic spline fit. The retention times were normalized using the calibration curve to glucose units (GU). The data obtained was processed and analyzed with the UNIFI Biopharmaceutical software platform (version 1.8).

Data analytics

Qualitative protocol: Sciex 32Karat version 10.1 included a GU Value calculation component (FastGlycan) that was used to identify glycans in the acquired data. It is based on the triple standard approach previously described [10,11]. Glycans were matched to an GU-CE APTS database by finding the closest GU value in the database to the observed GU value. For UPLC-MS the data obtained was processed and analyzed with the UNIFI Biopharmaceutical software platform (version 1.8) where glycans were matched using the internal UNIFI RFMS GU-mass database and corresponding functions.

Quantitative protocol: Our quantitative approach consists of two software components: HappyTools previously described [16] and our in-house clustering algorithm. HappyTools was first used to calibrate/align the electropherograms. HappyTools performed calibration by examining user defined calibrant peak list consisting of: the third bracketing standard DP15, consistently highly abundant Anti-HER-2 glycan peaks such FA2, FA1 and FA2G2. This gave a good spread of calibration peaks across the electropherograms. For details on the calibration algorithm see [16]. The calibrated electropherograms were then clustered. The clustering algorithm was implemented in-house using the SciPy python package. The clustering step consists of hierarchical clustering using a single linkage algorithm and forms flat clusters using the inconsistency method with a cut-off threshold of 0.7, which was determined as achieving the same discrimination as manual classification on some test electropherograms. The data points presented to the clustering algorithm were an array of continuous signal intensities between migration time 2.9 and 4.1 (i.e., the Anti-HER-2 peaks). The pairwise similarity between any two electropherograms was calculated using Euclidean distance metric. The clustering algorithm is presented in Supporting Information File 1, Figure S3. After clustering, each cluster contained N electropherograms and each peak's central migration time (CRT) and window (ΔW) were defined by visualizing the N electropherograms as an overlay. Then, each electropherogram was quantitated by supplying CRT and ΔW for all peaks via the HappyTools analysis file. Peak area was calculated using the Riemann sum

$$P = \sum_{t=CRT-\Delta W+1}^{CRT+\Delta W} I_i (r_t - r_{t-1})$$

where I_i is the peak intensity and rt is the migration time at t in the electropherogram.

For quantitation comparison, HappyTools Gaussian fitting and the Sciex 32Karat version 10.1 default peak area functionality was also calculated.

Supporting Information

Supporting Information File 1

Additional tables and figures.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-176-S1.pdf]

Supporting Information File 2

Variability standard deviation data.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-176-S2.xlsx]

Acknowledgements

Authors would like to thank Gavin Teo for helping to generate some of the data.

Funding

The authors thank the Agency for Science, Technology and Research (A*STAR), Singapore for supporting this study (SSF Project Grant A1818g0025).

ORCID® iDs

Matthew S. F. Choo - https://orcid.org/0000-0002-1376-3352
Terry Nguyen-Khuong - https://orcid.org/0000-0002-5852-542X

References

1. Pereira, N. A.; Chan, K. F.; Lin, P. C.; Song, Z. *mAbs* **2018,** *10,* 693–711. doi:10.1080/19420862.2018.1466767
2. Chung, S.; Quarmby, V.; Gao, X.; Ying, Y.; Lin, L.; Reed, C.; Fong, C.; Lau, W.; Qiu, Z. J.; Shen, A.; Vanderlaan, M.; Song, A. *mAbs* **2012,** *4,* 326–340. doi:10.4161/mabs.19941
3. Raju, T. S. *Curr. Opin. Immunol.* **2008,** *20,* 471–478. doi:10.1016/j.coi.2008.06.007
4. Aoyama, M.; Hashii, N.; Tsukimura, W.; Osumi, K.; Harazono, A.; Tada, M.; Kiyoshi, M.; Matsuda, A.; Ishii-Watabe, A. *mAbs* **2019,** *11,* 826–836. doi:10.1080/19420862.2019.1608143
5. Goetze, A. M.; Liu, Y. D.; Zhang, Z.; Shah, B.; Lee, E.; Bondarenko, P. V.; Flynn, G. C. *Glycobiology* **2011,** *21,* 949–959. doi:10.1093/glycob/cwr027
6. Schiestl, M.; Stangler, T.; Torella, C.; Čepeljnik, T.; Toll, H.; Grau, R. *Nat. Biotechnol.* **2011,** *29,* 310–312. doi:10.1038/nbt.1839
7. Ehret, J.; Zimmermann, M.; Eichhorn, T.; Zimmer, A. *Biotechnol. Bioeng.* **2019,** *116,* 816–830. doi:10.1002/bit.26904
8. Costa, A. R.; Rodrigues, M. E.; Henriques, M.; Oliveira, R.; Azeredo, J. *Crit. Rev. Biotechnol.* **2014,** *34,* 281–299. doi:10.3109/07388551.2013.793649
9. Reusch, D.; Haberger, M.; Kailich, T.; Heidenreich, A.-K.; Kampe, M.; Bulau, P.; Wuhrer, M. *mAbs* **2014,** *6,* 185–196. doi:10.4161/mabs.26712
10. Jarvas, G.; Szigeti, M.; Chapman, J.; Guttman, A. *Anal. Chem. (Washington, DC, U. S.)* **2016,** *88,* 11364–11367. doi:10.1021/acs.analchem.6b03596

11. Jarvas, G.; Szigeti, M.; Guttman, A. *Electrophoresis* **2015,** *36,* 3094–3096. doi:10.1002/elps.201500397
12. Zhao, S.; Walsh, I.; Abrahams, J. L.; Royle, L.; Nguyen-Khuong, T.; Spencer, D.; Fernandes, D. L.; Packer, N. H.; Rudd, P. M.; Campbell, M. P. *Bioinformatics* **2018,** *34,* 3231–3232. doi:10.1093/bioinformatics/bty319
13. Jarvas, G.; Szigeti, M.; Campbell, M. P.; Guttman, A. *Glycobiology* **2020,** *30,* 362–364. doi:10.1093/glycob/cwz102
14. Borza, B.; Szigeti, M.; Szekrenyes, A.; Hajba, L.; Guttman, A. *J. Pharm. Biomed. Anal.* **2018,** *153,* 182–185. doi:10.1016/j.jpba.2018.02.021
15. Abrahams, J. L.; Taherzadeh, G.; Jarvas, G.; Guttman, A.; Zhou, Y.; Campbell, M. P. *Curr. Opin. Struct. Biol.* **2020,** *62,* 56–69. doi:10.1016/j.sbi.2019.11.009
16. Jansen, B. C.; Hafkenscheid, L.; Bondt, A.; Gardner, R. A.; Hendel, J. L.; Wuhrer, M.; Spencer, D. I. R. *PLoS One* **2018,** *13,* e0200280. doi:10.1371/journal.pone.0200280
17. Guttman, A.; Szigeti, M.; Lou, A.; Gutierrez, M. *Sciex Appl. Note* 2017.
18. Hilliard, M.; Alley, W. R., Jr.; McManus, C. A.; Yu, Y. Q.; Hallinan, S.; Gebler, J.; Rudd, P. M. *mAbs* **2017,** *9,* 1349–1359. doi:10.1080/19420862.2017.1377381
19. Harvey, D. J.; Merry, A. H.; Royle, L.; Campbell, M. P.; Dwek, R. A.; Rudd, P. M. *Proteomics* **2009,** *9,* 3796–3801. doi:10.1002/pmic.200900096
20. Varki, A.; Cummings, R. D.; Aebi, M.; Packer, N. H.; Seeberger, P. H.; Esko, J. D.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; Prestegard, J. J.; Schnaar, R. L.; Freeze, H. H.; Marth, J. D.; Bertozzi, C. R.; Etzler, M. E.; Frank, M.; Vliegenthart, J. F.; Lütteke, T.; Perez, S.; Bolton, E.; Rudd, P.; Paulson, J.; Kanehisa, M.; Toukach, P.; Aoki-Kinoshita, K. F.; Dell, A.; Narimatsu, H.; York, W.; Taniguchi, N.; Kornfeld, S. *Glycobiology* **2015,** *25,* 1323–1324. doi:10.1093/glycob/cwv091

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

GlypNirO: An automated workflow for quantitative *N*- and *O*-linked glycoproteomic data analysis

Toan K. Phung[‡1], Cassandra L. Pegg[‡1] and Benjamin L. Schulz[*1,2,§]

Full Research Paper

Open Access

Address:
[1]School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, Queensland, 4072, Australia and [2]ARC Training Centre for Biopharmaceutical Innovation, Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, St. Lucia, QLD 4072, Australia

Email:
Benjamin L. Schulz[*] - b.schulz@uq.edu.au

* Corresponding author ‡ Equal contributors
§ Telephone: +61 7 336 54875

Keywords:
glycoproteomics; mass spectrometry; N-glycosylation; O-glycosylation; Python

Beilstein J. Org. Chem. **2020,** *16,* 2127–2135.
https://doi.org/10.3762/bjoc.16.180

Received: 16 June 2020
Accepted: 20 August 2020
Published: 01 September 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editors: N. H. Packer and F. Lisacek

Abstract

Mass spectrometry glycoproteomics is rapidly maturing, allowing unprecedented insights into the diversity and functions of protein glycosylation. However, quantitative glycoproteomics remains challenging. We developed GlypNirO, an automated software pipeline which integrates the complementary outputs of Byonic and Proteome Discoverer to allow high-throughput automated quantitative glycoproteomic data analysis. The output of GlypNirO is clearly structured, allowing manual interrogation, and is also appropriate for input into diverse statistical workflows. We used GlypNirO to analyse a published plasma glycoproteome dataset and identified changes in site-specific *N*- and *O*-glycosylation occupancy and structure associated with hepatocellular carcinoma as putative biomarkers of disease.

Introduction

Glycosylation is a key post-translational modification critical for protein folding and function in eukaryotes [1-3]. Diverse types of glycosylation are known, all involving modification of specific amino acid residues with complex carbohydrate structures. *N*-Linked glycosylation of asparaginies and *O*-linked glycosylation of serines and threonines are the most widely encountered and well studied in eukaryotes. A key feature of glycosylation critical to its biological functions and important for its analysis is its high degree of heterogeneity [4]. This heterogeneity can take the form of variable occupancy, also known as macroheterogeneity – the presence or absence of modification at a particular site in a protein, due to inefficient transfer of the initial glycan structure [5]. In addition, the non-template-driven synthesis of glycan structures means that there

can be multiple different glycan structures attached at the same site in a pool of mature glycoproteins [6]. This structural heterogeneity is also known as microheterogeneity. This heterogeneity in glycan structure and occupancy can be influenced by many genetic and environmental factors. As such, protein glycosylation is often regulated in response to physiological or pathological conditions [7]. Accurately profiling the site-specific occupancy and structural heterogeneity of glycosylation across glycoproteomes can therefore provide insight into the biology of healthy and diseased states [8].

The current state-of-the art technology for the characterisation, identification, and quantification of the glycome or glycoproteome is liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS) [9]. Popular and powerful glycoproteomic workflows typically involve standard proteomic sample preparation and protease digestion, coupled with depletion of abundant proteins or enrichment of glycopeptides to enable their measurement. There have also been several advances in glycopeptide quantification strategies including chemical labelling, label-free and data-independent acquisition methods [10]. Progress in MS technology in particular has enabled deep and sensitive measurement of highly complex glycoproteomes, generating large amounts of high quality data [11]. With that comes the need for robust and automated workflows for extracting meaningful results. Numerous software packages have been developed for analysis of outputs from MS technology to automate the process of transformation of raw MS data into ion intensities and matching them with appropriate glycan and peptide sequence databases for glycopeptide identification (reviewed in [12-16]). However, there are few efficient, robust, and automated workflows for glycopeptide quantification. There are several freely available software programs for quantitative label-free glycoproteomics using MS1 or data-dependent acquisition. These include LaCyTools [17], MassyTools [18], and GlycoSpectrumScan [19], which use a predefined list of analytes and masses to interrogate MS1 data, and I-GPA [20], GlycopeptideGraphMS [21], GlycoFragwork [22], and GlycReSoft [23], which integrate identification and abundance/intensity information for glycopeptides (a recent review is provided in [10]). Importantly, the complexity of glycan heterogeneity requires that downstream analysis often involves manual processing in addition to standard informatics workflows.

Here, we developed and used GlypNirO, an automated bioinformatic workflow for label-free quantitative *N*- and *O*-glycoproteomics that focuses on improving robustness and throughput. Our workflow uses a collection of scripts built on an in-house sequence string handling library and the scientific Python data handling package pandas [24], and integrates outputs of two commonly used software packages in glycoproteomic MS data analysis, Proteome Discoverer (Thermo Fisher) and Byonic (Protein Metrics), to extract occupancy and glycoform abundancy of all identified glycopeptides from LC–MS/MS datasets. We applied the workflow to a published dataset comparing the plasma glycoproteomes of liver cancer patients (heptatocellular carcinoma, HCC) and healthy controls [20]. Our analysis revealed differences in occupancy and glycan compositions of several proteins as potential HCC tumor biomarkers.

Results and Discussion

GlypNirO

Byonic is powerful software that allows identification of glycopeptides and peptides from complex glyco/proteomic LC–MS/MS datasets but does not perform quantification. Proteome Discoverer allows robust and facile measurement of peptide abundances using MS1 peptide area under the curve (AUC) information. We developed GlypNirO to integrate the outputs from Byonic and Proteome Discoverer to improve the efficiency, ease, and robustness of quantitative glycoproteomic data analysis. GlypNirO takes Byonic and Proteome Discover output files, and user-defined sample information and processing parameters, performs a series of automated data integration and computational steps, and provides informative and intuitive output files with site or peptide-specific glycoform abundance data. Glyco/peptide identifications from Byonic are linked with AUC data from Proteome Discoverer by matching the experimental scan number. The sites of glycosylation within each peptide assigned by Byonic are identified and clearly labelled. While identification of glycopeptides based on peptide sequence and glycan monosaccharide composition is comparatively reliable with modern LC–MS/MS and data analytics, it is much more difficult to unambiguously assign the precise site of modification within a glycopeptide. GlypNirO therefore provides two options for analysis: site-specific or peptide-specific. If the user trusts Byonic's site-specific assignment, then all peptide variants that contain that site are included in calculations of its occupancy and glycoform distribution. If the user prefers to perform a peptide-specific analysis, then each proteolytically unique peptide form is treated separately. GlypNirO then calculates the occupancy and proportion of each glycoform at each site/peptide, and provides output files with the protein name, site and/or peptide information, and occupancy and glycoform abundance. Full details are provided in the Experimental section.

To provide a proof-of-concept use of GlypNirO, we performed an exploratory reanalysis of a previously published dataset [20] obtained from the ProteomeXchange Consortium via the MassIVE repository (PXD003369, MSV000079426). This study performed glycoproteomic LC–MS/MS analysis of whole plasma or plasma depleted of six abundant proteins from liver

cancer (hepatocellular carcinoma (HCC)) patients and healthy controls. We identified glycopeptides and peptides in the datafiles from these samples using Byonic, searching separately for *O*- and *N*-glycopeptides (Supporting Information File 1, Tables S1–S24), and processed the files with Proteome Discoverer (Supporting Information File 1, Tables S25–S36). We then used GlypNirO to process these results files. This analysis was able to identify and measure 851 *N*-glycopeptides (site-specific) from 150 proteins and 301 *O*-glycopeptides (peptide level) from 89 proteins (Supporting Information File 1, Tables S37–S40).

Several changes in site-specific glycosylation associated with HCC had been previously identified [20]. We benchmarked the performance of our workflow using GlypNirO with these previously reported changes. Consistent with previous analysis, we found that agalactosylated *N*-glycans on IgG were increased in abundance in HCC (Figure 1a), and the relative abundance of the HexNAc(5)Hex(6)NeuAc(3) composition at multiple sites on alpha-1-antichymotrypsin was decreased in HCC (Figure 1b).

N-Glycoproteome analysis

To extend our analysis, we next investigated the full suite of *N*-glycosylation sites that we were able to identify and measure with GlypNirO. Comparing the site-specific glycoform relative abundance and occupancy, we identified 111 unique glycopeptides with increased and 128 with decreased abundance in HCC compared with healthy controls in depleted plasma, and 93 increased and 67 decreased in HCC in non-depleted plasma ($P < 0.05$, Figure 2a and 2b). This analysis with GlypNirO of site-specific relative glycoform abundance confirmed that HCC was associated not only with changes in glycoprotein abundance in plasma, but with changes in the proportions of different glycan structures at specific sites in diverse glycoproteins.

Examining the data in more detail identified several sites with multiple glycoforms with statistically significant changes in abundance. Specifically, HCC patients had decreased abundance of disialylated *N*-glycans at alpha-1-antitrypsin N271 and haptoglobin N184 (Figure 3a and 3b), with increased abundance of non-sialylated *N*-glycans at fibrinogen N78

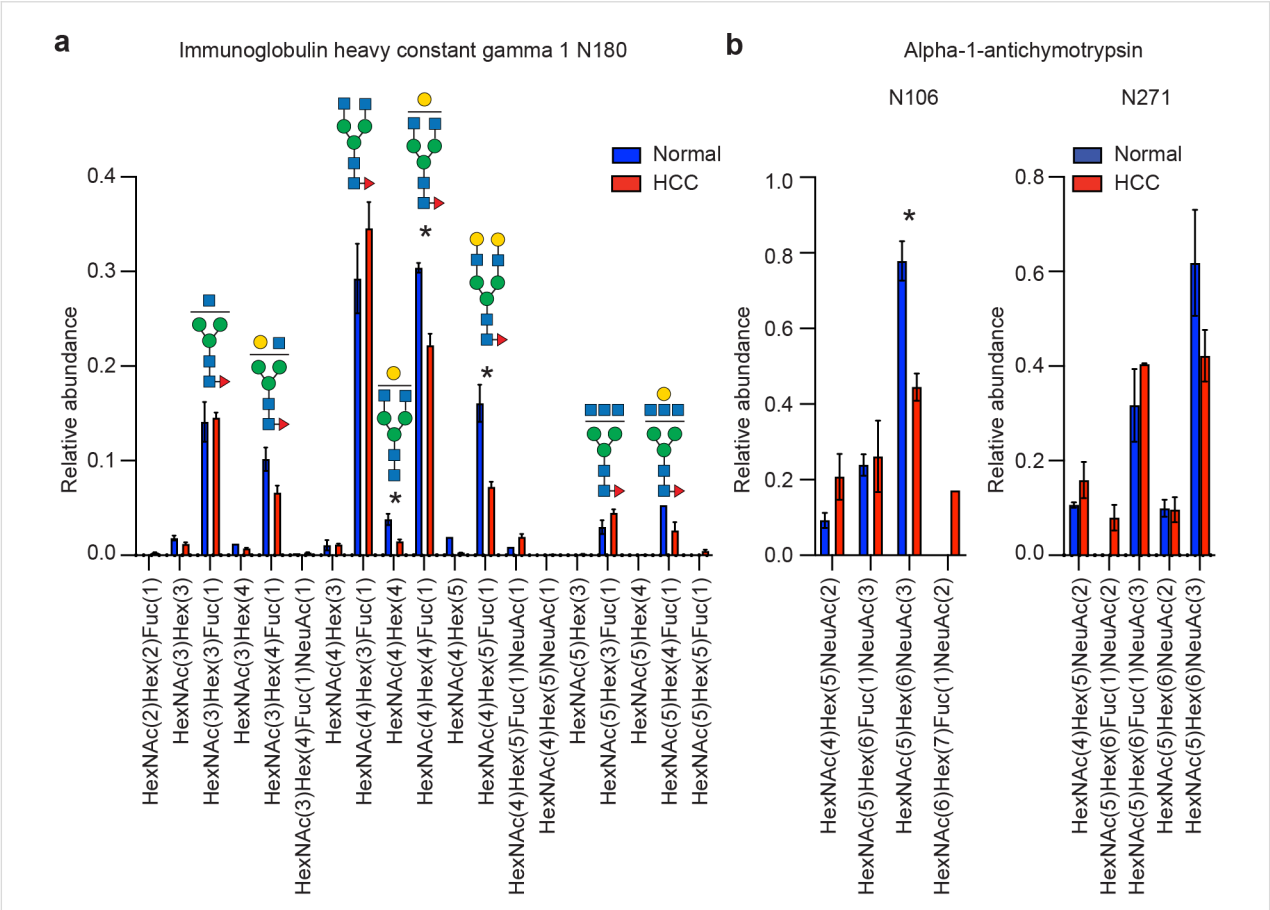


Figure 1: Evaluation of GlypNirO site-specific *N*-glycosylation profiling. Site-specific relative glycoform abundance in HCC and healthy controls at (a) immunoglobulin heavy constant gamma 1 (IgG1) N180, and (b) alpha-1-antichymotrypsin N106 and N271.

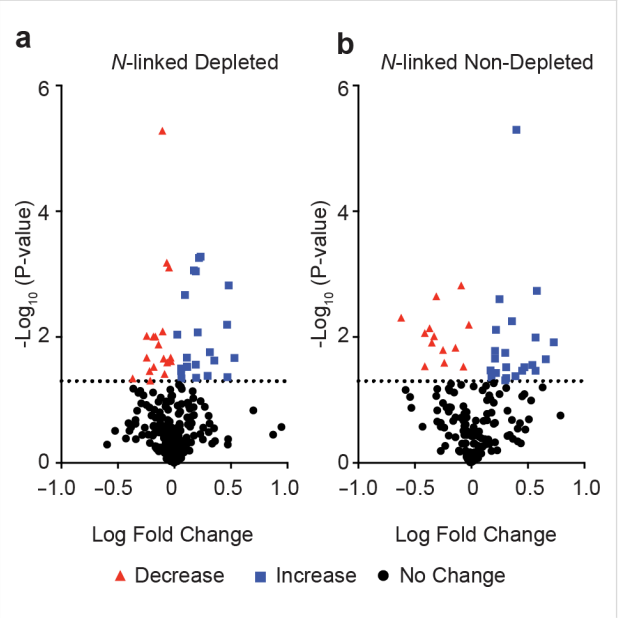


Figure 2: *N*-Glycoproteome profiling with GlypNirO. Volcano plots of site-specific *N*-glycoform relative abundance in HCC patients versus healthy controls in (a) depleted, and (b) non-depleted plasma.

(Figure 3c), and decreased abundance of trisialylated *N*-glycans at alpha-2-HS-glycoprotein N176 (Figure 3d). Together, this suggests an overall decrease in sialylation of *N*-glycans across the plasma glycoproteome in HCC.

O-Glycoproteome analysis

The plasma *O*-glycoproteome is perhaps somewhat neglected [25], despite the importance of *O*-glycosylation to diverse aspects of fundamental biology, health, and disease. We therefore investigated all *O*-glycosylation sites that we were able to identify and measure with GlypNirO. Because there are often multiple potential sites of *O*-glycosylation within a tryptic peptide and site-specific assignment is challenging with CID or HCD fragmentation information, we used peptide-centric analysis of the plasma *O*-glycoproteome. Comparing peptide-specific glycoform relative abundance and occupancy, we identified 41 unique *O*-glycopeptides with increased and 27 with decreased abundance in HCC compared with healthy controls in depleted plasma, and 17 increased and 26 decreased in HCC in non-depleted plasma ($P < 0.05$, Figure 4a and 4b). As the dataset we analysed measured enriched glycopeptides, it is likely that unglycosylated peptides forms are underrepresented.

We could identify both changes in peptide-specific *O*-glycan compositions and in *O*-glycan occupancy. HCC patients had increased glycan occupancy and decreased abundance of monosialylated *O*-glycan on fibrinogen alpha chain G_{272}GSTSYGTGSETESPR (Figure 5a). HCC patients showed a relative decrease in disialylated and an increase in monosialylated *O*-glycan abundance on both plasma protease C1 inhibitor V_{45}AATVISK and histidine-rich glycoprotein S_{271}STTKPPFKPHGSR (Figure 5b and 5c). Together, and consistent with our *N*-glycoproteome analyses, this suggests that HCC is associated with an overall decrease in sialylation of *N*- and *O*-glycans across the plasma glycoproteome.

Conclusion

GlypNirO is an automated software pipeline that integrates glyco/peptide identification from Byonic and quantification from Proteome Discoverer, and provides output that is appropriate for both manual inspection and further statistical analyses. We note that all glycopeptide identification and quantification workflows will include false positive and negative results, and users should ensure data is appropriately searched and curated before processing with GlypNirO. Additionally, modern LC–MS/MS glycoproteomics cannot fully structurally characterise glycans and often struggles to confidently assign the precise sites of modification; ambiguities which may confound quantification workflows. Our proof-of-principle analysis of a plasma glycoproteome dataset demonstrated that GlypNirO can be used to detect changes in site-specific glycosylation occupancy and structure of *N*- and *O*-glycosylation in complex glycoproteomes. Specifically, we found that HCC was associated with decreased sialylation of both *N*- and *O*-glycans at specific sites on selected plasma glycoproteins. GlypNirO will be a useful tool for enabling robust high-throughput quantitative glycoproteomics.

Experimental

Byonic and Proteome Discoverer analysis

We identified glycopeptides and peptides using Byonic (Protein Metrics, v. 3.8.13) searching all DDA files ($n = 12$) downloaded from a previously published dataset [20] obtained from the ProteomeXchange Consortium via the MassIVE repository (PXD003369, MSV000079426). Two searches were conducted on each file, one *N*-linked and one *O*-linked. A human protein database was used (UniProt UP000005640, downloaded April 20, 2018 with 20,303 reviewed proteins) [26]. Cleavage specificity was set as C-terminal to Arg/Lys with a maximum of one missed cleavage. The precursor mass tolerance was 10 ppm and fragment ion mass tolerances for CID and HCD were 0.5 Da and 20 ppm, respectively. Carbamidomethylation of cysteines was set as a fixed modification, and dynamic modifications included deamidation of asparagine, monooxidised methionine, and the formation of pyroglutamate at *N*-terminal glutamic acid and glutamine. All variable modifications were set as “Common 1” allowing each modification to be present once on a peptide. For *N*-linked searches (N-X-S/T) a database of 164 *N*-glycans was used (Supporting Information File 1, Table S41) and for the

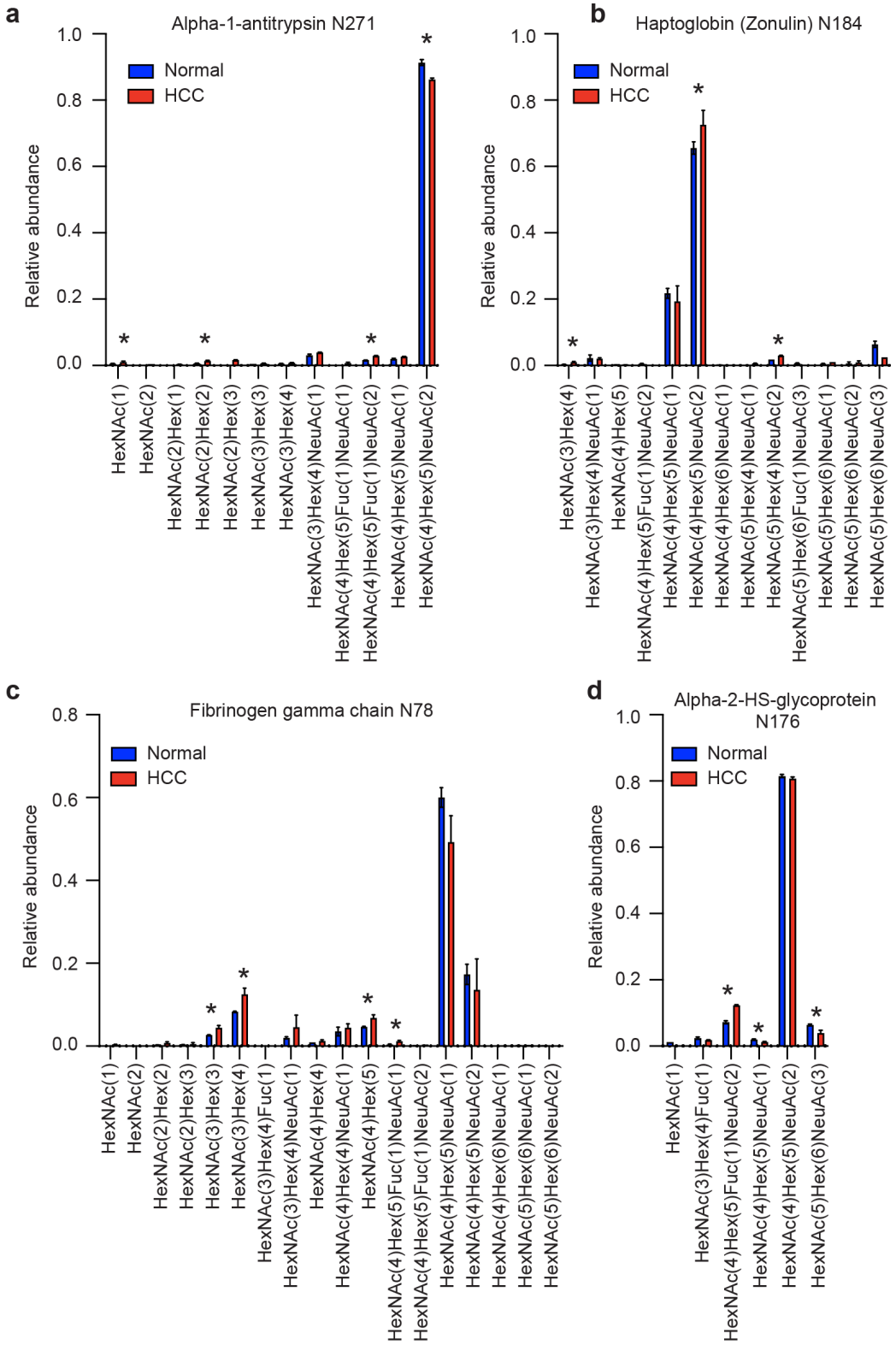


Figure 3: Site-specific *N*-glycopeptide profiling with GlypNirO. Site-specific relative glycoform abundance in HCC patients and healthy controls at (a) alpha-1-antitrypsin N271, (b) haptoglobin N184, (c) fibrinogen gamma chain N78, and (d) alpha-2-HS-glycoprotein N176. $N = 3$; values show mean; error bars show standard error of the mean; *, $P < 0.05$.

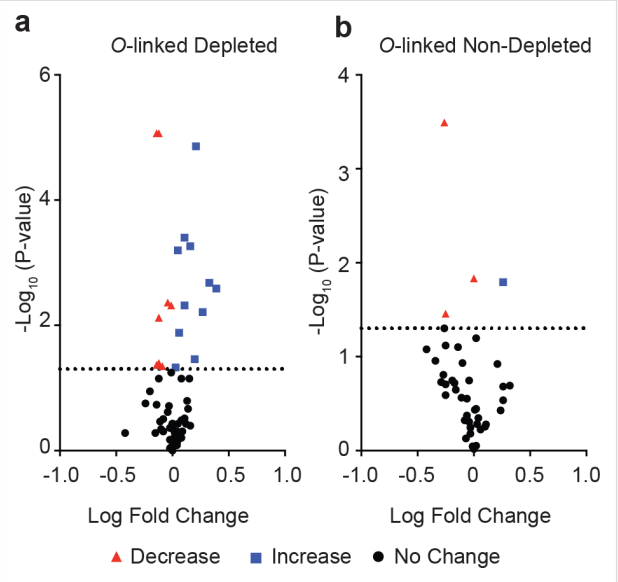


Figure 4: *O*-Glycoproteome profiling with GlypNirO. Volcano plots of site-specific *O*-glycoform relative abundance in HCC patients versus healthy controls in (a) depleted, and (b) non-depleted plasma.

O-linked searches (at any S/T) a database of 49 *O*-glycans (Supporting Information File 1, Table S42) was used. All glycan modifications were set as "Rare 1" allowing each modification to be present once on a peptide. A maximum of two common modifications and one rare modification were allowed per peptide. A protein false discovery rate cut-off of 1% was applied along with the peptide automatic score cutoff [27]. Precursor peak areas were calculated using the Precursor Ions Area Detector node in Proteome Discoverer (v. 2.0.0.802 Thermo Fisher Scientific). Text output files from Proteome Discoverer and Byonic were then used in GlypNirO (https://github.com/bschulzlab/glypniro and Supporting Information File 3).

Output combination and preprocessing

GlypNirO was built and used in Python 3.8.3 with backward compatibility tested up to Python 3.6. Each Byonic output file was first iteratively prepared for linking with AUC information from the Proteome Discoverer output. Using a regular expression pattern provided by UniProtKB, the UniProtKB accession ID of each protein from the *Protein Name* column of the Byonic output was parsed and saved into a new temporary *master id* column. If a UniProtKB accession ID could not be matched, the entire protein name was saved into the *master id* column. *Reverse* (decoy) sequences and *Common contaminant proteins* were filtered and removed from the dataset.

To combine data from different isoforms of the same protein, the Byonic output was grouped by accession ID in the *master id* column. From the *Scan number* column, the numeric scan number associated with a PSM was extracted into a temporary *Scan number* column. Area Under the Curve (AUC) information from the *First Scan* column from the Proteome Discoverer output text file was assigned to Byonic data at each corresponding scan number, in the *Area* column. Entries with no AUC value and those not meeting a user-defined Byonic score cutoff (200 here) were removed from the data set.

Using the *Glycans NHFAGNa* and *Modification Type(s)* column, the script obtained the monosaccharide composition of the attached glycan. In the standard Byonic output, only the Δ mass of the modification is directly indicated on the modified peptide sequence, with no direct indication of the identity of the corresponding modification. The script therefore calculated the theoretical mass of the glycan from the *Glycans NHFAGNa* column, and matched this to the corresponding amino acid in the peptide. This allowed the unambiguous assignment of each site of glycosylation from the Byonic output. Options were provided to either include Byonic assignments of site-specificity, or not, in calculation for the final output.

Unique PSM selection and glycoform AUC calculation

The compiled dataset as a whole was sorted based on two levels in descending order, first by *Area* and then by *Score*. Two options were available for glyco/peptide grouping: site-specific analysis, or peptide-specific analysis. For site-specific analysis, the site-specificity of glycosylation assigned by Byonic was trusted, and all peptide variants that contained that site were included in calculations of its occupancy and glycoform distribution. PSMs with identical unmodified peptide sequence, glycan monosaccharide composition, calculated *m/z*, and site of glycosylation were grouped. For each group, the PSM precursor *m/z* value with the highest associated *Area* was selected as the unique PSM. The *Area* associated with each unique PSM was used for the calculation of the total AUC of each glycoform at each identified glycosylation site.

For peptide-specific analysis, the precise site of glycosylation within a peptide as assigned by Byonic was ignored, and each proteolytically unique peptide form was treated separately. PSMs with identical unmodified peptide sequence, glycan monosaccharide composition, and calculated *m/z* were grouped. As with site-specific analysis, for each group, the PSM with the highest *Area* was selected as its unique PSM. The *Area* of each unique PSM was used for the calculation of the total AUC of each glycoform for each unique proteolytic peptide.

Proportional data analysis and final output

In order to allow comparisons of site-specific glycoform abundance and occupancy between different samples, the proportion of each glycoform was calculated with and without inclusion of

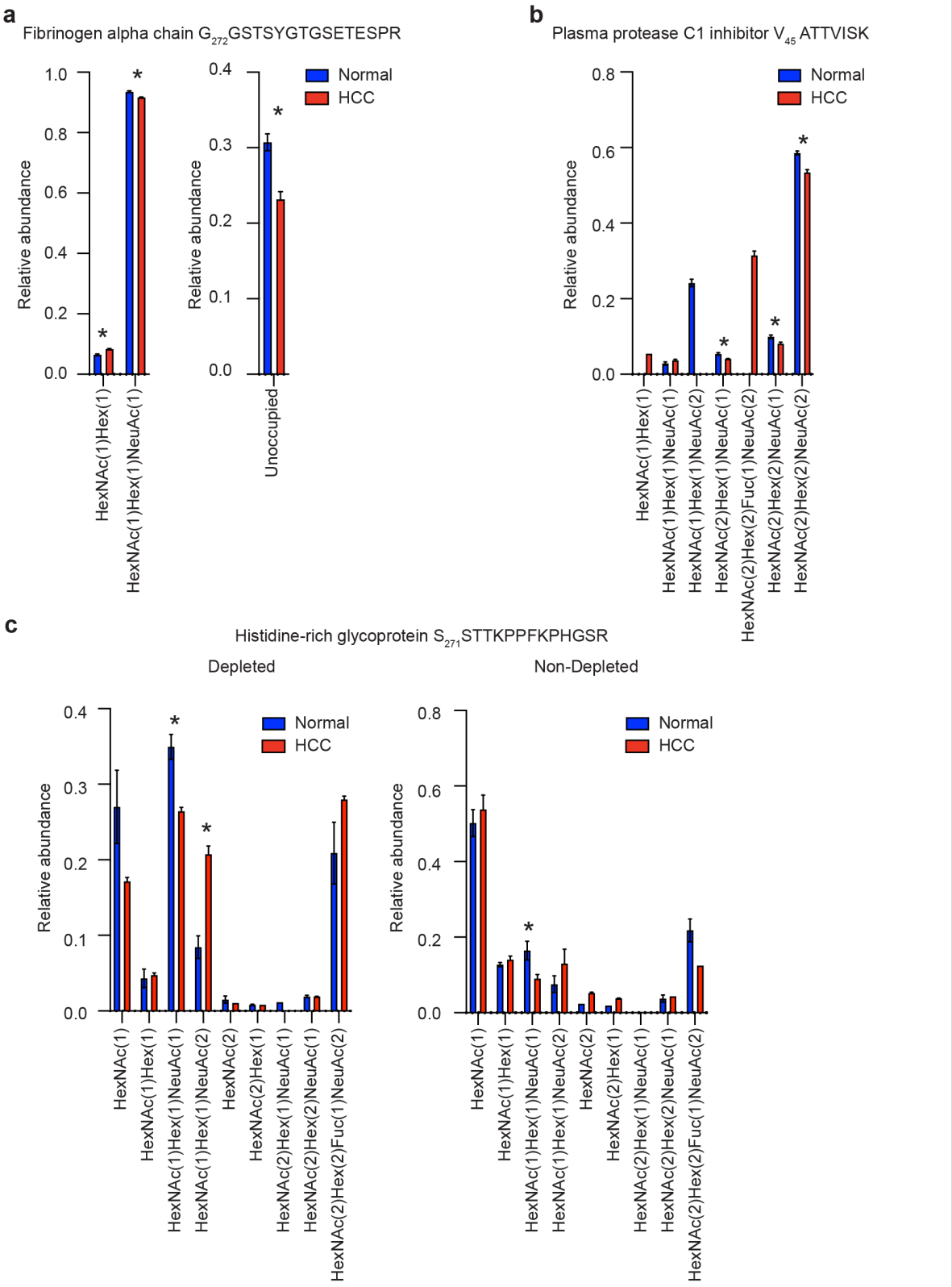


Figure 5: Peptide-specific *O*-glycosylation profiling with GlypNirO. Peptide-specific relative glycoform abundance in HCC patients and healthy controls on (a) fibrinogen alpha chain G_{272}GSTSYGTGSETESPR, (b) plasma protease C1 inhibitor V_{45}AATVISK, and (c) histidine-rich glycoprotein S_{271}STTKPPFKPHGSR. N = 3; values show mean; error bars show standard error of the mean; *, $P < 0.05$.

unglycosylated peptides. For calculation of proportion, glycosylation status was assumed to not quantitatively affect detection. These results were concatenated into the final output file, where columns are the different samples and rows are the different peptide and glycoforms that have been analysed. The protein name of each glycosylated protein detected in the analysis was also included, parsed from the online UniProtKB database using an inhouse Python library.

Statistical analyses

Significant differences in glycoform abundances between healthy and diseased samples were evaluated using an unpaired two-tailed t-test without corrections for multiple comparisons. Missing values were not imputed. Spectra were manually validated for glycoforms of interest.

Supporting Information

Supporting Information File 1

Supplementary Tables S1–S42.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-180-S1.zip]

Supporting Information File 2

GlypNirO workflow overview.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-180-S2.pdf]

Supporting Information File 3

Glypniro-master; automated script for processing and combining Byonic and PD standard output.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-180-S3.zip]

Funding

This work was funded by an Australian Research Council Discovery Project DP160102766 to BLS, an Australian Research Council Industrial Transformation Training Centre IC160100027 to BLS, and a National Health and Medical Research Council Ideas Grant APP1186699 to BLS and CLP.

ORCID® iDs

Toan K. Phung - https://orcid.org/0000-0002-2964-6070
Cassandra L. Pegg - https://orcid.org/0000-0002-6080-7047
Benjamin L. Schulz - https://orcid.org/0000-0002-4823-7758

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.1101/2020.06.15.153528

References

1. Aebi, M. *Biochim. Biophys. Acta, Mol. Cell Res.* **2013,** *1833,* 2430–2437. doi:10.1016/j.bbamcr.2013.04.001
2. Joshi, H. J.; Narimatsu, Y.; Schjoldager, K. T.; Tytgat, H. L. P.; Aebi, M.; Clausen, H.; Halim, A. *Cell* **2018,** *172,* 632–632.e2. doi:10.1016/j.cell.2018.01.016
3. Shental-Bechor, D.; Levy, Y. *Curr. Opin. Struct. Biol.* **2009,** *19,* 524–533. doi:10.1016/j.sbi.2009.07.002
4. Riley, N. M.; Hebert, A. S.; Westphall, M. S.; Coon, J. J. *Nat. Commun.* **2019,** *10,* 1311. doi:10.1038/s41467-019-09222-w
5. Zacchi, L. F.; Schulz, B. L. *Glycoconjugate J.* **2016,** *33,* 359–376. doi:10.1007/s10719-015-9641-3
6. Bertozzi, C. R.; Rabuka, D. Structural basis of glycan diversity. In *Essentials of Glycobiology;* Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E., Eds.; Cold Spring Harbor Laboratory Press: New York, NY, USA, 2009.
7. Cobb, B. A. *Glycobiology* **2020,** *30,* 202–213. doi:10.1093/glycob/cwz065
8. Rabinovich, G. A.; van Kooyk, Y.; Cobb, B. A. *Ann. N. Y. Acad. Sci.* **2012,** *1253,* 1–15. doi:10.1111/j.1749-6632.2012.06492.x
9. Wuhrer, M.; Catalina, M. I.; Deelder, A. M.; Hokke, C. H. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2007,** *849,* 115–128. doi:10.1016/j.jchromb.2006.09.041
10. Delafield, D. G.; Li, L. *Mol. Cell. Proteomics* **2020,** mcp.R120.002095. doi:10.1074/mcp.r120.002095
11. Thaysen-Andersen, M.; Packer, N. H.; Schulz, B. L. *Mol. Cell. Proteomics* **2016,** *15,* 1773–1790. doi:10.1074/mcp.o115.057638
12. Hu, H.; Khatri, K.; Klein, J.; Leymarie, N.; Zaia, J. *Glycoconjugate J.* **2016,** *33,* 285–296. doi:10.1007/s10719-015-9633-3
13. Cao, W.; Liu, M.; Kong, S.; Wu, M.; Zhang, Y.; Yang, P. *Mol. Cell. Proteomics* **2020,** mcp.R120.002090. doi:10.1074/mcp.r120.002090
14. Chen, Z.; Huang, J.; Li, L. *TrAC, Trends Anal. Chem.* **2019,** *118,* 880–892. doi:10.1016/j.trac.2018.10.009
15. Abrahams, J. L.; Taherzadeh, G.; Jarvas, G.; Guttman, A.; Zhou, Y.; Campbell, M. P. *Curr. Opin. Struct. Biol.* **2020,** *62,* 56–69. doi:10.1016/j.sbi.2019.11.009
16. Hu, H.; Khatri, K.; Zaia, J. *Mass Spectrom. Rev.* **2017,** *36,* 475–498. doi:10.1002/mas.21487
17. Jansen, B. C.; Falck, D.; de Haan, N.; Hipgrave Ederveen, A. L.; Razdorov, G.; Lauc, G.; Wuhrer, M. *J. Proteome Res.* **2016,** *15,* 2198–2210. doi:10.1021/acs.jproteome.6b00171
18. Jansen, B. C.; Reiding, K. R.; Bondt, A.; Hipgrave Ederveen, A. L.; Palmblad, M.; Falck, D.; Wuhrer, M. *J. Proteome Res.* **2015,** *14,* 5088–5098. doi:10.1021/acs.jproteome.5b00658
19. Deshpande, N.; Jensen, P. H.; Packer, N. H.; Kolarich, D. *J. Proteome Res.* **2010,** *9,* 1063–1075. doi:10.1021/pr900956x
20. Park, G. W.; Kim, J. Y.; Hwang, H.; Lee, J. Y.; Ahn, Y. H.; Lee, H. K.; Ji, E. S.; Kim, K. H.; Jeong, H. K.; Yun, K. N.; Kim, Y.-S.; Ko, J.-H.; An, H. J.; Kim, J. H.; Paik, Y.-K.; Yoo, J. S. *Sci. Rep.* **2016,** *6,* 21175. doi:10.1038/srep21175
21. Choo, M. S.; Wan, C.; Rudd, P. M.; Nguyen-Khuong, T. *Anal. Chem. (Washington, DC, U. S.)* **2019,** *91,* 7236–7244. doi:10.1021/acs.analchem.9b00594
22. Mayampurath, A.; Yu, C.-Y.; Song, E.; Balan, J.; Mechref, Y.; Tang, H. *Anal. Chem. (Washington, DC, U. S.)* **2014,** *86,* 453–463. doi:10.1021/ac402338u

23. Klein, J.; Carvalho, L.; Zaia, J. *Bioinformatics* **2018,** *34,* 3511–3518. doi:10.1093/bioinformatics/bty397
24. McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference (SciPy 2010),* Austin, Texas, USA, June 28–July 3, 2010; 2010; pp 56–61. doi:10.25080/majora-92bf1922-00a
25. Darula, Z.; Medzihradszky, K. F. *Mol. Cell. Proteomics* **2018,** *17,* 2–17. doi:10.1074/mcp.mr117.000126
26. The UniProt Consortium. *Nucleic Acids Res.* **2019,** *47,* D506–D515. doi:10.1093/nar/gky1049
27. Bern, M. W.; Kil, Y. J. *J. Proteome Res.* **2011,** *10,* 5296–5301. doi:10.1021/pr200780j

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Tools for generating and analyzing glycan microarray data

Akul Y. Mehta*, Jamie Heimburg-Molinaro and Richard D. Cummings*

Review

Open Access

Address:
Department of Surgery, Beth Israel Deaconess Medical Center, National Center for Functional Glycomics, Harvard Medical School, Boston, MA, 02215, USA

Email:
Akul Y. Mehta* - aymehta@bidmc.harvard.edu;
Richard D. Cummings* - rcummin1@bidmc.harvard.edu

* Corresponding author

Keywords:
data analysis; glycan binding; glycan microarray; glycomics; informatics

Beilstein J. Org. Chem. **2020,** *16,* 2260–2271.
https://doi.org/10.3762/bjoc.16.187

Received: 10 April 2020
Accepted: 19 August 2020
Published: 10 September 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: K. F. Aoki-Kinoshita

Abstract

Glycans are one of the major biological polymers found in the mammalian body. They play a vital role in a number of physiologic and pathologic conditions. Glycan microarrays allow a plethora of information to be obtained on protein–glycan binding interactions. In this review, we describe the intricacies of the generation of glycan microarray data and the experimental methods for studying binding. We highlight the importance of this knowledge before moving on to the data analysis. We then highlight a number of tools for the analysis of glycan microarray data such as data repositories, data visualization and manual analysis tools, automated analysis tools and structural informatics tools.

Introduction

Glycans represent a major type of biomolecule in all living things, along with DNA, RNA, lipids and proteins [1]. In mammals, glycans commonly occur as post-translational modifications of proteins (glycoproteins), but they are also linked to lipids (glycolipids) and occur as free molecules. Such glycomolecules have vital roles in a wide range of physiological functions and also participate in many pathologic conditions [2]. Some classic examples of important glycans include the blood group antigens (A, B, O), which are glycan structures found on blood cells and tissues that play a critical role in determining transfusion compatibility during blood and organ donation [3], sialyl-Lewis^A antigen, known more commonly as CA19-9, which is a known tumor marker for pancreatic cancer, and could possibly promote cancer [4], and the *O*-glycan of PSGL-1 which is recognized by P- and L-selectin, which is critical for leukocyte recruitment [5,6]. Other roles of glycans (including glycosaminoglycans/proteoglycans) and glycan binding proteins (GBPs) (including lectins and antibodies) in biological systems have been discovered with respect to cancer, infectious diseases, and genetic disorders [7-15].

As a technology to study glycan recognition by GBPs, glycan microarrays offer an invaluable tool, and permit examination of all types of lectins, along with antibodies. Glycans are recognized by many pathogens, including viruses and bacteria, and glycan microarrays are commonly now used to explore

pathogen recognition of glycans [16-21]. Conversely, glycans from pathogens are also recognized by proteins in the human body and even produce an immune response [13,22,23]. An overview of a typical glycan microarray experiment is provided in Figure 1. The protocol involves the chemical covalent conjugation or noncovalent attachment of glycans (usually 20 up to

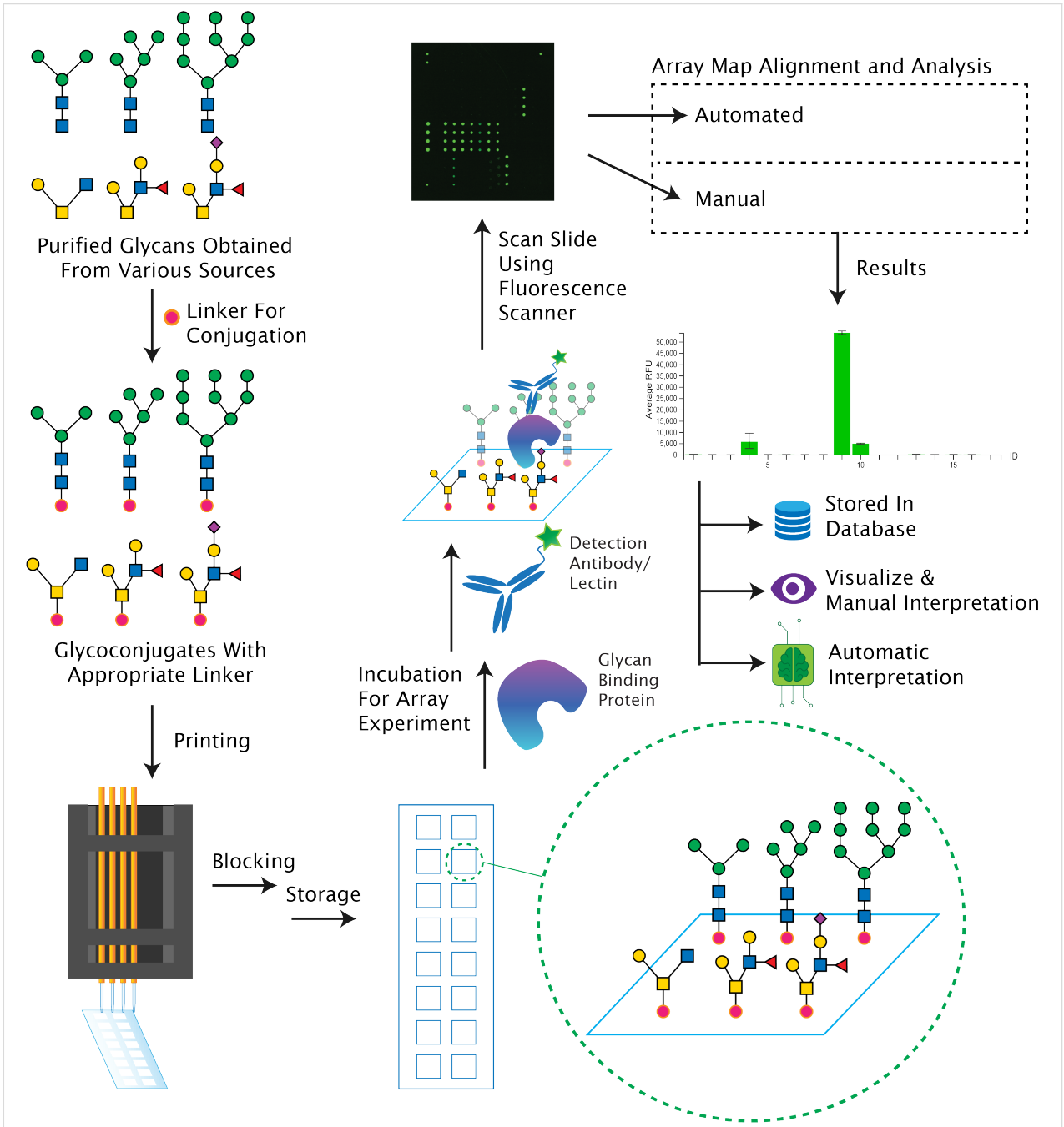


Figure 1: Overview of a typical glycan microarray workflow, beginning with the obtention of glycans to analysis of binding data. Briefly, glycans are chemically or enzymatically synthesized, or isolated and purified from either source materials, and then conjugated with a linker which is appropriate for the printing surface. The glycoconjugates are then printed upon appropriately functionalized slides, followed by blocking; the printed slides are stored under ideal conditions prior to experiments. Many arrays can be printed on a single slide, termed sub-arrays. The slides can then be used in a glycan microarray experiment where they are incubated with a glycan binding protein (GBP), such as lectin, antibody, or serum, virus, etc., followed by addition of a detection reagent, if the primary analyte was not fluorescently labeled, for example a fluorescent secondary antibody or streptavidin. After washing the slide to remove unbound material, the bound material is then identified and measured by scanning using a fluorescence microarray scanner. The image produced can then be analyzed using automated or manual methods to generate the array results. These results can in turn be stored in a database, or interpreted either manually or by automatic algorithms. The glycan structures in the figure were produced using GlycoGlyph [29].

≈700 glycans) in multiple replicates and often at varying concentrations to a slide surface which is appropriately functionalized [24,25]. Such a slide can then be used to probe GBPs or pathogens using an ELISA-like sandwich assay at microscale. This enables a high-throughput screening of glycan-mediated interactions. In this review we describe how glycan microarrays are generated, how a typical glycan microarray experiment is carried out, the type of data generated, as well as the informatic tools either currently available or being developed, for the important but complex step of analyzing glycan microarray data. While parts of this review are specific for sequence defined glycan microarrays, which are the major type of glycan microarrays, there are other sophisticated approaches such as shotgun arrays and beam search array technologies for glycans from natural sources [26-28].

Review

Preparation of glycan microarrays

Decisions and steps before preparation

The selection of glycan microarray surface and linker is a reciprocal process, involving the preparation of the glycans in the context of appropriate surface to which the glycans are desired to be attached. Several types of functionalized slide surfaces are available such as NHS, epoxy, nitrocellulose and PVDF (Table 1); each utilizes a different mechanism of binding the ligands to the surface. Choosing an appropriate surface often depends on the type of glycoconjugates to be printed, as well as the GBP and detection wavelengths used. For example, a nitrocellulose surface has an intrinsic high background when scanned at 488 nm wavelength, thus making the surface incompatible with detection reagents which rely on this wavelength. While it is possible to use nitrocellulose slides at 488 nm wavelength with lower detector sensitivity (PMT setting) and lower scan power (laser power), it might be more advisable to check other specialized surface types (e.g., nitrocellulose PATH® slides) which have lower background signals at this wavelength. A decision chart is provided in Figure 2 which can help to decide which surface would be best for a variety of situations.

A more detailed discussion of recent glycan linkers and surfaces was recently reported by Gao et al. [30] and McQuillan et al. [31]. Once the surface of choice is selected, the glycans need to be conjugated to an appropriate aglycone to form a glycoconjugate which can be used to link the glycan to the surface of the slide. If the glycans are already conjugated to an aglycone (for example, a glycolipid obtained directly by extraction from a natural origin), an appropriate surface needs to be selected which is compatible with the glycoconjugate.

Printing methods

Once the decision about the surface to which they are to be attached is made, the glycans are dissolved in appropriate printing buffers depending upon the surface chemistry involved. A microarray printer, either contact-type or non-contact-type (e.g., piezoelectric dispensing), is used to dispense small drops of 50–100 μm in diameter each onto the slide surface in a rapid manner. Usually each sample is printed as ≥4 spots/array and a single slide can have 1 or many arrays (also called sub-arrays). The concentration of the glycan in the printing buffer depends upon the efficiency of the mechanism of linking the glycan to the surface. The spots are separated from each other by a given

Table 1: Summary of slide surfaces commercially available for microarrays.

Surface type	Corresponding glycoconjugate properties required	Example commercial sources
covalent conjugation		
N-hydroxysuccinimide (NHS)	amino functional group (primary amine preferred)	Nexterion® Slide H (Schott)
epoxy	amino functional group (primary amine preferred)	Nexterion® Slide E (Schott)
aldehyde	primary amino functional group	Nexterion® Slide AL (Schott)
noncovalent adsorption		
streptavidin	biotin functionalization	SuperStreptavidin (Arrayit®)
nitrocellulose – porous type	hydrophobic, i.e., protein or lipid conjugated	ONCYTE® SuperNOVA (Grace Bio-labs), UniSart® 3D Nitro slides (Sartorius)
nitrocellulose – non-porous type	hydrophobic, i.e., protein or lipid conjugated	PATH® (Grace Bio-labs)
PVDF	protein conjugate	SuperPVDF (Arrayit®)

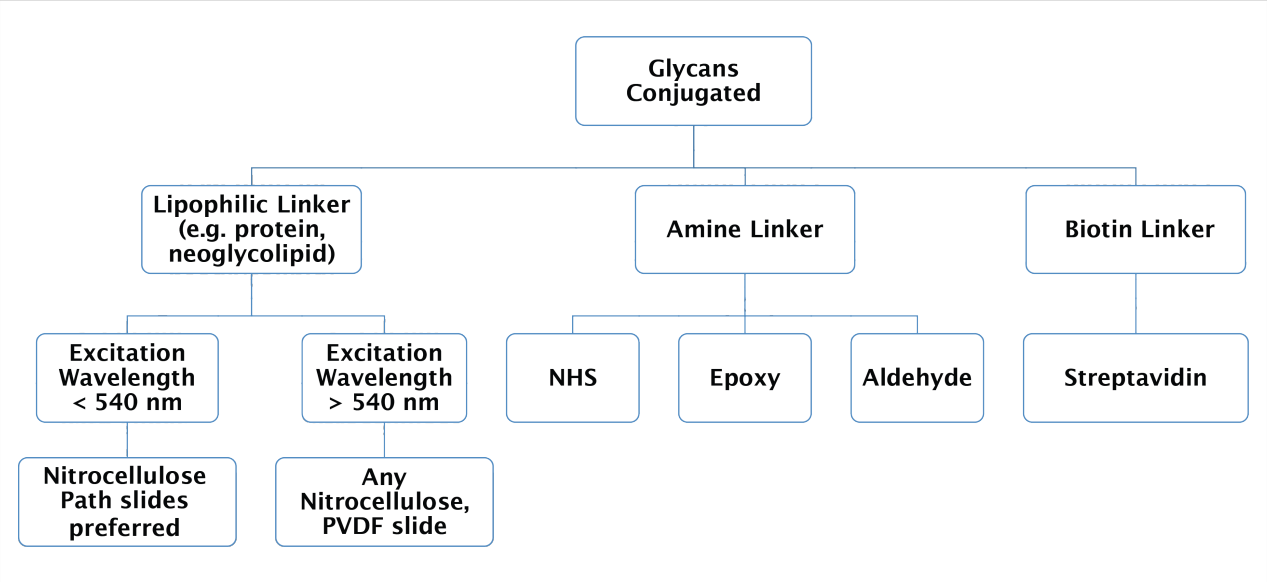


Figure 2: A decision tree to determine which type of surface to use depending on glycoconjugate and linker type, and scanning requirements. These linker types and surfaces are the most commonly used and there are more specialized surfaces and linker chemistries that are commercially available.

space called the "spot pitch" (usually 2–4 times the diameter of the spots) to form distinct spots for analysis and to avoid merging of spots. Depending on the printer throughput, the number of slides/arrays to be printed and number of glycans to be printed, this process can take several hours or an entire day, and requires monitoring for accurate printing. Between the printing of each probe (glycan), there is usually a wash step which is performed to prevent any carryover for the next ligand, and this would need to be determined empirically depending on the glycoconjugates used. After the entire print run, the slides are left for incubation either in a humidified chamber at room temperature or in a cold room for several hours (or overnight), depending upon the linking mechanism. Following this incubation time, the rest of the reactive slide surface is blocked using an appropriate blocking solution and the slides are dried for storage.

Slide storage and handling

Slides should be stored stably under appropriate conditions, depending upon the slide surface type and the linking stability, for many months. Glycan microarray slides are typically stored under vacuum sealed conditions at a cold temperature (−20 to 4 °C). When the slides are to be used, it is advised to let the slides come to room temperature without external warming in a vacuum desiccator prior to use.

Glycan microarray experiments

If the slide is composed of several sub-arrays, the multi-well chamber method is used. If the slide is composed of 1 large array, the coverslip method can be used.

Multi-well chamber method

Several multi-well chambers of different array layouts are available: 8 × 3 (24-well), 8 × 2 (16-well), 8 × 1 (8-well), or 4 × 1 (4-well) chambers. The choice of the multi-well chamber to be used depends on the print layout of the arrays, which in turn would depend upon the number of glycans to be printed. The larger the number of glycans printed per array, the larger the printed area is, and hence lower number of arrays per slide. The chambers are usually made of plastic with a silicone rubber gasket that fits on top of the slide (for example, ProPlate type of multi-well chambers sold by Grace Biolabs). Such chambers allow the complete separation of a single slide into multiple wells each containing a separate sub-array that can be incubated with a sample. This enables testing of multiple samples/experiments on a single slide simultaneously with minimal sample volumes. Although usually inert to biological samples, a precautionary test should be performed to ensure that the GBP sample is compatible with the gasket/chamber materials. The chambers can be covered with parafilm or plastic plate covers to further isolate each well and prevent evaporation during assay incubation.

Coverslip method

When using an entire slide, the use of a multi-well chamber is impractical as it would require a large amount of GBP sample to fill the chamber. As a result an alternative coverslip method is used. In this method, a small volume of sample is placed on slide (≈70 µL), and a coverslip is placed on top to spread the sample evenly across the array surface of the slide and helps to prevent evaporation during the assay.

Detailed protocols for either method of microarray experiment and that cover different types of samples and detection methods are available on the National Center for Functional Glycomics website (https://ncfg.hms.harvard.edu/protocols), for the multi-well chamber method (e.g., NCFG slides) and for the coverslip method (e.g., CFG slides).

Data acquisition

Once the slides are dried post sample incubation, the slides are scanned using a microarray scanner (for example, Genepix 4400A). Microarray scanners are fluorescence scanners which utilize laser technology, such that the excitation wavelength is generated by specific lasers. Commonly used laser wavelengths are 488 nm, 532 nm, 594 nm and 635 nm, which match with usual fluorophore labels on detecting reagents (e.g., labeled GBP, antibody or streptavidin). The emission wavelength of the fluorophore determines the filter used by the scanner before the intensities are measured by a photomultiplier tube (PMT) or by CCD camera. Currently, CCD camera systems are less sensitive as compared to PMT type detectors and therefore PMT systems are preferred for more accurate measurements. In addition, newer LED-based excitation systems are being developed, but are still not as sensitive and therefore laser scanners are still used. Microarray scanners scan at pixel resolutions ranging from 2.5–100 μm/pixel. This means that each pixel obtained in the final image corresponds to 2.5–100 μm on the slide depending on the resolution selected during the scan. The lower the pixel resolution value, the higher the resolution of the final image and the more data points are obtained for each spot on the slide. Thus, high-resolution images (2.5–10 μm/pixel) yield adequate data points for glycan microarray spots to provide lower standard deviations between replicate spots. The fluorescence intensity of the spots can be fine-tuned by controlling the laser power (also called LP) and the photomultiplier tube gain (also called PMT Gain). The image produced is saved as a TIFF image, usually with headers which describe the scanner settings used to acquire the image, and the intensity at each pixel is saved as the relative fluorescence units (RFU) for those scan settings.

Spot alignment and data processing

Once the image is acquired, the image is aligned to an array map (for GenePix scanners this is called the GenePix Array List file or .gal file), which indicates the coordinates of the various spots by (row, column) numbers correlated to the material which was printed at those positions. The alignments are usually done by hand with assistance from the scanner software, which usually offers partial alignment algorithms based on background intensities. The spot diameters are also adjusted as some glycans just form smaller spots in comparison to others. Spots where information is not reliable due to extraneous factors such as poor printing due to a flaw in the surface, spot overlap/fusion with adjacent spot or presence of dust particles are flagged with a “bad” (or some numeric value) flag, so as to be disregarded in the data processing. The software then provides a results file as an output (for GenePix scanners this is the GenePix Results file or .gpr file), which contains the information of the alignment file along with spot intensity information (such as mean and median fluorescence intensity for the spot) along with information about the local background around the spot (such as mean and median background intensities). This results file is processed using Microsoft Excel in a variety of ways which usually involve consolidating data from multiple replicates of spots to provide the average background subtracted intensities in RFUs for each individual compound printed on the surface, i.e., the average of (mean spot intensity – mean background intensity) for all the replicate spots of the material for a particular excitation wavelength. This is provided along with standard deviation (SD) and coefficient of variation (SD ÷ mean) as a percentage (%CV). The results are usually presented as bar graphs with the mean or average fluorescence intensity (RFUs) on the *y*-axis with the glycan/print material id number on the *x*-axis, while the error bars represent the SD or standard error of the mean (SEM).

Glycan microarray reporting guidelines

In order to report glycan microarray experiment and results one should follow the MIRAGE (Minimum Information Required for A Glycomics Experiment) guidelines for microarray data [32]. These guidelines cover some of the important aspects mentioned above in order for anyone to be able to reproduce the glycan microarray experiment. The current version published in June 2016 can be found at: https://www.beilstein-institut.de/download/1458/mirage_glycan_array_guidelines_version_1.0__22_june_2016.pdf.

Tools for glycan microarray data

A) Repositories of glycan microarray data

1. CFG Database:

Status: Available.

Address: http://www.functionalglycomics.org/glycomics/publicdata/primaryscreen.jsp.

Description: The Consortium for Functional Glycomics (CFG) database of glycan array data (Figure 3A) is one of the largest archival resources for glycan microarray data. The CFG was founded in 2001 and was funded for about 10 years, during which time the main website and database were created [33]. The microarray work was then taken over by the National Center for Functional Glycomics (NCFG) which was established in 2013.

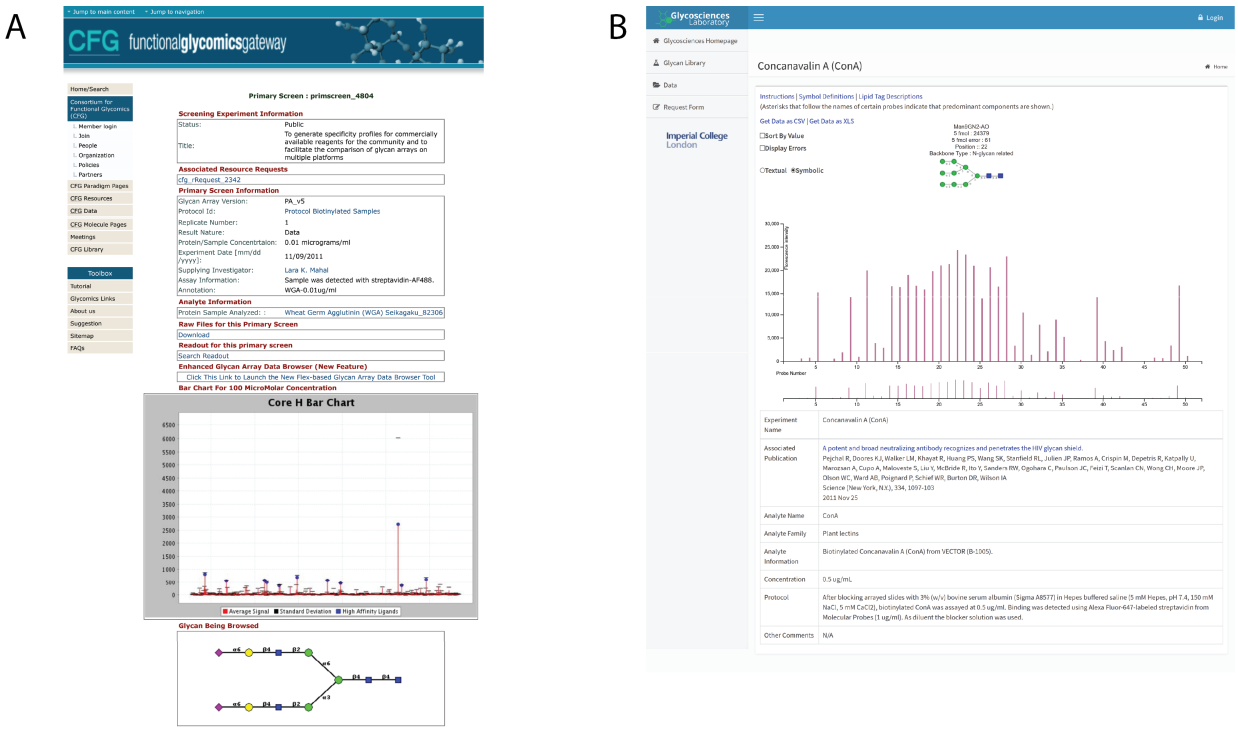


Figure 3: Screenshot of microarray databases: (A) Screenshot of an example of CFG glycan array data; (B) screenshot of an example of Imperial College microarray data online portal.

The CFG database has over 3000 experimental results files with 2 major types of arrays, the mammalian glycan printed array (noted as "Mammalian Printed Array" on the website and not to be confused with the "mammalian plate array" also present on the same page, which is not a microarray platform) and the pathogen array [34-36]. There are multiple versions of the mammalian glycan printed array (i.e., from v1 to v5.2) where each version differs by the addition or removal of some glycans from the list. The mammalian glycan arrays contain between 200–611 different glycan compounds printed per microarray. The pathogen array consists of glycans/polysaccharide isolated from bacterial sources and currently has over 300 compounds printed, with an earlier version containing fewer compounds. The database has samples consisting of animal GBPs (e.g., C-type lectins, siglecs, galectins), plant lectins, antibodies, serum samples, pathogens and microbial proteins, cells, and organisms. The website contains data available as downloadable .xlsx format along with metadata associated with the experiment, including sample and assay information.

Newer data has been challenging to add to the database, due to lack of funding support, and the use of outdated technologies which make it difficult to upgrade the current CFG database. As a result, the eventual aim is to move to a centralized microarray repository which is utilizing more modern web technologies such as that in development by GlyGen (see below) so that new data can be made easily accessible to the public.

2. Imperial College microarray data online portal:
Status: Available.
Address: https://glycosciences.med.ic.ac.uk/data.html.
Description: Imperial College Microarray Data Online Portal of glycan microarray data (Figure 3B) consists of ≈160 experiments (from ≈36 publications) on a variety of microarray platforms composed of different glycans. The database includes data on antibodies, animal and plant lectins, viruses and virus-like particles, virus proteins, microbial proteins. The website is designed using newer protocols in comparison to the CFG, however, the data was classically stored in an MS Office-based platform [37] and is now being upgraded to the newer CarbArrayART database (see below).

3. CarbArrayART:
Status: Development (available for testing upon request).
Description: CarbArrayART is a software in development for the storage, processing and presentation of microarray data [38,39]. It is based on the GRITS Toolbox (classically used for

mass spectrometry data) [40]. Features of CarbArrayART include storage of glycan array data from different array formats. This includes the results along with any array-specific metadata such as experiment protocol, array geometry etc. CarbArrayART will offer presentation of data with filtering and sorting functions, and generation of reports. More recently, the flexibility of the system has been improved by introducing several new input functions for sample information, experiment information and array geometries with multiple glyco-probe layouts.

4. GlyGen microarray repository:
Status: Development.
Description: GlyGen (https://www.glygen.org) is a growing resource for the inclusion of data from multiple sources for glycoinformatics [41]. A component of the project involves the creation of a glycan microarray data repository, whereby anyone can go to a website and deposit and view glycan microarray data, along with the metadata associated with the microarray experiment. Such a centralized database would greatly help the glycoscience community and help develop newer software for glycan microarray data analysis.

B) Data visualization/manual data analysis

1. GLAD:
Status: Available.
Address: https://glycotoolkit.com/GLAD/.
Description: Traditionally, glycan array data was shared only as excel files which are non-interactive and often troublesome to visualize glycan structure alongside data visualizations such as bar charts. Yet manual data analysis is still widely used to deduce most information based on glycan microarray data. The GLycan Array Dashboard (GLAD) is a tool to visualize, analyze, compare and mine glycan array data. The tool allows users to visualize glycan array data alongside the structures using bar charts, heatmaps, calendar heatmaps, force directed graphs and correlation plots. In addition, the tool also couples some data mining features such as the ability to filter glycans by name, fragments, IDs, cutoff threshold or by rank. It also has features to sort data in particular order, normalize data and discard data points which are not present between datasets so as to get a more uniform view. All charts produced by GLAD are interactive to show the glycan structure provided the glycans are labeled using the CFG linear nomenclature system. This makes it particularly useful for manual data analysis. The plots and structures produced can directly be saved as SVG vector graphic files which can be used by most illustration software to create publication quality images [16,30,31] (Figure 4). GLAD also allows users to save and reload sessions using JSON formatted text file, which makes it easy to share the data as a GLAD session file.

C) Automated analysis

Automating analysis of glycan microarray data can be challenging due to the intricacies involved at multiple levels. Unlike DNA and proteins, glycans are neither linear nor template driven, making their structures computationally taxing. In addition, the complexity is compounded by the sample and experimental meta data associated with the glycan microarray experiment. The following software offer a means to address the need for automating analysis, yet it must be kept in mind that most of these software do not take into account the afore-mentioned meta data. Hence, while these software might be useful in simple use cases, even today, experts in the field prefer to use a manual methods of analysis which may be supported by some of these automated tools.

1. GlycoPattern:
Status: Currently Unavailable – Transitioning to new host.
Address: N/A.
Description: GlycoPattern is a web based resource to support the analysis of glycan array data [42]. Under the hood, it utilizes the GlycanMotifMiner (GLYMMR) (https://github.com/sagravat/glymmr) algorithm [43] to perform frequent subtree mining in order to identify binding motifs. It was one of the first automated software to help mine glycan array data. GlycoPattern and GLYMMR were designed to work with CFG microarray data and hence their applicability to other datasets remains questionable. While the software was freely available earlier, at the time of this writing, the software was unavailable due to lack of funding and costs of maintaining it on university servers [44]. However, steps are being taken to try to bring the program back as an important resource to the glycoscience community.

2. MotifFinder:
Status: Available.
Address: https://haablab.vai.org/tools/.
Description: MotifFinder is a graphical user interface (GUI) driven software which is able to mine glycan array data using predefined motif lists. It is a semi-automatic software in which the motifs need to be defined using MotifSpeak language which is an extension of the CFG linear nomenclature. The software is freely available for noncommercial uses, and needs to be installed in order to run. It has extensive documentation in the manual on how to perform different analysis either manually, or automated. The results output a motif table with a list of motifs, the number of glycans on the array which consist of that motif and the P-value. It also comes with other data visualizations such as box-plots, motif intensity maps, motif family membership map, list of motif glycan examples, all concentration plots, and a model structure. The software was designed to be useful for lectin and enzyme analysis. It has been used to

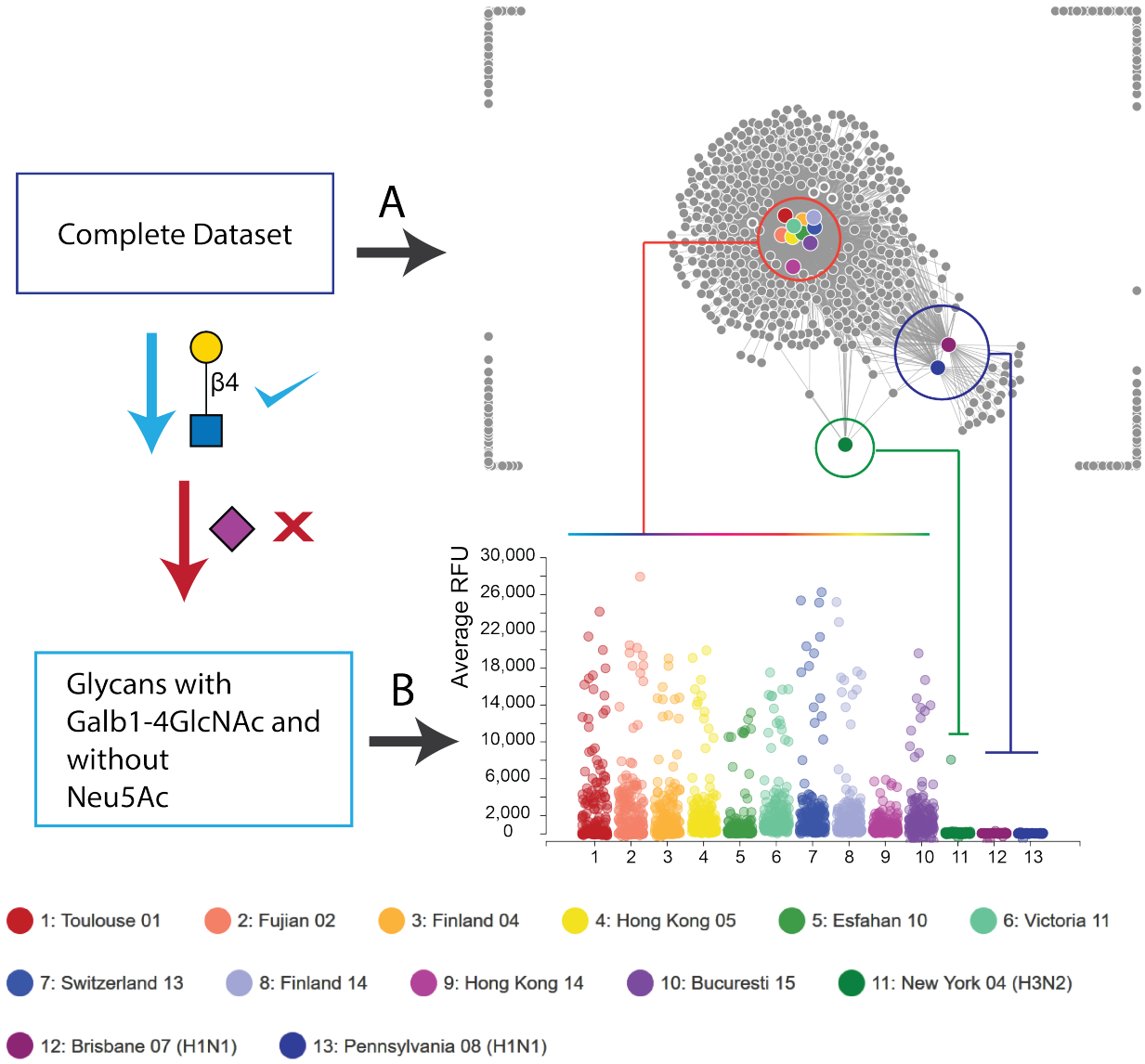


Figure 4: A demonstration of glycan array data visualization with GLAD. The dataset used is from Byrd-Leotis et al. [16], and is provided as a GLAD session file in the manuscript. The dataset contains data on the CFG microarray for various drift H3N2 strains of influenza (#1–10) in comparison to the distinct New York H3N2 strain (#11) and two H1N1 strains (#12 and #13). Details regarding the strains and the experiment can be found in the original paper. The legend for the colors is provided at the bottom of the figure. (A) The complete dataset can be visualized using the force graph where each strain of virus is represented by the colored circle, while each glycan is represented by a grey circle. The links between the virus to the glycan nodes are determined by the intensity of binding observed on the microarray, i.e., the higher the intensity the shorter is the link length. A short simulation is run to optimize the distances between the nodes so as to accommodate this link length and as a result of this nodes which bind similar glycans cluster together in comparison to those which bind others. Thus, the drift strains (#1–10) cluster together in the red cluster, in comparison to the New York H3N2 strain in green, all of which are separated from the purple cluster of the H1N1 strains. (B) To visualize specific binding to lactosamine (Galb1-4GalNAc) in the absence of sialic acid, the dataset is filtered for lactosamine containing glycan data and data for glycans with Neu5Ac are eliminated from the pool. A Bubble Plot of this filtered data shows clearly that the drift strains (#1–10) bind better to the lactosamine containing glycans without Neu5Ac as compared to the New York H3N2 strain and the two H1N1 strains. This analysis clearly shows the difference of the drift H3N2 strains which have acquired ability to bind lactosamine containing glycans.

discover fine specificities of lectins (AAL, SNA) and glycosidase enzymes (α1-2-fucosidase and an α2-3,6,8-neuraminidase) [45].

3. SignalFinder-Microarray:
Status: Available.
Address: https://haablab.vai.org/tools/.

Description: SignalFinder-Microarray is an image analysis tool which allows automation to begin one step before. The software is free to use for noncommercial uses. Using this tool a user can input simply the image obtained from the scanner (.tif file) and the array map file (.gal file) and it will automatically identify and align the spots. To do this, the software uses a segment and fit thresholding algorithm, which is also useful for immunofluorescence images [46,47]. The user then has the ability to override or flag any spots to be ignored in the analysis. The software then processes the image to yield the final results either as an output which can be used with MotifFinder (see above) or as traditional average data (i.e., with mean and CV).

4. MCAW-DB:
Status: Available.
Address: https://mcawdb.glycoinfo.org.
Description: MCAW-DB offers a ready-made analysis of over 1000 glycan array datasets from the CFG database (up to v5.1) via a web interface [48]. In this tool, rather than using predefined motifs, it utilizes Multiple Carbohydrate Alignment with Weights (MCAW) algorithm to align glycan structures as sequences based on their monosaccharide and linkages, and assigns each node weights depending on their binding to the ligand [49]. MCAW-DB offers a unique perspective to glycan array binding results and even takes into account gaps in structures. The tool has parameters (such as weighting) which may need to be optimized to work with other datasets, but the defaults work well with certain sample data such as lectins [48].

5. CCARL:
Status: Available.
Address: https://github.com/andrewguy/CCARL.
Description: CCARL is a very new method of identifying motifs from glycan microarray experiments [50]. Previous subtree mining approaches would not account for terminal motifs. CCARL customizes the frequent subtree mining approach by extending the glycan notation to include terminal node information by including additional nodes in the graph representation to indicate the absence or presence of linkage at particular backbone carbon positions. This enables identification of terminal residues (i.e. those with all backbone carbon positions without linkages except one). In addition, it uses a new algorithm termed minimum-redundance, maximum-relevance (mRMR) to perform the subtree mining, yielding more fine-tuned results. The authors have shown the utility of CCARL on lectin data extensively. CCARL, however, does not currently have a web or GUI interface making it only possible to use it programmatically. Like other automated methods, however, the authors accept that the parameters of this method would need to be fine-tuned depending on the dataset.

D) Structural information tools

1. GlyMDB:
Status: Available.
Address: http://www.glycanstructure.org/glymdb/.
Description: GlyMDB is a web-based database which links glycan microarray binding data from the CFG database to protein structures (PDB) [51]. A user can select a dataset from the CFG dataset available and set thresholds for binding versus nonbinding. The application can then show you motifs which make a significant binding contribution on the microarray. In addition it allows you to quickly search for PDB files with sequence identity matching to the protein sample put on the microarray along with glycan ligand length parameters. GlyMDB then retrieves the protein crystal structures for those PDB ids with protein matching the sample and glycans in the PDB structure. It allows you to view the structure in the browser to see how the glycan binds to the protein. GlyMDB thus provides a unique one-stop solution to cross referencing glycan microarray data alongside protein structure.

2. Gly-Spec (Grafting):
Status: Available.
Address: http://glycam.org/djdev/grafting/.
Description: Gly-Spec (Grafting) uses structural data to predict glycan microarray binding [52]. In this software a user uploads a glycan binding protein complexed to a carbohydrate fragment in PDB format. This need not be a co-crystal structure, and can be a modeled structure as well. The application then finds glycans that contain this fragment which are present on the CFG microarray data and predicts if the protein will be able to bind them. The current limitation is that it has data only from the CFG database. Thus, building glycan array databases which are easily accessible to multiple tools can further improve the development of tools like this and help grow a field of predictive glycobiology.

Conclusion

Glycan microarray technologies provide a wealth of information for functional glycomics, and in an efficient and decipherable manner. Once the microarrays are fabricated, the experiments can be performed within a few hours and the data analysis can be done in a day. Results from glycan microarray experiments provide needed information to develop new hypotheses about glycan recognition and function. In this article, we highlight some of the nuances of how fabrication of the glycan microarrays is done, along with tools currently available or in development to help with analysis and comparison of glycan microarray data. We identify a variety of glycan microarray repositories where interested readers can find microarray data to build new software. We also highlight manual and automated data analysis tools. One must always be aware that the intrinsic

complexity of the multistep process of glycan microarray experiments means that none of the tools currently available are foolproof and each approach and technology has strengths and weaknesses. Often, automated tools miss out on important patterns of binding which might be readily apparent to a trained individual using a manual mining approach. It is also possible that the results from automated software could be confounding rather than illuminating, if parameters of the experiment or even array fabrication (such as surface chemistry, etc.) are changed. In fact, the various chemical linkers through which glycans are attached to the array surface may dramatically change the way the glycan is presented on the microarray surface [53], thus, potentially indirectly affecting binding results. Hence, we advise experimentalists to carefully consider all of the metadata associated with each glycan array experiment. Since none of the current automation tools are flawless, the need for new tools for the analysis and reporting of glycan microarray data is ever-present.

Acknowledgements

The authors would like to thank Zachary Klamer, Johnathan Hall and Brian Haab (MotifFinder, SignalFinder) along with Lachlan Coff and Andrew Guy (CCARL), for help with their software.

Funding

This work was supported by National Institutes of Health Grants P41GM103694, U01GM125267, and R24GM137763 to R.D.C.

ORCID® iDs

Akul Y. Mehta - https://orcid.org/0000-0002-3450-3461
Jamie Heimburg-Molinaro - https://orcid.org/0000-0003-4987-3016
Richard D. Cummings - https://orcid.org/0000-0002-8918-5034

References

1. Stern, R.; Jedrzejas, M. J. *Chem. Rev.* **2008,** *108,* 5061–5085. doi:10.1021/cr078240l
2. Colley, K. J.; Varki, A.; Kinoshita, T. Cellular Organization of Glycosylation. In *Essentials of Glycobiology;* Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, U.S.A., 2015; pp 41–49.
3. Milland, J.; Sandrin, M. S. *Tissue Antigens* **2006,** *68,* 459–466. doi:10.1111/j.1399-0039.2006.00721.x
4. Engle, D. D.; Tiriac, H.; Rivera, K. D.; Pommier, A.; Whalen, S.; Oni, T. E.; Alagesan, B.; Lee, E. J.; Yao, M. A.; Lucito, M. S.; Spielman, B.; Da Silva, B.; Schoepfer, C.; Wright, K.; Creighton, B.; Afinowicz, L.; Yu, K. H.; Grützmann, R.; Aust, D.; Gimotty, P. A.; Pollard, K. S.; Hruban, R. H.; Goggins, M. G.; Pilarsky, C.; Park, Y.; Pappin, D. J.; Hollingsworth, M. A.; Tuveson, D. A. *Science* **2019,** *364,* 1156–1162. doi:10.1126/science.aaw3145
5. Leppänen, A.; White, S. P.; Helin, J.; McEver, R. P.; Cummings, R. D. *J. Biol. Chem.* **2000,** *275,* 39569–39578. doi:10.1074/jbc.m005005200
6. Hicks, A. E. R.; Leppänen, A.; Cummings, R. D.; McEver, R. P.; Hellewell, P. G.; Norman, K. E. *FASEB J.* **2002,** *16,* 1461–1462. doi:10.1096/fj.02-0075fje
7. Ju, T.; Cummings, R. D. *Nature* **2005,** *437,* 1252. doi:10.1038/4371252a
8. Holst, S.; Wuhrer, M.; Rombouts, Y. *Adv. Cancer Res.* **2015,** *126,* 203–256. doi:10.1016/bs.acr.2014.11.004
9. Yin, B. W. T.; Lloyd, K. O. *J. Biol. Chem.* **2001,** *276,* 27371–27375. doi:10.1074/jbc.m103554200
10. Magnani, J. L.; Steplewski, Z.; Koprowski, H.; Ginsburg, V. *Cancer Res.* **1983,** *43,* 5489–5492.
11. Pinho, S. S.; Reis, C. A. *Nat. Rev. Cancer* **2015,** *15,* 540–555. doi:10.1038/nrc3982
12. Munkley, J. *Oncol. Lett.* **2019,** *17,* 2569–2575. doi:10.3892/ol.2019.9885
13. Stowell, S. R.; Arthur, C. M.; McBride, R.; Berger, O.; Razi, N.; Heimburg-Molinaro, J.; Rodrigues, L. C.; Gourdine, J.-P.; Noll, A. J.; von Gunten, S.; Smith, D. F.; Knirel, Y. A.; Paulson, J. C.; Cummings, R. D. *Nat. Chem. Biol.* **2014,** *10,* 470–476. doi:10.1038/nchembio.1525
14. Stowell, S. R.; Arthur, C. M.; Dias-Baruffi, M.; Rodrigues, L. C.; Gourdine, J.-P.; Heimburg-Molinaro, J.; Ju, T.; Molinaro, R. J.; Rivera-Marrero, C.; Xia, B.; Smith, D. F.; Cummings, R. D. *Nat. Med.* **2010,** *16,* 295–301. doi:10.1038/nm.2103
15. Comstock, L. E.; Kasper, D. L. *Cell* **2006,** *126,* 847–850. doi:10.1016/j.cell.2006.08.021
16. Byrd-Leotis, L.; Gao, C.; Jia, N.; Mehta, A. Y.; Trost, J.; Cummings, S. F.; Heimburg-Molinaro, J.; Cummings, R. D.; Steinhauer, D. A. *J. Virol.* **2019,** *93,* e01178-19. doi:10.1128/jvi.01178-19
17. Bourdoulous, S.; Lemichez, E. *Nat. Microbiol.* **2018,** *3,* 124–126. doi:10.1038/s41564-018-0107-9
18. Poole, J.; Day, C. J.; von Itzstein, M.; Paton, J. C.; Jennings, M. P. *Nat. Rev. Microbiol.* **2018,** *16,* 440–452. doi:10.1038/s41579-018-0007-2
19. Yu, Y.; Lasanajak, Y.; Song, X.; Hu, L.; Ramani, S.; Mickum, M. L.; Ashline, D. J.; Prasad, B. V. V.; Estes, M. K.; Reinhold, V. N.; Cummings, R. D.; Smith, D. F. *Mol. Cell. Proteomics* **2014,** *13,* 2944–2960. doi:10.1074/mcp.m114.039875
20. Kubota, M.; Matsuoka, R.; Suzuki, T.; Yonekura, K.; Yanagi, Y.; Hashiguchi, T. *J. Virol.* **2019,** *93,* e00344-19. doi:10.1128/jvi.00344-19
21. Lonardi, E.; Moonens, K.; Buts, L.; de Boer, A. R.; Olsson, J. D. M.; Weiss, M. S.; Fabre, E.; Guérardel, Y.; Deelder, A. M.; Oscarson, S.; Wuhrer, M.; Bouckaert, J. *Biology (Basel, Switz.)* **2013,** *2,* 894–917. doi:10.3390/biology2030894
22. Mickum, M. L.; Prasanphanich, N. S.; Song, X.; Dorabawila, N.; Mandalasi, M.; Lasanajak, Y.; Luyai, A.; Secor, W. E.; Wilkins, P. P.; Van Die, I.; Smith, D. F.; Nyame, A. K.; Cummings, R. D.; Rivera-Marrero, C. A. *Infect. Immun.* **2016,** *84,* 1371–1386. doi:10.1128/iai.01349-15
23. Geissner, A.; Reinhardt, A.; Rademacher, C.; Johannssen, T.; Monteiro, J.; Lepenies, B.; Thépaut, M.; Fieschi, F.; Mrázková, J.; Wimmerova, M.; Schuhmacher, F.; Götze, S.; Grünstein, D.; Guo, X.; Hahm, H. S.; Kandasamy, J.; Leonori, D.; Martin, C. E.; Parameswarappa, S. G.; Pasari, S.; Schlegel, M. K.; Tanaka, H.; Xiao, G.; Yang, Y.; Pereira, C. L.; Anish, C.; Seeberger, P. H. *Proc. Natl. Acad. Sci. U. S. A.* **2019,** *116,* 1958–1967. doi:10.1073/pnas.1800853116

24. Song, X.; Heimburg-Molinaro, J.; Smith, D. F.; Cummings, R. D. *Glycoconjugate J.* **2015,** *32,* 465–473. doi:10.1007/s10719-015-9584-8
25. Song, X.; Heimburg-Molinaro, J.; Cummings, R. D.; Smith, D. F. *Curr. Opin. Chem. Biol.* **2014,** *18,* 70–77. doi:10.1016/j.cbpa.2014.01.001
26. Smith, D. F.; Cummings, R. D.; Song, X. *Biochem. Soc. Trans.* **2019,** *47,* 1–11. doi:10.1042/bst20170487
27. Song, X.; Lasanajak, Y.; Xia, B.; Heimburg-Molinaro, J.; Rhea, J. M.; Ju, H.; Zhao, C.; Molinaro, R. J.; Cummings, R. D.; Smith, D. F. *Nat. Methods* **2011,** *8,* 85–90. doi:10.1038/nmeth.1540
28. Li, Z.; Gao, C.; Zhang, Y.; Palma, A. S.; Childs, R. A.; Silva, L. M.; Liu, Y.; Jiang, X.; Liu, Y.; Chai, W.; Feizi, T. *Mol. Cell. Proteomics* **2018,** *17,* 121–133. doi:10.1074/mcp.ra117.000285
29. Mehta, A. Y.; Cummings, R. D. *Bioinformatics* **2020,** *36,* 3613–3614. doi:10.1093/bioinformatics/btaa190
30. Gao, C.; Wei, M.; McKitrick, T. R.; McQuillan, A. M.; Heimburg-Molinaro, J.; Cummings, R. D. *Front. Chem. (Lausanne, Switz.)* **2019,** *7,* 833. doi:10.3389/fchem.2019.00833
31. McQuillan, A. M.; Byrd-Leotis, L.; Heimburg-Molinaro, J.; Cummings, R. D. *Front. Mol. Biosci.* **2019,** *6,* 88. doi:10.3389/fmolb.2019.00088
32. Liu, Y.; McBride, R.; Stoll, M.; Palma, A. S.; Silva, L.; Agravat, S.; Aoki-Kinoshita, K. F.; Campbell, M. P.; Costello, C. E.; Dell, A.; Haslam, S. M.; Karlsson, N. G.; Khoo, K.-H.; Kolarich, D.; Novotny, M. V.; Packer, N. H.; Ranzinger, R.; Rapp, E.; Rudd, P. M.; Struwe, W. B.; Tiemeyer, M.; Wells, L.; York, W. S.; Zaia, J.; Kettner, C.; Paulson, J. C.; Feizi, T.; Smith, D. F. *Glycobiology* **2017,** *27,* 280–284. doi:10.1093/glycob/cww118
33. Raman, R.; Venkataraman, M.; Ramakrishnan, S.; Lang, W.; Raguram, S.; Sasisekharan, R. *Glycobiology* **2006,** *16,* 82R–90R. doi:10.1093/glycob/cwj080
34. Blixt, O.; Head, S.; Mondala, T.; Scanlan, C.; Huflejt, M. E.; Alvarez, R.; Bryan, M. C.; Fazio, F.; Calarese, D.; Stevens, J.; Razi, N.; Stevens, D. J.; Skehel, J. J.; van Die, I.; Burton, D. R.; Wilson, I. A.; Cummings, R.; Bovin, N.; Wong, C.-H.; Paulson, J. C. *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101,* 17033–17038. doi:10.1073/pnas.0407902101
35. Blixt, O. *Nat. Prec.* **2007**. doi:10.1038/npre.2007.12.1
36. Blixt, O.; Hoffmann, J.; Svenson, S.; Norberg, T. *Glycoconjugate J.* **2008,** *25,* 27–36. doi:10.1007/s10719-007-9045-0
37. Stoll, M.; Feizi, T. Software Tools for Storing, Processing and Displaying Carbohydrate Microarray Data. In *Proceedings of the Beilstein Symposium on Glyco-Bioinformatics - Bits 'n' Bytes of Sugars,* Hicks, M. G.; Kettner, C.; Seeberger, P. H., Eds.; Beilstein Institut: Frankfurt, Germany, 2009; pp 123–140.
38. Akune, Y.; Arpinar, S.; Silva, L. M.; Stoll, M.; Palma, A. S.; Liu, Y.; Ranzinger, R.; Feizi, T. *CarbArrayART: Carbohydrate Array Analysis and Reporting Tool New software for glycan array for data processing, storage and presentation;* Oxford University Press Inc.: Oxford, U.K., 2018; pp 1034–1035.
39. Akune, Y.; Arpinar, S.; Silva, L. M.; Stoll, M.; Palma, A. S.; Liu, Y.; Ranzinger, R.; Feizi, T. *An Update to a Software Tool for Glycan Array Data: CarbArrayART. Ready for Beta Testing*. Time-proof perspectives on Glycoscience – Beilstein Glyco-Bioinformatics Symposium 2019, Limburg, Germany, June 25–27, 2019; Beilstein Institut: Frankfurt, Germany, 2019; pp 40–41.
40. Weatherly, D. B.; Arpinar, F. S.; Porterfield, M.; Tiemeyer, M.; York, W. S.; Ranzinger, R. *Glycobiology* **2019,** *29,* 452–460. doi:10.1093/glycob/cwz023
41. York, W. S.; Mazumder, R.; Ranzinger, R.; Edwards, N.; Kahsay, R.; Aoki-Kinoshita, K. F.; Campbell, M. P.; Cummings, R. D.; Feizi, T.; Martin, M.; Natale, D. A.; Packer, N. H.; Woods, R. J.; Agarwal, G.; Arpinar, S.; Bhat, S.; Blake, J.; Castro, L. J. G.; Fochtman, B.; Gildersleeve, J.; Goldman, R.; Holmes, X.; Jain, V.; Kulkarni, S.; Mahadik, R.; Mehta, A.; Mousavi, R.; Nakarakommula, S.; Navelkar, R.; Pattabiraman, N.; Pierce, M. J.; Ross, K.; Vasudev, P.; Vora, J.; Williamson, T.; Zhang, W. *Glycobiology* **2020,** *30,* 72–73. doi:10.1093/glycob/cwz080
42. Agravat, S. B.; Saltz, J. H.; Cummings, R. D.; Smith, D. F. *Bioinformatics* **2014,** *30,* 3417–3418. doi:10.1093/bioinformatics/btu559
43. Cholleti, S. R.; Agravat, S.; Morris, T.; Saltz, J. H.; Song, X.; Cummings, R. D.; Smith, D. F. *OMICS* **2012,** *16,* 497–512. doi:10.1089/omi.2012.0013
44. Emory Comprehensive Glycomics Core - Our Core Facility Resources Page. https://web.archive.org/web/20200227160435/https://www.cores.emory.edu/ecgc/resources/ (accessed Feb 27, 2020).
45. Klamer, Z.; Staal, B.; Prudden, A. R.; Liu, L.; Smith, D. F.; Boons, G.-J.; Haab, B. *Anal. Chem. (Washington, DC, U. S.)* **2017,** *89,* 12342–12350. doi:10.1021/acs.analchem.7b04293
46. Ensink, E.; Sinha, J.; Sinha, A.; Tang, H.; Calderone, H. M.; Hostetter, G.; Winter, J.; Cherba, D.; Brand, R. E.; Allen, P. J.; Sempere, L. F.; Haab, B. B. *Anal. Chem. (Washington, DC, U. S.)* **2015,** *87,* 9715–9721. doi:10.1021/acs.analchem.5b03159
47. Barnett, D.; Hall, J.; Haab, B. *Am. J. Pathol.* **2019,** *189,* 1402–1412. doi:10.1016/j.ajpath.2019.03.011
48. Hosoda, M.; Takahashi, Y.; Shiota, M.; Shinmachi, D.; Inomoto, R.; Higashimoto, S.; Aoki-Kinoshita, K. F. *Carbohydr. Res.* **2018,** *464,* 44–56. doi:10.1016/j.carres.2018.05.003
49. Hosoda, M.; Akune, Y.; Aoki-Kinoshita, K. F. *Bioinformatics* **2017,** *33,* 1317–1323. doi:10.1093/bioinformatics/btw827
50. Coff, L.; Chan, J.; Ramsland, P. A.; Guy, A. J. *BMC Bioinf.* **2020,** *21,* 42. doi:10.1186/s12859-020-3374-4
51. Cao, Y.; Park, S.-J.; Mehta, A. Y.; Cummings, R. D.; Im, W. *Bioinformatics* **2020,** *36,* 2438–2442. doi:10.1093/bioinformatics/btz934
52. Grant, O. C.; Tessier, M. B.; Meche, L.; Mahal, L. K.; Foley, B. L.; Woods, R. J. *Glycobiology* **2016,** *26,* 772–783. doi:10.1093/glycob/cww020
53. Grant, O. C.; Smith, H. M.; Firsova, D.; Fadda, E.; Woods, R. J. *Glycobiology* **2014,** *24,* 17–25. doi:10.1093/glycob/cwt083

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Computational tools for drawing, building and displaying carbohydrates: a visual guide

Kanhaya Lal[‡1,2], Rafael Bermeo[‡1,2] and Serge Perez[*1]

Review

Open Access

Address:
[1]Univ. Grenoble Alpes, CNRS, CERMAV, 38000 Grenoble, France and [2]Dipartimento di Chimica, Università Degli Studi di Milano, via Golgi 19, I-20133, Italy

Email:
Serge Perez[*] - spsergeperez@gmail.com

* Corresponding author ‡ Equal contributors

Keywords:
bioinformatics; carbohydrate; glycan; glycobiology; nomenclature; oligosaccharide; polysaccharide; representation; structure

Beilstein J. Org. Chem. **2020,** *16,* 2448–2468.
https://doi.org/10.3762/bjoc.16.199

Received: 18 June 2020
Accepted: 17 September 2020
Published: 02 October 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: F. Lisacek

Abstract

Drawing and visualisation of molecular structures are some of the most common tasks carried out in structural glycobiology, typically using various software. In this perspective article, we outline developments in the computational tools for the sketching, visualisation and modelling of glycans. The article also provides details on the standard representation of glycans, and glycoconjugates, which helps the communication of structure details within the scientific community. We highlight the comparative analysis of the available tools which could help researchers to perform various tasks related to structure representation and model building of glycans. These tools can be useful for glycobiologists or any researcher looking for a ready to use, simple program for the sketching or building of glycans.

Introduction

Glycoscience is a rapidly surfacing and evolving scientific discipline. One of its current challenges is to keep up and adapt to the increasing levels of data available in the present scientific environment. Indeed, the rise of accessible experiment data has changed the landscape of how research is performed. The accessibility of this information, coupled with the emergence of new platforms and technologies, has benefitted glycoscience to the point of enabling the detection and high-resolution determination and representation of complex glycans [1]. Increasing numbers of carbohydrate sequences have accumulated throughout extensive work in areas of chemical and biochemical fragmentations followed by analysis using mass spectroscopy, nuclear magnetic resonance, crystallography and computational modelling. There have been some initiatives by independent research groups worldwide, that pushed the development of visual tools to improve some aspects of glycan identification, quantification and visualisation, some of which will be further developed throughout this article.

Biological molecules express their function throughout their three-dimensional structures. For this reason, structural biology places great emphasis on the three-dimensional structure as a central element in the characterisation of biological function. An adequate understanding of biomolecular mechanisms inherently requires our ability to model and visualise them. Visualisation of molecular structures is thus one of the most common tasks performed by structural biologists. As an essential part of the research process, data visualisation allows not only to communicate experimental results but also is a crucial step in the integration of multiple data derived resources, such as thermodynamics and kinetic analysis, glycan arrays, mutagenesis, etc. Data visualisation remains a challenge in glycoscience for both the developers and the end-users even for the simple task of describing molecular structures. Progress in this area allows to translate a static visualisation of single molecules into dynamic views of complex interacting large macromolecular assemblies, which increases our understanding of biological processes.

Representing the structures of carbohydrates has historically been considered to be a complicated task. Starting from the linear form of the Fischer projection, which is certainly not a realistic representation of a carbohydrate structure, there has been a continuous development and evolution of the description of monosaccharides [2]. Glycans are puzzles to many chemists, and biologists as well as bioinformaticians. This complexity occurs at different levels (which makes it incremental). Amongst the most recognisable "sugars", glucose is merely one of 60+ monosaccharides, all of which are, in truth, pairs of mirror-image enantiomers (L and D).

Moreover, monosaccharides occur as two forms: 5-atom ring (furanose) and 6-atom ring (pyranose). With the occurrence of a statistically rarer "open form," we obtain at least 6 "correct" representations of glucose. And yet, monosaccharides are only the chemical units and the individual building blocks of much more complex molecules; the carbohydrates, also referred to as glycans. The glycan family can be grouped in the following categories: (i) oligosaccharides (comprising two to ten monosaccharides linked together either linearly or branched); (ii) polysaccharides (for glycan chains composed of more than ten monosaccharides); (iii) glycoconjugates (where the glycan chains are covalently linked to proteins (glycoproteins), lipids (glycolipids). The complexity of glycans is a consequence of their branched structure and the range of building blocks available. Other levels of complexity include the nature of the glycosidic linkage (anomeric configuration, position and angles), the number of repeating units (polysaccharides) as well as the substitutions of the monosaccharides. Regardless of the different nomenclatures available to describe each monosaccharide, representing and encoding a glycan structure into a file is required for communication among scientists as well as for data processing.

As a consequence, glycobiologists have proposed different graphical representations, with symbols or chemical structures replacing monosaccharides. The description of carbohydrate structures using standard symbolic nomenclature enables easy understanding and communication within the scientific community. Research groups working on carbohydrates have developed schematic depictions with symbols [3] and expansions with greyscale colouring as the so-called Oxford nomenclature (UOXF) [4,5], and even fully coloured schemes later on. Among these, some of the proposed representation forms have been accepted and implemented by several groups and initiatives, namely the Consortium for Functional Glycomics (CFG) [6]. Whereas the initial versions of such representation were limited to mammalian glycans, an extension of the graphical representation of glycans, called SNFG Symbol Nomenclature for Glycans (SNFG) [7,8] resulted from a joint international agreement. The newly proposed nomenclature covers 67 monosaccharides aptly represented in eleven shapes and ten colours. There is the hope that it will cope better with the rapidly growing information on the structure and functions of glycans and polysaccharides from microbes, plants and algae. The rendering of glycan drawing and symbol representations motivated the development of several computer applications using a standardised notation. The earliest glycan editors allowed manual drawing similar to ChemDraw or used input files with glycan sequence KCF (KEGG Chemical Function) [9] in text format for similarity search against other structures deposited in the databases. Later developments supported the construction and representation of glycan structures in symbolic form by computational tools like GlycanBuilder [10]. Since then, several advancements have been made to allow the user to both draw glycans manually or by importing and exporting the structure files in different text formats [11].

Along the same line, the development of various other applications allowed the users to sketch 2D-glycan structures by dragging and dropping monosaccharides to canvas to generate 3D structures for further usages. These depictions comply with protein data bank (PDB) [12] format, or in the form of images [13,14]. Besides, these tools for representing glycans in 2D and 3D shape [15] allowed the integration of glycans into protein structures or complexes. The tools developed in the last few years have automated the sketching of glycans and glycopeptides, allowing rapid display of structures using IUPAC format [16] as input. This article explores and illustrates the concepts of "sketching", "building" and "viewing" glycans (Figure 1). It provides a descriptive analysis of the tools available for such

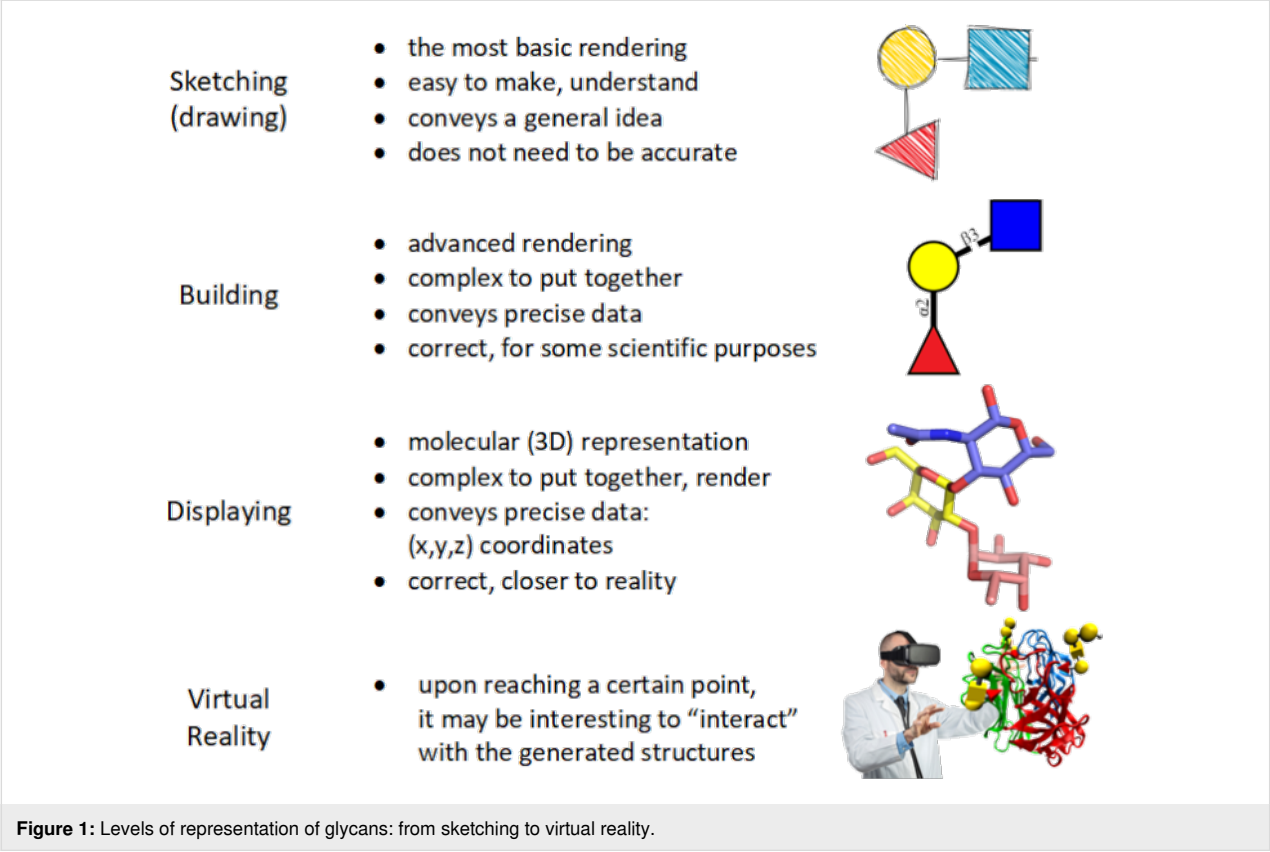


Figure 1: Levels of representation of glycans: from sketching to virtual reality.

activities, which can be useful for researchers looking for a ready-to-use simple program for sketching, building and 3D structure analysis of glycans and glycoconjugates. The scope of this work is relevant to N- and O-linked glycans, glycolipids, proteoglycans and glycosaminoglycans, lipopolysaccharides, plant, algal and bacterial polysaccharides.

Review

Methods

To facilitate glycoscience research, we have identified the tools and databases that are freely available on the internet and are regularly updated and improved [1]. The variety and complexity of glycan structures make their interpretation challenging. Consequently, in the past few years, several sketching, building and visualisation tools have been developed to depict better and understand the complex glycan structures. In this study, the freely available tools have been visited (April 2020) and analysed to highlight their core features but also explore their unique advancements to facilitate glycan research. Each of the computational tools was inspected for general features related to sketching, representing and model building, all of which could be further used as input for translation into other formats, search from glycan databases or complex calculations such as molecular simulations. Several tools feature an interactive interface which allows for manual editing of the structures. Examples of such tools are DrawRINGS [17], KegDraw [18], Glycano (available at http://glycano.cs.uct.ac.za/), GlycoEditor [19], GlycanBuilder [20], etc. These tools (except KegDraw) are provided with the list of CFG symbols to freely build glycan structures using the mouse on the canvas. In addition to manual sketching, some of these tools also can import text formats including IUPAC-condensed, GlycoCT and KEGG Chemical Function (KCF) format to display the glycan structures. Some applications also facilitate glycan search in various databases. Another category of tools included in this study involved glycan viewers which can only depict structures using the IUPAC three letter code or IUPAC-condensed nomenclature as input. These tools convert the input into a 2D image or 3D representation using SNFG symbols or 3D-SNFG illustration. Additionally, 3D representation of structures is provided by tools such as Visual Molecular Dynamics (VMD) [21] , and LiteMol [22], which allow for quick analysis of structural features in 3D space. All the tools mentioned were evaluated against a set of pre-selected criteria relating to ease of use, scientific precision and content, among others.

Table S1 (Supporting Information File 1) schematically summarises how these criteria are fulfilled. The analysis of the tools for input and output formats also provided information about their versatility to convert results into the standard or

desired format. The tools have been attributed to categories such as “sketcher”, “builder” and “viewer”, with eventual overlaps. A brief analysis of each application ordered by category is given in the next section.

Sketching with the free hand

As a preview of the following parts of this study, we performed an initial test of the tools available for the representation of a simple disaccharide: lactose (β-D-Gal*p*-(1→4)-D-Glc*p*).

Figure 2 shows how different web-available platforms rendered it. On the one hand, thanks to the unified nomenclature, there is no ambiguity regarding the nature of the carbohydrate represented. On the other hand, small differences between sketches appear. Such variations will multiply with the increasing complexity of the carbohydrates. It is, therefore, essential to choose which tools to use before starting an hour-long “drawing-spree”. The variations of the colour code used to represent the monosaccharides show striking differences across platforms even though the appropriate colours to be used are strictly defined (https://www.ncbi.nlm.nih.gov/glycans/snfg.html#tab2). The colour discrepancy observed here means that some of the tools do not conform to SNFG standards. For some purposes, this conformity might not be a strict necessity. Another pronounced disparity concerns the representations of the glycosidic linkage. Across sketches the length/width of the linkage varies, which will result in either compact or extended images, to be taken into account when considering the size available for the intended figures. Finally, the sketches provide further information about the linkage type: anomericity and position. These details can be either useful or superfluous depending on what is the intended use for the finished design. The main characteristic of a helpful sketching tool should be its adaptability. By allowing to modify colours, sizes, lengths/widths and turn some features on/off, a “sketcher” would allow maximum flexibility to depict carbohydrates in any desired or necessary form, size, orientation. However, this adaptability should become available without hampering the sketching effort. The perfect sketching tool would, therefore, combine flexibility and high usability.

Building with scientific accuracy

The necessity for precision is what, at some point, turns carbohydrate sketching into building. What defines this turning point (besides a certain level of accuracy) is the intended purpose for the produced figures/images. Scientific communication, comparison between similar yet different structures, or merely showcasing the complexity of carbohydrates: all three cases cannot rely on a sketching tool to convey their message. Consequently, a new set of considerations appears. The requirement for accurate depiction comes from the complexity mentioned

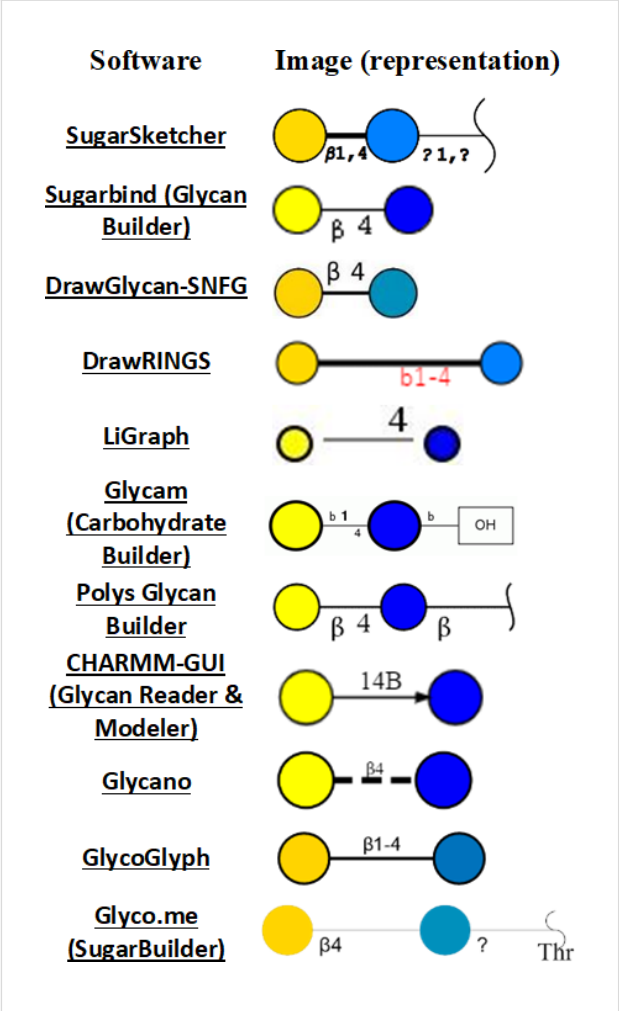


Figure 2: Depiction of lactose by various glycan sketching tools.

above of carbohydrates: anomeric configuration, substitution, glycosidic bond position, and repeating units (as well as tethering to larger macromolecules, and more). For the sake of accuracy, only the right combination of characteristics should be depicted, leaving no ambiguity: every relevant piece of data should be detailed. The glycosidic linkage is a perfect example to illustrate the necessity for accuracy in building, as opposed to sketching. While a simple line is enough to link two monosaccharides, it is necessary to define the linkage as alpha or beta (or unknown) and to state the positions of the glycosyl acceptor and even donor. Cellulose and amylose are two glucose-based polysaccharides that differ only in the nature of their glycosidic bond, and yet they have entirely different shapes and so, biological roles. For the sake of completion, the full description of a monosaccharide should obey the following rules: *<anomeric prefix><prefix for absolute configuration><the monosaccharide code><prefix for ring configuration>[<O-ester and O-ether substitutions and positions>]*. It is thus necessary to include such information when depicting carbohydrates, but

such features are simply absent in most of the existing glycan sketching tools.

Another feature that may become essential when the carbohydrate at hand is a polysaccharide is the possibility of building repeating units. Without this option, it would be simply impossible to build the required depiction. It emerges that an efficient carbohydrate builder must offer a wide array of options to characterise and personalise each monosaccharide. This would, in turn, entail a multitude of buttons, switches, etc.; which would result in a very complex interface. Consequently, unless the interface is rather straightforward and the building dynamic is well-designed, the software would be too difficult to use effectively. The ideal carbohydrate builder pick would also allow liberty for the user in terms of levels of precision since it has to fit every level of complexity above sketching. Lastly, once the building process is complete, a good builder must not only render all the provided data in the form of a precise figure but also allow the transfer of the data to other platforms (for example, by exporting the generated code).

Force fields for carbohydrates, 3D model building and beyond

Carbohydrates present various challenges to the development of force fields [23]. The tertiary structures of monosaccharides usually have a high number of chiral centres which increases the structural diversity and complexity. The structural diversity changes the electrostatic landscape of molecules; thus, it provides challenges in the development of force fields for accurate modelling of such variations in charge distributions. The monosaccharides can further form a large number of oligosaccharides which can enormously increase the conformational space, due to a high number of rotatable bonds. Nonetheless, recent developments in carbohydrate force fields enable to model and reproduce the energies associated with minute geometrical changes. The currently available force fields which are parameterised for carbohydrates are also capable of carrying out simulations of the oligosaccharides containing additional groups like sulfates, phosphates etc. [24] Generally used force fields for the Molecular Mechanics (MD) simulation of carbohydrates are CHARMM [25], GLYCAM [26], and GROMOS [27]. The structural complexity increases the computational cost, which makes simulations of large systems more challenging. Therefore, coarse-grained models [28] for carbohydrates are generally used for molecular modelling of large systems.

In terms of 3D model building, the complex topologies of glycans require dedicated molecular building procedures to convert sequence information into reliable 3D models. These tools generally use 3D molecular templates of monosaccharides to reconstruct a 3D model. Energy minimisation methods can further refine the models. These models are essential for structure-based studies and complex calculations like Molecular Dynamics simulations. Therefore, the accurate model building requires the use of reliable databases to generate atomic coordinates and topology to provide an acceptable model. Some of the computational tools usually contain atom coordinates of generally used monosaccharides (as templates) and also use libraries of bond and angle parameters from various force fields dedicated for carbohydrates. The accurately predicted oligosaccharide conformations are good starting points for further investigations. Of particular interest are the evaluations of the dynamics of glycans and their interactions with proteins which is a most significant concern in glycoscience. The joint need to better perceive and manipulate the three-dimensional objects that make up molecular structures is leading to a rapid appropriation of techniques of Virtual Reality (VR) by the molecular biology community. Generic definitions describe VR as being immersion in an interactive virtual reactive world. The computer-generated graphics provide a realistic rendering of an immersive and dynamic environment that responds to the user's requests. One finds in these definitions the three pillars that define VR: Immersion, Interaction, Information. Although it is difficult to extract a single, simple definition of VR, the main idea is to put the user at the centre of a dynamic and reactive VR environment, artificially created and which will supplant the real world for the time of the experiment.

Input and output for sketching, building and displaying applications

The variety and complexity of carbohydrate structures hamper the use of a unique nomenclature. The choice of notation depends on whether the study is focused on chemistry or has a more biological approach. The IUPAC-IUBMB (International Union for Pure and Applied Chemistry and International Union for Biochemistry and Molecular Biology) terminologies, in their extended and condensed forms [16], govern the naming of the primary structure or sequence.

Further down the line, the complexity of the existing nomenclatures for carbohydrate-containing molecules remains a significant hurdle to their practical use and exchanges within and outside the glycoscience cenacle. The linearisation of the description of the structure is a way to cope with the description of the structural complexity. The proposed formats provide rules to extract the structure of the branches and create a unique sequence for the carbohydrate. The most commonly used formats are IUPAC [16], GlycoCT [29], KCF [9], and WURCS [30].

The sketching of carbohydrates using computational tools generally requires the textual input and output in at least one of

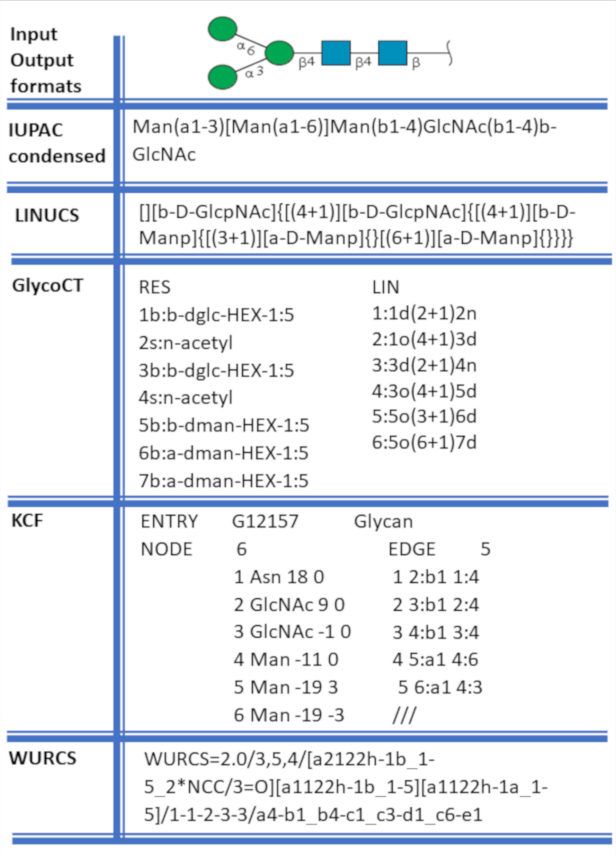


Figure 3: Examples of different glycan structure text formats for the same glycan. Data in these formats are generally used as input/output in glycan drawing and 3D structure building tools.

these formats (Figure 3). An alternate input method involves manual sketching of 2D glycan structures by dragging and dropping monosaccharide symbols on canvas (with or without grids) to connect them further. This method makes the sketching tools more friendly and interactive as it does not require large text code as input. Both input methods are compliant to the Symbol Nomenclature for Glycans (SNFG). Another symbolic representation that could clearly distinguish monosaccharides in monochrome colours is the Oxford notation [5]. In this method, dashed and solid lines represent the alpha and beta glycosidic linkages, respectively. There are few tools which have implemented this method while other tools use text to highlight this information in the structures. In addition to sketching tools, some applications, specific to the field of carbohydrates, provide the possibility to visualise and display 3D structures. These visualisation tools accept strings or files in text formats (GlycoCT, IUPAC-condensed, KCF) to display the structure via a graphical user interface. For instance, the DrawGlycan-SNFG [31] tool uses IUPAC-condensed nomenclature for input string and converts it into a 2D image represented in SNFG symbols. At the same time, the 3D-SNFG [15] can generate glycan structures by incorporating SNFG symbols in 3D space for further visualisation using the computational tools like visual molecular dynamics (VMD) [21] LiteMol [22] and Sweet Unity Mol [32].

Glycan sketchers

SugarSketcher. SugarSketcher [14] is a JavaScript interface module currently included in the tool collection of Glycomics@ExPASy (available at https://glycoproteome.expasy.org/sugarsketcher/) for online drawing of glycan structures. The interactive graphical interface (Figure 4, top) allows glycan drawing by glycobiologists and non-expert users. In particular, a "Quick Mode" helps users with limited knowledge of glycans to build up a structure quickly as compared to the normal mode, which offers options related to the structural features of complex carbohydrates (for example additional monosaccharides, isomers, ring types, etc.). The building of glycan structures uses mouse and proceeds via a selection of monosaccharides, substituents and linkages from the list of symbols. However, some wrong combinations of choices can block the interface, resulting in the need to re-start the process (SugarSketcher is on version beta 1.3). Alternatively, SugarSketcher also uses GlycoCT or a native template library as an input. A list of pre-built core N- and O-linked carbohydrate moieties, which are usually present in glycoproteins structures, can be used as a template for further modification. A shortlist of glycan epitopes is also included providing templates for drawing more complex molecules. The software uses the Symbol Nomenclature for Glycans (SNFG) notation for structure representation and exports the obtained sketch to text format (GlycoCT) or image (.svg) files. The software SugarSketcher is featured in the web portal GlyCosmos (https://glycosmos.org/glytoucans/graphic) [33]. Under the name "SugarDrawer", it provides an interface for generating carbohydrate structures to query the database included in GlyCosmos: GlyTouCan [34].

GlyCosmos is a web portal that integrates resources linking glycosciences with life sciences. Besides elements such as "SugarDrawer" and GlyTouCan (carbohydrate database), the platform GlyCosmos assembles data resources ranging from glycoscience standard ontologies to pathologies associated with glycans. GlyCosmos is recognized as the official portal of the Japanese Society for Carbohydrate Research and provides information about genes, proteins, lipids, pathways and diseases.

GlyTouCan (Figure 5) is a repository for glycans which is freely available for the registry of glycan structures. The repository can register structures ranging from monosaccharide compositions to fully defined structures of glycans. It assigns a unique accession number to any glycan to identify its structure and even allows to know its ID number in other databases. Al-

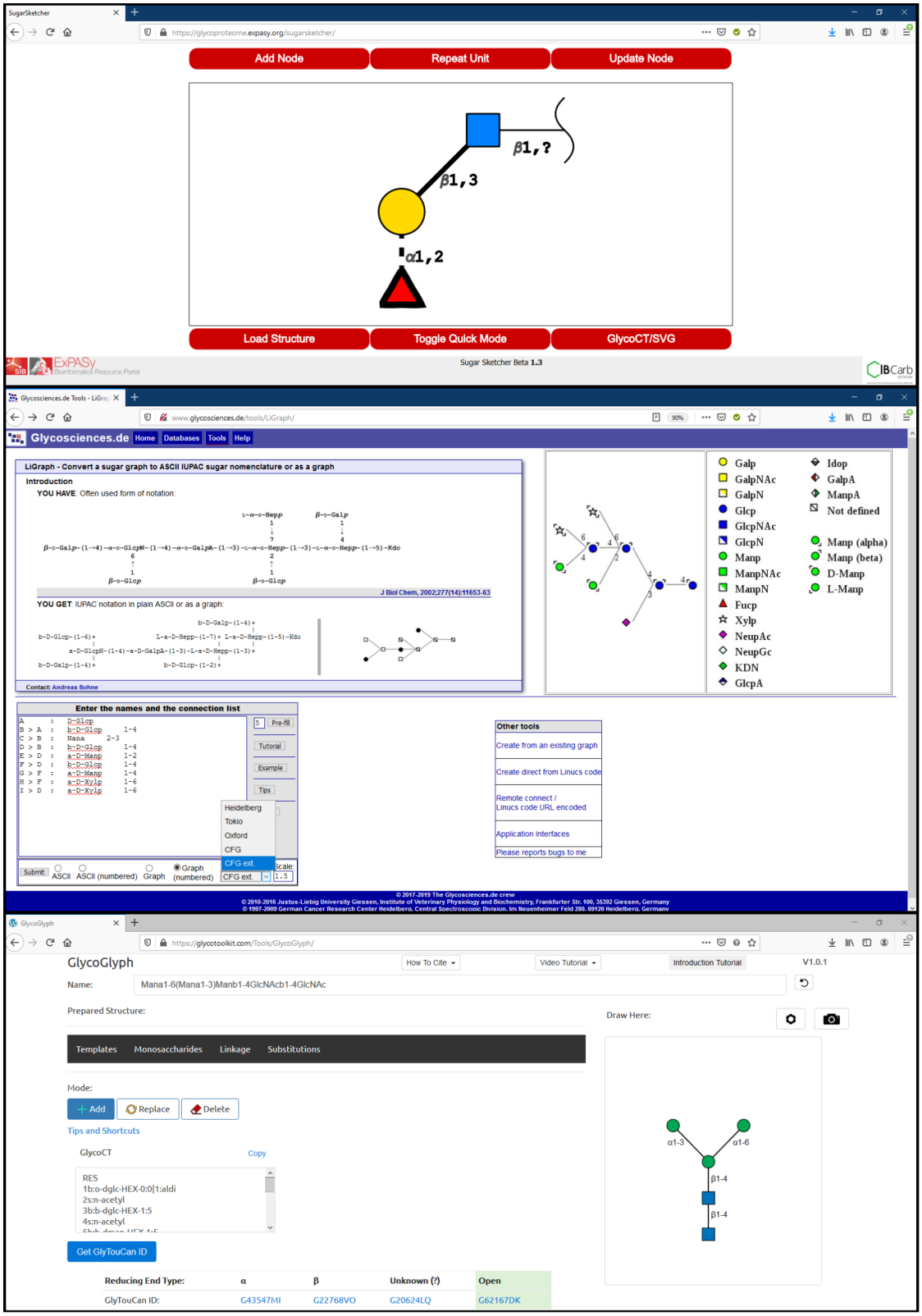


Figure 4: From top to bottom: SugarSketcher [36] interface with a glycan structure drawn using the “Quick Mode”. LiGraph interface showing input and output options for glycan structure representation. GlycoGlyph [37] interface with a text input (modified IUPAC condensed) converted into its glycan image.

Figure 5: GlyTouCan [38] interface allows to search for glycans structures in the database. Data contained in GlyCosmos portal (https://glycosmos.org/) and in GlyTouCan repository home page (https://glytoucan.org/), including their logos, are licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

ternatively, users can search and retrieve information about the glycan structures and motifs that have been already registered into the repository. The structures can be searched simply by browsing through the list of already registered glycans or by specifying a particular sub-structure to retrieve structurally similar glycans (https://glytoucan.org/Structures/graphical). The software tool featured in the GlyTouCan website is called GlycanBuilder and is presented in a later section of our analysis.

Recapitulating, SugarSketcher can be an efficient tool for non-glycobiologists or glycobiologists to sketch glycans. However, it does not accept different input or output formats like IUPAC, WURCS (Web3 Unique Representation of Carbohydrate Structures), which would make the tool more versatile.

LiGraph. LiGraph [35] (http://www.glycosciences.de/tools/LiGraph/) is an online tool based on the concept of schematic drawings of oligosaccharides to display glycan structures. This tool also renders images of glycans in different notation using a text input. The input for the carbohydrate structure consists of a list of names and connections. The glycan structure is output in the specified notation: either ASCII IUPAC sugar nomenclature or a graph which can be rendered in different themes which include Heidelberg, Oxford, Tokyo, CFG and extended CFG (Figure 4, middle). The output images for the glycan structure and the legends can be saved and downloaded in .svg format. This tool is useful for glycan sketching using text templates, but its shortcomings include a limited number of monosaccharide symbols and restricted compatibility with other input file formats.

GlycoGlyph. GlycoGlyph [39] is a web-based application (available at https://glycotoolkit.com/Tools/GlycoGlyph/) built using JavaScript which allows users to draw structures using a graphical user interface or via text string in the CFG linear (also known as modified IUPAC condensed) nomenclature dynamically. The interface (Figure 4, bottom) is equipped with templates for N- and O-linked glycans and terminals. Also, it provides 80+ monosaccharide (SNFG) symbols and a selection for substituents. The selected template or text string (in CFG linear nomenclature) input directly gets converted into an image in canvas and also appears as text in GlycoCT format. The output can be saved as a .svg file or as GlycoCT text. The interface also provides additional options to add, replace or delete each monosaccharide, modify the sizes of symbols and text

fonts, and turn off the linkage annotations or change their orientation; all of which increases the usability of the software. The input structure can be further used to search the GlyTouCan [34] database to explore the literature details related to the input structure.

GlycoGlyph is an efficient tool for sketching or building glycans with a highly usable interface that can significantly help researchers to improve the uniformity in glycan formats in literature/manuscripts. It can also be a tool of choice for text mining for the query structure.

GlycanBuilder2. GlycanBuilder2 [40] is a Java-based glycan drawing tool which runs locally as an application on different platforms including Windows, macOS and Linux. It is freely available for downloading at http://www.rings.t.soka.ac.jp/downloads.html. GlycanBuilder2 is a newer version of GlycanBuilder [20] with additional features. This version is capable of supporting various ambiguous glycans consisting of monosaccharides from plants and bacteria. The tool uses the SNFG notation to display glycan structures. Moreover, this updated version can convert a drawn structure into WURCS sequences for further use as a query for glycan search or registration in databases like GlyTouCan. GlycanBuilder2 provides an excellent interface (Figure 6, top) for glycan drawing. Glycan structures can be drawn manually using the mouse or by importing text input files. The interface provides a list of templates: N- O-glycans, glycosphingolipids, glycosaminoglycans(GAGs). Rows of CFG notations for monosaccharides assist with glycan structure drawing on canvas. The application also supports the glycan symbol notations for the University of Oxford (UOXF) format. The input complies with various linear sequence and text formats. They include GlycoCT¸ GLYcan structural Data Exchange using Connection Tables (GLYDE-II), Bacterial Carbohydrate Structures DataBase (BCSDB) [41], carbohydrate sequence markup language (CabosML) [42], CarbBank [43], LinearCode [44], LINUCS, IUPAC-condensed and GlycosuiteDB [45]. The output yields structures in the following formats: GlycoCT, LinearCode, GLYDE-II and LINUCS. Thus, GlycanBuilder2 is a versatile tool which can be used for glycan sketching or building and also as a glycan sequence converter from one format to another.

Original GlycanBuilder. GlycanBuilder [10,20] was originally part of the GlycoWorkbench platform [49]. This interface is integrated in most tools of the Glycomics@ExPASy collection that require a drawing interface to query data. GlycanBuilder is written in Java Programming language and can be used as standalone or as an applet for embedding in web pages for glycan search. For example, GlycanBuilder is integrated in SugarbindDB [50] to draw glycan structures and search the database (https://sugarbind.expasy.org/builder), and in GlycoDigest or GlyS3 [10,20] to define the input of these tools.

Technically, the tool provides an interactive interface which allows an automated glycan rendering using a library of individual monosaccharides or pre-built template structures (Figure 6, middle). GlycanBuilder provides access to 41 templates. They include N- and O-linked glycans, GAGs (glycosaminoglycans), glycosphingolipids and milk oligosaccharides. It also contains 68 entries from MonosaccharideDB (http://www.monosaccharidedb.org/) including monosaccharides, modifications (e.g. deoxy) and substituents. The tool provides options to modify a monosaccharide by adding substituents and alterations. Free movement of the monosaccharides is allowed through movement and orientation buttons. GlycanBuilder offers multiple options for glycan notation which include CFG, CFG colour, UOXF, UOXF colour and text only. GlycanBuilder can also calculate the masses of glycan structures according to the options selected by the user. GlycanBuilder is a versatile tool for building carbohydrates, with multiple options for exporting the generated structures in the form of text format (GlycoCT, LINUCS, Glycominds, Glyde II) or image (.svg, .png, .jpg, .bmp, .pdf, etc.) files.

DrawRINGS. DrawRINGS [17] is a Java-based applet for rendering glycan structures on canvas (http://www.rings.t.soka.ac.jp/drawRINGS-js/). The different drawing features in an interactive interface (Figure 6, bottom) can be selected with the mouse by surfing the buttons and scroll-down menus. Alternatively, KCF files or KCF text format can be used as input. The free movement of the monosaccharides allows drawings with flexible geometry, for example, for schematic studies of carbohydrates. The drawn glycan structure can be exported in the KCF or IUPAC text format or saved in .png format. The drawn structure can further be used as a query for the search in glycan databases; using match percentage (Similarity) or by the number of components matched (Matched) criteria. Four predefined score matrices are available, named: N-glycans, O-glycans, Sphingolipids and Link_similarity. The "Link_similarity" matrix is based on glycosidic linkages and monosaccharides that may be more highly substituted with other glycosidic linkages and monosaccharides, respectively. There is a query to search the generated structure in the RINGS or GlycomeDB databases (or both). The former compiles data from the KEGG GLYCAN and GLYCOSCIENCES.de databases. DrawRINGS is an efficient tool for sketching glycan figures as well as translating to (and from) the KCF and IUPAC text formats.

DrawGlycan-SNFG. DrawGlycan-SNFG [31] is an open-source program available with a web interface (Figure 7, top) at

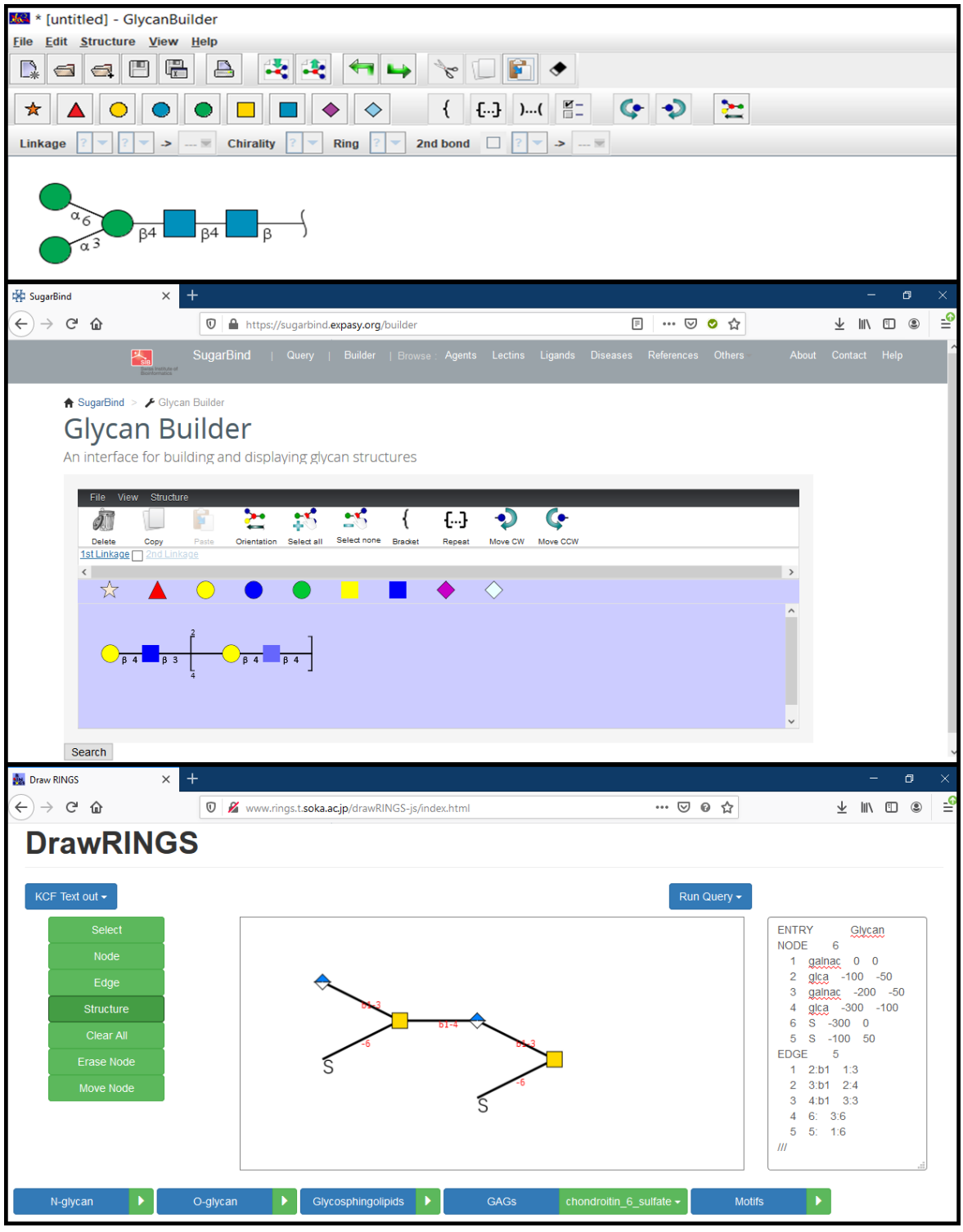


Figure 6: From top to bottom: GlycanBuilder2 [46] interface with a glycan image in SNFG notation. Original GlycanBuilder [47] interface with some of the available templates rendered as images. DrawRINGS [48] interface featuring a glycan and its KCF text output.

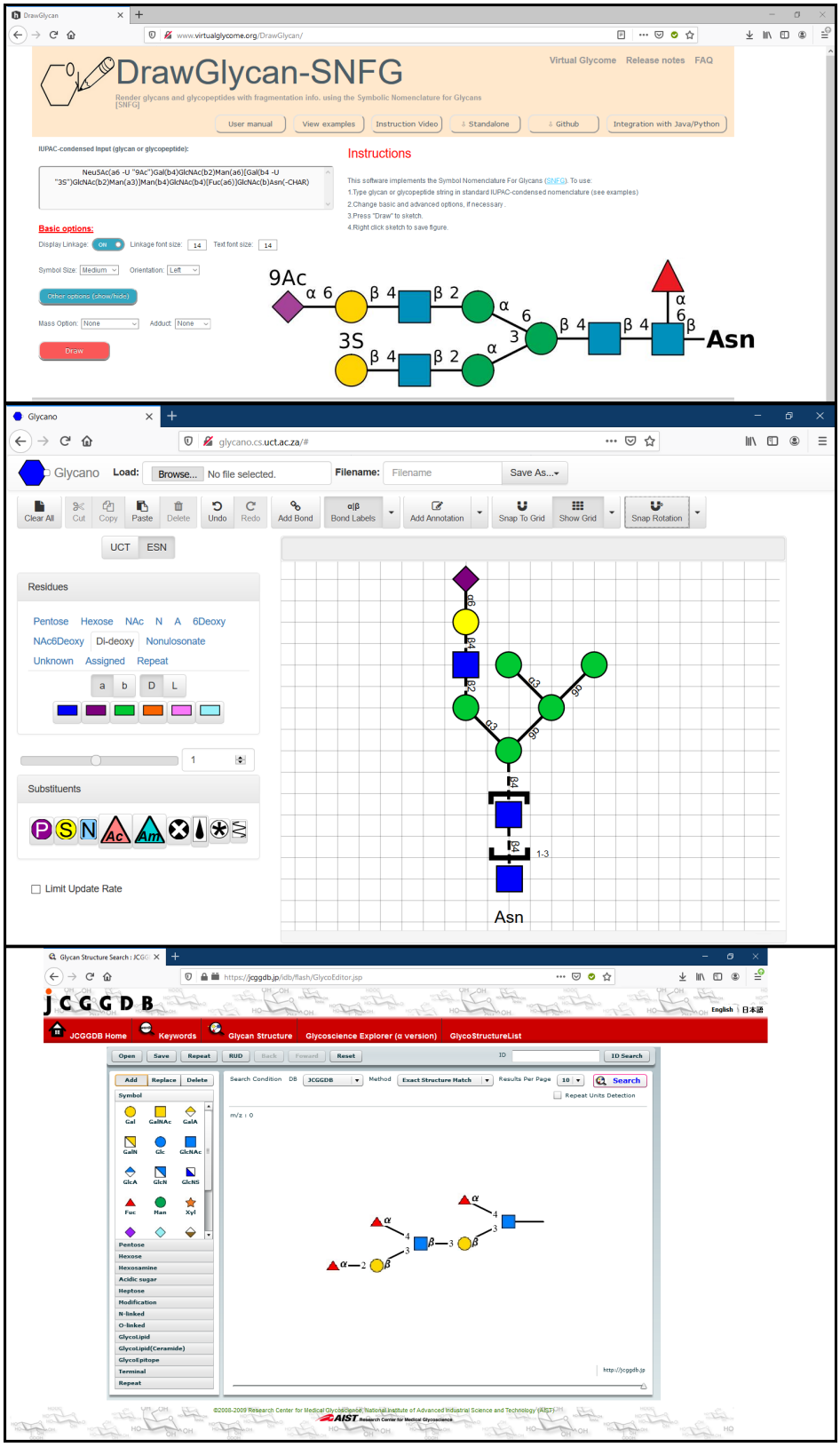


Figure 7: From top to bottom: DrawGlycan-SNFG [51] web interface with a glycan text input and the resulting image output. Glycano [52] interface with a glycan structure. GlycoEditor [53] interface, linkage selection is triggered by adding a new monosaccharide.

http://www.virtualglycome.org/DrawGlycan. The same web page gives access to a downloadable, standalone Graphical User Interface (GUI) version of this tool with additional functionality. It can be launched from different platforms including Windows, Mac or Linux. The program can be used to render glycans and glycopeptides using SNFG and uses IUPAC-condensed text inputs. The DrawGlycan-SNFG version with command-line operations makes it more versatile as it allows integration of multiple features of the program using custom scripts. The tool uses automatic operations for the majority of the drawing, which could meet the needs of researchers, but additional intervention may sometimes be required to get the desired output. For example, manual input in IUPAC-condensed language allows to generate, among others: repeating units, adducts, tethering to other structures (represented by text), and complex branching (the examples section showcases these options). The drawn glycan structure can be saved as .jpg image and modified through parameters such as symbol and text size, the thickness of lines, orientation of drawing and spacing. This software provides all the guidance and tools needed to generate high-quality pictures. DrawGlycan-SNFG is a reliable choice for building glycans.

In addition to glycan structure drawing, DrawGlycan-SNFG (version 2) [54] is equipped with a wide range of options to enhance the usability of the original code [32]. The new version is capable to accommodate the latest updates to the SNFG [7]. This tool has been particularly upgraded for MS spectrum annotation by adding an intuitive interface with additional features. The upgraded version can depict bond fragmentation, repeating structural unit anomeric groups, adduct ions, different types of glycosidic linkages etc. These advanced features make this tool ideal for integrated use with various glycoinformatics software and also for applications in glycoproteomics, glycomics and mass spectrometry (MS). One of the illustrations involves combined use with the gpAnnotate application, dedicated to score and annotate MS/MS glycopeptide spectrums in different fragmentation modes [54].

Glycano. Glycano (available at http://glycano.cs.uct.ac.za) is a software tool for drawing glycans. This tool is based on JavaScript, which can be used without the requirement of any server or browser dependency. The interactive interface allows sketching via the drag-and-drop method on canvas (with or without grid). The software is provided with "UCT" and "ESN", interchangeable interfaces (Figure 7, middle) with different symbols for monosaccharides. These names (UCT and ESN) correspond to the University of Cape Town, South Africa, where Glycano was developed, and to the "Essentials of Glycobiology Symbol Nomenclature", precursor of the SNFG symbol set [55]. The interface provides a wide choice of monosaccharides and substituents represented in SNFG symbols but lacks the standard colour scheme. The user can easily modify the structure with by click and drag, which allows to either cut/copy, delete or move a portion of the structure. The drawn structure can be saved in text format, in .gly format or as an image (PNG and SVG formats). A drawback to note is that linking the monosaccharides at specific positions is only possible in the UCT mode, which means that back-and-forth between the two symbol systems is necessary to define the linkages correctly. Despite some drawbacks, this is an excellent tool due to its ease-of-use, tenable degree of freedom, and functionalities/options for sketching and building glycan structures.

GlycoEditor. GlycoEditor [19] (available at https://jcggdb.jp/idb/flash/GlycoEditor.jsp) is an online software for drawing glycans. Through a straightforward interface, three ways of input are possible: by JCGGDB ID, through a library of common oligosaccharides and by direct input. A list of most common monosaccharides is presented, and the rest can be found categorised by family. The click and drag addition of new monosaccharides trigger the selection of linkage-type and configuration (Figure 7, bottom). The tool provides an option to create repeating units. Additionally, several functionalisation options are also available. Once the structure is ready, the user can save it as an .xml file. GlycoEditor allows searching a given structure across many databases in four ways: exact structure match (with or without anomer and linkage specifics) and the same for substructure match. The central database featured is the JCGGDB, to which can be added, among others: Glaxy, GlycomeDB, GlycoEpitope, GMDB, KEGG, etc. Searching by ID is also possible. GlycoEditor is a now dated tool that allows efficiently building glycans and performing databases searches.

GLYCO.ME (SugarBuilder). Glyco.me-SugarBuilder (available at https://beta.glyco.me/sugarbuilder) is online software for drawing glycans. The interface leads to rapid carbohydrate construction. A panel of monosaccharide templates complements the drawing interface (one pre-built oligosaccharide is available (Figure 8, top). The user can start a chain from amino acid residues: Asn, Ser or Thr, then structure building is limited by to a set of "rules" (limiting building options to known carbohydrates). These rules may be deactivated with a switch button to draw freely. A list of 13 monosaccharides is deployed, and sequential clicking allows their addition to the existing structure and definition of the associated glycosidic bond (the relative sizes of the options available related to their real statistical value for that particular linkage). Upon building some specific motifs, if they are recognised, an option for repeating units appears. Other switch buttons allow the user to change the orientation of the drawing, show/hide linkage information etc. The Oxford notation can be enabled for glycosidic bonds only. The

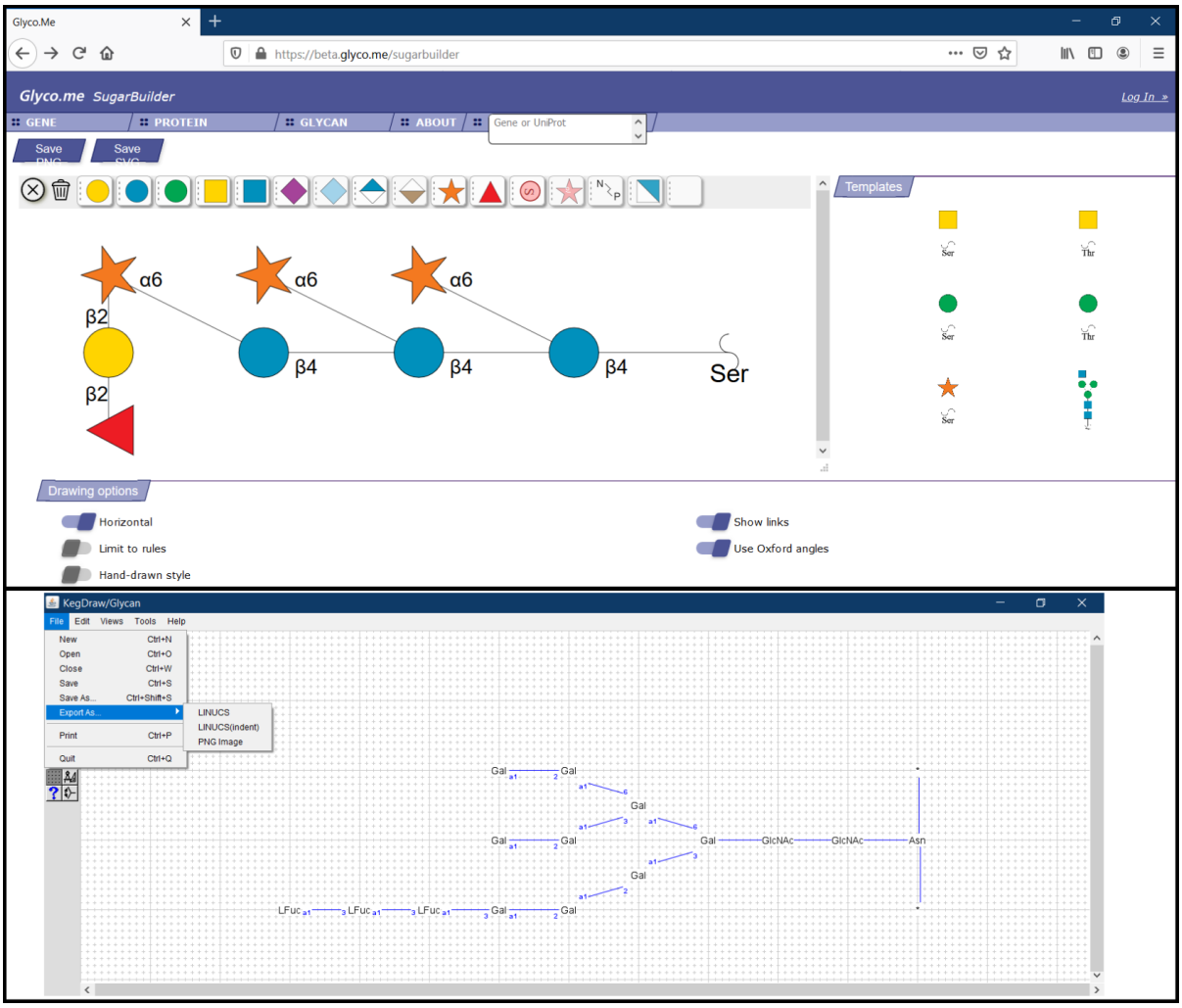


Figure 8: From top to bottom: Glyco.me SugarBuilder [56] interface with a glycan structure showing options to define anomericity and monosaccharide linkage position. KegDraw [57] interface with a glycan structure and available options to save the structure file in different formats.

structure obtained can be rendered as .png or .svg images. Glyco.me-SugarBuilder is still under development: more monosaccharides/substitutions/templates will complete an already very functional platform. The quick and easy options put forward offer natural building and liberty for tailoring the rendered image.

KegDraw. KegDraw (https://www.kegg.jp/kegg/download/kegtools.html) is a freely available Java application for rendering glycan structures. It can be downloaded and installed locally as a platform-independent tool. This tool can be used in two different modes: "Compound mode" which can be used for drawing small molecules (similarly to any chemical structure drawing software), and "Glycan mode" which is dedicated for rendering glycan structures using different monosaccharide units. The simplest method for drawing involves a selection of monosaccharides and glycosidic linkages from an available list to generate a glycan structure. Alternatively, a text box option provides a way to draw uncommon types of monosaccharides. The tool also contains templates from KEGG GLYCAN and their importation using their accession number. Besides, input files in KCF can be used while the output can be saved in LINUCS, KCF or an image in PNG format (Figure 8, bottom). The glycan structure in text format can be further used as a query for search in KEGG GLYCAN and CarbBank databases. Hence, KegDraw can be an option for the freely available tool for drawing and querying chemical structures. However, there are similar tools already available for glycan drawing with more advanced and acceptable notations.

Glycan builders

Sweet II. Sweet [58] is a web-based program for constructing 3D models of glycans from a sequence using standard nomenclature accessible at http://www.glycosciences.de/modeling/

sweet2/doc/index.php (Figure 9, top). This tool is available as a part of the glycosciences.de website, which also provides other options for analysing glycans in three-dimensional space. This program uses a glycan sequence in a standard format and generates a 3D model in the form of a .pdb file. The glycan input can come from a library of relevant oligosaccharides, available through one of the sub-menus. Alternatively, manual input is possible in three platforms adapted for increasing complexity. The model can be further minimised using MM2 [59] and MM3 [60] methods. The 3D models can be viewed using molecular viewers like JMol, WebMol-applet, Chemis3D-applet, etc. Besides, the program also generates additional files which can be used for molecular mechanics and molecular dynamics using molecular modelling tool like Tinker [61]. This tool is as a versatile tool for generating a 3D model for glycans.

GLYCAM-web (Carbohydrate Builder). Carbohydrate builder [65] is an online tool (at http://glycam.org/) for carbohydrate structure drawing and subsequent 3D structure building. With a flexible interface, it uses three methods for glycan building. The first method is manual building ("Carbohydrate Builder" button). It allows selection of monosaccharide, as well as defining linkages, branching and substitution (Figure 9, middle). The second method involves the use of a template library (using "Oligosaccharide libraries" button) containing commonly relevant structures (http://glycam.org/Pre-builtLibraries.jsp). The third option (direct input from a text sequence) becomes relevant when the glycan structure does not exist in the library or challenging to build due to structural complexity. In this case, a text for the oligosaccharide in GLYCAM-Web's condensed notation can be entered as an input to create the glycan structure. Once the glycan is generated, the options include the solvation of the structure and the manual input of the glycosidic linkages. The tool allows structure minimisation and generates rotamers which can be visualised using JSmol viewer. Information about the force field that is used to build the structure is also provided. The multiple structures can be downloaded compressed as .tar, .gz or .zip files containing .pdb files. Similarly, the 2D image can be saved in GIF format. GLYCAM-web- Carbohydrate Builder can be used to prepare the system for MD simulation as it solvates the glycans and also generates the topology and coordinate files. In addition to its carbohydrate builder, Glycam-web consists of additional tools like glycoprotein builder and glycosaminoglycans (GAG) builder.

CHARMM-GUI (Glycan Reader and Modeler). The CHARMM-GUI (http://www.charmm-gui.org) is a web-based graphical user interface which provides various functional modules to prepare complex biomolecular systems and input files for molecular simulations. Glycan Reader and Modeler [65-67] is a part of CHARMM-GUI (Figure 9, bottom) and available as a freely accessible online tool at http://charmm-gui.org/input/glycan. It can read input files in PDB, PDBx/mmCIF and CHARMM formats containing glycans and automatically detects the carbohydrate molecules and glycosidic linkage information. Alternatively, it can also read a glycan sequence (GRS format) to generate a 3D model and input files for MD simulation of the carbohydrate-only system. GRS carbohydrate sequences can be made through a straightforward interface: monosaccharides (20+ options) and their linkages are added incrementally from drop-down menus. A useful feature of this tool is the real-time rendering of the carbohydrate image: each added monosaccharide and modified linkage is directly reported to the image as well as to a text (GRS) format. Option for numerous chemical modifications is also available.

On the other hand, the Glycan Modeler allows in silico N-/O-glycosylation for glycan-protein complexes and generates a "most relevant" glycan structure through Glycan Fragment Database (GFDB) [68] search which gives proper orientations relative to the target protein. In the absence of target glycan sequence in GFDB, the structures are generated by using the valid internal coordinate information (averaged phi, psi, and omega glycosidic torsion angles) in the CHARMM force field. Input files for CHARMM can be generated for the purpose of MD simulation. Amongst other possible outputs, 3D representations of the glycans are available as .pdb files. This tool can be helpful for researchers to generate 2D depictions of a glycan and then obtain the corresponding 3D representation, which can be useful for modelling studies of glycans and glycoconjugates.

doGlycans. doGlycans [69] is a compilation of tools dedicated for preparing carbohydrate structures for atomistic simulations of glycoproteins, carbohydrate polymers and glycolipids using GROMACS [70,71] In the form of Python scripts; the tools are used to prepare the system, which generally includes the processing of a.pdb file using the *pdb2gmx* tool. Subsequently, a glycosylation model can be prepared for carbohydrate polymer simulation using the *prepreader.py* script. Similarly, the *doglycans.py* script can be used to develop models for glycoproteins and glycolipids. Together, these tools are called doGlycans toolset. Although doGlycans is highly flexible, it only uses the sugar units that are defined in GLYCAM. The topologies generated for glycosylated proteins and glycolipids are compatible with the OPLS [72] and AMBER [73] force fields. The topology for carbohydrate polymers is based on the GLYCAM force field. The user needs to provide the ceramide topology as input to generate the topologies for glycolipids. The tools contained in doGlycans create 3D models and simulation files as a starting point for more complex molecular simulation studies.

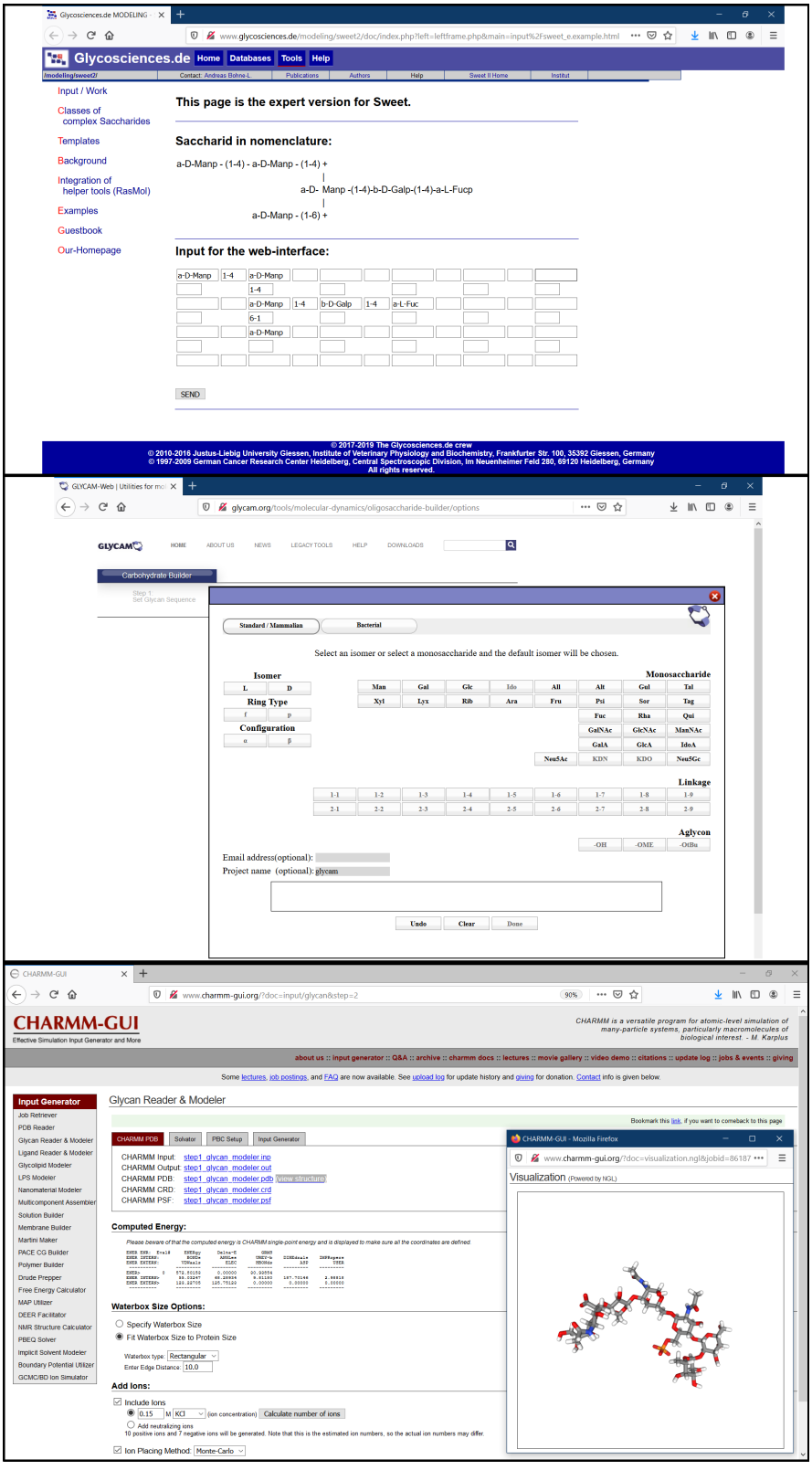


Figure 9: From top to bottom: Sweet II [62] web-interface with a text input to generate a 3D model. GLYCAM Carbohydrate Builder [63] interface which accepts a text input for glycans and generates 3D models. CHARMM-GUI (Glycan reader and Modeler) [64] interface with a 3D structure output generated using a glycan sequence as input.

RosettaCarbohydrate. Rosetta is a software suite for macromolecular modelling as an extensive collection of computer code mostly written in C++ and Python languages. Rosetta is available to academic and commercial researchers through a license available at https://www.rosettacommons.org/software/license-and-download. The licence is free for academic users. The tool runs best on Linux or macOS platforms only. It can be installed on a multiprocessor computing cluster to increase efficiency. RosettaCarbohydrate [74,75] tool provides the methods for general modelling and docking applications for glycans and glycoconjugates. The application accepts the standard PDB, GLYCAM, and GlycoWorkbench (.gws) file formats and the available utilities (codes) helps with the general problems in sampling, scoring, and nomenclature related to glycan modelling. It samples glycosidic bonds, ring forms, side-chain conformations, and utilises a glycan-specific term within its scoring function. The tool also consists of utilities for virtual glycosylation, protein–glyco-ligand docking, and glycan "loop" modelling. This tool is best for the researcher with basic knowledge and skills to work with a command-line interface (Linux).

PolysGlycanBuilder. PolysGlycanBuilder [76] is a web-based tool (http://glycan-builder.cermav.cnrs.fr/) with an interactive and more usable interface (Figure 10). The software translates a glycan sequence or polysaccharide repeat unit into the coordinate set of the corresponding tertiary structure, in one or several of its low energy conformations. The construction follows an intuitive scheme which is as close as possible to the way glycoscientists draw the sequence of their structures. The simplest method for model building involves dragging and dropping monosaccharide units to the canvas or workspace grid. The software displays rows of monosaccharides in the form of standard SNFG symbols with 3D information (furanose/pyranose shape, configuration, anomericity, and ring conformation). Glycosidic linkages can be easily defined, as the values of the dihedral angles (Φ, Ψ, Ω). They can be manually set or extracted from a database of low energy conformations of 600 disaccharide segments. The monosaccharides have been subjected to geometry optimisation using molecular mechanics approach. For a given input sequence, the corresponding 3D coordinates are generated at the PDB format. Within the process of construction, the structure is displayed via the LiteMol and eventually optimised to remove any steric clashes. The image for the glycan can be downloaded and saved in SVG format. Keeping the glycan/polysaccharide structure in text format (condensed IUAPC, GlycoCT, SNFG and INP) offers several ways to connect to other applications. Other than drag and drop method, PolysGlycan-Builder also accepts input of files in INP, IUPAC and GlycoCT formats. An interactive interface accompanies the application, which makes it more versatile for glycan drawing and 3D model building.

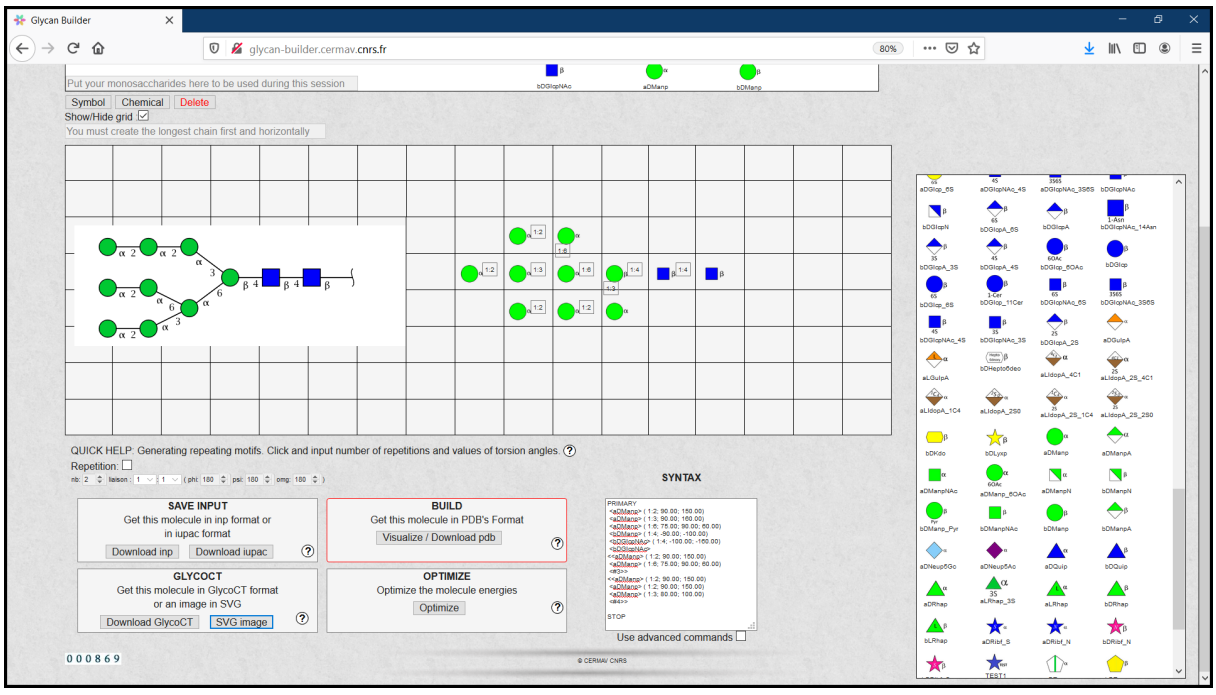


Figure 10: PolysGlycanBuilder [77] interface illustrating glycan drawing using SNFG symbols. The glycan can be further converted into a 3D model.

Displaying 3D structures of glycans

3D-SNFG VMD interface and visualisation algorithms. The recently introduced 3D-Symbol Nomenclature for Glycans (3D-SNFG) [15] allows the representation of carbohydrates in an unusual way: the SNFG symbols are added to a three-dimensional structure. The 3D-SNFG script must be integrated into the visual molecular dynamics (VMD) [21,78] viewer software to enable the representation of glycans as large SNFG-matching3D shapes that can either replace the molecular monosaccharides or stay lodged at the geometric centre of the cycle (Figure 11, top left). Upon the input of a glycan-containing structure (in PDB format), the integrated script in VMD automatically recognises the common monosaccharide names and generates the 3D shapes. The embedded script also enables shortcuts keys from keyboard to quickly change between large and small 3D-SNFG shapes and also label the reducing terminus. The 3D structure displayed in VMD can be saved as a .bmp image file. Thanks to 3D-SNFG, the standardised representation of glycan structures can finally take a step into the 3D space. The obtained images can become very useful for quick assessment of 3D glycan models.

In addition to the 3D-SNFG script, *PaperChain* and *Twister* [83] are two visualisation algorithms available with the Visual Molecular Dynamics (VMD) package. These algorithms are useful to visualize complex cyclic molecules and multi-branched polysaccharides.{Cross, 2009 #69} *PaperChain* displays rings in a molecular structure with a polygon and colours them according to the ring pucker. The other algorithm (*Twister)* traces glycosidic bonds in a ribbon representation that twists and changes its orientation according to the relative position of following sugar residues, hence provides an important conformational detail in polysaccharides. Combination of these algorithms with other visualisation features available in VMD can enhance the flexibility of displaying structural details of glycoconjugate, glycoprotein and cyclic structures.

LiteMol. The LiteMol [22] viewer is a freely available web application (Figure 11, top right) for 3D visualisation of macromolecules and other related data. LiteMol enables standard visualisation of macromolecules in different representation modes like surface, cartoons, ball-and-stick, etc. The software can be accessed at v.litemol.org and also available for integration in a webpage from the github (https://github.com/dsehnal/LiteMol). LiteMol is compatible with all modern browsers without the support of additional plugins. The viewer automatically depicts any carbohydrate residues and displays 3D structures of carbohydrates with 3D-SNFG symbols, which allows the viewer to identify the monosaccharides readily. The presented structure can be saved as a.png image file. Any monosaccharide with a residue name in PDB can be visualised using 3D-SNFG in LiteMol. However, a significant portion of the carbohydrates may contain some form of error in annotation, which would result in either no symbol or an incorrect symbol. Although LiteMol is an efficient and rapid 3D viewer for glycans, 3D representation does not provide any information about the glycosidic linkage type (e.g. α1-3 or β1-4). Also, it does not display any information about connection and configuration. If this information is required, returning to the classic molecular representation is possible.

PyMOL- Azahar plugin. Azahar [84] is a plugin in PyMOL [85] which enables building, visualization and analysis of glycans and glycoconjugates. This tool is based on Python and provides additional computing environment within the PyMOL package. The tool is provided with a template list of saccharide structures to facilitate structure building and visualisation. The interface provides three option menus to assist glycan structure building. The two first options help to specify residues to be connected from a list of available templates, and the third one allows selection of the chemical bond between the residues. The visualisation using PyMOL includes three cartoon-like representations. These display modes provided in the tool simplify the representation of glycan structures in cartoon, wire and bead representations. In cartoon and wire representations, the rings in sugars are shown as non-flat polygons connected by rods while in the bead representation mode, these cycles are represented as a sphere. In addition of visualization of static structures, the tool also allows analysis of trajectories of MD simulations. The tool can be used for conformational search using a Monte Carlo approach [86]. The conformational search is done by perturbing a torsional angle, followed by an energy minimization using the MMFF94 force field. Azahar is freely accessible from http://www.pymolwiki.org/index.php/Azahar.

UnityMol/SweetUnityMol. Sweet UnityMol [32] is a molecular structure viewer (Figure 11, middle) developed from the game engine Unity3D. The software is available for free download (https://sourceforge.net/projects/unitymol/files/UnityMol_1.0.37/) from the SourceForge project website. It can be installed in Mac, Windows and Linux platforms. The program reads files in PDB, mmCIF, Mol2, GRO, XYZ, and SDF formats, OpenDX potential maps and XTC trajectory files. It efficiently displays specific structural features for the simplest to the most complex carbohydrate-containing biomolecules. Sweet UnityMol displays 3D carbohydrate structures with different modes of representation, such as: liquorice, ball-and-stick, hyperBalls, RingBlending, hydrophilic/hydrophobic character of sugar face etc. The most recent version is fully compatible with the SNFG colour coding, which also uses acceptable pictorial representation, generally used in carbohydrate chemistry, biochemistry and glycobiology.

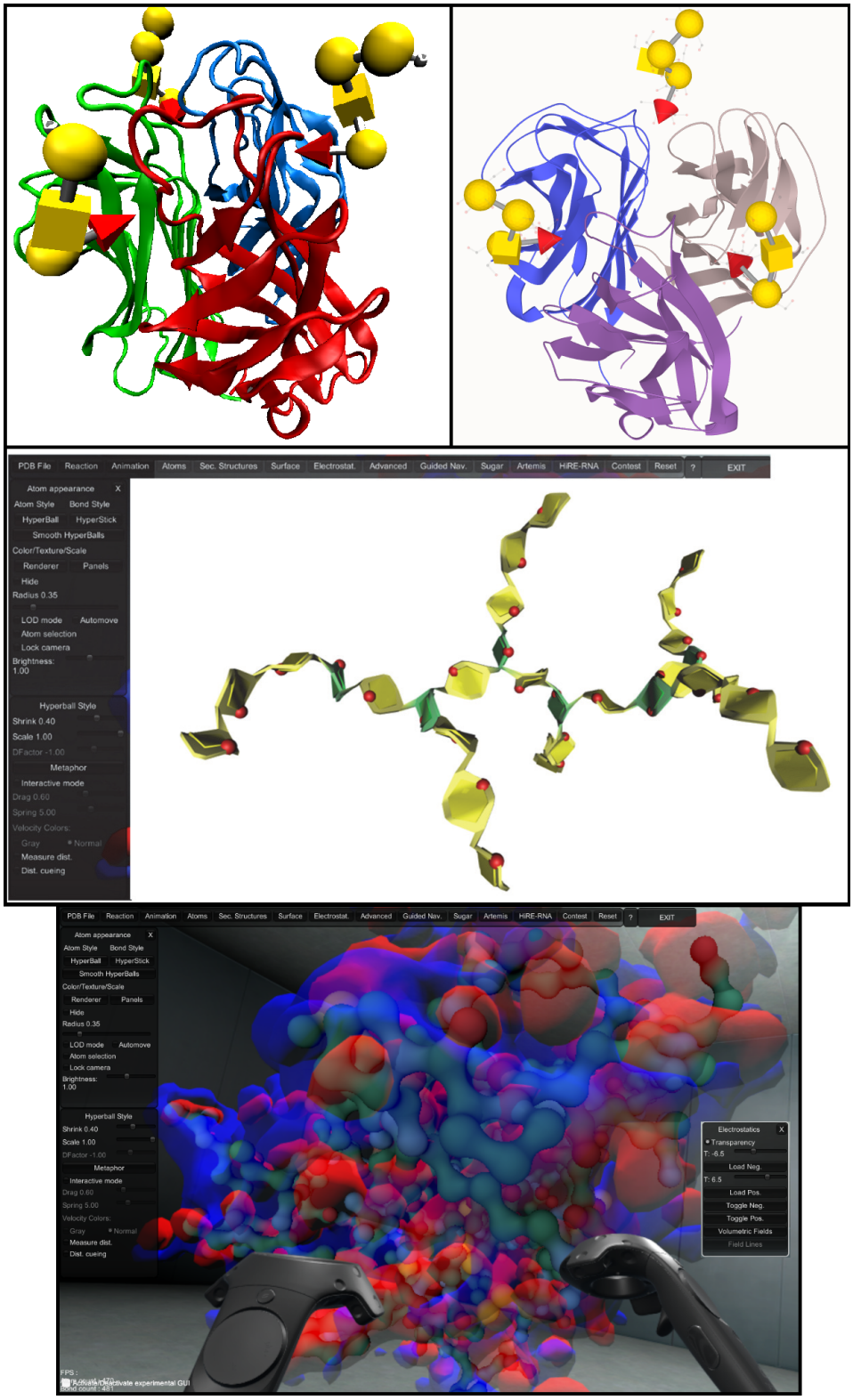


Figure 11: From top to bottom: 3D-SNFG representation of glycan using 3D-SNFG script integrated VMD [79]. LiteMol [80] interface with 3D-SNFG representation of glycan in a protein–glycan complex. SweetUnityMol [81] among the several types of representations a ribbon-like display of polysaccharide ribbons maintains the SNFG colour coding of monosaccharides. UnityMol [82] within an immersive virtual reality context.

SweetUnityMol provides a continuum from the conventional ways to depict the primary structures of complex carbohydrates all the way to visualising their 3D structures. Several options are offered to the user to select the most relevant type of depictions, including new features, such as "Coarse-Grain" representation while keeping the option to display the details of the atomic representations. Powerful rendering methods produce high-quality images of molecular structures, bio-macromolecular surfaces and molecular interactions.

A recently developed version of UnityMol has been implemented with the immersive Virtual Reality context using head-mounted displays [87]. It offers high-quality visual representations, ease of interactions with multiple molecular objects, powerful tools for visual manipulations, accompanied by the evaluation of intermolecular interactions. Consequently, simultaneous investigations of multiple objects such as macromolecular interactions gain in efficiency and accuracy. (Figure 11, bottom).

Conclusion

The set of computational tools presented above illustrates the rich contributions of a community devoted to enabling the accurate representation of complex carbohydrates via the development and implementation of a versatile informatics toolbox. These legitimate efforts aim at facilitating communication within the scientific community. To establish a comparative analysis of the several available applications, we evaluated 17 selected items that characterise best their availability, implementation, maintenance and field of use. The comparative analysis of tools could be useful for glycobiologists or any researcher looking for a ready to use, simple application for the sketching, building and display of glycans.

This article provides an overview of the computational tools and resources available for glycan sketching, building and representing. It also provides a descriptive analysis of the recently developed software tools dedicated explicitly to glycans and glycoconjugates. The newly developed tools are more advanced and use the standard nomenclature and symbols for glycan representation. These tools can further help to standardise the description of glycans in research, communication and databases.

Supporting Information

Supporting Information File 1

Features of glycan sketchers, builders and viewers.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-199-S1.pdf]

Acknowledgements

Appreciation is extended to Drs. A. Imberty, A. Varrot, L. Belvisi and A. Bernardi for their support.

Funding

This research was performed within the framework of the PhD4GlycoDrug Innovative Training Network and was funded from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765581. The work was supported by the Cross-Disciplinary Program Glyco@Alps, within the framework "Investissement d'Avenir" program [ANR-15IDEX-02].

ORCID® iDs

Kanhaya Lal - https://orcid.org/0000-0001-8555-7948
Rafael Bermeo - https://orcid.org/0000-0002-4451-878X
Serge Perez - https://orcid.org/0000-0003-3464-5352

References

1. Alocci, D.; Lisacek, F.; Perez, S. *A Traveler's Guide to Complex Carbohydrates in the Cyber Space*. http://www.glycopedia.eu/IMG/pdf/traveler_s_guide_to_cyber_space.pdf
2. Perez, S.; Aoki-Kinoshita, K. F. *Development of Carbohydrate Nomenclature and Representation;* Springer, 2017; pp 7–25. doi:10.1007/978-4-431-56454-6_2
3. Kornfeld, S.; Li, E.; Tabas, I. *J. Biol. Chem.* **1978,** *253,* 7771–7778.
4. Royle, L.; Dwek, R. A.; Rudd, P. M. *Curr. Protoc. Protein Sci.* **2006,** *43,* 12.6.1–12.6.45. doi:10.1002/0471140864.ps1206s43
5. Harvey, D. J.; Merry, A. H.; Royle, L.; Campbell, M. P.; Dwek, R. A.; Rudd, P. M. *Proteomics* **2009,** *9,* 3796–8301. doi:10.1002/pmic.200900096
6. Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Marth, J. D.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E. *Proteomics* **2009,** *9,* 5398–5399. doi:10.1002/pmic.200900708
7. Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; Varki, A.; Woods, R. J.; Darvill, A.; Dell, A.; Henrissat, B.; Bertozzi, C.; Hart, G.; Narimatsu, H.; Freeze, H.; Yamada, I.; Paulson, J.; Prestegard, J.; Marth, J.; Vliegenthart, J. F. G.; Etzler, M.; Aebi, M.; Kanehisa, M.; Taniguchi, N.; Edwards, N.; Rudd, P.; Seeberger, P.; Mazumder, R.; Ranzinger, R.; Cummings, R.; Schnaar, R.; Perez, S.; Kornfeld, S.; Kinoshita, T.; York, W.; Knirel, Y. *Glycobiology* **2019,** *29,* 620–624. doi:10.1093/glycob/cwz045
8. Varki, A.; Cummings, R. D.; Aebi, M.; Packer, N. H.; Seeberger, P. H.; Esko, J. D.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; Prestegard, J. J.; Schnaar, R. L.; Freeze, H. H.; Marth, J. D.; Bertozzi, C. R.; Etzler, M. E.; Frank, M.; Vliegenthart, J. F. G.; Lütteke, T.; Perez, S.; Bolton, E.; Rudd, P.; Paulson, J.; Kanehisa, M.; Toukach, P.; Aoki-Kinoshita, K. F.; Dell, A.; Narimatsu, H.; York, W.; Taniguchi, N.; Kornfeld, S. *Glycobiology* **2015,** *25,* 1323–1324. doi:10.1093/glycob/cwv091
9. Aoki, K. F.; Yamaguchi, A.; Ueda, N.; Akutsu, T.; Mamitsuka, H.; Goto, S.; Kanehisa, M. *Nucleic Acids Res.* **2004,** *32,* W267–W272. doi:10.1093/nar/gkh473

10. Ceroni, A.; Dell, A.; Haslam, S. M. *Source Code Biol. Med.* **2007,** *2,* No. 3. doi:10.1186/1751-0473-2-3
11. Damerell, D.; Ceroni, A.; Maass, K.; Ranzinger, R.; Dell, A.; Haslam, S. M. Annotation of Glycomics MS and MS/MS Spectra Using the GlycoWorkbench Software Tool. In *Glycoinformatics. Methods in Molecular Biology;* Lütteke, T.; Frank, M., Eds.; Humana Press: New York, NY, 2015; Vol. 1273, pp 3–15. doi:10.1007/978-1-4939-2343-4_1
12. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000,** *28,* 235–242. doi:10.1093/nar/28.1.235
13. Engelsen, S. B.; Hansen, P. I.; Pérez, S. *Biopolymers* **2014,** *101,* 733–743. doi:10.1002/bip.22449
14. Alocci, D.; Suchánková, P.; Costa, R.; Hory, N.; Mariethoz, J.; Vařeková, R.; Toukach, P.; Lisacek, F. *Molecules* **2018,** *23,* 3206. doi:10.3390/molecules23123206
15. Thieker, D. F.; Hadden, J. A.; Schulten, K.; Woods, R. J. *Glycobiology* **2016,** *26,* 786–787. doi:10.1093/glycob/cww076
16. McNaught, A. D. *Adv. Carbohydr. Chem. Biochem.* **1997,** *52,* 44–177. doi:10.1016/s0065-2318(08)60090-6
17. Akune, Y.; Hosoda, M.; Kaiya, S.; Shinmachi, D.; Aoki-Kinoshita, K. F. *OMICS* **2010,** *14,* 475–486. doi:10.1089/omi.2009.0129
18. Hashimoto, K.; Goto, S.; Kawano, S.; Aoki-Kinoshita, K. F.; Ueda, N.; Hamajima, M.; Kawasaki, T.; Kanehisa, M. *Glycobiology* **2006,** *16,* 63R–70R. doi:10.1093/glycob/cwj010
19. Maeda, M.; Fujita, N.; Suzuki,, Y.; Sawaki, H.; Shikanai, T.; Narimatsu, H. JCGGDB: Japan Consortium for Glycobiology and Glycotechnology Database. In *Glycoinformatics. Methods in Molecular Biology;* Lütteke, T.; Frank, M., Eds.; Humana Press: New York, NY, 2015; Vol. 1273, pp 161–179. doi:10.1007/978-1-4939-2343-4_12
20. Damerell, D.; Ceroni, A.; Maass, K.; Ranzinger, R.; Dell, A.; Haslam, S. M. *Biol. Chem.* **2012,** *393,* 1357–1362. doi:10.1515/hsz-2012-0135
21. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996,** *14,* 33–38. doi:10.1016/0263-7855(96)00018-5
22. Sehnal, D.; Grant, O. C. *J. Proteome Res.* **2019,** *18,* 770–774. doi:10.1021/acs.jproteome.8b00473
23. Foley, B. L.; Tessier, M. B.; Woods, R. J. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012,** *2,* 652–697. doi:10.1002/wcms.89
24. Mallajosyula, S. S.; Guvench, O.; Hatcher, E.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2012,** *8,* 759–776. doi:10.1021/ct200792v
25. Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2009,** *5,* 2353–2370. doi:10.1021/ct900242e
26. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008,** *29,* 622–655. doi:10.1002/jcc.20820
27. Lins, R. D.; Hünenberger, P. H. *J. Comput. Chem.* **2005,** *26,* 1400–1412. doi:10.1002/jcc.20275
28. Molinero, V.; Goddard, W. A. *J. Phys. Chem. B* **2004,** *108,* 1414–1427. doi:10.1021/jp0354752
29. Herget, S.; Ranzinger, R.; Maass, K.; Lieth, C.-W. v. d. *Carbohydr. Res.* **2008,** *343,* 2162–2171. doi:10.1016/j.carres.2008.03.011
30. Tanaka, K.; Aoki-Kinoshita, K. F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2014,** *54,* 1558–1566. doi:10.1021/ci400571e
31. Cheng, K.; Zhou, Y.; Neelamegham, S. *Glycobiology* **2017,** *27,* 200–205. doi:10.1093/glycob/cww115
32. Perez, S.; Tubiana, T.; Imberty, A.; Baaden, M. *Glycobiology* **2015,** *25,* 483–491. doi:10.1093/glycob/cwu133
33. Yamada, I.; Shiota, M.; Shinmachi, D.; Ono, T.; Tsuchiya, S.; Hosoda, M.; Fujita, A.; Aoki, N. P.; Watanabe, Y.; Fujita, N.; Angata, K.; Kaji, H.; Narimatsu, H.; Okuda, S.; Aoki-Kinoshita, K. F. *Nat. Methods* **2020,** *17,* 649–650. doi:10.1038/s41592-020-0879-8
34. Tiemeyer, M.; Aoki, K.; Paulson, J.; Cummings, R. D.; York, W. S.; Karlsson, N. G.; Lisacek, F.; Packer, N. H.; Campbell, M. P.; Aoki, N. P.; Fujita, A.; Matsubara, M.; Shinmachi, D.; Tsuchiya, S.; Yamada, I.; Pierce, M.; Ranzinger, R.; Narimatsu, H.; Aoki-Kinoshita, K. F. *Glycobiology* **2017,** *27,* 915–919. doi:10.1093/glycob/cwx066
35. Lütteke, T.; Bohne-Lang, A.; Loss, A.; Goetz, T.; Frank, M.; von der Lieth, C.-W. *Glycobiology* **2006,** *16,* 71R–81R. doi:10.1093/glycob/cwj049
36. *Sugar Sketcher.* https://glycoproteome.expasy.org/sugarsketcher/ (accessed April 2020).
37. *GlycoGlyph.* https://glycotoolkit.com/Tools/GlycoGlyph/ (accessed April 2020).
38. *GlyTouCan*. https://glytoucan.org/ (accessed April 2020).
39. Mehta, A. Y.; Cummings, R. D. *Bioinformatics* **2020,** *36,* 3613–3614. doi:10.1093/bioinformatics/btaa190
40. Tsuchiya, S.; Aoki, N. P.; Shinmachi, D.; Matsubara, M.; Yamada, I.; Aoki-Kinoshita, K. F.; Narimatsu, H. *Carbohydr. Res.* **2017,** *445,* 104–116. doi:10.1016/j.carres.2017.04.015
41. Toukach, P. V.; Egorova, K. S. *Nucleic Acids Res.* **2016,** *44,* D1229–D1236. doi:10.1093/nar/gkv840
42. Kikuchi, N.; Kameyama, A.; Nakaya, S.; Ito, H.; Sato, T.; Shikanai, T.; Takahashi, Y.; Narimatsu, H. *Bioinformatics* **2005,** *21,* 1717–1718. doi:10.1093/bioinformatics/bti152
43. Doubet, S.; Bock, K.; Smith, D.; Darvill, A.; Albersheim, P. *Trends Biochem. Sci.* **1989,** *14,* 475–477. doi:10.1016/0968-0004(89)90175-8
44. Banin, E.; Neuberger, Y.; Altshuler, Y.; Halevi, A.; Inbar, O.; Nir, D.; Dukler, A. *Trends Glycosci. Glycotechnol.* **2002,** *14,* 127–137. doi:10.4052/tigg.14.127
45. Cooper, C. A.; Harrison, M. J.; Wilkins, M. R.; Packer, N. H. *Nucleic Acids Res.* **2001,** *29,* 332–335. doi:10.1093/nar/29.1.332
46. *GlycanBuilder2*. Downloaded from http://www.rings.t.soka.ac.jp/downloads.html (accessed April 2020).
47. *SugarBind GlycanBuilder.* https://sugarbind.expasy.org/builder (accessed April 2020).
48. *DrawRINGS*. http://www.rings.t.soka.ac.jp/drawRINGS-js/ (accessed April 2020).
49. Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. *J. Proteome Res.* **2008,** *7,* 1650–1659. doi:10.1021/pr7008252
50. Mariethoz, J.; Khatib, K.; Alocci, D.; Campbell, M. P.; Karlsson, N. G.; Packer, N. H.; Mullen, E. H.; Lisacek, F. *Nucleic Acids Res.* **2016,** *44,* D1243–D1250. doi:10.1093/nar/gkv1247
51. *DrawGlycan-SNFG*. http://www.virtualglycome.org/DrawGlycan/ (accessed April 2020).
52. *Glycano.* http://glycano.cs.uct.ac.za/ (accessed April 2020).
53. *GlycoEditor*. https://jcggdb.jp/idb/flash/GlycoEditor.jsp (accessed April 2020).
54. Cheng, K.; Pawlowski, G.; Yu, X.; Zhou, Y.; Neelamegham, S. *Bioinformatics* **2019**. doi:10.1093/bioinformatics/btz819

55. Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H., *Essentials of Glycobiology [Internet].* 3 ed.; Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press: 2015–2017.
56. *Glyco.me SugarBuilder.* https://beta.glyco.me/sugarbuilder (accessed April 2020).
57. *KegDraw.* Downloaded from https://www.kegg.jp/kegg/download/kegtools.html (accessed April 2020).
58. Bohne, A.; Lang, E.; von der Lieth, C. W. *Bioinformatics* **1999,** *15,* 767–768. doi:10.1093/bioinformatics/15.9.767
59. Allinger, N. L. *J. Am. Chem. Soc.* **1977,** *99,* 8127–8134. doi:10.1021/ja00467a001
60. Lii, J. H.; Allinger, N. L. *J. Am. Chem. Soc.* **1989,** *111,* 8566–8575. doi:10.1021/ja00205a002
61. Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. *J. Chem. Theory Comput.* **2018,** *14,* 5273–5289. doi:10.1021/acs.jctc.8b00529
62. *Sweet.* http://www.glycosciences.de/modeling/sweet2/doc/index.php (accessed April 2020).
63. *GLYCAM Web*. (2005–2020) Complex Carbohydrate Research Center, University of Georgia, Athens, GA. (http://glycam.org).
64. *CHARMM-GUI Glycan Reader & Modeler*. http://www.charmm-gui.org/?doc=input/glycan (accessed April 2020).
65. Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008,** *29,* 1859–1865. doi:10.1002/jcc.20945
66. Jo, S.; Song, K. C.; Desaire, H.; MacKerell, A. D., Jr.; Im, W. *J. Comput. Chem.* **2011,** *32,* 3135–3141. doi:10.1002/jcc.21886
67. Park, S.-J.; Lee, J.; Qi, Y.; Kern, N. R.; Lee, H. S.; Jo, S.; Joung, I.; Joo, K.; Lee, J.; Im, W. *Glycobiology* **2019,** *29,* 320–331. doi:10.1093/glycob/cwz003
68. Jo, S.; Im, W. *Nucleic Acids Res.* **2013,** *41,* D470–D474. doi:10.1093/nar/gks987
69. Danne, R.; Poojari, C.; Martinez-Seara, H.; Rissanen, S.; Lolicato, F.; Róg, T.; Vattulainen, I. *J. Chem. Inf. Model.* **2017,** *57,* 2401–2406. doi:10.1021/acs.jcim.7b00237
70. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008,** *4,* 435–447. doi:10.1021/ct700301q
71. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005,** *26,* 1701–1718. doi:10.1002/jcc.20291
72. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. *J. Chem. Theory Comput.* **2016,** *12,* 281–296. doi:10.1021/acs.jctc.5b00864
73. *AMBER 2018;* University of California: San Francisco, 2018.
74. Labonte, J. W.; Adolf-Bryfogle, J.; Schief, W. R.; Gray, J. J. *J. Comput. Chem.* **2017,** *38,* 276–287. doi:10.1002/jcc.24679
75. Frenz, B.; Rämisch, S.; Borst, A. J.; Walls, A. C.; Adolf-Bryfogle, J.; Schief, W. R.; Veesler, D.; DiMaio, F. *Structure* **2019,** *27,* 134–139.e3. doi:10.1016/j.str.2018.09.006
76. Perez, S.; Rivet, A., *Methods in Molecular Biology, Glycoinformatics, Methods and Protocols.* 2nd ed.; 2020, (in press).
77. *PolysGlycanBuilder.* http://glycan-builder.cermav.cnrs.fr/ (accessed April 2020).
78. Kuttel, M.; Gain, J.; Burger, A.; Eborn, I. *J. Mol. Graphics Modell.* **2006,** *25,* 380–388. doi:10.1016/j.jmgm.2006.02.007
79. *3D-SNFG.* Downloaded from http://glycam.org/3d-snfg (accessed April 2020).
80. *LiteMol.* https://v.litemol.org/ (accessed April 2020).
81. *SweetUnityMol.* https://sourceforge.net/projects/unitymol/files/OtherVersions/UnityMol-r676-SweetUnityMol/ (accessed April 2020).
82. *UnityMol.* https://sourceforge.net/projects/unitymol/files/ (accessed April 2020).
83. Cross, S.; Kuttel, M. M.; Stone, J. E.; Gain, J. E. *J. Mol. Graphics Modell.* **2009,** *28,* 131–139. doi:10.1016/j.jmgm.2009.04.010
84. Arroyuelo, A.; Vila, J. A.; Martin, O. A. *J. Comput.-Aided Mol. Des.* **2016,** *30,* 619–624. doi:10.1007/s10822-016-9944-x
85. *PyMOL: An open-source molecular graphics tool;* DeLano Scientific, 2002, http://www.pymol.org.
86. Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U. S. A.* **1987,** *84,* 6611–6615. doi:10.1073/pnas.84.19.6611
87. Martinez, X.; Chavent, M.; Baaden, M. *Biochem. Soc. Trans.* **2020,** *48,* 499–506. doi:10.1042/bst20190621

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Leveraging glycomics data in glycoprotein 3D structure validation with Privateer

Haroldas Bagdonas[1], Daniel Ungar[2] and Jon Agirre[*1]

Full Research Paper

Open Access

Address:
[1]York Structural Biology Laboratory, Department of Chemistry, University of York, Wentworth Way, York, YO10 5DD, UK and [2]Department of Biology, University of York, Wentworth Way, York, YO10 5DD, UK

Email:
Jon Agirre[*] - jon.agirre@york.ac.uk

* Corresponding author

Keywords:
electron cryomicroscopy; glycoinformatics; glycomics; Privateer; X-ray crystallography

Beilstein J. Org. Chem. **2020,** *16,* 2523–2533.
https://doi.org/10.3762/bjoc.16.204

Received: 18 July 2020
Accepted: 06 October 2020
Published: 09 October 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: N. H. Packer

Abstract

The heterogeneity, mobility and complexity of glycans in glycoproteins have been, and currently remain, significant challenges in structural biology. These aspects present unique problems to the two most prolific techniques: X-ray crystallography and cryo-electron microscopy. At the same time, advances in mass spectrometry have made it possible to get deeper insights on precisely the information that is most difficult to recover by structure solution methods: the full-length glycan composition, including linkage details for the glycosidic bonds. The developments have given rise to glycomics. Thankfully, several large scale glycomics initiatives have stored results in publicly available databases, some of which can be accessed through API interfaces. In the present work, we will describe how the Privateer carbohydrate structure validation software has been extended to harness results from glycomics projects, and its use to greatly improve the validation of 3D glycoprotein structures.

Introduction

Glycosylation-related processes are prevalent in life. The attachment of carbohydrates to macromolecules extends the capabilities of cells to convey significantly more information than what is available through protein synthesis and the expression of the genetic code alone. For example, glycosylation is used as a switch to modulate protein activity [1]; glycosylation plays a crucial part in folding/unfolding pathways of some proteins in cells [2,3]; the level of *N*-glycan expression regulates the adhesiveness of a cell [4]; glycosylation also plays a role in immune function [5] and cellular signalling [5,6]. At the forefront, glycosylation plays a significant role in influencing protein–protein interactions. For example, the influenza virus uses the haemagglutinin glycoprotein to recognise and bind sialic acid decorations of human cells in the respiratory tract [7]. Glycosylation is also used by pathogens to evade the host's immune system via glycan shields [8-10], and thereby to delay

an immune response [11]. The structural study of these glycan-mediated interactions can provide unique insight into the molecular interplay governing these processes. In addition, it can provide structural snapshots in atomistic detail that can be used to generate molecular dynamics simulations describing a wider picture underpinning glycan and protein interactions [12]. Unfortunately, significant challenges have affected the determination of glycoprotein structures for decades and have had a detrimental impact on the quality and reliability of the produced models. Anomalies have been reported regarding carbohydrate nomenclature [13], glycosidic linkage stereochemistry [14] and torsion [15,16], and most recently, ring conformation [17]. Most of these issues have now been addressed as part of ongoing efforts to provide better software tools for structure determinations of glycoproteins, although the most difficult cases remain hard to solve. Chiefly among these is the scenario where the experimentally resolved electron density map provides evidence of glycosylation, without enough resolution to derive definite and comprehensive details about the structural composition of the oligosaccharides (Figure 1). Glycan microheterogeneity and the lack of carbohydrate-specific modelling tools have often been named as the principal causes for these issues [18].

Heterogeneity of glycoproteins

Unlike protein synthesis, which is encoded in the genome and follows a clear template, glycan biosynthesis is not template-directed. A single glycoprotein will exist in multiple possibilities of products that can emerge from the glycan biosynthesis pathways, and these are known as glycoforms [22]. More specifically, the variation can appear in terms of which potential glycosylation sites are occupied at any time – macroheterogeneity – or variations in the compositions of the glycans added to specific glycosylation sites – microheterogeneity. This variation in the microheterogeneous composition patterns arises due to the competition of glycan-processing enzymes in biosynthesis pathways [23].

Implications for the structure determination of glycoproteins

Several experimental techniques can be used to obtain 3D structures of glycoproteins: X-ray crystallography (MX, which stands for macromolecular crystallography), nuclear magnetic resonance spectroscopy (NMR) and electron cryomicroscopy (cryo-EM). As of publication date, the overwhelming majority of glycoprotein structures have been solved using MX [24,25].

The biggest bottleneck in MX is the formation of crystals of the target macromolecule or complex. The quality of the crystal directly determines the resolution – a measure of the detail in the electron density map. Homogenous samples at high concentrations are required to produce well-diffracting crystals [26]. Samples containing glycoprotein molecules do not usually fulfill this criterion. More often than not, MX falls short at elucidating carbohydrate features in glycoproteins due to glycosylated proteins being inherently mobile and heterogeneous [22]. Moreover, oligosaccharides often significantly interfere with the formation of crystal contacts that allow the formation of well-diffracting crystals. Because of this, glycans are often truncated in MX samples to aid crystal formation [27].

In cryo-EM, samples of glycoproteins are vitrified at extremely low temperatures rather than crystallised, as in MX. The rapid cooling of the sample allows to capture snapshots of the molecules at their various conformational states, and thus potentially maintaining glycoprotein states more closely to their native environments in comparison to crystallography [28]. Nevertheless, cryo-EM is still not an end-all solution to solving glycoprotein structures: the flexible and heterogeneous nature of glycans still has an adverse effect on the quality of the data,

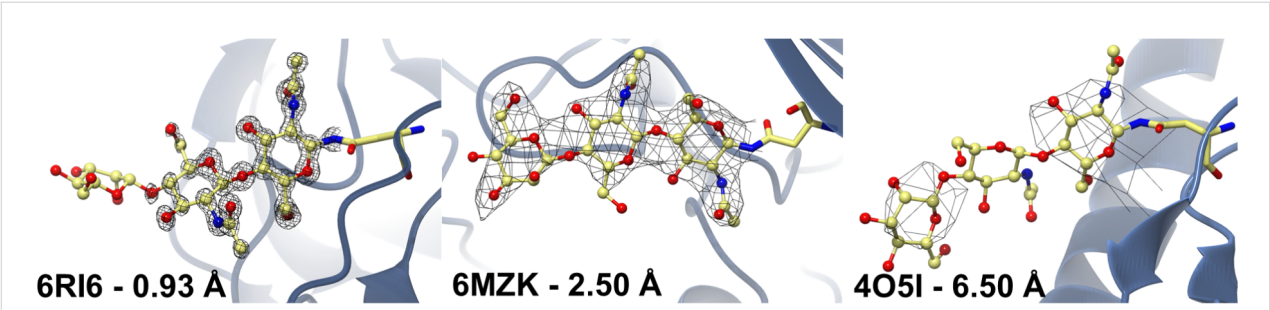


Figure 1: Comparison of the glycan features in electron density maps over a range of resolutions from selected glycoprotein structures (PDB entries: 6RI6 [19]; 6MZK [20]; 4O5I [21]). The electron density maps were obtained with X-ray crystallography. The data resolution and PDB entry IDs associated with the structures have been directly annotated on the structure. Left: A high-resolution example where monosaccharides and the conformations can be elucidated; middle: A medium resolution example where the identification starts to become difficult; right: A low-resolution example for which all prior knowledge must be used. Despite coming from different glycoprotein structures, the glycan has the same composition, and thus is assigned a unique GlyTouCan ID of G15407YE.

affecting the image reconstruction [29]. Moreover, due to the low signal-to-noise ratio, the technique works more easily with samples of a high molecular weight; this situation, however, is evolving rapidly, with reports of sub-100 kDa structures becoming more frequent lately [30,31]. Crucially, MX and cryo-EM can complement each other to counteract issues that both face individually [32].

The two techniques produce different information – electron density (MX) or electron potential (cryo-EM) maps – but the practical considerations in terms of the atomistic interpretation hold true for both: provided that at least the secondary structural features can be resolved in a 3D map, a more or less complete atomic model will be expected as the final result of the study. Modelling of carbohydrates into 3D maps can be more complex than modelling proteins [33], although recent advances in software are closing the gap [34-36]. However, to date it remains true that most model building software is protein-centric [15]. As a consequence, the glycan chains in glycoprotein models that have been elucidated before recent developments in carbohydrate validation and modelling software tend to contain a significant amount of errors: wrong carbohydrate nomenclature [13], biologically implausible glycosidic linkage stereochemistry [14], incorrect torsion [15,16], and unlikely high-energy ring conformations [17]. Early efforts in the validation of carbohydrate structures saw the introduction of online tools such as PDB-CARE [37] and CARP [16]; more recently, we released the Privateer software [24], which was the first carbohydrate validation tool available as part of the CCP4i2 crystallographic structure solution pipeline [38]. In its first release, Privateer was able to perform stereochemical and conformational validation of pyranosides, analyse the glycan fit to electron density map and offered tools for restraining a monosaccharide minimal-energy conformation.

While these features were recognised to address some long-standing needs in carbohydrate structure determination [39,40], significant challenges remain, particularly in the scenario where the glycan composition cannot be ascertained solely from the three-dimensional map. Unfortunately, this problematic situation happens frequently, especially in view of the fact that the median resolution for glycoproteins (2.4 Å) is lower than that of non-glycosylated – potentially including fully deglycosylated – proteins (2.0 Å) [41]. To date, only one publicly available model building tool has attacked this issue: the Coot software offers a module that will build some of the most common *N*-linked glycans in a semiautomated fashion [34]. Indeed, the Coot module was built around the suggestion that only the most probable glycoforms should be modelled unless prior knowledge of an alternative glycan composition exists in the form of, e.g., mass spectrometry data [14].

Harnessing glycomics and glycoproteomics results to inform glycan model building

Current methods used to obtain accurate atomistic descriptions of molecules fall short in dealing with the heterogeneity of glycoproteins. However, there are other methods that have been proven to successfully tackle the challenges posed by glycan heterogeneity, with mass spectrometry emerging as the one with the most relevance due to the ability to elucidate the complete composition descriptions of individual oligosaccharide chains on glycoproteins [42].

The mass spectrometric analysis of glycosylated proteins can be with (glycomics) or without (glycoproteomics) the release of oligosaccharides from the glycoprotein. Usually, glycomics and glycoproteomics experiments are carried out together to obtain a complete description of the glycoprotein profile. Glycomics experiments are required to distinguish stereoisomers and the linkage information in order to obtain a full structural description about a glycan, whereas glycoproteomics are required to establish the glycan variability and occupancy at the glycosylation sites of the protein [43]. Typically, these analyses are based on mass spectrometry techniques, such as electrospray ionization mass spectrometry (ESIMS) and matrix-assisted laser desorption ionization MS (MALDIMS) [43]. Mass spectrometry techniques are best suited for the determination of the composition of monosaccharide classes and the chain length. However, the in-depth analysis of a glycan typically requires the integration of complementary analytic techniques, such as nuclear magnetic resonance (NMR) and capillary electrophoresis (CE). Nevertheless, depending on the sample, advanced mass spectrometry techniques can be used to counteract the need for complementary analytic techniques. One of the examples of this is tandem mass spectrometry, where the glycan fragmentation is controlled to obtain the identification of the glycosylation sites and a complete description of the glycan structure compositions, including linkage and sequence information [44]. Moreover, recent advances in ion mobility mass spectrometry can now also be used for a complete glycan analysis [45].

The analysis and interpretation of mass spectrometry spectra produced by glycans is a challenge. Most significantly, in MS outputs, glycans appear in their generalized composition classes, i.e., Hex, HexNAc, dHex, NeuAc, etc. The identity elucidation of generalized unit classes into specific monosaccharide units (such as Glc, Gal, Man, GalNAc, etc.) requires prior knowledge of the glycan biosynthetic pathways [46]. Additional sources of prior knowledge are bioinformatics databases that have been curated through the deposition of experimental data. Bioinformatics databases contain detailed descriptions of the glycan compositions and

m/*z* values of specific glycans, and therefore aiding the process of glycan annotation [47]. Such bioinformatics databases can usually be interrogated using textual or graphical notations that describe the glycan sequence. However, due to the glycan complexity and the incremental nature of the different glycomics projects, numerous notations have been developed over the years – e.g., CarbBank [48] utilized CCSD [48] and EuroCarbDB [49] and GlycomeDB [50] used GlycoCT [51] (Table 1).

Thankfully, data from discontinued glycomics projects are not lost but were integrated into newer platforms, often with novel notations. One such example is GlyTouCan [53], which uses both GlycoCT [54] and WURCS [53] as notation languages. As a result, tools that interconvert between notations were developed to successfully integrate old data into new platforms. Additionally, the introduction of tools such as GlycanFormatConverter [55] to convert WURCS notations into more human-readable formats has eased the interpretation of glycan databases.

Significantly, the GlyTouCan project aims to create a public repository of known glycan sequences by assigning them unique identification tags. Each identification tag describes a glycan sequence in the WURCS notation, and this allows to link specific glycans to other databases, such as GlyConnect [56], UniCarb-DB [57] and others, any of which are tailored to specific flavours of glycomics and glycoproteomics experiments. Ideally, this implementation ends up requiring the user to be familiar with a single notation – WURCS – used to represent sequences of glycans.

From glycomics/glycoproteomics to carbohydrate 3D model building and validation in Privateer

Many fields, for example pharmaceutical design and engineering [58], molecular dynamics simulations [59] and protein interaction studies [60], rely upon structural biology to produce accurate atomistic descriptions of glycoproteins. However, due to clear limitations of elucidating carbohydrate features in MX/cryo-EM electron-density maps, structural biologists are likely to make mistakes. This introduces the possibility of modelling wrong glycan compositions in glycoprotein models, going as far as not conforming with general glycan biosynthesis knowledge. Model building pipelines would therefore greatly benefit from the ability to validate against the knowledge of glycan compositions elucidated via glycomics/glycoproteomics experiments. This warrants the need for new tools that are able to link these methodologies, through an intermediate interconversion library.

A foundation for such interconversion libraries exists in the form of the carbohydrate validation software Privateer. The program is able to compute individual monosaccharide conformations from a glycoprotein model, check whether the modelled carbohydrates atomistic definitions match dictionary standards as well as output multiple helper tools to aid the processes of refinement and model building [24]. Most importantly, Privateer already contains methods that allow the extraction of carbohydrate atomistic definitions to create abstract definitions of glycans in memory, and thus already laying a foundation for the generation of unique WURCS notations and providing a straightforward access to bioinformatics databases that are integrated in the GlyTouCan project.

Table 1: A comparison of the structural information storage capabilities of different sequence formats used in glycobioinformatics.[a]

notation	multiple connections	repeating units	alternative residues	linear notation	atomic ambiguity
CCSD(CarbBank)	–	+	–	+	–
LINUCS	–	+	–	+	–
GlycoSuite	–	–	+	+	–
BCSDB	(+)	(+)	+	+	–
LinearCode	–	–	+	+	–
KCF	+	+	–	–	–
GlycoCT	+	+	+	–	–
Glyde-II	+	+	–	–	–
WURCS 2.0	+	+	+	+	+

[a]"+" Denotes that information can be stored directly without any significant issues, "(+)" denotes that information can be stored indirectly, or that there are some issues and "–" denotes that information description in the particular sequence format is unavailable. This table is a simplified version of the one originally published by Matsubara et al. [52].

Methods

The algorithm used to generate the WURCS notation in Privateer is based on the description published in Tanaka et al. [61], with required updates applied from Matsubara et al. [52]. WURCS was designed to deal with the incomplete descriptions of glycan sequences emerging from glycomics/glycoproteomics experiments (i.e., undefined linkages, undefined residues and ambiguous structures in general). However, the lack of this detail is unlikely to be supported in "pdb" or "mmCIF'" format files, which are a standard in structural biology. As a result, the "atomic ambiguity" capability (Table 1) is not supported in Privateer's implementation. Moreover, Privateer's implementation of WURCS relies on a manually compiled dictionary that translates the PDB Chemical Component Dictionary [62] three-letter codes of carbohydrate monomer definitions found in the structure files into WURCS definitions of unique monomers (described as "UniqueRES" [52]).

The WURCS notations are generated for all detected glycans that are linked to protein backbones in the input glycoprotein model. For every glycan chain in the model, the algorithm computes a list of all detected monosaccharides that are unique and stores that information internally in memory. Then, the algorithm calculates the unit counts in a glycan chain – how many unique monosaccharides are modelled in the glycan chain, the total length of the glycan chain and computes the total number linkages between monosaccharides. After the composition calculations are carried out, the algorithm begins the generation of the notation by printing out the unit counts. Then, the list of unique monosaccharide definitions in the glycan chain are printed out by converting the three-letter PDB codes into WURCS-compliant definitions. Afterwards, each individual monosaccharide of the glycan is assigned a numerical ID according to its occurrence in the list of unique monosaccharides. Finally, the linkage information between monosaccharide pairs are generated by assigning individual monosaccharides a unique letter ID according to their position in the glycan chain. Alongside a unique letter ID, a numerical term is added that describes a carbon position from which the bond is formed to another carbohydrate unit. Crucially, the linkage detection in Privateer does not rely at all on metadata present in the structure file. Instead, linkages are identified based on the perceived chemistry of the input model: which atoms are close enough – but not too close – to be plausibly linked.

The generated WURCS string can then be used to search whether an individual glycan chain has been deposited in GlyTouCan. The scan of the repository occurs internally within the Privateer software, as all the data is stored in a single structured data file written in JSON format that is distributed together with Privateer. If the existence of a glycan in the database is confirmed, then the software can attempt to find records about the sequence on other, more specialised databases (currently only GlyConnect) to obtain information such as the source organism, the type of glycosylation and the glycan core to carry out further checks in the glycoprotein model (Figure 2).

Availability and performance of the algorithm

This new version of Privateer (MKIV) will be released as an update to CCP4 7.1. To demonstrate the capabilities of the computational bridge integrated in the newest version of Privateer (for standalone bundles, please refer to privateer branch "privateerMKIV_noccp4" of GitHub repository with the installation instructions provided in the README.md file [63]), it was run on all *N*-glycosylated structures in the PDB solved using MX and cryo-EM. The list of structures used in this demonstration was obtained from Atanasova et al. [18]. The computational analysis of the demonstration revealed a relatively small proportion of deposited glycoprotein models containing glycan chains that do not have a unique GlyTouCan accession ID assigned, raising questions about the provenance of their structures. Importantly, the majority of the glycan chains that do have a unique GlyTouCan accession ID assigned (except for single residues linked to protein backbones), have also been successfully matched on the GlyConnect database (Table 2).

Results

Examples of use

As observed in previous studies, glycoprotein models deposited in the PDB feature flaws ranging from minor irregularities to gross modelling errors [14,17,41,64]. The automated validation of minor irregularities was already possible with automated tools such as pdb-care [37], CARP [65], and Privateer [24]. However, the automated detection of gross modelling errors is currently a challenge due to the lack of publicly available tools. Our newly developed computational bridge between structural biology and glycomics databases makes the detection of gross modelling errors easier, as demonstrated by the following examples.

Example 1 – 2H6O

The glycoprotein model (PDB code 2H6O) proposed by Szakonyi et al. [66] contains 12 glycans, as detected by Privateer. The model became infamous after it sparked the submission of a critical correspondence published by Crispin et al. [14]. The article contained a discussion about the proposed model containing glycans that were previously unreported and inconsistent with glycan biosynthetic pathways. In particular, the model contained oligosaccharide chains with Man-(1→3)-GlcNAc and GlcNAc-(1→3)-GlcNAc linkages, β-galactosyl

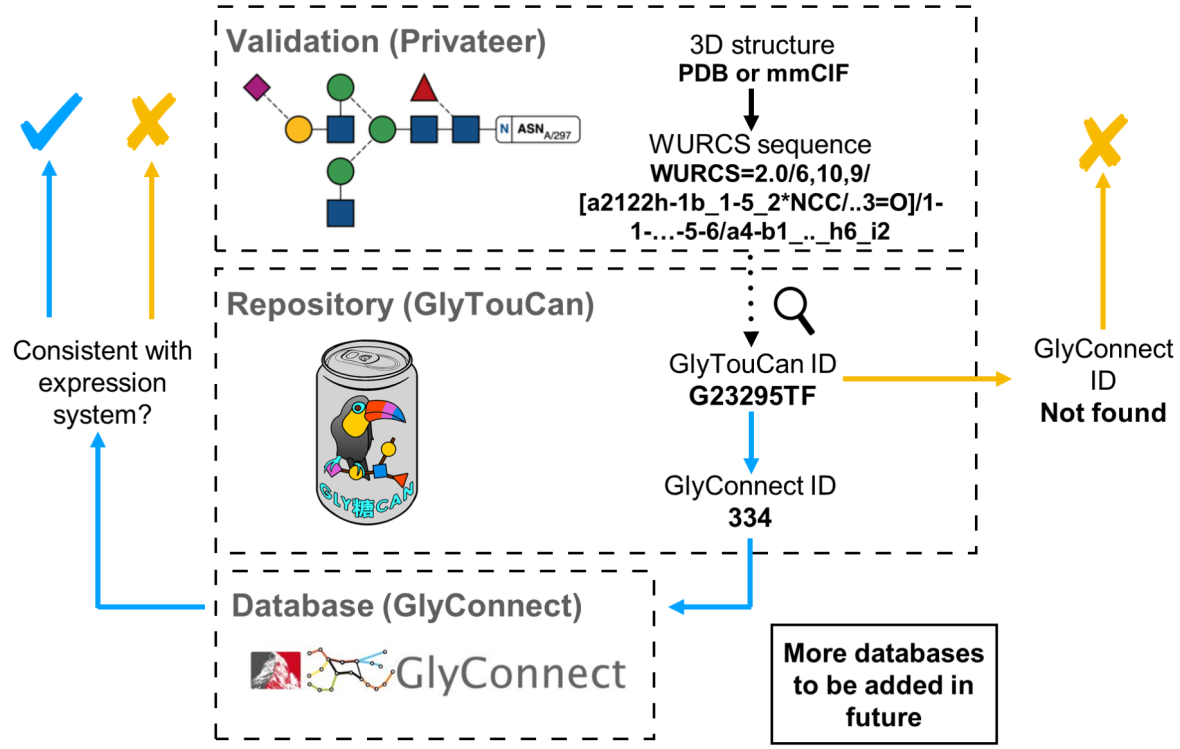


Figure 2: A roadmap of the software development project that allows structural biologists to quickly obtain detailed information about specific glycans in glycoprotein models from glycomics/glycoproteomics databases. The GlyTouCan (https://glytoucan.org/) and GlyConnect (https://glyconnect.expasy.org/) logos have been reproduced here under explicit permission from their respective authors.

Table 2: Comparison of the successful glycan matches detected by Privateer in the GlyTouCan and the GlyConnect database.[a]

experimental technique	glycan chain length	GlyTouCan ID found	GlyTouCan ID not found	% of GlyTouCan in GlyConnect	total glycan chains
MX	1	16797	0	1%	16797
MX	2	5870	5	90%	5875
MX	3	2550	17	71%	2567
MX	4	1012	21	80%	1033
MX	5	834	72	74%	906
MX	6	460	85	69%	545
MX	7	345	55	77%	400
MX	8	235	25	85%	260
MX	9	164	16	81%	180
MX	10	118	5	92%	123
MX	11	20	5	85%	25
MX	12	8	4	75%	12
MX	13	0	1	0%	1
MX	14	0	0	0%	0
MX	15	2	0	0%	2
MX	16	0	1	0%	1
cryo-EM	1	2080	0	3%	2080
cryo-EM	2	1081	0	98%	1081
cryo-EM	3	439	0	96%	439
cryo-EM	4	143	0	93%	143

Table 2: Comparison of the successful glycan matches detected by Privateer in the GlyTouCan and the GlyConnect database.[a] (continued)

cryo-EM	5	146	2	85%	148
cryo-EM	6	70	1	97%	71
cryo-EM	7	45	0	100%	45
cryo-EM	8	26	0	88%	26
cryo-EM	9	15	1	100%	16
cryo-EM	10	16	0	100%	16
cryo-EM	11	4	0	100%	4
cryo-EM	12	1	0	100%	1
cryo-EM	13	1	0	0%	1

[a]Glycans obtained from the glycoprotein models were elucidated by X-ray crystallography and cryo-EM.

motifs capping oligomannose-type glycans and hybrid-type glycans containing terminal Man-(1→3)-GlcNAc [14]. Moreover, the proposed model contained systematic errors in the anomer annotations and carbohydrate stereochemistry. To this day, there is still no experimental evidence reported for these types of linkages and capping in an identical context.

The new version of Privateer was run on the proposed model. WURCS notations were successfully generated for all glycans, with only 1 glycan chain out of 12 successfully returning a GlyTouCan ID. Under further manual review of the one glycan and with help from other validation tools contained in Privateer, it was found to contain anomer mismatch errors (the three letter code denoting one anomeric form did not match the anomeric form reflected in the atomic coordinates). After the anomer mismatch errors were corrected, the oligosaccharide chain also failed to return GlyTouCan and GlyConnect IDs. The other 11 chains that failed to return a GlyTouCan ID also contained flaws, as described previously (Figure 3).

The analysis of this PDB entry highlights the kind of cross-checks that could be done by Protein Data Bank annotators upon validation and deposition of a new glycoprotein entry. It should be recognised that PDB annotators might not necessarily be experts in structural glycobiology. The fact that these glycans could not be matched to standard database entries should be enough to raise the question with depositors, and at the very least write a caveat on a deposited entry where glycans could not be correctly identified. Furthermore, despite the example showing just *N*-glycosylation, other kinds of glycosylation are searchable as well, and therefore this tool could shed much needed light on the validity of models representing more obscure types of modifications.

Example 2 – 2Z62

Successfully matching the WURCS string to a GlyTouCan ID, should not be a sole measure of a structure validity. GlyTouCan is a repository of all potential glycans collected from a set of databases, with the entries often representing glycans. Therefore, the correctness of the composition should be critically validated against the information provided in specialized and high-quality databases such as GlyConnect [56] and UniCarbKB [67]. The computational bridge provides direct search of entries stored in GlyConnect, with plans to expand this to more databases in the near future.

An example where the sole reliance on the detection of a glycan in GlyTouCan would not be sufficient is rebuilding of the 2Z62 glycoprotein structure [68] to improve the model quality [41] (Figure 4). The analysis of the original model generated the GlyTouCan ID G28454KX, which could not be detected in GlyConnect. The automated tools used by PDB-REDO slightly improved the model by renaming one of the fucose residues from FUL to FUC due to an anomer mismatch between the three letter code and the actual coordinates of the monomer. The new model thus generated the GlyTouCan ID G21290RB, which in turn could be matched to the GlyConnect ID 54. Under further manual review of mFo-DFc difference density map, a (1→3)-linked fucose was added, along with additional corrections to the coordinates of the molecule [41]. The newly generated WURCS notation for the model returned a GlyTouCan ID of G63564LA, with a GlyConnect ID of 145. The iterative steps taken to rebuild the glycoprotein model have been portrayed (Figure 4). Because the data in GlyConnect is approximately 70% manually curated by experts in the field [56], a match of a specific glycan in this database is likely a valid confirmation of a specific oligosaccharide composition and linkage pattern found in nature.

Conclusion

The mirrors of GlyConnect and GlyTouCan were obtained thanks to the public access to the API commands, which allowed to create scripts that automated the query of the entries stored in the databases with relative ease. However, the integration of additional databases might require support from the developers of those databases. Support for lipopolysaccharides

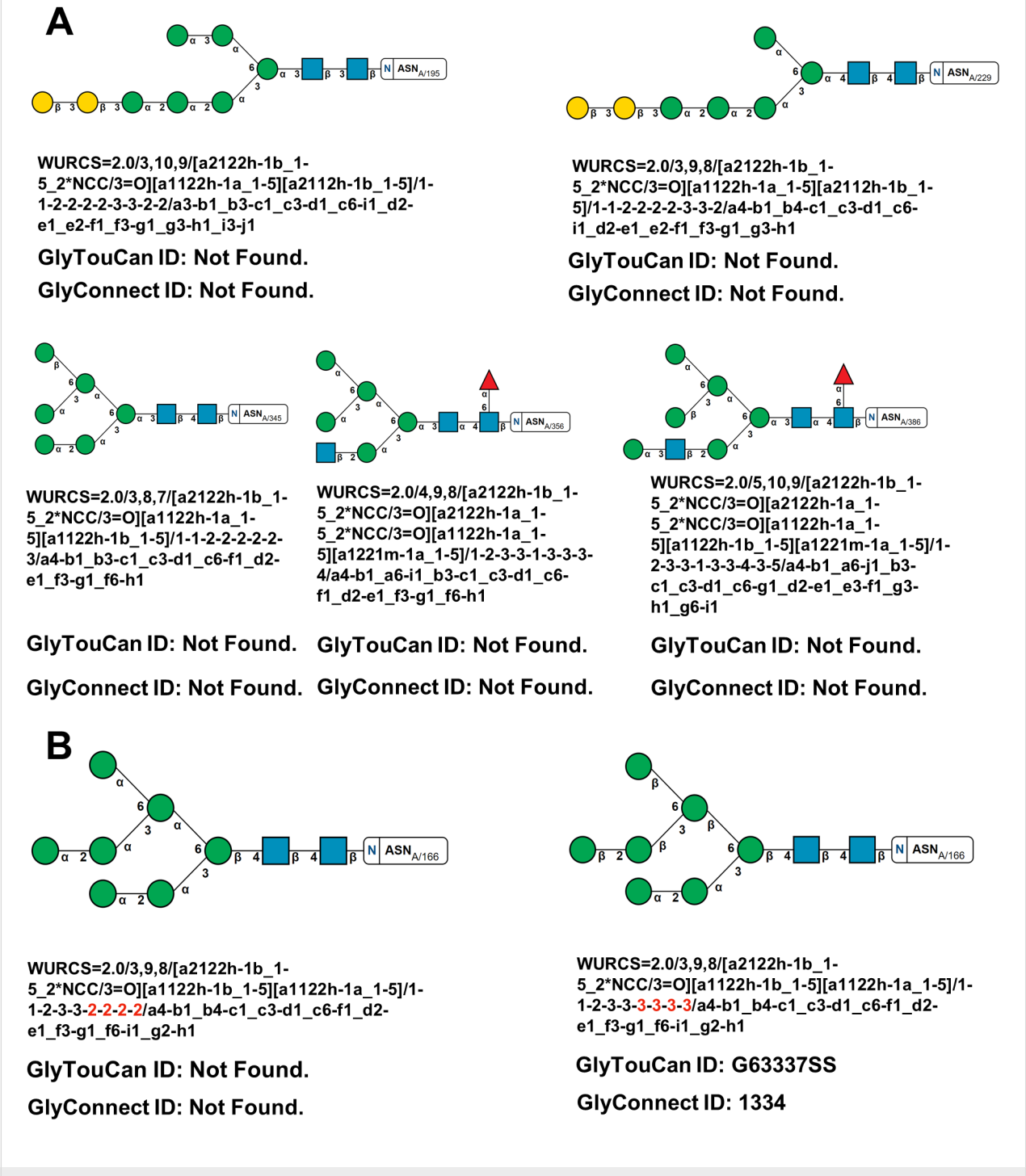


Figure 3: N-Linked glycans in Epstein Barr virus major envelope glycoprotein (PDB entry: 2H6O [66]). A) A selection of the glycan chains that failed to return database IDs with their WURCS sequences extracted from the Privateer CCP4i2 report. B) Glycan chain (right) for which a GlyTouCan and GlyConnect ID have successfully been matched with the modelling errors present in the model. After manual fixing (left), the WURCS sequence for the glycan failed to return database IDs. Highlighting in red depicts the locations in WURCS notation where both glycans differ.

and polysaccharides may be added in future, too, owing to the general purpose of the integrated databases – i.e., they are not limited to protein glycosylation.

Currently, the generated WURCS strings are matched against an identical sequence in the database. This means that if a glycan model has a single modelling mistake, for example, at

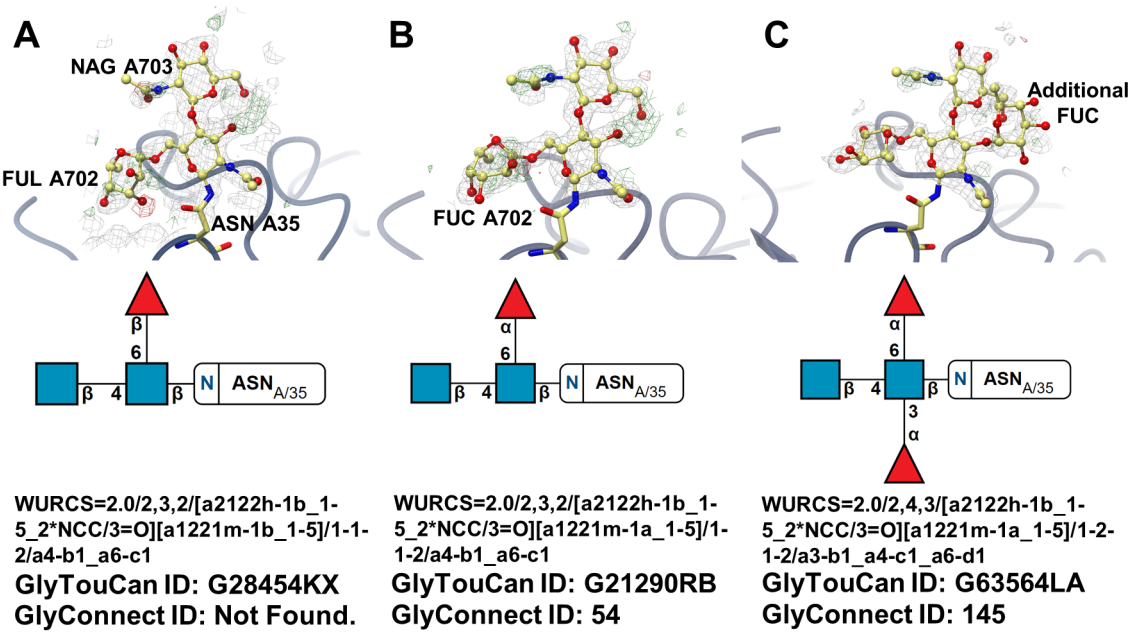


Figure 4: An *N*-linked glycan attached to Asn35 of human Toll-like receptor 4 (A: PDB entry 2z62 [68]). Model iteratively rebuilt by PDB-Redo as shown in steps B and C [41]. Pictures at the top depict glycoprotein models of the region of interest and electron-density maps of the glycan chain (grey: 2mFo DFc map, green and red: mFo DFc difference density map). Pictures at the bottom depict the SNFG representations of glycan chains, their WURCS sequence and accession IDs to relevant databases (taken directly from Privateer's CCP4i2 report).

one end of the chain but is correct elsewhere, the current version of the software would still fail to return a match. This issue has been solved in the development version by the incorporation of a subtree matching algorithm, which will reveal modelling mistakes at specific positions of the glycans, and report these to the user.

Currently, all the developments outlined in this work are accessible exclusively through the Privateer command line interface and through Coot scripts. In order to facilitate the interaction with users, a graphical interface to the new functionality will be provided through the CCP4i2 [38] framework. This new version of the interface is at the testing stage at the time of publication.

Acknowledgements

We would also like to acknowledge the support of the Departments of Chemistry and Biology at the University of York.

Funding

Haroldas Bagdonas is funded by The Royal Society [grant number RGF/R1/181006]. Jon Agirre is the Royal Society Olga Kennard Research Fellow [award number UF160039]. The work in Daniel Ungar's group is supported by the BBSRC [grant number BB/M018237/1].

ORCID® iDs

Haroldas Bagdonas - https://orcid.org/0000-0001-5028-4847
Daniel Ungar - https://orcid.org/0000-0002-9852-6160
Jon Agirre - https://orcid.org/0000-0002-1086-0253

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2020.83.v1

References

1. Rohne, P.; Prochnow, H.; Wolf, S.; Renner, B.; Koch-Brandt, C. *Cell. Physiol. Biochem.* **2014,** *34,* 1626–1639. doi:10.1159/000366365
2. Wyss, D. F.; Choi, J. S.; Li, J.; Knoppers, M. H.; Willis, K. J.; Arulanandam, A. R.; Smolyar, A.; Reinherz, E. L.; Wagner, G. *Science* **1995,** *269,* 1273–1278. doi:10.1126/science.7544493
3. Mitra, N.; Sharon, N.; Surolia, A. *Biochemistry* **2003,** *42,* 12208–12216. doi:10.1021/bi035169e
4. Gu, J.; Isaji, T.; Xu, Q.; Kariya, Y.; Gu, W.; Fukuda, T.; Du, Y. *Glycoconjugate J.* **2012,** *29,* 599–607. doi:10.1007/s10719-012-9386-1
5. Lyons, J. J.; Milner, J. D.; Rosenzweig, S. D. *Front. Pediatr.* **2015,** *3,* 54. doi:10.3389/fped.2015.00054
6. Boscher, C.; Dennis, J. W.; Nabi, I. R. *Curr. Opin. Cell Biol.* **2011,** *23,* 383–392. doi:10.1016/j.ceb.2011.05.001
7. Russell, R. J.; Kerry, P. S.; Stevens, D. J.; Steinhauer, D. A.; Martin, S. R.; Gamblin, S. J.; Skehel, J. J. *Proc. Natl. Acad. Sci. U. S. A.* **2008,** *105,* 17736–17741. doi:10.1073/pnas.0807142105

8. Crispin, M.; Ward, A. B.; Wilson, I. A. *Annu. Rev. Biophys.* **2018,** *47,* 499–523. doi:10.1146/annurev-biophys-060414-034156
9. Watanabe, Y.; Raghwani, J.; Allen, J. D.; Seabright, G. E.; Li, S.; Moser, F.; Huiskonen, J. T.; Strecker, T.; Bowden, T. A.; Crispin, M. *Proc. Natl. Acad. Sci. U. S. A.* **2018,** *115,* 7320–7325. doi:10.1073/pnas.1803990115
10. Pinger, J.; Nešić, D.; Ali, L.; Aresta-Branco, F.; Lilic, M.; Chowdhury, S.; Kim, H.-S.; Verdi, J.; Raper, J.; Ferguson, M. A. J.; Papavasiliou, F. N.; Stebbins, C. E. *Nat. Microbiol.* **2018,** *3,* 932–938. doi:10.1038/s41564-018-0187-6
11. Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. *Cell* **2020,** *181,* 281–292. doi:10.1016/j.cell.2020.02.058
12. Wood, N. T.; Fadda, E.; Davis, R.; Grant, O. C.; Martin, J. C.; Woods, R. J.; Travers, S. A. *PLoS One* **2013,** *8,* e80301. doi:10.1371/journal.pone.0080301
13. Lütteke, T.; von der Lieth, C. W. Data mining the PDB for Glyco-related data. In *Glycomics. Methods in Molecular Biology;* Packer, N. H.; Karlsson, N. G., Eds.; Humana Press: Totowa, NJ, USA, 2009; Vol. 534, pp 293–310. doi:10.1007/978-1-59745-022-5_21
14. Crispin, M.; Stuart, D. I.; Jones, E. Y. *Nat. Struct. Mol. Biol.* **2007,** *14,* 354–355. doi:10.1038/nsmb0507-354a
15. Agirre, J.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. *Curr. Opin. Struct. Biol.* **2017,** *44,* 39–47. doi:10.1016/j.sbi.2016.11.011
16. Frank, M.; Lutteke, T.; von der Lieth, C.-W. *Nucleic Acids Res.* **2007,** *35,* 287–290. doi:10.1093/nar/gkl907
17. Agirre, J.; Davies, G.; Wilson, K.; Cowtan, K. *Nat. Chem. Biol.* **2015,** *11,* 303. doi:10.1038/nchembio.1798
18. Atanasova, M.; Bagdonas, H.; Agirre, J. *Curr. Opin. Struct. Biol.* **2020,** *62,* 70–78. doi:10.1016/j.sbi.2019.12.003
19. Polyakov, K. M.; Gavryushov, S.; Fedorova, T. V.; Glazunova, O. A.; Popov, A. N. *Acta Crystallogr., Sect. D: Struct. Biol.* **2019,** *75,* 804–816. doi:10.1107/s2059798319010684
20. Dai, Y. N.; Fremont, D. H. PDB ID 6MZK; Crystal structure of hemagglutinin from influenza virus A/Pennsylvania/14/2010 (H3N2). https://www.wwpdb.org/pdb?id=pdb_00006mzk (accessed Oct 5, 2020). doi:10.2210/pdb6mzk/pdb
21. Lee, P. S.; Ohshima, N.; Stanfield, R. L.; Yu, W.; Iba, Y.; Okuno, Y.; Kurosawa, Y.; Wilson, I. A. *Nat. Commun.* **2014,** *5,* 3614. doi:10.1038/ncomms4614
22. Rudd, P. M.; Dwek, R. A. *Crit. Rev. Biochem. Mol. Biol.* **1997,** *32,* 1–100. doi:10.3109/10409239709085144
23. Fisher, P.; Thomas-Oates, J.; Wood, A. J.; Ungar, D. *Front. Cell Dev. Biol.* **2019,** *7,* 157. doi:10.3389/fcell.2019.00157
24. Agirre, J.; Iglesias-Fernández, J.; Rovira, C.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. *Nat. Struct. Mol. Biol.* **2015,** *22,* 833–834. doi:10.1038/nsmb.3115
25. Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H. *Essentials of Glycobiology,* 3rd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2016.
26. Geerlof, A.; Brown, J.; Coutard, B.; Egloff, M.-P.; Enguita, F. J.; Fogg, M. J.; Gilbert, R. J. C.; Groves, M. R.; Haouz, A.; Nettleship, J. E.; Nordlund, P.; Owens, R. J.; Ruff, M.; Sainsbury, S.; Svergun, D. I.; Wilmanns, M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006,** *62,* 1125–1136. doi:10.1107/s0907444906030307
27. Stura, E. A.; Nemerow, G. R.; Wilson, I. A. *J. Cryst. Growth* **1992,** *122,* 273–285. doi:10.1016/0022-0248(92)90256-i
28. Cheng, Y.; Grigorieff, N.; Penczek, P. A.; Walz, T. *Cell* **2015,** *161,* 438–449. doi:10.1016/j.cell.2015.03.050
29. Serna, M. *Front. Mol. Biosci.* **2019,** *6,* 33. doi:10.3389/fmolb.2019.00033
30. Fan, X.; Wang, J.; Zhang, X.; Yang, Z.; Zhang, J.-C.; Zhao, L.; Peng, H.-L.; Lei, J.; Wang, H.-W. *Nat. Commun.* **2019,** *10,* 2386. doi:10.1038/s41467-019-10368-w
31. Herzik, M. A., Jr.; Wu, M.; Lander, G. C. *Nat. Commun.* **2019,** *10,* 1032. doi:10.1038/s41467-019-08991-8
32. Wang, H.-W.; Wang, J.-W. *Protein Sci.* **2017,** *26,* 32–39. doi:10.1002/pro.3022
33. Agirre, J. *Acta Crystallogr., Sect. D: Struct. Biol.* **2017,** *73,* 171–186. doi:10.1107/s2059798316016910
34. Emsley, P.; Crispin, M. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018,** *74,* 256–263. doi:10.1107/s2059798318005119
35. Croll, T. I. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018,** *74,* 519–530. doi:10.1107/s2059798318002425
36. Frenz, B.; Rämisch, S.; Borst, A. J.; Walls, A. C.; Adolf-Bryfogle, J.; Schief, W. R.; Veesler, D.; DiMaio, F. *Structure* **2019,** *27,* 134–139. doi:10.1016/j.str.2018.09.006
37. Lütteke, T.; von der Lieth, C.-W. *BMC Bioinf.* **2004,** *5,* 69. doi:10.1186/1471-2105-5-69
38. Potterton, L.; Agirre, J.; Ballard, C.; Cowtan, K.; Dodson, E.; Evans, P. R.; Jenkins, H. T.; Keegan, R.; Krissinel, E.; Stevenson, K.; Lebedev, A.; McNicholas, S. J.; Nicholls, R. A.; Noble, M.; Pannu, N. S.; Roth, C.; Sheldrick, G.; Skubak, P.; Turkenburg, J.; Uski, V.; von Delft, F.; Waterman, D.; Wilson, K.; Winn, M.; Wojdyr, M. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018,** *74,* 68–84. doi:10.1107/s2059798317016035
39. Gristick, H. B.; Wang, H.; Bjorkman, P. J. *Acta Crystallogr., Sect. D: Struct. Biol.* **2017,** *73,* 822–828. doi:10.1107/s2059798317013353
40. Joosten, R. P.; Lütteke, T. *Curr. Opin. Struct. Biol.* **2017,** *44,* 9–17. doi:10.1016/j.sbi.2016.10.010
41. van Beusekom, B.; Lütteke, T.; Joosten, R. P. *Acta Crystallogr., Sect. F: Struct. Biol. Commun.* **2018,** *74,* 463–472. doi:10.1107/s2053230x18004016
42. Nakahara, Y.; Miyata, T.; Hamuro, T.; Funatsu, A.; Miyagi, M.; Tsunasawa, S.; Kato, H. *Biochemistry* **1996,** *35,* 6450–6459. doi:10.1021/bi9524880
43. Shajahan, A.; Heiss, C.; Ishihara, M.; Azadi, P. *Anal. Bioanal. Chem.* **2017,** *409,* 4483–4505. doi:10.1007/s00216-017-0406-7
44. Liu, H.; Zhang, N.; Wan, D.; Cui, M.; Liu, Z.; Liu, S. *Clin. Proteomics* **2014,** *11,* 14. doi:10.1186/1559-0275-11-14
45. Hofmann, J.; Pagel, K. *Angew. Chem., Int. Ed.* **2017,** *56,* 8342–8349. doi:10.1002/anie.201701309
46. Leymarie, N.; Zaia, J. *Anal. Chem. (Washington, DC, U. S.)* **2012,** *84,* 3040–3048. doi:10.1021/ac3000573
47. Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. *J. Proteome Res.* **2008,** *7,* 1650–1659. doi:10.1021/pr7008252
48. Albersheim, P. Technical Report of CarbBank: A structural and bibliographic data base. USA, 1989; https://www.osti.gov/biblio/5715461-m7GJFJ/ (accessed Oct 5, 2020). doi:10.2172/5715461

49. 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.; Frank, M.; Geyer, H.; Geyer, R.; Harrison, M. J.; Henrick, K.; Herget, S.; Hull, W. E.; Ionides, J.; Joshi, H. J.; Kamerling, J. P.; Leeflang, B. R.; Lütteke, T.; Lundborg, M.; Maass, K.; Merry, A.; Ranzinger, R.; Rosen, J.; Royle, L.; Rudd, P. M.; Schloissnig, S.; Stenutz, R.; Vranken, W. F.; Widmalm, G.; Haslam, S. M. *Glycobiology* **2011,** *21,* 493–502. doi:10.1093/glycob/cwq188
50. Ranzinger, R.; Herget, S.; Wetter, T.; von der Lieth, C.-W. *BMC Bioinf.* **2008,** *9,* 384. doi:10.1186/1471-2105-9-384
51. Herget, S.; Ranzinger, R.; Maass, K.; Lieth, C.-W. v. d. *Carbohydr. Res.* **2008,** *343,* 2162–2171. doi:10.1016/j.carres.2008.03.011
52. Matsubara, M.; Aoki-Kinoshita, K. F.; Aoki, N. P.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2017,** *57,* 632–637. doi:10.1021/acs.jcim.6b00650
53. Tiemeyer, M.; Aoki, K.; Paulson, J.; Cummings, R. D.; York, W. S.; Karlsson, N. G.; Lisacek, F.; Packer, N. H.; Campbell, M. P.; Aoki, N. P.; Fujita, A.; Matsubara, M.; Shinmachi, D.; Tsuchiya, S.; Yamada, I.; Pierce, M.; Ranzinger, R.; Narimatsu, H.; Aoki-Kinoshita, K. F. *Glycobiology* **2017,** *27,* 915–919. doi:10.1093/glycob/cwx066
54. Aoki-Kinoshita, K.; Agravat, S.; Aoki, N. P.; Arpinar, S.; Cummings, R. D.; Fujita, A.; Fujita, N.; Hart, G. M.; Haslam, S. M.; Kawasaki, T.; Matsubara, M.; Moreman, K. W.; Okuda, S.; Pierce, M.; Ranzinger, R.; Shikanai, T.; Shinmachi, D.; Solovieva, E.; Suzuki, Y.; Tsuchiya, S.; Yamada, I.; York, W. S.; Zaia, J.; Narimatsu, H. *Nucleic Acids Res.* **2016,** *44,* D1237–D1242. doi:10.1093/nar/gkv1041
55. Tsuchiya, S.; Yamada, I.; Aoki-Kinoshita, K. F. *Bioinformatics* **2019,** *35,* 2434–2440. doi:10.1093/bioinformatics/bty990
56. Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.; Packer, N. H.; Lisacek, F. *J. Proteome Res.* **2019,** *18,* 664–677. doi:10.1021/acs.jproteome.8b00766
57. Hayes, C. A.; Karlsson, N. G.; Struwe, W. B.; Lisacek, F.; Rudd, P. M.; Packer, N. H.; Campbell, M. P. *Bioinformatics* **2011,** *27,* 1343–1344. doi:10.1093/bioinformatics/btr137
58. Congreve, M.; Murray, C. W.; Blundell, T. L. *Drug Discovery Today* **2005,** *10,* 895–907. doi:10.1016/s1359-6446(05)03484-7
59. Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. *Curr. Opin. Struct. Biol.* **2015,** *31,* 64–74. doi:10.1016/j.sbi.2015.03.007
60. Aloy, P.; Russell, R. B. *Proc. Natl. Acad. Sci. U. S. A.* **2002,** *99,* 5896–5901. doi:10.1073/pnas.092147999
61. Tanaka, K.; Aoki-Kinoshita, K. F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2014,** *54,* 1558–1566. doi:10.1021/ci400571e
62. Westbrook, J. D.; Shao, C.; Feng, Z.; Zhuravleva, M.; Velankar, S.; Young, J. *Bioinformatics* **2015,** *31,* 1274–1278. doi:10.1093/bioinformatics/btu789
63. GitHub repository of Privateer. United Kingdom, 2020; https://github.com/glycojones/privateer (accessed Oct 5, 2020).
64. Lütteke, T.; Frank, M.; von der Lieth, C.-W. *Carbohydr. Res.* **2004,** *339,* 1015–1020. doi:10.1016/j.carres.2003.09.038
65. Lütteke, T.; Frank, M.; von der Lieth, C.-W. *Nucleic Acids Res.* **2005,** *33* (Suppl. 1), D242–D246. doi:10.1093/nar/gki013
66. Szakonyi, G.; Klein, M. G.; Hannan, J. P.; Young, K. A.; Ma, R. Z.; Asokan, R.; Holers, V. M.; Chen, X. S. *Nat. Struct. Mol. Biol.* **2006,** *13,* 996–1001. doi:10.1038/nsmb1161
67. Campbell, M. P.; Peterson, R.; Mariethoz, J.; Gasteiger, E.; Akune, Y.; Aoki-Kinoshita, K. F.; Lisacek, F.; Packer, N. H. *Nucleic Acids Res.* **2014,** *42,* D215–D221. doi:10.1093/nar/gkt1128
68. Kim, H. M.; Park, B. S.; Kim, J.-I.; Kim, S. E.; Lee, J.; Oh, S. C.; Enkhbayar, P.; Matsushima, N.; Lee, H.; Yoo, O. J.; Lee, J.-O. *Cell* **2007,** *130,* 906–917. doi:10.1016/j.cell.2007.08.002

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Comparative ligand structural analytics illustrated on variably glycosylated MUC1 antigen–antibody binding

Christopher B. Barnett*[1], Tharindu Senapathi[1] and Kevin J. Naidoo*[1,2]

Full Research Paper

Open Access

Address:
[1]Scientific Computing Research Unit and Department of Chemistry, University of Cape Town, Rondebosch, 7701, South Africa and [2]Infectious Disease and Molecular Medicine, Faculty of Health Science, University of Cape Town, Rondebosch, 7701, South Africa

Email:
Christopher B. Barnett* - Chris.Barnett@uct.ac.za; Kevin J. Naidoo* - Kevin.Naidoo@uct.ac.za

* Corresponding author

Keywords:
binding; conformation; Galaxy; glycoprotein; in silico

Beilstein J. Org. Chem. **2020,** *16,* 2540–2550.
https://doi.org/10.3762/bjoc.16.206

Received: 09 June 2020
Accepted: 30 September 2020
Published: 13 October 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editors: N. H. Packer, F. Lisacek and N. Karlsson

Abstract

When faced with the investigation of the preferential binding of a series of ligands against a known target, the solution is not always evident from single structure analysis. An ensemble of structures generated from computer simulations is valuable; however, visual analysis of the extensive structural data can be overwhelming. Rapid analysis of trajectory data, with tools available in the Galaxy platform, can be used to understand key features and compare differences that inform the preferential ligand structure that favors binding. We illustrate this informatics approach by investigating the in-silico binding of a peptide and glycopeptide epitope of the glycoprotein Mucin 1 (MUC1) binding with the antibody AR20.5. To study the binding, we performed molecular dynamics simulations using OpenMM and then used the Galaxy platform for data analysis. The same analysis tools are applied to each of the simulation trajectories and this process was streamlined by using Galaxy workflows. The conformations of the antigens were analyzed using root-mean-square deviation, end-to-end distance, Ramachandran plots, and hydrogen bonding analysis. Additionally, RMSF and clustering analysis were carried out. These analyses were used to rapidly assess key features of the system, interrogate the dynamic structure of the ligand, and determine the role of glycosylation on the conformational equilibrium. The glycopeptide conformations in solution change relative to the peptide; thus a partially pre-structuring is seen prior to binding. Although the bound conformation of peptide and glycopeptide is similar, the glycopeptide fluctuates less and resides in specific conformers for more extended periods. This structural analysis which gives a high-level view of the features in the system under observation, could be readily applied to other binding problems as part of a general strategy in drug design or mechanistic analysis.

Introduction

A typical sequence of events in research and discovery is noticing a critical biological interaction, searching for structural data, and then searching for the molecular rationale. This is the connection between biology, chemical biology, and chemistry. The Galaxy project is a popular open web-based platform for accessible, reproducible, and transparent computational

research [1]. Originally built to support bioinformatics, Galaxy now supports a much more expansive community including proteomics [2], metabolomics [3], cheminformatics [4], glycoinformatics [5], and chemistry [6]. Of value to these communities are the broad range of tools and ways to connect tools (workflows) in Galaxy that enable diverse, multidisciplinary research. In this paper, we show how an informatics approach provides a high-level overview, thus enabling rapid observations of changes in molecular details pertinent to the system under investigation. We apply this approach to the binding of glycosylated molecules for the well-known system of mucin binding to the AR20.5 murine antibody.

The binding of glycosylated biomolecules is of increasing interest as glycans are found to be involved in cellular functioning and messaging. The mucins, which are cell surface-associated glycoproteins, are found in mucous secretions and are heavily O-glycosylated [7]. Mucins serve several functions: including protecting the body from pathogens by forming chemical barriers and cellular signaling. Mucin 1 (MUC1) is tethered to the cellular membrane and is found to be aberrantly glycosylated and overexpressed in several epithelial cancers [8]. Further, it is thought to participate in the hyperactivation of selected intracellular signal transduction pathways that promote tumorigenicity [9]. MUC1 is a cancer biomarker that can be detected by serum biomarker assays (such as the CA15-3 test [10,11]). The mode of binding between MUC1 and antibodies has received much attention, and the specificity of this interaction is of interest in improving the performance of these biomarker assays [12,13].

The extracellular domain of MUC1 contains a variable number of tandem repeats (VNTR). The VNTR region is comprised of a repeating sequence of 20 amino acids (–His-Gly-Val-**Thr**-**Ser**-Ala-Pro-Asp-**Thr-**Arg-Pro-Ala-Pro-Gly-**Ser**-**Thr**-Ala-Pro-Pro-Ala–)$_n$, and there are five sites where O-glycosylation may occur (indicated in bold). In cancerous cells, the glycans tend to be truncated or have additional sialylation [14]. For example, in mammary epithelial cells, the mixture of O-glycans that glycosylate mucins are *extended core* 2 structures, while in breast cancer cells, O-glycan mass decreases (hypoglycosylation), and there is an increase in abundance of *sialylated core* 1 [15]. The upregulation of Tn (αGalNAc) and STn (αNeuAc-2,6-αGalNAc) antigens are commonly associated with cancerous cells [14].

Movahedin et al. confirmed that the glycosylation of MUC1 influences its binding to the AR20.5 murine antibody [16], specifically the Tn-antigen binds more strongly than the nonglycosylated antigen. AR20.5 is known to bind a specific epitope within the MUC1 VNTR domain. Thus, a synthetic 8-amino acid peptide (APDTRPAP) and the corresponding Tn glycopeptide were synthesized. It was found from the co-crystallization of the AR20.5 antigen-binding fragment (Fab) with the MUC1 peptide and glycopeptide that the glycan moiety of the glycopeptide did not bind to the antibody (Figure 1 and PDB ID:5T6P, 5T78). This is unusual considering that in previous experiments of murine antibody SM3 that Brooks [17] found the glycan forms part of the epitope and binds directly to the antibody. Movahedin et al. hypothesized that the glycan modulates the conformation of the peptide portion of the antigen and does not bind directly.

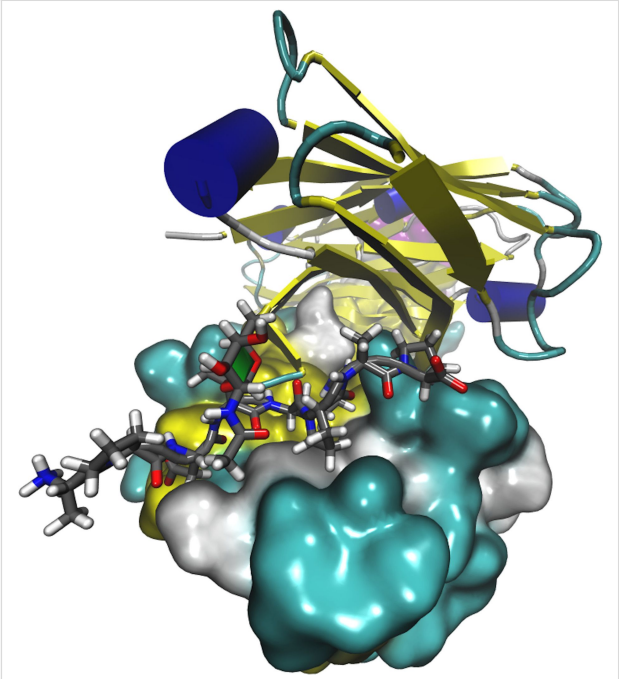


Figure 1: A representation of mucin glycopeptide bound to AR20.5 antibody. Chain A is represented as a molecular surface colored by secondary structure, chain B is represented in cartoon and colored by secondary structure. The mucin peptide is represented as licorice. The Tn glycan (*N*-acetylgalactosamine) is represented as licorice, and the sugar ring is highlighted with the paper chain representation [18,19].

Previous studies have shown that O-glycosylation may provide increased physical stability [20], rigid conformations for protein stability [21], induce the formation of stiff and extended peptide conformations [22], and may affect peptide conformations near the glycosylation site and at distant sites [23]. In glycopeptide enkephalin analogs, the only observed conformational effects due to O-glycosylation were on the residue of attachment and its neighboring residue [24]. While for prion peptides, the O-glycosylation (α-GalNAc) is able to affect the structural transition and suppresses the formation of amyloid fibril formation [25]. The solution structure of O-glycosylated prion peptide was not shifted significantly, with only minor shifts seen in the vicinity of the glycosylation site. Yet there is a

stabilization of the β-structure relative to the random coil and the effects of the glycosylation were hypothesized to relate to the conformational properties of the peptides in solution (as opposed to their equilibrium structures in solution) [25].

A comprehensive structural study of the O-glycosylation-induced changes in a mucin octapeptide showed that the peptide conformation depended on the extent of glycosylation. Glycosylation induces small changes in protein structure and shifts it from a random to a more turn-like structure [26]. Kirnasky et al. noted that O-glycosylation slightly affected the conformational equilibrium of the peptide backbone near the glycosylated residue for a 15-residue mucin peptide. The APDTRP fragment resembled an S-shaped bend and a clustering of low-energy conformations revealed structural similarities between glycosylated and nonglycosylated peptides [23].

The work by Movahedin et al. and others [14,16] provides a foundation for further investigation into the binding of glycopeptide antigens to antibodies using computational modeling. Molecular dynamics (MD) simulations and analysis thereof are a well-known ingredient of the in-silico process for mechanistic screening of glycopeptide fragment binding to antibodies. In this work, the peptide only antigen (Ala-Pro-Asp-Thr-Arg-Pro-Ala-Pro, APDTRPAP) and the Tn glycosylated antigen (APDT(Tn)RPAP) are considered in solution and complex with the AR20.5 antibody. The Tn-antigen is of interest as it is often found upregulated in breast cancer [11,13]. We use MD simulations to investigate the conformational behavior of (glyco)peptide antigens bound to the AR20.5 antibody and to investigate the hypothesis that the glycan modulates the conformation of the peptide portion of the antigen. Primarily showcasing a structural analytics approach, we aim to use the tools and workflows available as part of the Galaxy project to analyze MD simulations to find out if the sugar moiety of the Tn-antigen binds directly to the antibody. Further, if the sugar does not bind directly (as found previously), then we will use these analyses to observe how the sugar modulates binding.

Methods

The inputs, simulation scripts, Galaxy workflows (a series of tools and dataset actions that run in sequence), and data for these simulations are available at https://github.com/chrisbarnettster/bjoc-paper-2020-sm.

Simulation

There is an increasing number of software available to assist with the building up of glycosylated biomolecular systems. As opposed to manual preparation, there are glycan-specific tools and toolkits such as doGlycans [27], Glycosylator [28], and online platforms such as GLYCAM-WEB [29] and CHARMM-GUI [30]. In this work, the CHARMM-GUI server [30] which includes several helper tools (PDB Manipulator [31] and Glycan Reader [32,33]), was used to build these systems and generate input files [34] for use with OpenMM.

Five systems were built in CHARMM-GUI based on initial structures from the Protein Data bank (PDB ID:5T6P, 5T78). The assumption was made that the Tn-antigen binds as per the PDB structure, and other modes of binding are not possible. The solvated receptor, solvated antigens (both the nonglycosylated and Tn-antigen), and a solvated complex (with both antigens) were built in 0.15 M KCl aqueous solution at 310.15 K (physiological temperature). Missing amino acid residues were added. Energy minimization and MD (equilibration and production) simulations were performed using OpenMM [35] and the CHARMM36 force field [36] using the OpenCL platform with mixed precision. Equilibration and production dynamics were carried out as per the scripts provided with CHARMM-GUI, except for adjustments to the time step and number of iterations. The calculations were carried out using Nvidia V100 GPUs.

The equilibration step included 5000 steps of minimization follows by 25000 steps of NVT dynamics (constant volume and temperature) with a time step of 0.001 ps. The particle mesh Ewald (PME) method was used. Nonbonded interactions were cut-off using the force-switching method from 10 Å to 12 Å, and hydrogen bonding constraints applied. During equilibration, the protein backbone and side chains were restrained (force constants of 400.0 and 40.0 kJ mol^{-1} nm^{-2} were used, respectively). The production dynamics were simulated using an *NpT* ensemble and using a time step of 0.002 ps. The antigen–antibody complex in solution was run for 210 ns, while the antigen was run for 500 ns. The antibody was run for 100 ns.

Analysis

The majority of the analyses was carried out using Galaxy, the popular open web-based platform for bioinformatics and computational data analysis, which enables the creation of repeatable analysis pipelines (workflows). There are several well-known molecular dynamics analysis packages (MDAnalysis [37], Bio3D [38] and MDTraj [39]) which are available as computational chemistry analysis tools in Galaxy [6], and these were used to analyze the molecular dynamics trajectories.

The root-mean-square deviation (RMSD) is calculated to measure the stability and conformation of a set of selected atoms. The RMSD is a standard measure of the structural distance between coordinate sets that measures the average distance between a group of atoms [40]. The peptide portion of the antigens was selected for analysis. The root-mean-square-fluc-

tuation (RMSF) represents the deviation at a reference position over time and was calculated in order to measure the variability of the carbon backbone (C-α atoms were selected) of the peptide portion of the antigen (Figure 2).

The end-to-end distance (displacement length) was used as a metric to understand the mobility and conformation of the peptide portion of the antigen throughout the simulation. This is defined as the carbon–nitrogen distance between the first and last amino acid residues of the antigen. A time-series analysis provides some insight, while a histogram provides a clearer understanding of the most populated conformations (Figure 3).

A Ramachandran plot [41] is a well-known method for investigating the φ–ψ (phi–psi dihedral angle) preferences around protein backbones (Figure 4). All φ–ψ angles for the peptide portion of the antigens were measured for each frame of the simulation and aggregated per residue. The glycosidic-linkage dihedral angles of the Tn-antigen (in solution and bound to antibody) were also measured. A standard hydrogen-bonding analysis using MDAnalysis and VMD was carried out with the default angle cut-off and distance cut-off.

A cluster analysis of the peptide portion of the antigen was carried out (Figure 5) using TTClust [42]. The clusters were chosen automatically based on the carbon backbone of the peptide portion of the antigen and clustered using the Ward algorithm.

Results

The antigens were simulated in solution to understand the innate flexibility prior to binding to the antibody, and then also simulated in the complex with AR20.5 MUC1 antibody to understand the effect of glycosylation on antigen conformation during binding. With the rationale that a high-level overview can be used to understand the molecular changes, various analyses were considered: root-mean-square, end-to-end distance, clustering, φ–ψ backbone dihedral angles, and hydrogen-bonding interactions. These analyses focused primarily on the antigen as the antibody conformation does not change significantly in the time frame of the simulation. The peptide-only antigen will be referred to as the 'antigen' while the Tn-glycosylated antigen will be referred to as the 'Tn-antigen'.

Root-mean-square-analysis

In solution (unbound), the RMSD (Figure 2) has a broad spread and a similar center for both the antigen and Tn-antigen. It is readily apparent that the glycosylated antigen has a bimodal distribution (secondary peak at 5.7 Å), indicating at least one other interesting conformation. On consideration of the RMSD for the bound antigens, a narrowing in the distributions is noted. Bound Tn-antigen (Figure 2F) has the narrowest distribution, with a spread from 0.8 Å to 1.6 Å; this unimodal distribution is centered at 1.25 Å. There is no longer a secondary peak, indicating that there is restricted movement on binding. The bound antigen (Figure 2E), instead has a bimodal distribution with a significant population centered at 1.25 Å, a minor population centered at 2.25 Å, and a broad tail that extends to 3.5 Å. While there is restricted movement on binding, the antigen shows unexpected flexibility and a secondary peak at 2.25 Å. From the RMSD, we can infer there is a much tighter range of structures for both antigens when bound than in solution (this should be apparent as there is restricted motion due to the binding of the antigen to the antibody) and the bound Tn-antigen has a more defined and stable conformation.

The RMSFs of the two antigens in solution (Figure 2C and D) have a similar trend with fluctuations ranging between 1.4 Å and 3 Å. Both have large fluctuations, especially for Ala1, Pro6, and Pro8. The Tn-antigen RMSF fluctuates more than the antigen especially for Ala1, Thr4, and Pro8, respectively. When bound, both antigens show restricted fluctuations (Figure 2G and H), with the Tn-antigen showing less fluctuation about each residue. The first and last residues still fluctuate but all RMSF values are less than 1.1 Å indicating relatively minor fluctuations occur for the C-α carbons of the peptide backbone. Another noticeable change is the shift in Pro6, which fluctuated significantly in solution, and now does not. The antigen fluctuates most at the first residue, Ala1, and least at Asp3 and Thr4, while the Tn-antigen fluctuates most for the first and last residues, Ala1 and Pro8, and least for Asp3 and Thr4.

End-to-end analysis

In solution (Figure 3A and B), the displacement lengths of the antigens have a similar range (3.0 Å to 25.0 Å vs. 6.5 Å to 25.0 Å), and both antigens adopt a wide range of conformations with a preference for extended structures. There is a tendency for the Tn-antigen to also prefer a compact conformation, as per the sampling seen at 9.5 Å in the histogram (Figure 3B). The antigen has a left-skewed distribution centered at 19.5 Å, while the Tn-antigen could be bimodal (see the sampling at 9.5 Å and 19.5 Å) or a left-skewed unimodal distribution centered at 19.5 Å.

In contrast, the bound antigens have a much narrower spread (Figure 3C and D). The end-to-end distance for the antigen ranges from 12.5 Å to 22.5 Å, with a distribution centered at 18.9 Å; while the Tn-antigen end-to-end distance ranges from 16.0 Å to 22.0 Å and is centered at 19.5 Å. This is a short peptide so the head and tail regions do fluctuate which could explain the significant spread in the end-to-end distance even

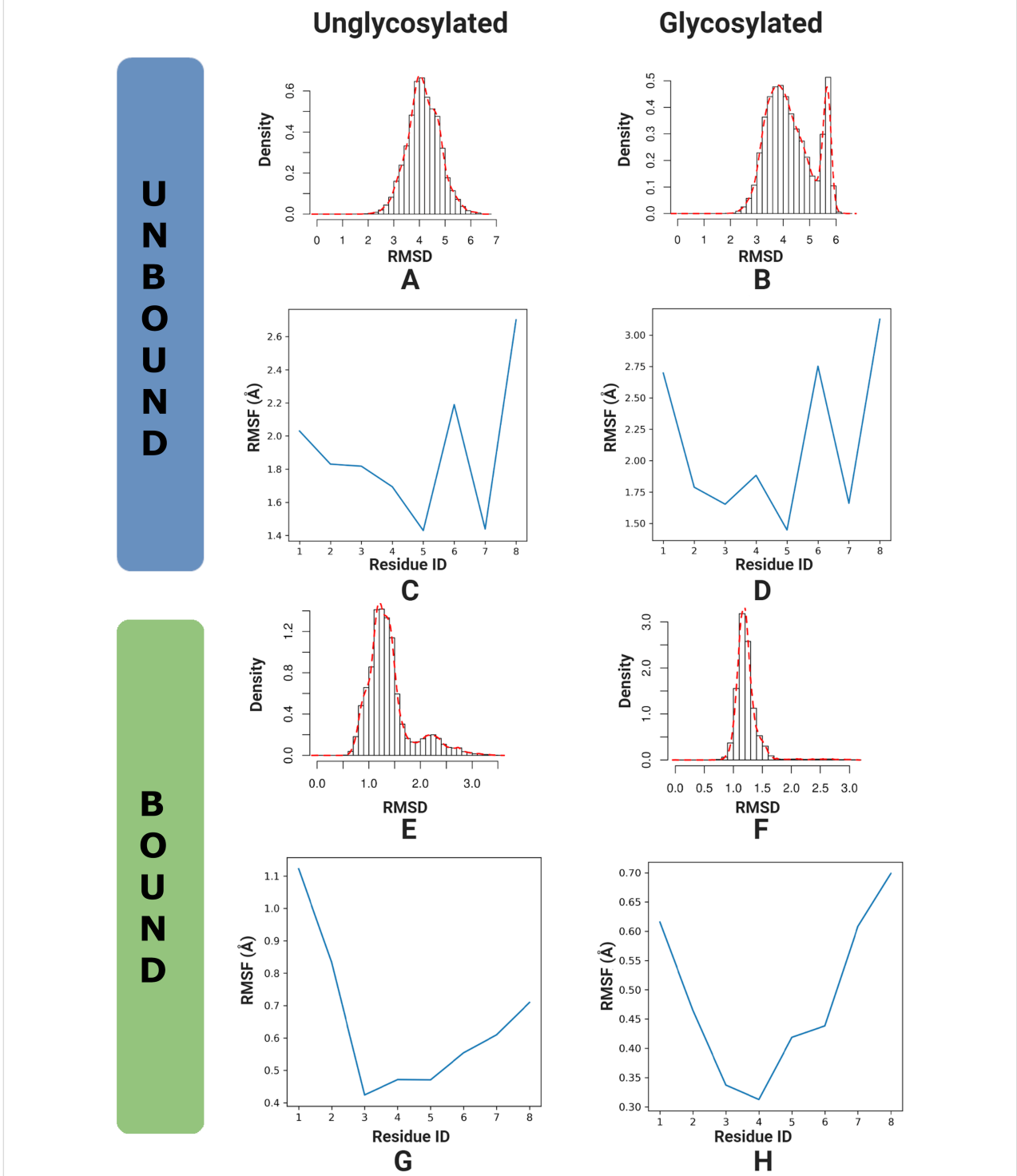


Figure 2: A comparison of root mean analyses for the antigen and Tn-antigen in solution (unbound) and in antibody (bound). RMSD histograms in solution (A, B) and antibody (E, F). RMSF's in solution (C, D) and antibody (G, H). The graphs on the left are for the antigen and those in the right panel are for the Tn-antigen.

though the antigen is bound to the antibody. Nevertheless, the Tn-antigen shows a slightly narrower spread and a more compact ensemble of structures, but otherwise, the end-to-end distance is very similar for both bound antigens.

Ramachandran analysis

The φ–ψ angles of the antigens are considered using a Ramachandran plot. Figure 4 shows a Ramachandran plot for two key amino acids, the glycosylated threonine (Thr4) and

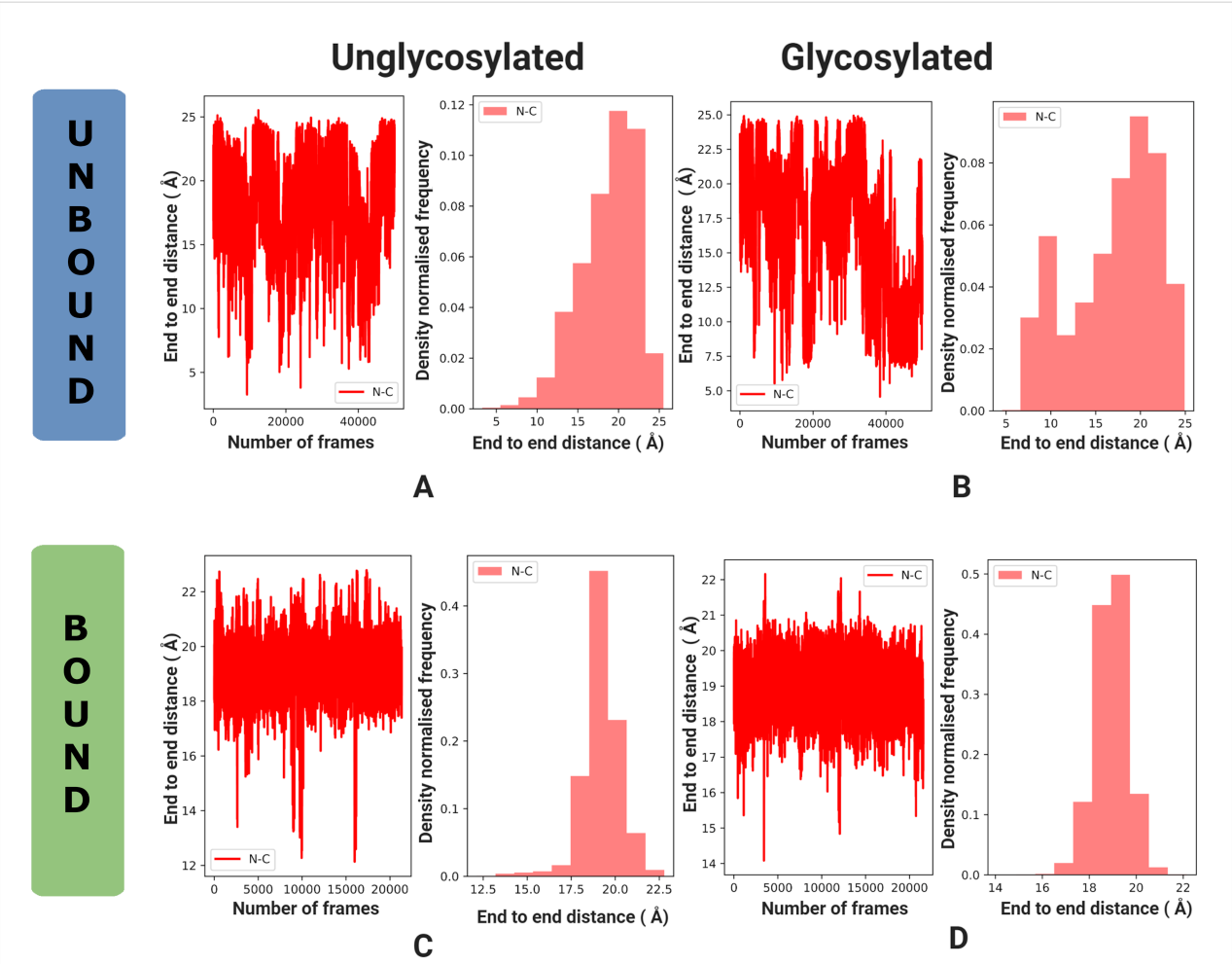


Figure 3: End-to-end time series and histogram for the antigen and Tn-antigen in solution (A, B) and the antibody (C, D). Plots for the antigen are in (A, C) and for the Tn-antigen in (B, D).

neighboring aspartate (Asp3), and considers the φ–ψ angles over all frames of the simulation grouped for these residues. Detailed Ramachandran plots are available (Figures S1 and S2 in Supporting Information File 1) for all residues that can be measured (residues 2–8).

Ramachandran plots show that the φ–ψ distribution for the antigens differs in solution but is the same when bound to the antibody. This is a prestructuring effect and is likely an important contributor to the improved binding affinities seen for the Tn-antigen.

In solution, the third residue (Asp3) prefers (−60°, 135°) for the antigen (Figure 4A and B) with some sampling at (−60°, −40°) and minimal sampling at (60°, 60°). When glycosylated, the ψ sampling shifts to become a balanced bimodal distribution (Figure 4C and D) with similar sampling at (−60°, 135°) and (−60°, −40°), and minimal sampling seen at (60°, 160°) and (60°, −170°). Note that the probability distribution gives the best indication of relevant regions. The fourth residue (Thr4) shows multimodal sampling in φ and a bimodal distribution in ψ, with conformers at (−100°, 0°) and (−60°, 130°) being preferred for the antigen (Figure 4E and F). However, when glycosylated the sampling of Thr is restricted (Figure 4G and H), with a strong preference for (−120°, 120°) and the ψ distribution is effectively unimodal.

The antibody prefers that both antigens adopt a particular shape to fit, and this is seen in the φ–ψ distributions, which shift for Asp3 and Thr4. When bound, both antigens have an almost identical φ–ψ distribution except that the peaks are slightly narrower for the Tn-antigen. In some cases, the preference stays the same and reduced flexibility is observed, for example, Pro2 (Figure S1 and S2 in Supporting Information File 1). In other cases, the conformational preferences shift on binding but this shows no correlation to the effect of glycosylation, for example,

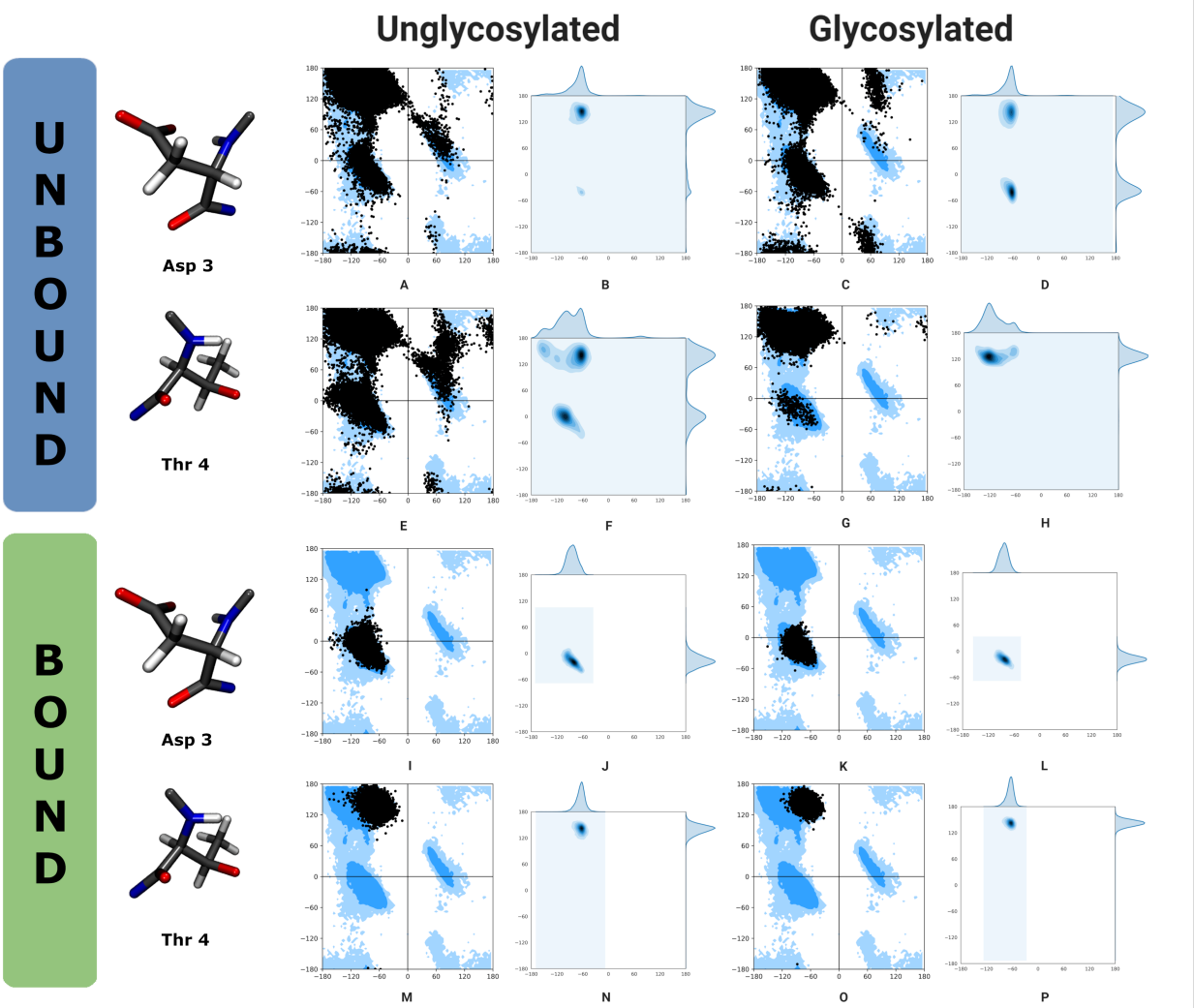


Figure 4: A comparison of Ramachandran analyses for two key amino acids, Asp3 and Thr4. The first row (A–D) illustrates the φ–ψ angles for amino acid 3 of the peptide, aspartate, with a scatter plot showing the allowed φ–ψ regions highlighted in blue (A), and a probability density Ramachandran plot (B) for the antigen, and a scatter plot (C) and probability density Ramachandran plot (D) for the Tn-antigen. While the second row (E–H) illustrates the φ–ψ angles for amino acid 4 of the peptide, threonine. The left panel of Ramachandran plots are for the antigen, and the right panel are for the Tn-antigen. The first two rows are for the antigens in solution (unbound, A–H), while the final two rows are for the antigens bound to the antibody (I–P).

Pro6, Ala7 (Figure S1 and S2 in Supporting Information File 1), and finally, the conformational preference seen for glycosylation in solution aligns with the preference seen for both bound antigens, for example, Asp3 and Thr4 (Figure 4I–P).

For Asp3, the φ–ψ preference for both bound antigens is (−60°, −40°), which correlates with the shift seen on glycosylation in solution where the φ–ψ preference moved from (−60°, 135°) to sample an additional region of phase space and a combination of conformations at (−60°, −40°) and (−60°, 135°). For Thr4, the φ–ψ preference for both bound antigens is (−65°, 140°) which correlates with the shift seen on glycosylation in solution where the φ–ψ preference moved from (−100°, 0°) and (−60°, 130°) to (−120°, 120°). The antibody binds both glycosylated and unglycosylated antigen with the same conformational preference at Asp3 and Thr4 which correlates with the preferred states seen for the glycosylated antigen in solution. There is some evidence of a pre-structuring or pre-organization effect, where O-glycosylation shifts the conformational equilibrium of the peptide towards conformations that are pre-organized for antibody binding.

A Ramachandran plot can be used to understand the role of the sugar moiety, by comparison of the dihedral angle distribution of the glycosidic linkage between the glycan and peptide portion of the Tn-antigen (Figure S3 in Supporting Information File 1). In solution, there is a preference for (70°, 100°) with limited sampling observed in the negative regions of the ψ dis-

tribution. On binding, this preference is limited and changes slightly to (70°, 120°) with no sampling observed in the negative regions of the φ distribution.

Cluster analysis

A cluster analysis of the solution structures yields 5 clusters for the antigen and 4 clusters for the Tn-antigen (Figure 5A and B). The predominant conformer in both antigens is the extended form (Figure 5C), while for the Tn-antigen, the fourth cluster exhibits a more compact conformation (a transparent green conformer in Figure 5C) as noted in previous analysis.

A cluster analysis of the bound antigens yields 5 clusters for the antigen and 4 clusters for the Tn-antigen (Figure 5D and E). The predominant conformer in both antigens is similar (Figure 5F), as noted in previous analysis. For the antigen, the

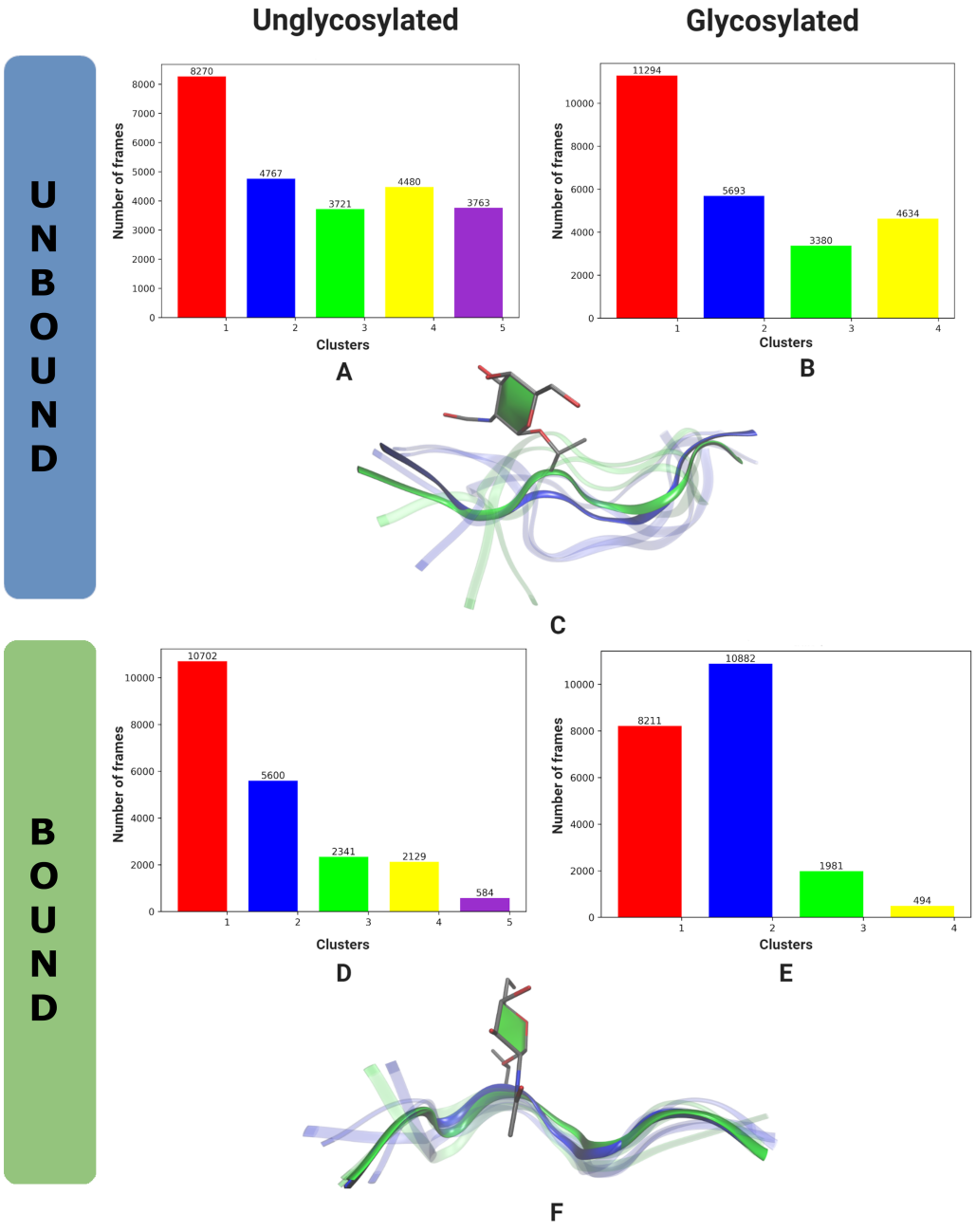


Figure 5: Distribution of clusters, found using TTClust, for the antigen and Tn-antigen in solution (A, B) and when bound to the antibody (D, E). The conformation of the clusters for solution (C) and bound (F) where the antigen is drawn as a blue ribbon with first cluster conformation in opaque blue. The Tn-antigen is drawn as a green ribbon with first cluster conformation in opaque green. The sugar is drawn without hydrogens in licorice and the sugar ring is highlighted with the paper chain representation [18,19].

first cluster dominates (43%) with the second cluster about half as many members (22%), and the third cluster accounting for 9% of all conformations analyzed. For the Tn-antigen, the first and second clusters dominate accounting for respectively 33% and 44% of all conformations analyzed.

The cluster analysis indicates key conformations of the antigens seen in solution and when bound. In solution, the Tn-antigen can adopt a compact conformation while both antigens adopt extended structures when bound to the antibody. When considering the population count (Figure 5D and E) and residence time of the clusters (Figures S4 and S5 in Supporting Information File 1), the bound Tn-antigen is able to stay resident in the dominant conformation without regularly flipping to other conformations.

Hydrogen bonding

The specifics of intermolecular interactions can also be considered, and here we utilized a hydrogen-bonding analysis to consider how the sugar moiety could interact with the antibody (Tables S1–S7 in Supporting Information File 1).

In solution, hydrogen bonds occur within the antigen between Arg5–Asp3 and Arg5–Pro8 (in order donor–acceptor) with occupancies of 31.83% and 14.32% (and 13.67%). For the Tn-antigen, the peptide portion has hydrogen bonds between Arg5–Pro8 (26.69% and 26.58%), Arg5–Asp3 (12.45%), an Arg5–Pro2 interaction is observed with an occupancy of 7.13%, and an intramolecular hydrogen bond between the C3 alcohol and the carbonyl of the *N*-acetyl moiety of the GalNAc has an occupancy of 6.92%. A shift in hydrogen-bonding populations on glycosylation and the appearance of the Arg5–Pro2 (7.13%) interaction aligns with the compact structure noted previously for the Tn-antigen.

When bound, additional intramolecular hydrogen bonds are observed for the Tn-antigen with interactions between the GalNAc–Thr4 (NH of the acetyl group to carbonyl group) and GalNAc–GalNAc (NH of the acetyl group and the C3 alcohol with the carbonyl of the *N*-acetyl moiety), which occur with occupancies of 23.04% and 29.08%, respectively. These two hydrogen bonds may play a crucial role in maintaining the conformation of the Tn-antigen. There are no intramolecular hydrogen bonds between the peptide moiety of the antigens; these are replaced by hydrogen-bonding between the antigen and chain A of the antibody. The following hydrogen bonds occur between the antigen and antibody: Arg5–Glu39 (141.21%, above 100% as counting both acceptor sites on Arg), Lys58–Asp3 (44.44%), Tyr37–Pro2 (42.55%), Arg55–Asp3 (38.11%), and Tyr54–Asp3 (28.51%). The following hydrogen bonds occur between the Tn-antigen and chain A of the antibody: Arg5–Glu39 (137.49%, above 100% as counting both acceptor sites on Arg), Lys58–Asp3 (42.80%), Tyr37–Pro2 (46.73%), Arg55–Asp3 (37.77%), and Tyr54–Asp3 (31.44%). A hydrogen bond (0.15%) was observed between the hydroxy group of Tyr100 of chain B of the antibody and the 6-hydroxy group of the GalNAc. While seemingly short-lived, it occurs with some frequency throughout the simulation (see Figure S6 in Supporting Information File 1). Movahedin et al. hypothesized that the glycan modulates the conformation of the peptide portion of the Tn-antigen and does not bind directly, noting that in the crystal structure GalNAc is positioned 4 Å away from the side chain of Tyr100, and indicating that any dispersion interactions would be insufficient to explain a 20-fold increase in affinity. It is unlikely that this hydrogen bond explains a 20-fold increase in affinity yet note that the mobility of the glycan moiety allows the hydrogen-bond interaction to occur. The hydrogen-bonding preferences and occupancies between the antigens and the antibody are very similar.

Discussion

RMSD, RMSF, end-to-end distance, and Ramachandran analyses support that the Tn-antigen has slightly less conformational play than the nonglycosylated antigen when bound to the antibody. The analysis of the φ–ψ preference showed that the antibody binds both glycosylated and unglycosylated antigen with the same conformational preference (at Asp3 and Thr4) as that of the glycosylated antigen in solution. There is some evidence of a prestructuring or preorganization effect, where O-glycosylation shifts the conformational equilibrium of the peptide towards conformations that are preorganized for antibody binding. This should decrease the overall entropic penalty upon binding, and therefore would explain an increased binding affinity for the glycosylated antigen.

A cluster analysis showed that the dominant conformation for the bound antigens are similar. Intramolecular hydrogen-bonding interactions within GalNAc were more dominant in the antibody (have a higher occupation) than in solution. An intramolecular hydrogen bond within the Tn-antigen between the GalNAc–Thr4 (NH of the acetyl group to carbonyl group) may be responsible for maintaining the conformation of the Tn-antigen. The role of the sugar in excluding water was not investigated. A short-lived intermolecular hydrogen bond (0.15%) was observed between Tyr100 and GalNAc, and this is unlikely to be significant. These results correlated with the hypothesis put forward previously that glycosylation alters the conformational equilibrium of the antigen.

Conclusion

We have shown how an informatics approach can be used to rapidly obtain key indicators of structural features for under-

standing the molecular level behavior of a system. We illustrated this informatics approach for the binding of glycosylated molecules, in particular for variably glycosylated mucin in solution and when bound to an antibody. RMSD, end-to-end distance, Ramachandran analysis, and hydrogen-bonding analyses were carried out using the Galaxy platform. Additionally, RMSF and cluster analysis were carried out. These analyses were used to gain rapid insight into the behavior of the system. The solution conformations of the Tn-antigen and the antigen were generally extended, yet the Tn-antigen was found to sample a more compact conformation. When bound to the antibody, both antigens had considerably less freedom than when in solution, as expected, and the Tn-antigen had less conformational play. However, this was not the result of hydrogen-bonding interactions between the glycan and the antibody or significantly different interactions between the peptide portion of the Tn-antigen and the antibody. Instead, contributing factors included an intramolecular hydrogen-bonding interaction between GalNAc and Thr4, and a preorganization effect (seen from Ramachandran analysis), where O-glycosylation shifted the conformational equilibrium of the peptide towards conformations that are preorganized for antibody binding. The results agreed with previous findings that glycosylation may affect peptide conformations near the glycosylation site and correlated with the hypothesis that glycosylation alters the conformational equilibrium of the antigen. This structural analysis which gives a high-level view of the features in the system under observation, could be readily applied to other binding problems as part of a general strategy in drug design or mechanistic analysis.

Supporting Information

Supporting Information File 1

Additional molecular dynamics analyses.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-206-S1.pdf]

Acknowledgements

We would like to acknowledge the Galaxy community, the Galaxy Europe team, and the Galaxy computational chemistry team on GitHub. We thank the University of Cape Town eResearch (for support and use of the ilifu data centre) and the Centre for High Performance Computing (for the use of their GPU cluster).

Funding

We thank the University of Cape Town Research Committee and the National Research Foundation of South Africa (Grant Numbers 115215 and 116362) for funding.

ORCID® iDs

Christopher B. Barnett - https://orcid.org/0000-0002-1467-5741
Tharindu Senapathi - https://orcid.org/0000-0002-3277-4022
Kevin J. Naidoo - https://orcid.org/0000-0002-9898-3708

References

1. Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B. A.; Guerler, A.; Hillman-Jackson, J.; Hiltemann, S.; Jalili, V.; Rasche, H.; Soranzo, N.; Goecks, J.; Taylor, J.; Nekrutenko, A.; Blankenberg, D. *Nucleic Acids Res.* **2018,** *46,* W537–W544. doi:10.1093/nar/gky379
2. Stewart, P. A.; Kuenzi, B. M.; Mehta, S.; Kumar, P.; Johnson, J. E.; Jagtap, P.; Griffin, T. J.; Haura, E. B. The Galaxy Platform for Reproducible Affinity Proteomic Mass Spectrometry Data Analysis. In *Mass Spectrometry of Proteins: Methods and Protocols;* Evans, C. A.; Wright, P. C.; Noirel, J., Eds.; Methods in Molecular Biology; Springer: New York, NY, USA, 2019; pp 249–261. doi:10.1007/978-1-4939-9232-4_16
3. Davidson, R. L.; Weber, R. J. M.; Liu, H.; Sharma-Oates, A.; Viant, M. R. *GigaScience* **2016,** *5,* s13742-016-0115-8. doi:10.1186/s13742-016-0115-8
4. Bray, S. A.; Lucas, X.; Kumar, A.; Grüning, B. A. *J. Cheminf.* **2020,** *12,* 40. doi:10.1186/s13321-020-00442-7
5. Barnett, C. B.; Aoki-Kinoshita, K. F.; Naidoo, K. J. *Bioinformatics* **2016,** *32,* 3005–3011. doi:10.1093/bioinformatics/btw341
6. Senapathi, T.; Bray, S.; Barnett, C. B.; Grüning, B.; Naidoo, K. J. *Bioinformatics* **2019,** *35,* 3508–3509. doi:10.1093/bioinformatics/btz107
7. Brockhausen, I.; Schachter, H.; Stanley, P. O-GalNAc Glycans. In *Essentials of Glycobiology;* Varki, A.; Cummings, R. D.; Esko, J. D., Eds.; Cold Spring Harbor Laboratory Press: New York, NY, USA, 2009.
8. Kufe, D. W. *Oncogene* **2013,** *32,* 1073–1081. doi:10.1038/onc.2012.158
9. Nath, S.; Mukherjee, P. *Trends Mol. Med.* **2014,** *20,* 332–342. doi:10.1016/j.molmed.2014.02.007
10. Al-azawi, D.; Kelly, G.; Myers, E.; McDermott, E. W.; Hill, A. D. K.; Duffy, M. J.; Higgins, N. O. *BMC Cancer* **2006,** *6,* 220. doi:10.1186/1471-2407-6-220
11. Williams, K. A.; Terry, K. L.; Tworoger, S. S.; Vitonis, A. F.; Titus, L. J.; Cramer, D. W. *PLoS One* **2014,** *9,* e88334. doi:10.1371/journal.pone.0088334
12. Ricardo, S.; Marcos-Silva, L.; Pereira, D.; Pinto, R.; Almeida, R.; Söderberg, O.; Mandel, U.; Clausen, H.; Felix, A.; Lunet, N.; David, L. *Mol. Oncol.* **2015,** *9,* 503–512. doi:10.1016/j.molonc.2014.10.005
13. Teramoto, K.; Ozaki, Y.; Hanaoka, J.; Sawai, S.; Tezuka, N.; Fujino, S.; Daigo, Y.; Kontani, K. *Ther. Adv. Med. Oncol.* **2017,** *9,* 147–157. doi:10.1177/1758834016678375
14. Song, W.; Delyria, E. S.; Chen, J.; Huang, W.; Lee, J. S.; Mittendorf, E. A.; Ibrahim, N.; Radvanyi, L. G.; Li, Y.; Lu, H.; Xu, H.; Shi, Y.; Wang, L.-X.; Ross, J. A.; Rodrigues, S. P.; Almeida, I. C.; Yang, X.; Qu, J.; Schocker, N. S.; Michael, K.; Zhou, D. *Int. J. Oncol.* **2012,** *41,* 1977–1984. doi:10.3892/ijo.2012.1645
15. Brockhausen, I. *EMBO Rep.* **2006,** *7,* 599–604. doi:10.1038/sj.embor.7400705
16. Movahedin, M.; Brooks, T. M.; Supekar, N. T.; Gokanapudi, N.; Boons, G.-J.; Brooks, C. L. *Glycobiology* **2017,** *27,* 677–687. doi:10.1093/glycob/cww131

17. Brooks, C. L.; Schietinger, A.; Borisova, S. N.; Kufer, P.; Okon, M.; Hirama, T.; MacKenzie, C. R.; Wang, L.-X.; Schreiber, H.; Evans, S. V. *Proc. Natl. Acad. Sci. U. S. A.* **2010,** *107,* 10056–10061. doi:10.1073/pnas.0915176107
18. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996,** *14,* 33–38. doi:10.1016/0263-7855(96)00018-5
19. Cross, S.; Kuttel, M. M.; Stone, J. E.; Gain, J. E. *J. Mol. Graphics Modell.* **2009,** *28,* 131–139. doi:10.1016/j.jmgm.2009.04.010
20. Chaffey, P. K.; Guan, X.; Chen, C.; Ruan, Y.; Wang, X.; Tran, A. H.; Koelsch, T. N.; Cui, Q.; Feng, Y.; Tan, Z. *Biochemistry* **2017,** *56,* 2897–2906. doi:10.1021/acs.biochem.7b00195
21. Steen, P. V. d.; Rudd, P. M.; Dwek, R. A.; Opdenakker, G. *Crit. Rev. Biochem. Mol. Biol.* **1998,** *33,* 151–208. doi:10.1080/10409239891204198
22. Jentoft, N. *Trends Biochem. Sci.* **1990,** *15,* 291–294. doi:10.1016/0968-0004(90)90014-3
23. Kirnarsky, L.; Prakash, O.; Vogen, S. M.; Nomoto, M.; Hollingsworth, M. A.; Sherman, S. *Biochemistry* **2000,** *39,* 12076–12082. doi:10.1021/bi0010120
24. Kriss, C. T.; Lou, B.-S.; Szabò, L. Z.; Mitchell, S. A.; Hruby, V. J.; Polt, R. *Tetrahedron: Asymmetry* **2000,** *11,* 9–25. doi:10.1016/s0957-4166(99)00544-3
25. Chen, P.-Y.; Lin, C.-C.; Chang, Y.-T.; Lin, S.-C.; Chan, S. I. *Proc. Natl. Acad. Sci. U. S. A.* **2002,** *99,* 12633–12638. doi:10.1073/pnas.192137799
26. Wu, W.-g.; Pasternack, L.; Huang, D.-H.; Koeller, K. M.; Lin, C.-C.; Seitz, O.; Wong, C.-H. *J. Am. Chem. Soc.* **1999,** *121,* 2409–2417. doi:10.1021/ja983474v
27. Danne, R.; Poojari, C.; Martinez-Seara, H.; Rissanen, S.; Lolicato, F.; Róg, T.; Vattulainen, I. *J. Chem. Inf. Model.* **2017,** *57,* 2401–2406. doi:10.1021/acs.jcim.7b00237
28. Lemmin, T.; Soto, C. *BMC Bioinf.* **2019,** *20,* 513. doi:10.1186/s12859-019-3097-6
29. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008,** *29,* 622–655. doi:10.1002/jcc.20820
30. Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008,** *29,* 1859–1865. doi:10.1002/jcc.20945
31. Jo, S.; Cheng, X.,; Islam, S. M.; Huang, L.; Rui, H.; Zhu, A.; Lee, H. S.; Qi, Y.; Han, W.; Vanommeslaeghe, K.; MacKerell, A. D., Jr.; Roux, B.; Im, W. CHARMM-GUI PDB Manipulator for Advanced Modeling and Simulations of Proteins Containing Nonstandard Residues. In *Advances in Protein Chemistry and Structural Biology;* Karabencheva-Christova, T., Ed.; Biomolecular Modelling and Simulations, Vol. 96; Academic Press, 2014; pp 235–265. doi:10.1016/bs.apcsb.2014.06.002
32. Park, S.-J.; Lee, J.; Qi, Y.; Kern, N. R.; Lee, H. S.; Jo, S.; Joung, I.; Joo, K.; Lee, J.; Im, W. *Glycobiology* **2019,** *29,* 320–331. doi:10.1093/glycob/cwz003
33. Park, S.-J.; Lee, J.; Patel, D. S.; Ma, H.; Lee, H. S.; Jo, S.; Im, W. *Bioinformatics* **2017,** *33,* 3051–3057. doi:10.1093/bioinformatics/btx358
34. Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; Jo, S.; Pande, V. S.; Case, D. A.; Brooks, C. L., III; MacKerell, A. D., Jr.; Klauda, J. B.; Im, W. *J. Chem. Theory Comput.* **2016,** *12,* 405–413. doi:10.1021/acs.jctc.5b00935
35. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. *PLoS Comput. Biol.* **2017,** *13,* e1005659. doi:10.1371/journal.pcbi.1005659
36. Huang, J.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2013,** *34,* 2135–2145. doi:10.1002/jcc.23354
37. Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. *J. Comput. Chem.* **2011,** *32,* 2319–2327. doi:10.1002/jcc.21787
38. Skjærven, L.; Yao, X.-Q.; Scarabelli, G.; Grant, B. J. *BMC Bioinf.* **2014,** *15,* 399. doi:10.1186/s12859-014-0399-6
39. McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. *Biophys. J.* **2015,** *109,* 1528–1532. doi:10.1016/j.bpj.2015.08.015
40. Bray, S. A.; Senapathi, T.; Barnett, C. B.; Grüning, B. A. *bioRxiv* **2020,** 2020.05.08.084780. doi:10.1101/2020.05.08.084780
41. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *J. Mol. Biol.* **1963,** *7,* 95–99. doi:10.1016/s0022-2836(63)80023-6
42. Tubiana, T.; Carvaillo, J.-C.; Boulard, Y.; Bressanelli, S. *J. Chem. Inf. Model.* **2018,** *58,* 2178–2182. doi:10.1021/acs.jcim.8b00512

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

A consensus-based and readable extension of *Li*near *Co*de for *R*eaction *R*ules (LiCoRR)

Benjamin P. Kellman‡, Yujie Zhang‡, Emma Logomasini, Eric Meinhardt, Karla P. Godinez-Macias, Austin W. T. Chiang, James T. Sorrentino, Chenguang Liang, Bokan Bao, Yusen Zhou, Sachiko Akase, Isami Sogabe, Thukaa Kouka, Elizabeth A. Winzeler, Iain B. H. Wilson, Matthew P. Campbell, Sriram Neelamegham, Frederick J. Krambeck, Kiyoko F. Aoki-Kinoshita and Nathan E. Lewis*

Commentary

Open Access

Address:
See end of main text.

Email:
Nathan E. Lewis* - nlewisres@ucsd.edu

* Corresponding author ‡ Equal contributors

Keywords:
glycoinformatics; linear code; systems glycobiology

Beilstein J. Org. Chem. **2020,** *16,* 2645–2662.
https://doi.org/10.3762/bjoc.16.215

Received: 01 June 2020
Accepted: 17 September 2020
Published: 27 October 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: F. Lisacek

Abstract

Systems glycobiology aims to provide models and analysis tools that account for the biosynthesis, regulation, and interactions with glycoconjugates. To facilitate these methods, there is a need for a clear glycan representation accessible to both computers and humans. Linear Code, a linearized and readily parsable glycan structure representation, is such a language. For this reason, Linear Code was adapted to represent reaction rules, but the syntax has drifted from its original description to accommodate new and originally unforeseen challenges. Here, we delineate the consensuses and inconsistencies that have arisen through this adaptation. We recommend options for a consensus-based extension of Linear Code that can be used for reaction rule specification going forward. Through this extension and specification of Linear Code to reaction rules, we aim to minimize inconsistent symbology thereby making glycan database queries easier. With a clear guide for generating reaction rule descriptions, glycan synthesis models will be more interoperable and reproducible thereby moving glycoinformatics closer to compliance with FAIR standards. Here, we present Linear Code for Reaction Rules (LiCoRR), version 1.0, an unambiguous representation for describing glycosylation reactions in both literature and code.

Introduction

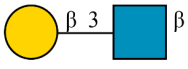
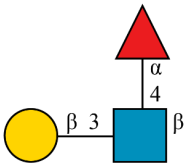
Glycans are predominantly synthesized through the serial addition of monosaccharides to form large polysaccharides. To build computational models of glycan synthesis, the biochemical reactions involved must be defined and described mathematically in a form that can be interpreted by computers [1-3]. Several groups have created such models using a variety of

strategies, including mechanistic and nonlinear [4-12], linear probabilistic [13,14], machine learning [15], formal-grammar [16], and substructural [17]. Unfortunately, most of these approaches use slightly different expressions of the building blocks, the reaction rules, therefore, model comparison is more challenging than it needs to be, with certain inconsistencies remaining to be resolved.

In the past few decades, substantial efforts made in the construction of these models of glycan synthesis were mostly focused on defining reaction rules that benefit from an unambiguous representation with human readability. For example, graphical denotation is one of the most human-understandable representations to describe reaction rules [18-20]. While graphical representations are intuitive and extremely accessible to a human reader, they are not computationally accessible due to ambiguities in their representations. There are already efforts to create computationally transmissible rule sets in XML-type representations like BioPAX [21], CellML [22], and SBML [7,8] which are readily interoperable and reusable. However, the XML-type model representations are not designed to be human-readable or included in the main text of a manuscript confining many design details to the supplement of a publication. As systems glycobiology develops, there is a need to develop a standard nomenclature for unambiguous and readable reaction rules to facilitate development, exchange, extension, and validation of glycosylation models and analysis tools.

Here we bring explicit attention to the concerns we raise above, we provide a focused, text-based representation of reaction rules that have been introduced for the purpose of formalizing these communications. GlycoCT [23] and WURCS [24,25] are two popular glycan nomenclatures in use today. GlycoCT was designed to maximize the descriptive specificity of the experimentally derived glycan structures data. WURCS, on the other hand, focuses on the uniqueness of a linear representation which promises efficient lookup in database queries. Both GlycoCT and WURCS produce unambiguous representations and are thereby invaluable for many applications, ranging from systems biology analyses [17] to an international glycan structure repository [26-29]. GlycoCT and WURCS provide a high degree of unambiguous detail; however, they are limited in their human-readability. The glycan extension to IUPAC, on the other hand, is more human-readable [30]. It specifies the linkage and branch information in an intuitive and linear manner. In the hopes of mitigating the inconsistent application of IUPAC and inconvenient illustrations, Linear Code described a simplified version of IUPAC nomenclature [31]. Specifically, Linear Code is a syntax for representing glycoconjugates and their associated molecules in a simple linear fashion. While keeping the linkage and branch information, Linear Code removes the hyphens between monosaccharides and abbreviates the glycan symbols, thereby simplifying the representation without limiting flexibility. Given its readability and parsability, Linear Code has become a popular choice for representing reaction rules in computational models of glycan synthesis (Table 1). However,

Table 1: The reaction rule Ab3GNb → Ab3(Fa4)GNb represented in Symbol Nomenclature for Glycans [18], Linear Code, IUPAC, GlycoCT, and WURCS separately. Linear Code provides the most straightforward and succinct representation.

	Reactant	Product
Structure plot (with link info)		
Linear Code	Ab3GNb	Ab3(Fa4)GNb
IUPAC-extended	β-D-Gal*p*-(1-3)-β-D-Glc*p*2NAc	β-D-Gal*p*-(1-3)-[α-L-Fuc*p*-(1-4)]β-D-Glc*p*2NAc
IUPAC-condensed	Gal(β1-3)GlcNAc(β1-	Gal(β1-3)[Fuc(α1-4)]GlcNAc(β1-
glycoCT	RES 1b:b-dglc-HEX-1:5 2s:n-acetyl 3b:b-dgal-HEX-1:5 LIN 1:1d(2+1)2n 2:1o(3+1)3d	RES 1b:b-dglc-HEX-1:5 2s:n-acetyl 3b:b-dgal-HEX-1:5 4b:a-lgal-HEX-1:5|6:d LIN 1:1d(2+1)2n 2:1o(3+1)3d 3:1o(4+1)4d
WURCS	WURCS=2.0/2,2,1/[a2122h-1b_1-5_2*NCC/3=O] [a2112h-1b_1-5]/1-2/a3-b1	WURCS=2.0/3,3,2/[a2122h-1b_1-5_2*NCC/3=O] [a2112h-1b_1-5][a1221m-1a_1-5]/1-2-3/a4-c1_a3-b1

with the rise of Linear Code adaptations to represent reaction rules, we have seen increasing diversity in the syntax, including branch constraints, duplicate monosaccharides omission, logical operators, etc.

Here we critically review reaction rule nomenclature. In doing so, we seek to promote the development of a standardized and unambiguous, readable, and computable reaction rule representation. First, we examine the original usage of Linear Code for reaction rule representation by discussing six major categories of syntax rules. Second, we discuss the various adaptations that have been introduced in the current usage of Linear Code to represent reaction rules. Third, we further discuss the apparent nomenclature ambiguity emerging in the adaptation of Linear Code to systems glycobiology. Finally, we demonstrate the depth of the nomenclature crisis through the minimal overlap in presumably similar networks. While many solutions to this nomenclature might be offered, we focus on six major recommendations to provide a unified representation of reaction rules that are likely to have a broad impact on minimizing change to the current adaptations.

Common lore at universities describes architects who, rather than "prescribe" ideal paths for students through the mall, waited to see where students would walk. They built their paths over the trampled grass of the "descriptive" paths chosen by the students. Similarly, we intend to extend the thoughtful "prescription" of Linear Code to "descriptive" extensions that will comfortably accommodate those currently working in systems glycobiology. We also provide some key definitions for ease of reading (Figure 1, Table 2).

Syntax Rules of the Original Linear Code

Linear Code rules can be separated into six categories of syntax rules (Table 3): Stereospecificity and ring structure rules (SRS), modification rules (MR), branch rules (BR), repetition rules (RR), glycoconjugate rules (GR), and uncertainty rules (UR). The saccharide unit (SU) refers to a structure with four elements: anomericity, position number, modifications, and monosaccharide (MS).

Stereospecificity and ring structure rules are set to differentiate the stereoisomers or distinct ring structures. A change from primary to secondary stereospecificity is denoted by " ' ", while a change to secondary ring structure is denoted " ^ ". A change to both secondary ring and stereospecificity is denoted " ~ ". For example, " G " represents glucopyranose, the pyranose con-

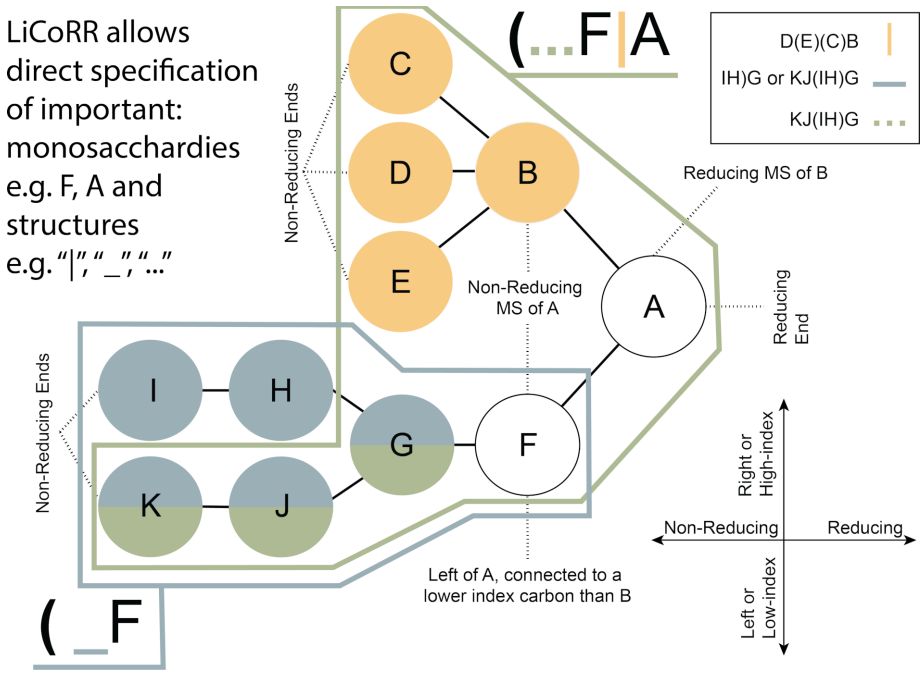


Figure 1: Common terminology and anatomy of a theoretical glycan, (KJ(IH)GF(D(E)(C)B)A. In this figure, we demonstrate some key terminology as well as the three primary uncertainty operators: branch (orange), continuation (blue), and ligand (green). The structures matching these terms are shown in matching colors, those matching both the continuation and ligand are shown in purple. A ligand can typically be removed with one cut. A continuation is a connection from a node to a root that can "continue" or "bypasses" other branch points. The paths from I to G or K to G represent one continuation; to represent both paths, a continuation is necessary because traversing from I to G requires the syntactic "bypass" of the KJ branch.

Table 2: Glossary of essential terms.

Term	Definition
saccharide unit (SU)	composed of a monosaccharide name, modifications (if any), anomericity (α or β configurations of the glycosidic bond), and the position it is bonded to a given SU.
monosaccharide (MS)	a sugar monomer.
lowest-carbon-index chain	the lowest carbon index branch corresponding to the non-reducing sugar connected to the lowest reducing-end carbon.
branch	any right branch, pictorially "right" of the reducing MS (Figure 1), where a non-reducing sugar is not connected to the lowest reducing-end carbon.
reducing and non-reducing ends	these are the MSs that appear "first" (closest to the glycoconjugate or first added in the synthesis) and "last" (leaves or terminal MS, those farthest from the "first" MS within a branch and have no linkage to a non-reducing MS). Typically, there is one reducing end and there are often multiple non-reducing ends.
reducing MS	closer to the first MS or the "reducing-end".
non-reducing MS	farther from the first MS and closer to a non-reducing end.

Table 3: Original Linear Code rules (Banin et al. [31]).[a]

	Rule description	Example
saccharide unit (SU)	1. see one-letter MS names in Table 4.	
	2. the anomer, where an α conformation is denoted as "a," and β as "b," follows the one-letter MS name.	Ga, Gb
	3. the carbon number by which the SU is attached follows the anomer.	Ga3, Gb2
	4. modifications. see modification rules for details, which follow after the carbon number.	see modification rules examples.
open form (OF)	1. open form notation. If a carbon is in its open-chain form, an "o" is attached to the end.	AbGo, AbG[P]o
stereospecificity and ring structures (SRS)	1. the less common stereoisomer (D or L) of an MS is indicated with apostrophes (‘).	D-Glc*p*: G L-Glc*p*: G’
	2. MSs with uncommon ring structures (e.g., furanose, pyranose) are indicated with a caret (^).	D-Glc*p*: G D-Glc*f*: G^
	3. MS that differ in both common stereospecificity and ring structure are indicated with a tilde (~).	D-Glc*p*: G L-Glc*f*: G~
modification rules (MR)	1. the modifications are represented by adding square brackets that include the connecting position of the modification to the SU, followed by the modification symbol (Table 5) in the form: [<number><symbol>] Exceptions include certain monosaccharides with common modifications (D-GalpNAc is AN instead of A[2N]). Anomericity (α or β) is expressed immediately after the modification.	β-D-Gal*p*(2P)-(1-3)-β-D-Glc*p*: A[2P]b3Gb
branch rules (BR)	1. when the non-reducing MSs are identical, the MS linked to the higher index carbon will branch (appear first in the written representation when read right to left, reducing to non-reducing end).	GNb2Ma3(NNa3Ab3GNb2Ma6)Mb4GNb
	2. when the non-reducing MSs are different, the less frequent non-reducing MS will branch (MS frequency Table 4).	Ab3ANb4(NNa3)Ab4Gb
repetition rules (RR)	1. repeating units are expressed inside parentheses, with an ‘n’ representing the number of repeats.	cellulose, which is a polymer of D-glucose residues joined by β-1,4 linkages are represented as {nGb4}
	2. if not the non-reducing end, the head of a repeated motif is expressed two dashes " - - "	{nGa6Ga4(-Ab3-)Ub2Ha3Ha3Ha3}
	3. if not the reducing end, the tail of a cyclic motif is expressed using the letter "c".	nGa6Ga4(-Ab3-)Ub2Ha3Hca3Ha3

Table 3: Original Linear Code rules (Banin et al. [31]).[a] (continued)

glycoconjugate rules (GR)	1. amino acid sequences are written after ';'. Lipid moieties are written after ':'. Other glycosides are written after '#'.	Ga;NY-S-C. Gb:C GNb3Ab#4-Trifluoroacetamidophenol
uncertainty rules (UR)	1. α or β linkage unknown, or connection position unknown: ?	AN?3G
	2. both linkage and connection position unknown: ??	AN??G
	3. an entire SU unknown: * * could match any whole SU.	ANb3*A
	4. when two possibilities are given for the identity of an SU element, use "/"	ANb3/4
	5. when two options are given for the identity of a complete SU, use "//"	Ab4//Ga2Aa3 represents Ab4Aa3 or Ga2Aa3
	6. for glycan fragments, use an index number + '%' as a variable for the fragment, and a '|' to separate the fragment from the core.	NNa6=1%|1%Ab4GNb2Ma3(1%Ab4GNb2Ma6)Mb4Gb denotes that Ab4GNb2Ma3(Ab4GNb2Ma6)Mb4Gb is the core, and that the linkage of the fragment NNa6 to the core is uncertain. % means uncertain, 1 is the index referring to the uncertain MS.

[a]"(#)" - Rules deprecated in LiCoRR.

formation of glucose, with D stereospecificity. Glucopyranose with L stereospecificity is written as " G' " (SRS1). Glucofuranose with D stereospecificity is written as " G^ " (SRS2), and glucofuranose with L specificity is written as " G~ " (SRS3). Similarly, galactofuranose, a common fungal monosaccharide, would be written "A^"

Open form rule indicates that if the MS at the reducing end is open – a linear rather than cyclic MS, then the final character to the right of the string should be "o". For example, lactose, galactose β-linked to glucose would be written as AbG if the reducing end glucose is closed and AbGo if the glucose is open; the open "o" takes the place of the linkage in this context. If the glucose is phosphorylated, this structure would be written AbG[P]o.

Modification rules specify a modification of a MS at certain positions (MR1). MS + " [" + modification + "] " is used to denote the modification. For example, "G[2S]" describes sulfation on the second carbon of a D-glucopyranose. The anomericity is expressed to the right of the modification (i.e., "G[2S]a"). Multiple modifications to the same MS are ordered based on the position number inside the same brackets; ascending order from left to right. For some common modifications like *N*-acetylgalactosamine, instead of "A[2N]," Linear Code uses "AN" directly. Table 4 includes syntaxes of MS in Linear Code and common modified MSs. Common modification names can be found in Table 5. Given multiple modifications, carbon numbers are written in ascending alphanumeric order. Therefore, dideoxy galactose, or abequose, is written "A[2,6D]" while *N*-aceytlfucosamine could be written "A[6D,2N]".

Branch rules specify which non-reducing saccharide unit (SU) should be in the branch and which SU should continue the lowest-carbon-index chain; branching is determined by the identity of the first MS in a chain. When the non-reducing MSs are identical, the MS and its substituent chain, linked to the higher carbon of the reducing MS, will branch while the MS and substituent chain, linked to the lower carbon position of the same reducing MS, remains in the lowest-carbon-index chain (BR1). Otherwise, if the non-reducing MSs are different, the chain with a less frequent non-reducing MS (lower rank in Table 4) is considered the branch (BR2). The MS frequency is specified in Table 4, decreasing from top to bottom. When there are more than two non-reducing MSs linked to the same reducing MS, they are ranked, first by frequency, then by linkage index. The highest frequency MS is ranked higher, further to the left when the expression is written. Any MSs with equal rank after the frequency rank – those that are the same MS – are ranked by their linkage index, the lowest linkage indexes are ranked higher. A higher rank means these MSs, and their associated chains, will remain on the lowest-carbon-index chain, while the lower rank MSs will branch.

Repetition rules specify the contraction syntax for succinctly describing repeating MS units. The repetition structure is denoted by curly brackets, with a prefix of repetition times inside the brackets. For example, cellulose, which is a polymer of D-glucose residues joined by β-1,4 linkages, is represented as

Table 4: Common monosaccharides and their Linear Codes (adapted from [31]). We have added NG as it has become a clearly important monosaccharide excluded from the original list. Full monosaccharide descriptions are recorded in IUPAC [18]; all terms can be found at https://www.qmul.ac.uk/sbcs/iupac/2carb/38.html.

Monosaccharides[a]	Linear Code	IUPAC
D-glucose	G	Glc
D-galactose	A	Gal
N-acetylglucosamine	GN	GlcNAc
N-acetylgalactosamine	AN	GalNAc
D-mannose	M	Man
N-acetylneuraminic acid	NN	Neu5Ac
**N*-glycolylneuraminic acid[b]	NG	Neu5Gc
neuraminic acid	N	Neu
2-keto-3-deoxynononic acid	K	KDN[c]
3-deoxy-D-manno-2 octulopyranosylonic acid	W	Kdo
D-galacturonic acid	L	GalA
L-iduronic acid	I	D-IdoA
L-rhamnose	H	Rha
L-fucose	F	Fuc
D-xylose	X	Xyl
D-ribose	B	Rib
L-arabinofuranose	R	Ara*f*
D-glucuronic acid	U	GlcA
D-allose	O	All
D-apiose	P	D-Api
D-fructofuranose	E	Fru*f*
*ascarylose[b]	C	Asc
*ribitol[b]	T	Rib-ol (Rbo)

[a]All the monosaccharides are in their pyranose form unless otherwise noted. [b]Asterisk ("*") represents an update from the original table. [c]KDN: 3-deoxy-D-glycero-D-galacto-nonulosonic acid. Kdn: 3-deoxy-D-glycero-D-galacto-nonulosonic acid.

Table 5: Common modifications and their Linear Code (from [31]).

Modification type	Linear Code	IUPAC
deacetylated *N*-acetyl	Q	N
phosphoethanolamine	PE	Pe
inositol	IN	In
methyl	ME	Me
N-acetyl	N	NAc
O-acetyl	T	Ac
phosphate	P	P
phosphocholine	PC	Pc
pyruvate	PYR	Pyr
sulfate	S	S
sulfide	SH	Sh
aminoethylphosphonate	EP	Ep
*deoxy[a]	D	d
*carboxylic acid[a]	CA	-oic
*amine[a]	A	-amine
*amide[a]	AO	-amide
*ketone[a]	K	-one

[a]Asterisk ("*") represents an update from the original table.

"{nGb4}" (RR1). If a ring structure is repeated and the repeating unit is not connected "head to tail," the MS where the repeating units are connected is marked between 2 dashes " - - " (RR2). An example is {nGa6Ga4(-Ab3-)Ub2Ha3Ha3Ha3}. Additionally, Banin et al. specify that a cyclic motif, a form of repetition, is expressed using the letter "c" [31]. While specification was limited in the original publication, we interpret "c" as denoting the "tail." (-X-) denotes the head if it is not the left end and "c" denotes the tail if it is not the right end of the string. For example, in the molecule nGa6Ga4(-Ab3-)Ub2Ha3Ha3Ha3, Ab3 connects to the reducing end, Ha3. But if Ab3 was connected to the second Ha3 from right instead, we can specify the point of the cycle using a "c," nGa6Ga4(-Ab3-)Ub2Ha3Hca3Ha3.

Glycoconjugate rules describe when a reducing end of a SU is connected to non-carbohydrate moieties, Glycoconjugate rules regulate that amino acid sequences are written after " ; ", lipid moieties are written after " : ", and other glycosides are written after " # " (GR1). For example, a glucose β-linked to a Ceramide is written as "Gb:C."

Uncertainty rules describe syntax for when certain features of the SU are unknown or have more than one possibility. If the anomericity of certain bonds is unknown, Linear Code uses " ? " (i.e., AN?3G) (UR1). If both linkage anomericity and position are unknown, Linear Code uses " ?? " (i.e., AN??G) (UR2). If an entire SU is unknown, " * " can be used instead. ANb3*A represents a three SU glycan, where the second SU is unknown (UR3). When two monosaccharides are possible for a given SU, Linear Code uses the forward slash to separate them. When SU ambiguity refers to anomericity, position number, modifications, or MS, a single " / " is used (i.e., ANb3/4) (UR4). Given two complete possible SUs, Linear Code uses " // " to separate them (i.e., Ab4//Ga2Aa3 represents Ab4Aa3 or Ga2Aa3) (UR5). When analyzing fragmented glycans, an "< index number>%" is used to store fragmented structures as a variable. For example, NNa6=1%|1%Ab4GNb2Ma3(1%Ab4GNb2Ma6)Mb4Gb is a glycan containing a terminal α-2,6-linked sialic acid (NNa6) whose linkage position is unknown. Here, the " | " is used to separate the fragment(s) and core structure components (UR6).

In the interest of demonstrating the reach of single letter LC monosaccharides (Table 4), we provide a monosaccharide

network suggesting demonstrating non-trivial functional-group (Table 5) relations between monosaccharides (Figure 2). We used RDKit, an open-source cheminformatics toolkit, to identify chiral centers and further determine stereochemical equivalence classes. Monosaccharides were clustered with an 80% stereo-similarity threshold (Figure 2A), and the maximum common substructure (MCS) of each cluster was obtained (Figure 2B). These MCS equivalence classes were used to group monosaccharides explicitly listed in Table 4 and connect them through addition or subtraction of functional groups in Table 5 (Figure 2C) to every major monosaccharide listed by SNFG (Figure 2D). Figure 2D shows some of these non-trivial paths (e.g., beyond GlcNac; G → GN or G[2N]) from Table 4 monosaccharides, to all listed SFNG monosaccharides via modifications from Table 5. We further provide a full network (Table S6, Supporting Information File 1) to facilitate the discovery of any monosaccharide–monosaccharide relation. For example, the fucose-galactose relation can be found in row 1479 of Table S6 (Supporting Information File 1). They differ by one hydroxy group therefore fucose could be represented as "A[6D]". Similarly, abequose, a dideoxy galactose, could be represented as "A[2,6D]" or "F[3D]". Through simple lookup in Table S6 of Supporting Information File 1, many noncanonical monosaccharides can be described thus mitigating the limitations of the single-letter monosaccharide representation.

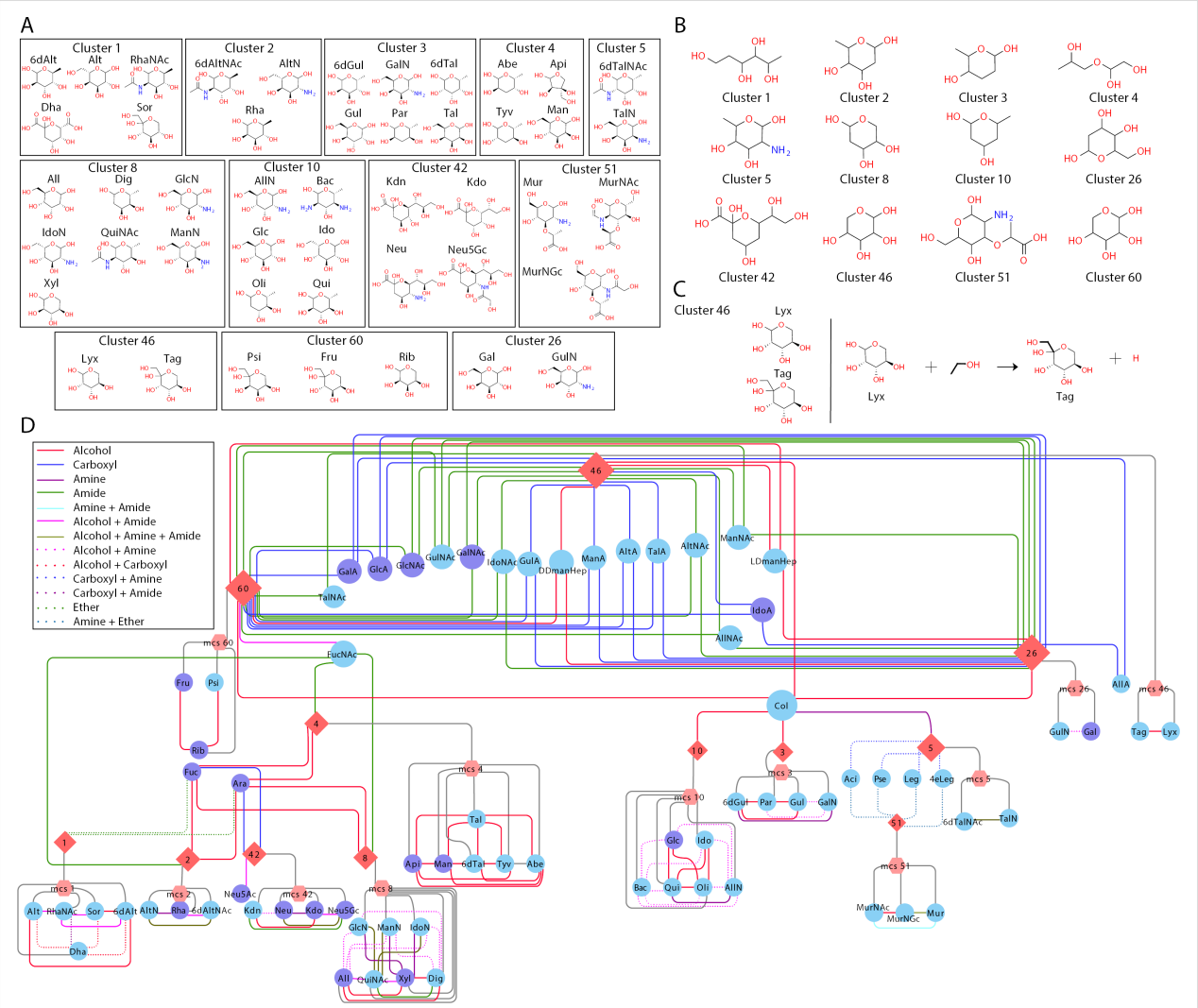


Figure 2: Monosaccharide reachability analysis. (A) Clusters contain monosaccharides with highly similar stereochemistry (>80%). (B) The maximum common substructure (MCS) associated with each cluster. (C) An example to illustrate the modifications needed to reach one monosaccharide to another, as identified by the complete monosaccharide reachability network (Table S6, Supporting Information File 1). (D) The monosaccharides reachability network, showing only connectivity for the least number of modifications needed, differentiated by color as stated in the legend, between monosaccharides (circle) and clusters (diamond). Additionally, the node size denotes the number of different possible paths taken for them to be reached. Please note that each edge is not a predicted or proposed feasible reaction. Edges denote functional groups that can be added or removed from one monosaccharide to represent another.

Current Usage of Linear Code to Represent Reaction Rules

Linear Code was first used to represent reaction rules in 2009. A reaction network, specifying glycans with condensed IUPAC and Linear Code, was trained on mass spectrometry abundance to learn biosynthetic enzyme activities [10]. Their reaction rules table contained four features: enzyme, reactant, product, and constraint. For their implementation, not all original Linear Code rules are adopted. Krambeck et al. [10] maintained the linkage information (Table 3: SU2), one-letter MS abbreviation (Table 4), and branch rules (Table 3: BR), which are the necessary conditions to denote a glycan with branches [10]. On the other hand, symbols " ~ ", " * ", " | " were defined with new meanings, though they already had their meanings in the original Linear Code rules (Table 3: SRS3, UR3, UR6, respectively). Instead, Krambeck et al. introduced several new symbols to convey logical relationships (" & ", " ~ ", "or") and structural ambiguity (" ... ", " _ ", " | ", " * "), all of which were used to specify constraints. For example, a constraint "Ma6 & Ma3" means the reaction will happen only if both Ma6 and Ma3 appear in the glycan; as an N-glycan, these are the terminal mannoses capping the chitobiose core. The "Ma6 or Ma3" constraint promotes the reaction if either Ma6 or Ma3 exists. "~Ma6" means the reaction will not happen if Ma6 is present in the glycan. The structure denotations are indicators of certain parts of the glycan. The entry " ... " can be replaced with either nothing or any polysaccharide with matched parenthesis. The entry " _ ", in Krambeck et al., can be replaced with either nothing or any polysaccharide where each left parenthesis is matched to a right parenthesis but where right parentheses are not necessarily matched. Entry " | " represents a possible branch. We expand on the distinctions between " … ", " _ " and " | " in a later section "Substring uncertainty operators" (Table 4). The asterisk " * " stands for the reaction site, which is the position where the new MS will be added or an MS is removed. Krambeck et al. also uses " # " to describe constraints around the number of MS that may appear in a glycan. For example, the constraint "#A = 0" means the reaction will happen only if there is no galactose. The Krambeck et al. adaptation is the most common adaptation of Linear Code to represent reaction rules [7,13,15,32].

Based on the Linear Code reaction rules framework Krambeck et al. created, later researchers introduced new attributions that specify and simplify the description of reactions. Bennun et al. and Spahn et al. include the amino acid at the end of the Linear Code attached by a semicolon " ; ". This suffix is exactly the syntax from the original Linear Code rules (Table 3: GR1). The reaction rules table generated by Spahn et al. also provided localization information, which is either *cis*, *trans*, or *medial* to denote the Golgi compartment where the reactions happen [13,14]. The subcellular localization of a reaction, in the endoplasmic reticulum, Golgi, cytoplasm (bacteria and archaea), or lysosome (degradation, Man-6-P dephosphorylation and lysosomal glycoprotein biosynthesis [33,34] or paucimannose recycling [35]), are important constraints on glycosylation [36], therefore, the addition of this information to the Linear Code reaction rules provides insights into the glycosylation types.

Some models of glycan synthesis generated reaction rule tables with an additional column Enzyme Commission number (EC number) [7,16,37]. The EC number system is a numerical classification scheme for enzyme-catalyzed reactions that provides an unambiguous accession to a cataloged reaction [38]. The inclusion of an EC number in the reaction rules table, therefore, promotes the clarity, interoperability, and reproducibility of the generated reaction model.

A common syntax used by most studies is the leftmost " (" to represent the terminal, non-reducing end of the glycan chain. It specifies whether the leftmost MS is the terminal MS both visually and computationally. For example, the reaction rule (GN → (Ab3GN applies to all reactions which add one galactose to a terminal *N*-acetylglucosamine. On the other hand, the reaction rule GN → Ab3GN applies to all reactions which add a galactose to an *N*-acetylglucosamine, but not necessarily the terminal one. The leftmost " (", therefore, can easily vary the glycan substrate substantially.

Though Linear Code was developed with parsability in mind, some have found it useful to make a specific computational implementation of the reaction rules to accommodate the syntactic constraint of programming languages. A human milk oligosaccharide metaglycome was constructed using a combination of linear code, glycan structures represented in XML and XPath queries [39]. Separately, Akune et al. generated a theoretical *N*-glycan database called UniCorn, based upon a Perl implementation of reactions on glycans represented in Linear Code [37]. Though Linear Code is computer-parsable, there is still substantial work necessary to implement that parsing because there is no standard representation for handling the wide variety of reactions possible, nor open-source software available to implement the parsing of such rules.

Representing reaction rules in Linear Code is not easy because of a few ambiguous cases not completely described in the initial Linear Code paper. Subsequent studies, therefore, have developed their own ways to idealize reaction rule implementations based on Linear Code. Using the framework Krambeck et al. built [10], new information like Golgi localization and EC numbers are added to specify and simplify the reaction rules.

Original Prescriptions for Substring Uncertainty Operators

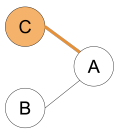
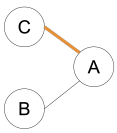
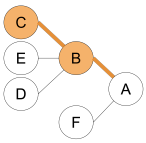
In its original conception [10], the adaptation of Linear Code to represent reaction rules aimed to describe how glycosylation enzymes change the structure of glycans in terms of how the Linear Code character string descriptions of the glycans are changed (Figure 1). In the simplest case, we can specify a substring of the substrate code to be replaced by a new substring to form the product code. In addition, there can be constraint and adjustment substrings whose presence or absence within the substrate string either restricts which glycans can be substrates of a particular enzyme or modifies the reaction rate parameters. Uncertainty operators have been developed to facilitate searching substrings for specific structural features of a glycan implied by the substrings.

The substring specifications for the substrate, product and adjustments can include any combination of characters included in the glycan codes in addition to uncertainty operators inserted within the directly specified characters. Each uncertainty operator is represented by one or more characters, such as "…" or "_" (Table 6). To perform substring matching of a glycan to a substring with uncertainty operators, we first identify the characters of the specified string immediately before and after the uncertainty operator. If found, we then test the substring of the glycan string between these two matched character strings and check for the appropriate uncertainty operator properties. In parsing the glycan code, an initial left parenthesis is always added to the complete glycan code so that the terminal end of every branch of the glycan is always a left parenthesis. Below, in defining the properties of substrings corresponding to an uncertainty operator, we use the symbol X to represent some monosaccharide with its connection, such as Ma3, GNb4, etc.

There are three types of uncertainty operators. The ligand " ... ", the continuation " _ ", and the branch " | ". Each has a specific syntactic match, but intuitively, the ligand is a chain that can contain branches, the continuation is a chain that can include branches terminated outside of the continuation, and the branch is either a complete branch or nothing. Functionally, " ... " indicates the "leftward" extension along the lowest-carbon-index chain, " | " indicates the "rightward" extension along the highest-carbon-index chain, and " _ " indicates an extension along either the left or right chain (Figure 1).

More specifically, (1) the ligand uncertainty operator indicates a chain of MSs that can include attached branches completely contained in the substring, (2) the continuation uncertainty operator indicates a chain of MSs that can include attached branches that may not be wholly contained in the substring, and

Table 6: The difference between " _ ", " … " and " | " with illustrations. These symbols were proposed by Krambeck et al. [10]. The initial names are ligand (" ... "), continuation (" _ "), and possible branch (" | "). Each uncertainty operator in the last four example columns can be replaced by the substring in red to achieve the behavior described in the column header. For a more comprehensive look at the usage of these uncertainty operators, see Supporting Information File 1, Table S1 for a manual collection of matches, and Table S4 (Supporting Information File 1) for an automated collection of matches.[a]

Symbols	Syntax	Meaning	Add a whole new branch **B*(C)*A** 	Initiating branch **B(C)A** 	Extending lowest-carbon-index chain **BCA**	Initiating nested branch **E(D*(C)B*)A** 
_	any string where every '(' has a matching ')'. Includes the empty string.	chain bypassing a branch to reach a reducing MS; A continuation cannot necessarily be removed by splitting one linkage (can contain branches)	B_A	B(C_A	B_A	E(D_A
...	any string with all matching parentheses. Includes the empty string.	chain to a reducing MS; A ligand can typically be removed by splitting one linkage (can contain branches)	B...A		B...A	
+ (formerly "|")	')' or '(...)' or ')(...)'. Or an empty string.	possible branch point.	B+A	B(C+A		

[a]A, B, C, D, E are abstract monosaccharides.

(3) the possible branch uncertainty operator indicates where a branch may be included in the substring. Due to the nuances of representing a glycan linearly, these are not complete definitions.

Ligand "…" – A ligand is a fragment of a larger molecule connected to the rest of the molecule at one point. Glycans are themselves ligands, as they are pieces of larger molecules. A substring is a valid "ligand" if each parenthesis in that substring is uniquely and appropriately matched; each left parenthesis must be followed by a corresponding right parenthesis and each right parenthesis must be preceded by a corresponding left parenthesis: ")(" are not matched parentheses. Any substring with all left and right parentheses matched, including an empty string, is considered a ligand. If we select a substring of the code representing a glycan it may or may not represent a ligand. For example, XXX)X, XX(XX, X)(XX are not valid ligands, while XX(XXX)XX is valid.

Functionally, ligands can serve as connectors between the left and right portions of a glycan a user would like to specify. A ligand is simply a chain of monosaccharides which may contain nested branches; the nested branches must also be ligands. However, there can be many chains or paths through a ligand, starting from one of the terminal monosaccharides and culminating at the root end; there are many ligands within most ligands. Any of these ligands can serve as a connector from the root (reducing) end of one ligand to a terminal (non-reducing) end of another. The key property of a ligand substring is that all the included branches of the ligand are completely contained in the substring.

Continuation "_" – As we parse from left to right through a substring, we may find left parenthesis (entering into a branch) and right parenthesis (exiting a branch). A ligand, with matched parentheses, indicates an equal number of branch initiations and completions. On the other hand, a substring with an unmatched right parenthesis, for example, XXX)XX or X)(XX)XX, indicates a net termination of branching; each right parenthesis indicates moving out of a branch towards a root. As long as all the left parentheses encountered are followed by right parentheses, we are following a path along a connected chain of the glycan structure. A substring where every left parenthesis can be matched with a following right parenthesis, but not necessarily vice versa, is a "continuation." Again, we include the empty string in this class of substrings. Note that any ligand is also a continuation.

The continuation uncertainty operator can be very useful in formulating rules that apply to specific monosaccharides connected by a chain of monosaccharides to a particular reducing monosaccharide of the glycan structure. For example, the iGnT enzyme adds a Gnb3 group to a terminal galactose group and has a preference for the two branches that are connected to the Ma6 of the root Ma3(Ma6)Mb4 structure. This leads to an adjustment rule based on the string Ma3|(*_Ma6)Mb4. Here the "|" uncertainty operator is used to allow for the possible presence of a bisecting GlcNAc on the root mannose: Ma3(GNb4)(…Ma6)Mb4. The "*" indicates the site of the enzyme action.

Possible branch "|" – As discussed, parsing a linear glycan from left to right, we can encounter matched parentheses indicative of a ligand or unmatched right parentheses indicative of a closing branch. We can leverage the branch closer offered by these symbols to mandate a possible branch. The definition of the "possible branch" is one of either: ") ", " (...) ", ")(…) " or an empty string. This uncertainty operator can be replaced with either a branch, the start of a branch or nothing. It allows the same specification string to work whether an additional branch is present at the position of the uncertainty operator or not, as in the above example.

Divergence in Current Implementations of Reaction Rules from Original Linear Code

Linear Code is a useful notation to succinctly describe glycan structures. It is thus useful to represent substrates and products. However, constraints on the glycan acceptor class where a new monosaccharide is added, was beyond the scope of the original Linear Code rules. Therefore, different adaptations are introduced throughout the literature.

We have identified four symbols that are prescribed with different meanings than they were originally assigned in the Linear Code rules:

Ambiguous symbol 1 – Originally, " ~ " following an MS name was used to denote the MS with different stereospecificity and ring structure from the common form (Table 3: SRS3). For example, while " G " represents D-glucopyranose, " G~ " represents L-glucofuranose, which is rarely seen. To represent reaction rules, " ~ " was used, instead, to convey logical negation [7,10,13,32]. For example, a constraint "~Ab" means the reaction will not happen if a β-galactose is present.

Ambiguous symbol 2 – " | ". For reaction rules, " | " is widely used to represent a potential branched structure in a substrate [7,10,13,32]. For example, "(GNb2|Ma3" represents a glycan structure with a potential branch on the mannose. However, " | " was originally designed to separate the certain and uncertain

parts in a fragmented glycan, where there is a possibility of different structures (UR6).

Ambiguous symbol 3 – " # ". Originally, " # " was designated to signify the starting point of glycosides that are not amino acids or lipid moieties (Table 3: GR1). For example, "GNb3Ab" connected to a "4-trifluoroacetamidophenol" is written as "GNb3Ab#4-Trifluoroacetamidophenol."

Ambiguous symbol 4 – " * ". Another ambiguous symbol is the asterisk " * ". In the original Linear Code context, " * " is used when an entire saccharide unit in the complex carbohydrate is unknown. In reaction rules representation, " * " marks the "reaction site", the position of the first difference between product and substrate strings in Linear Code form [7,10,13,32]. Note that the "reaction site" does not necessarily refer to the exact place that the reaction happens. For example, given the reaction "(...Ab4GNb → (Fa3(...Ab4)GNb," the constraint " (*Ab4 or (*Fa2Ab4" means that the reaction will happen if and only if the " … " in the reactant represents either nothing or "Fa2." In this case, " * " on the left of "Ab4" indicates where the reactant and the product differ from left to right in the Linear Code expression. However, the real reaction takes place at the "GNb," not "Ab4." Demonstrating the left-to-right specificity of " * ", consider the rule, (Ma2Ma → (Ma with constraint ~*2Ma3(…Ma6)Ma6. This constraint rules out removing the Ma2 on the middle branch (underlined) of the original M9 glycan, Ma2Ma2Ma3(Ma2Ma3(Ma2Ma6)Ma6)Mb4GNb4GN;Asn. If parsed from right to left, the constraint would be ~*Ma3(…Ma6)Ma6.

Linear Code is primarily a representation of glycan structure, and the formulation of reaction rules from Linear Code emerged as it was adapted for use with systems biology reaction networks. Specifically, when researchers aimed to define rules for reactions when building the networks, additional symbols were needed and, therefore, proposed. However, these now differ between studies.

In the first study, building reaction networks from Linear Code, Krambeck et al. defined " … ", " _ ", and " | " as uncertainty operators to indicate specific combinations or balanced or unbalanced (complete or incomplete) branches [7,10,32]. Spahn et al. used only two of the three symbols; " | " to indicate branching and " … " to represent continuation [13]. In this section, we will only focus on Krambeck et al. syntax. Syntactically, each of these symbols specifies whether or not the monosaccharides following the symbol, the first monosaccharide within the uncertainty operator replacement, appear within parentheses. If the monosaccharides appear within parentheses, it is "branching" off the lowest-carbon-index chain; otherwise, it is a "continuation" along the lowest-carbon-index chain. Each uncertainty operator describes a branching and/or continuation. Additionally, an uncertainty operator can require a complete phrase, with matched parentheses, or not. Finally, some uncertainty operators can be replaced with nothing (the empty string).

In the original Krambeck et al. implementation, multiple disjunctive constraints are connected by the logical disjunction "or." An example is "(*Ab4 or (*NNa3Ab4" (Table 7). In the Liang et al. adaptation, however, the "or" relationship is delineated by writing each reaction rule on separate lines. For example, the two constraints for the reaction rule "(...Ab4GNb → (Fa3(...Ab4)GNb" would simply be written on two lines (Table 8).

Table 7: The reaction rule (GN → (Ab4GN with four constraints written in the same cell.

Enzyme	Reactant	Product	Constraint
b4GalT	(GN	(Ab4GN	*...GNb2|Ma3 or *...GNb4|Ma3 or *...GNb2|Ma6 or *...GNb6|Ma6

Table 8: The reaction rule (GN → (Ab4GN with four constraints written on separate lines.

Enzyme	Reactant	Product	Constraint
b4GalT	(GN	(Ab4GN	*...GNb2|Ma3
b4GalT	(GN	(Ab4GN	*...GNb4|Ma3
b4GalT	(GN	(Ab4GN	*...GNb2|Ma6
b4GalT	(GN	(Ab4GN	*...GNb6|Ma6

Most adaptations of reaction rule implementations are more or less related to the earliest Krambeck et al. adaptation. Some symbols are only seen in the Krambeck et al. adaptation. Besides the " # " as the number symbol, Krambeck et al. also uses "Gnbis" to refer to the specific structure of bisecting GN, which is "Ma3(GNb4)(...Ma6)Mb4."

Several reaction rules for N-glycan biosynthesis are presented for direct comparison (Table 9, Table S5 in Supporting Information File 1). While there were several apparent divergences in the usage of terms, the rules are predominantly similar. The intent of this paper is to ensure the consistency of these rulesets going forward.

Table 9: Reaction rules from multiple N-glycan biosynthesis models in LiCoRR representation. This table describes select rules from Krambeck et al. [10] in LiCoRR and LiCoRRICE representation. Representations across multiple manuscripts can be found in Linear Code, LiCoRR and LiCoRRICE in Table S5 (Supporting Information File 1).

Enz.	Substrate	Product	Constraints (LiCoRR)	Constraints (LiCoRRICE)
ManI	(Ma2Ma	(Ma	!@2Ma3(…Ma6)Ma6 & !Ga3	nMan(a1-?)>4 & nMan(a1-?)<8 & !Man(a1-2)Man(a1-3)... Man(a1-6) & !Glc(a1-3)
ManI	(Ma3(Ma2Ma3(Ma6)Ma6)	(Ma3(Ma3(Ma6)Ma6)	!Ga3	!Glc(a1-3)
ManII	(Ma3(Ma6)Ma6	(Ma6Ma6	(GNb2+Ma3 & !Gnbis	!Gal(b1-?) & !GlcNAc(b1-4)...Man(b1-4) & GlcNAc(b1-2)Man(a1-3)
ManII	(Ma6Ma6	(Ma6	(GNb2+Ma3 & !Gnbis	
a6FucT	GNb4GN	GNb4(Fa6)GN	GNb2+Ma3 & #A=0 & !Gnbis	GlcNAc(b1-2)Man(a1-3)... Man(b1-4) & !GlcNAc(b1-4)...Man(b1-4) & !Fuc(a1-3)
GnTI	(Ma3(Ma3(Ma6)Ma6)Mb4	(GNb2Ma3(Ma3(Ma6)Ma6)Mb4		nMan(a1-?)=4
GnTII	(GNb2+Ma3(Ma6)Mb4	(GNb2+Ma3(GNb2Ma6)Mb4		nMan(a1-?)=2 & !GlcNAc(b1-4)...Man(b1-4) & !Fuc(a1-3) & !Gal(b1-?)
GnTIII	GNb2+Ma3	GNb2+Ma3(GNb4)	!Ab & !Gnbis	GlcNAc(b1-2)Man(a1-3)... Man(b1-4) & !Gal(b1-?)
GnTIV	(GNb2Ma3	(GNb2(GNb4)Ma3	!Gnbis	!Gal(b1-?) & !GlcNAc(b1-4)...Man(b1-4)
GnTV	(GNb2Ma6	(GNb2(GNb6)Ma6	!Gnbis	!Gal(b1-?) & !GlcNAc(b1-4)...Man(b1-4)
iGnT	(Ab4GN	(GNb3Ab4GN	!@_Ma3+Mb4	
b4GalT	(GN	(Ab4GN	!@GNb4)(...Ma6)Mb4	!Gal(b1-3)GlcNAc(b1-?) & !@GlcNAc(b1-4)...Man(b1-4)
b3GalT	(GN	(Ab3GN	!@GNb4)(...Ma6)Mb4	!Gal(b1-4)GlcNAc(b1-?) & !@GlcNAc(b1-4)...Man(b1-4)

Recommendations to Unify Descriptive Usages of Linear Code for Reaction Rules (*LiCoRR*)

Linear Code has shown its utility for the compact description of glycans and compatibility with efforts to define glycan reaction rules for systems biology models. A few ambiguities have emerged through different interpretations and implementations. Here we propose possible solutions as described by the original prescription for Linear Code, the consensus of the community, and our recommendation following this survey.

We have demonstrated the LiCoRR representation of all N-glycosylation reaction rules discussed in this paper in Table 9. Table 9 also includes an instance of these reaction rules written with IUPAC monosaccharides and linkages from GlycoEnzDB. Due to incomplete adoption and flexibility of Linear Code monosaccharides, we encourage users to accommodate both Linear Code and IUPAC monosaccharides when possible to facilitate interoperability; Linear Code monosaccharides may not be sufficient for every project while IUPAC-extended nomenclature [18] is actively maintained to ensure complete coverage of known sugars. If a user wants to specify that they are using LiCoRR with IUPAC monosaccharides, they can specify it as “LiCoRRICE” the LiCoRR-IUPAC Complement Expression. We also provide the matched constraints in Table 9 as Original Linear Code (Table S5 in Supporting Information File 1). It should be noted

that IUPAC uses square brackets, "[]", rather than parentheses, "()", to delineate branching. Therefore, the wildcards should recognize square brackets rather than parentheses. Additionally, IUPAC does not use deterministic branching. Therefore, specifying branch direction is not meaningful and the three branch-specific LiCoRR wildcards can be reduced to one, "...", in LiCoRRICE. With these small changes, LiCoRR can be extended to LiCoRRICE and, as such, gain access to its carefully curated and growing list of MS units and modifications.

The original Linear Code syntax contains eighteen specific regulations across seven categories, among which only five regulations are seen in reaction rule implementations. In fact, the five regulations include three SU elements (MS name, linkage-type, position number), denotations (Table 3: SU) and one branch rule (Table 3: BR1). BR1 dictates that when two branching MSs are identical, the MS linked to the higher index carbon will have its chain on the branch (Table 3: BR1). If we extend the condition for BR1 from identical MSs to all MSs, written glycan structures will still maintain their uniqueness since each position on the MS can only connect to a single MS. BR2 dictates that the least frequent MS of the pair will branch (Table 3: BR2). BR2 solves the case when there are more than two non-reducing MSs linked to the same reducing MS. However, if we applied the expanded BR1 and ordered the chains based on decreasing position numbers from right to left in multi-chain cases, BR2 would be redundant. For example, Ab4(GNb4GNb3)(GNb6)Ab4Gb will be written as GNb4GNb3(Ab4)(GNb6)Ab4Gb.

Among the logical relationships required for constraint specification, only "or" is seen in the original Linear Code rules. " / " was designed to separate two possibilities within an SU (Table 3: UR4) and " // " was used to separate two possible complete SU options (Table 3: UR5). It would cause unnecessary confusion if " / " and " // " are used to denote the "or" relationship between constraints. Therefore, the task to convey Boolean logic among constraints was left to emerge organically in its application to reaction rules.

Recommendation 1 – "Logical negation." The field chose to use the " ~ " to indicate logical negation (Table 10: a). Unfortunately, this choice conflicts with the ability to express uncommon stereospecificity, as prescribed in the original Linear Code (Table 3: SRS3). Though this is a rare necessity, and the original Linear Code tilde appears on the right of the monosaccharide, usage of a " ! " - as used in many common programming languages – to indicate logical negation would preserve the original meaning of the tilde in case it becomes necessary in a future notation.

Recommendation 2 – "And." Similarly, the field chose " & " to represent the conjunction relationship between constraints. We recommended preserving this symbol use since it is human and computer-readable and does not overlap with any notation in the original Linear Code.

Recommendation 3 – "Number." " # " was defined to combine glycans with glycosides other than amino acids and lipids (Table 3: GR1). Krambeck et al. use it to represent the number of times a certain MS appears, a common use of " # ". In LiCoRR, we deprecate the use of " # ", " ; ", and " : " to specify the glycoconjugate class. The number sigh " # " can be used to separate a glycan (on the left) from any conjugate (on the right). Colon and semicolon can therefore be reserved for other future uses. To specify a glycopeptide, users may also inscribe them directly in the peptide using the existing branching rules: "PEP(AG(LY)CAN)TIDE" would describe a biantennary glycan bound to the threonine of a peptide. Because the number sign is used to indicate a glycoconjugate, we recommend using "n." For example, "#A = 5" will then be written as "nA=5" (Table 10: i).

Recommendation 4 – "Splitting & 'or'." In addition to having several constraints split by "or," we can rewrite the rules several times with a single constraint for each rule, as done for the reaction rule b4GalT in [14]. Splitting disjunctions over multiple lines is similar to atomization, the first normal form of database normalization requiring the domain of each attribute to contain an indivisible element. In addition, the separate rules have the advantage that they can have different reaction rate parameters. This advantage can eliminate the need for separate adjustment rules for various cases. Depending on the circumstances, splitting disjunctions across multiple lines may be necessary, though it is often more succinct to condense them, separated by an "or" within a single rule.

Recommendation 5 – "Branch point." Many studies using Linear Code to define glycan synthesis networks assigned " | " as a possible branch point [7,10,13]. Our recommendation, however, is to use " + " instead of " | " as the branch point because " | " is already assigned within the original Linear Code. Additionally, we think " + " is more morphologically close to a branch.

Recommendation 6 – "Omission." Though " * " has been widely used by the systems glycobiology field to represent the reaction site, the original Linear Code rules actually specify " * " to stand for the omission of an entire saccharide unit (Table 3: UR3). We wish to minimize this inconsistency with the original statement of Linear Code [31]. Therefore, for " * ", we recommend preserving the meaning of the omission of one

Table 10: Symbols previously used by systems glycobiologists and our recommendations. Rows a–i are the functions implemented by published papers. Rows j–m are the functions prescribed in the original Linear Code rules. (A) Symbols to represent reaction rules across publications utilizing Linear Code. (B) Consensus and recommendation for reaction rule representation going forward.

	(A) Symbol used	OLC [31]	Kra [10]	Spa [13]	Lia [14]	Hou [7]	(B) Consensus adaptation of OLC to reaction rules	LiCoRR	Examples
a	logical negation		~	~	~	~	~	!	!Ma
b	and		&			&	&	&	!Ma & Ab3
c	or		or			or	or	separate rules, or	!Ma or Ab3
d	continuation (left parenthesis matched to right parenthesis.)		_	...	...	_	... or _	_	see Table 6
e	ligand (all parenthesis matched)		...			...	...	...	see Table 6
f	possible branch point		|	|	|	|	|	+	see Table 6
g	reaction site (Code change site)		*	*	*	*	*	@	!@…Ma2
h	possible modification						$	$	A$GN
i	number		#				#	n	nA=0 nA>2
j	divide certainty and uncertainty (Table 2: UR6)	|					nothing	nothing	
k	omission of an entire SU (Table 2: UR3)	*					nothing	*	ANb3*N
l	glycosides (Table 2: GR1)	;, :, #		;			nothing	; for amino acid, : for lipid moieties, # for other glycosides	Ga;NY-S-C Gb:C
m	MS with uncommon stereospecificity and ring structure (Table 2: SRS3)	~					nothing	~	L-Glcf: G~

Abbreviations: OLC (Original Linear Code [31]), Kra (Krambeck et al. [10]), Spa (Spahn [13]), Lia (Liang et al. [14]), Hou (Hou et al. [7]).

entire SU. In theory, according to Banin's definition, a saccharide unit can be specified as " ??? ". Using " * " to indicate a complete SU, would avoid using an unmanageable number of question marks to represent an ambiguous glycan. Question marks should still be used to indicate unknown elements of an SU (e.g., "Ab4Gb" without knowledge of "b4G" could be written as "A???b"), but there should never be four adjacent question marks. We propose a substitute for the reaction site in Recommendation 7.

Recommendation 7 – "Reaction site." The reaction site is the location of the first change to the glycan expression. Because " * " is already defined within Linear Code to indicate "omission," we choose " @ " to indicate the reaction site. The reaction site, in previous reaction rules as " * " and going forward as " @ ", is the position of the first difference between product and substrate strings in the Linear Code form.

Recommendation 8 – "Modification." As specified in the original Linear Code, we recommend using " [] " to represent known modifications (Table 3: MR1). For example, "A[2P]" represents a galactose with its second position modified by a phosphate. However, this specific modification may not always be known. Therefore, in addition to " [] " as exact modifications, we recommended using the " $ " sign to represent a possible modification site. For example, "A$GN" represents a

GlcNAc connected to a galactose that might be modified. The modification can be specified (e.g., phosphorylation on the 2nd carbon) in the typical way, with square brackets "A$[2P]GN".

Recommendation 9 – "Branching index." In LiCoRR we have deprecated the original linear code branching rules due to redundancy and default to a version of BR1: Regardless of whether the MSs are equivalent, the MS linked to the higher index carbon will branch (appear first in the written representation when read right to left, reducing to non-reducing end). This rule can be extended to glycopeptides providing a means of representing glycans directly embedded in a glycopeptide. "PEP(Gal[3S]b3(GNb6)AN)TIDE" would describe a trisaccharide O-glycan bound to the threonine of an eponymously named glycoprotein.

Overall, the consensus in these representations centers around the foundational work of the original Linear Code paper [31] and Krambeck et al. [10]. We have simply highlighted gaps in clarity that have resulted in colloquially small but computationally important divergences throughout the literature.

Conclusion

The field of systems glycobiology is poised to tackle increasingly complex glycan synthesis problems owing to the advent of a number of enabling computational modeling technologies. Linear Code is used to represent reaction rules of glycan synthesis thereby bringing both human-readability and computer-parsability to the glycoinformatics space. The utility of Linear Code in glycoinformatics has been extended by the inclusion of new symbols, relations, and attributes that accommodate the challenge of specifying reaction rules. Yet various implementations conflict with each other and the original Linear Code. Here, we have delineated the various adaptations made to accommodate reaction rule representation, the discordance between various implementations, and proposed a consensus for future representations called LiCoRR.

The adoption of a common reaction rule representation would increase FAIR (Findable, Accessible, Interoperable, Reusable) standards [40] compliance in glycoinformatics which will have far-reaching implications. As demonstrated by WURCS, a deterministic exemplar of glycan representation that can be used as a database key, "findability" can be improved by unifying data with metadata. While not fully deterministic, LiCoRR is a predictable representation for reaction rules thereby findability search through data-metadata unification. Towards improving the findability of glycans through data-metadata unification, we provide a parser (gRegex, see Supporting Information File 1) and a context-free grammar which should facilitate integration into several formal-language compatible glycoinformatics tools including glycologue [41], glypy [42], glycome-db [43]; adoption LiCoRR or these wildcards in other glycan representations could shift glycan-database search from monosaccharide count to substructure class specification. While computational tools exist to compare XML-type models directly [44], the verbosity of the models can challenge comprehension. While less descriptive, succinct human-readable and understandable LiCoRR expressions provide an opportunity for a human observer to manually compare and consider two related models. Ideally, succinct, readable, and comprehensible reaction rules sets will be sufficiently standardized, like XML-type representations, so that they will be "interoperable" across multiple modeling software so that models can be "reused," reproduced, validated, and extended across labs. Toward encouraging the reuse of LiCoRR, we would like to acknowledge the trademark held by a former company, Glycominds Ltd. As our work is an extension and consolidation of novel development throughout the public domain, and we have no intent to exploit the trademark for financial gain, it is our understanding that we may publish freely and dedicate LiCoRR to the public domain under a CC-BY free-use with attribution license. Increased readability and FAIRness through clarifying the nomenclature will help advance glycoinformatics technologies by making possible cross-platform and multi-omics integration and interpretation; interoperability may be enhanced through a community-endorsed vocabulary.

We further hope that the symbols described in this work, specifically the wildcards, will be used in other glycan representations and applications beyond biosynthesis modeling. The definition of glycan classes can be useful for efficiently and unambiguously describing the key elements of large complex glycans while only communicating the central information. Adoption of these symbols, now well-defined symbols, by more popular representations, such as IUPAC, could increase both the flexibility and succinctness of those representations. We believe the utility of these wild-cards extends beyond biosynthesis modeling (Table 9) and may be useful in the description of glycan-chemosynthetic procedures, lectin identification of glycan motifs, and any other purpose where a group of glycans (rather than an individual glycan) is being discussed or described. We hope to encourage that adoption through our LiCoRRICE examples.

Increased FAIRness will facilitate the validation and distribution of developing glycoinformatics toolkits. Easy-to-use glycoinformatics toolkits, made possible by the fluency of interoperability across tools, are one mechanism by which glycobiology can be shared with the broader community of biology.

Supporting Information

Supporting Information File 1

Supporting tables.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-215-S1.zip]

Funding

This work was supported by generous funding from the Novo Nordisk Foundation provided to the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517) and from NIGMS (R35 GM119850). Additional funding was provided by the Triton Research & Experiential Learning Scholars (TRELS) program at the University of California, San Diego, the NIDDK Information Network (dkNET) supported by NIH & NIDDK (2U24DK097771-06), and NIH (NIH HL103411).

Authors

Benjamin P. Kellman[1,2,3], Yujie Zhang[1,2], Emma Logomasini[1,2], Eric Meinhardt[4], Karla P. Godinez-Macias[1], Austin W. T. Chiang[1,2], James T. Sorrentino[1,2,3], Chenguang Liang[1,2], Bokan Bao[1,2], Yusen Zhou[5], Sachiko Akase[6], Isami Sogabe[6], Thukaa Kouka[6], Elizabeth A Winzeler[1], Iain B. H. Wilson[7,8], Matthew P. Campbell[7,9], Sriram Neelamegham[5,7], Frederick J. Krambeck[7,10,11], Kiyoko F. Aoki-Kinoshita[6,7,12] and Nathan E. Lewis[*,1,2,3,7,13]

Addresses

[1]Department of Pediatrics, University of California San Diego School of Medicine, 9500 Gilman Dr, La Jolla, 92093, California, USA, [2]Department of Bioengineering, University of California San Diego School of Engineering, 9500 Gilman Dr, La Jolla, 92093, California, USA, [3]Bioinformatics and Systems Biology Program, University of California San Diego School of Engineering, 9500 Gilman Dr, La Jolla, 92093, California, USA, [4]Department of Linguistics, University of California San Diego, 9500 Gilman Dr, La Jolla, 92093, California, USA, [5]Department of Chemical and Biological Engineering School of Engineering and Applied Sciences, State University of New York, University at Buffalo, 303 Furnas Hall, Buffalo, 14260-4200, New York, USA, [6]Graduate School of Engineering, Soka University, 1-236 Tangi-machi, Hachioji-shi, Tokyo, Japan, [7]Systems Glycobiology Consortium, [8]Institute of Biochemistry, University of Natural Resources and Life Sciences, Gregor-Mendel-Straße 33, 1180 Vienna, Austria, [9]Institute for Glycomics, Griffith University, Glycomics 1, G26/1 Parklands Dr, Southport QLD 4215, Australia, [10]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, 21218, Maryland, USA, [11]ReacTech Inc., 810 Cameron St, Alexandria, 22314, Virginia, USA, [12]Faculty of Science and Engineering, Soka University, 1-236 Tangi-machi, Hachioji-shi, Tokyo, Japan, and [13]Novo Nordisk Foundation Center for Biosustainability at the University of California San Diego School of Medicine, 9500 Gilman Dr. La Jolla, 92093, California, USA.

ORCID® iDs

Benjamin P. Kellman - https://orcid.org/0000-0002-0780-6096
Yujie Zhang - https://orcid.org/0000-0002-1017-0316
Thukaa Kouka - https://orcid.org/0000-0002-4442-2294
Iain B. H. Wilson - https://orcid.org/0000-0001-8996-1518
Matthew P. Campbell - https://orcid.org/0000-0002-9525-792X
Frederick J. Krambeck - https://orcid.org/0000-0003-3838-3775
Nathan E. Lewis - https://orcid.org/0000-0001-7700-3654

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: doi:10.1101/2020.05.31.126623

References

1. Kontoravdi, C.; Jimenez del Val, I. *Curr. Opin. Chem. Eng.* **2018,** *22,* 89–97. doi:10.1016/j.coche.2018.08.007
2. Spahn, P. N.; Lewis, N. E. *Curr. Opin. Biotechnol.* **2014,** *30,* 218–224. doi:10.1016/j.copbio.2014.08.004
3. Puri, A.; Neelamegham, S. *Ann. Biomed. Eng.* **2012,** *40,* 816–827. doi:10.1007/s10439-011-0464-5
4. Krambeck, F. J.; Bennun, S. V.; Andersen, M. R.; Betenbaugh, M. J. *PLoS One* **2017,** *12,* e0175376. doi:10.1371/journal.pone.0175376
5. Sou, S. N.; Jedrzejewski, P. M.; Lee, K.; Sellick, C.; Polizzi, K. M.; Kontoravdi, C. *Biotechnol. Bioeng.* **2017,** *114,* 1570–1582. doi:10.1002/bit.26225
6. del Val, I. J.; Polizzi, K. M.; Kontoravdi, C. *Sci. Rep.* **2016,** *6,* 28547. doi:10.1038/srep28547
7. Hou, W.; Qiu, Y.; Hashimoto, N.; Ching, W.-K.; Aoki-Kinoshita, K. F. *BMC Bioinf.* **2016,** *17* (Suppl. 7), 240. doi:10.1186/s12859-016-1094-6
8. Liu, G.; Neelamegham, S. *Wiley Interdiscip. Rev.: Syst. Biol. Med.* **2015,** *7,* 163–181. doi:10.1002/wsbm.1296
9. McDonald, A. G.; Hayes, J. M.; Bezak, T.; Głuchowska, S. A.; Cosgrave, E. F. J.; Struwe, W. B.; Stroop, C. J. M.; Kok, H.; van de Laar, T.; Rudd, P. M.; Tipton, K. F.; Davey, G. P. *J. Cell Sci.* **2014,** *127,* 5014–5026. doi:10.1242/jcs.151878
10. Krambeck, F. J.; Bennun, S. V.; Narang, S.; Choi, S.; Yarema, K. J.; Betenbaugh, M. J. *Glycobiology* **2009,** *19,* 1163–1175. doi:10.1093/glycob/cwp081
11. Liu, G.; Marathe, D. D.; Matta, K. L.; Neelamegham, S. *Bioinformatics* **2008,** *24,* 2740–2747. doi:10.1093/bioinformatics/btn515
12. Liu, G.; Puri, A.; Neelamegham, S. *Bioinformatics* **2013,** *29,* 404–406. doi:10.1093/bioinformatics/bts703
13. Spahn, P. N.; Hansen, A. H.; Hansen, H. G.; Arnsdorf, J.; Kildegaard, H. F.; Lewis, N. E. *Metab. Eng.* **2016,** *33,* 52–66. doi:10.1016/j.ymben.2015.10.007
14. Liang, C.; Chiang, A. W. T.; Hansen, A. H.; Arnsdorf, J.; Schoffelen, S.; Sorrentino, J. T.; Kellman, B. P.; Bao, B.; Voldborg, B. G.; Lewis, N. E. *Curr. Res. Biotechnol.* **2020,** *2,* 22–36. doi:10.1016/j.crbiot.2020.01.001

15. Spahn, P. N.; Hansen, A. H.; Kol, S.; Voldborg, B. G.; Lewis, N. E. *Biotechnol. J.* **2017,** *12,* 1600489. doi:10.1002/biot.201600489
16. McDonald, A. G.; Tipton, K. F.; Davey, G. P. *PLoS Comput. Biol.* **2016,** *12,* e1004844. doi:10.1371/journal.pcbi.1004844
17. Bao, B.; Kellman, B. P.; Chiang, A. W. T.; York, A. K.; Mohammad, M. A.; Haymond, M. W.; Bode, L.; Lewis, N. E. *bioRxiv* **2019,** 693507. doi:10.1101/693507
18. Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; Varki, A.; Woods, R. J.; The SNFG Discussion Group. *Glycobiology* **2019,** *29,* 620–624. doi:10.1093/glycob/cwz045
19. Mehta, A. Y.; Cummings, R. D. *Bioinformatics* **2020,** *36,* 3613–3614. doi:10.1093/bioinformatics/btaa190
20. Cheng, K.; Zhou, Y.; Neelamegham, S. *Glycobiology* **2017,** *27,* 200–205. doi:10.1093/glycob/cww115
21. Demir, E.; Cary, M. P.; Paley, S.; Fukuda, K.; Lemer, C.; Vastrik, I.; Wu, G.; D'Eustachio, P.; Schaefer, C.; Luciano, J.; Schacherer, F.; Martinez-Flores, I.; Hu, Z.; Jimenez-Jacinto, V.; Joshi-Tope, G.; Kandasamy, K.; Lopez-Fuentes, A. C.; Mi, H.; Pichler, E.; Rodchenkov, I.; Splendiani, A.; Tkachev, S.; Zucker, J.; Gopinath, G.; Rajasimha, H.; Ramakrishnan, R.; Shah, I.; Syed, M.; Anwar, N.; Babur, Ö.; Blinov, M.; Brauner, E.; Corwin, D.; Donaldson, S.; Gibbons, F.; Goldberg, R.; Hornbeck, P.; Luna, A.; Murray-Rust, P.; Neumann, E.; Ruebenacker, O.; Samwald, M.; van Iersel, M.; Wimalaratne, S.; Allen, K.; Braun, B.; Whirl-Carrillo, M.; Cheung, K.-H.; Dahlquist, K.; Finney, A.; Gillespie, M.; Glass, E.; Gong, L.; Haw, R.; Honig, M.; Hubaut, O.; Kane, D.; Krupa, S.; Kutmon, M.; Leonard, J.; Marks, D.; Merberg, D.; Petri, V.; Pico, A.; Ravenscroft, D.; Ren, L.; Shah, N.; Sunshine, M.; Tang, R.; Whaley, R.; Letovksy, S.; Buetow, K. H.; Rzhetsky, A.; Schachter, V.; Sobral, B. S.; Dogrusoz, U.; McWeeney, S.; Aladjem, M.; Birney, E.; Collado-Vides, J.; Goto, S.; Hucka, M.; Le Novère, N.; Maltsev, N.; Pandey, A.; Thomas, P.; Wingender, E.; Karp, P. D.; Sander, C.; Bader, G. D. *Nat. Biotechnol.* **2010,** *28,* 935–942. doi:10.1038/nbt.1666
22. Lloyd, C. M.; Halstead, M. D. B.; Nielsen, P. F. *Prog. Biophys. Mol. Biol.* **2004,** *85,* 433–450. doi:10.1016/j.pbiomolbio.2004.01.004
23. Herget, S.; Ranzinger, R.; Maass, K.; Lieth, C.-W. v. d. *Carbohydr. Res.* **2008,** *343,* 2162–2171. doi:10.1016/j.carres.2008.03.011
24. Tanaka, K.; Aoki-Kinoshita, K. F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2014,** *54,* 1558–1566. doi:10.1021/ci400571e
25. Matsubara, M.; Aoki-Kinoshita, K. F.; Aoki, N. P.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2017,** *57,* 632–637. doi:10.1021/acs.jcim.6b00650
26. Aoki-Kinoshita, K.; Agravat, S.; Aoki, N. P.; Arpinar, S.; Cummings, R. D.; Fujita, A.; Fujita, N.; Hart, G. M.; Haslam, S. M.; Kawasaki, T.; Matsubara, M.; Moreman, K. W.; Okuda, S.; Pierce, M.; Ranzinger, R.; Shikanai, T.; Shinmachi, D.; Solovieva, E.; Suzuki, Y.; Tsuchiya, S.; Yamada, I.; York, W. S.; Zaia, J.; Narimatsu, H. *Nucleic Acids Res.* **2016,** *44,* D1237–D1242. doi:10.1093/nar/gkv1041
27. Campbell, M. P.; Peterson, R.; Mariethoz, J.; Gasteiger, E.; Akune, Y.; Aoki-Kinoshita, K. F.; Lisacek, F.; Packer, N. H. *Nucleic Acids Res.* **2014,** *42,* D215–D221. doi:10.1093/nar/gkt1128
28. York, W. S.; Mazumder, R.; Ranzinger, R.; Edwards, N.; Kahsay, R.; Aoki-Kinoshita, K. F.; Campbell, M. P.; Cummings, R. D.; Feizi, T.; Martin, M.; Natale, D. A.; Packer, N. H.; Woods, R. J.; Agarwal, G.; Arpinar, S.; Bhat, S.; Blake, J.; Castro, L. J. G.; Fochtman, B.; Gildersleeve, J.; Goldman, R.; Holmes, X.; Jain, V.; Kulkarni, S.; Mahadik, R.; Mehta, A.; Mousavi, R.; Nakarakommula, S.; Navelkar, R.; Pattabiraman, N.; Pierce, M. J.; Ross, K.; Vasudev, P.; Vora, J.; Williamson, T.; Zhang, W. *Glycobiology* **2020,** *30,* 72–73. doi:10.1093/glycob/cwz080
29. Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.; Packer, N. H.; Lisacek, F. *J. Proteome Res.* **2019,** *18,* 664–677. doi:10.1021/acs.jproteome.8b00766
30. Panico, R.; Richer, J.-C., Eds. *A Guide to IUPAC Nomenclature of Organic Compounds: Recommendations 1993;* Blackwell Science, Inc., 1993.
31. Banin, E.; Neuberger, Y.; Altshuler, Y.; Halevi, A.; Inbar, O.; Nir, D.; Dukler, A. *Trends Glycosci. Glycotechnol.* **2002,** *14,* 127–137. doi:10.4052/tigg.14.127
32. Bennun, S. V.; Yarema, K. J.; Betenbaugh, M. J.; Krambeck, F. J. *PLoS Comput. Biol.* **2013,** *9,* e1002813. doi:10.1371/journal.pcbi.1002813
33. Varki, A.; Sherman, W.; Kornfeld, S. *Arch. Biochem. Biophys.* **1983,** *222,* 145–149. doi:10.1016/0003-9861(83)90511-8
34. Rohrer, J.; Kornfeld, R. *Mol. Biol. Cell* **2001,** *12,* 1623–1631. doi:10.1091/mbc.12.6.1623
35. Hare, N. J.; Lee, L. Y.; Loke, I.; Britton, W. J.; Saunders, B. M.; Thaysen-Andersen, M. *J. Proteome Res.* **2017,** *16,* 247–263. doi:10.1021/acs.jproteome.6b00685
36. Röttger, S.; White, J.; Wandall, H. H.; Olivo, J. C.; Stark, A.; Bennett, E. P.; Whitehouse, C.; Berger, E. G.; Clausen, H.; Nilsson, T. *J. Cell Sci.* **1998,** *111,* 45–60.
37. Akune, Y.; Lin, C.-H.; Abrahams, J. L.; Zhang, J.; Packer, N. H.; Aoki-Kinoshita, K. F.; Campbell, M. P. *Carbohydr. Res.* **2016,** *431,* 56–63. doi:10.1016/j.carres.2016.05.012
38. McDonald, A. G.; Boyce, S.; Tipton, K. F. *Nucleic Acids Res.* **2009,** *37,* D593–D597. doi:10.1093/nar/gkn582
39. Agravat, S. B.; Song, X.; Rojsajjakul, T.; Cummings, R. D.; Smith, D. F. *Bioinformatics* **2016,** *32,* 1471–1478. doi:10.1093/bioinformatics/btw048
40. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. *Sci. Data* **2016,** *3,* 160018. doi:10.1038/sdata.2016.18
41. McDonald, A. G.; Tipton, K. F.; Stroop, C. J.; Davey, G. P. *BMC Res. Notes* **2010,** *3,* 173. doi:10.1186/1756-0500-3-173
42. Klein, J.; Zaia, J. *J. Proteome Res.* **2019,** *18,* 3532–3537. doi:10.1021/acs.jproteome.9b00367
43. Alocci, D.; Mariethoz, J.; Horlacher, O.; Bolleman, J. T.; Campbell, M. P.; Lisacek, F. *PLoS One* **2015,** *10,* e0144578. doi:10.1371/journal.pone.0144578
44. Scharm, M.; Wolkenhauer, O.; Waltemath, D. *Bioinformatics* **2016,** *32,* 563–570. doi:10.1093/bioinformatics/btv484

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Semiautomated glycoproteomics data analysis workflow for maximized glycopeptide identification and reliable quantification

Steffen Lippold, Arnoud H. de Ru, Jan Nouta, Peter A. van Veelen, Magnus Palmblad, Manfred Wuhrer and Noortje de Haan*

Full Research Paper

Open Access

Address:
Center for Proteomics and Metabolomics, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, Netherlands

Email:
Noortje de Haan* - ndehaan@sund.ku.dk

* Corresponding author

Keywords:
bioinformatics; cysteine oxidation; glycoproteomics; immunoglobulins; mass spectrometry

Beilstein J. Org. Chem. **2020,** *16,* 3038–3051.
https://doi.org/10.3762/bjoc.16.253

Received: 19 July 2020
Accepted: 23 November 2020
Published: 11 December 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: N. H. Packer

Abstract

Glycoproteomic data are often very complex, reflecting the high structural diversity of peptide and glycan portions. The use of glycopeptide-centered glycoproteomics by mass spectrometry is rapidly evolving in many research areas, leading to a demand in reliable data analysis tools. In recent years, several bioinformatic tools were developed to facilitate and improve both the identification and quantification of glycopeptides. Here, a selection of these tools was combined and evaluated with the aim of establishing a robust glycopeptide detection and quantification workflow targeting enriched glycoproteins. For this purpose, a tryptic digest from affinity-purified immunoglobulins G and A was analyzed on a nano-reversed-phase liquid chromatography–tandem mass spectrometry platform with a high-resolution mass analyzer and higher-energy collisional dissociation fragmentation. Initial glycopeptide identification based on MS/MS data was aided by the Byonic software. Additional MS1-based glycopeptide identification relying on accurate mass and retention time differences using GlycopeptideGraphMS considerably expanded the set of confidently annotated glycopeptides. For glycopeptide quantification, the performance of LaCyTools was compared to Skyline, and GlycopeptideGraphMS. All quantification packages resulted in comparable glycosylation profiles but featured differences in terms of robustness and data quality control. Partial cysteine oxidation was identified as an unexpectedly abundant peptide modification and impaired the automated processing of several IgA glycopeptides. Finally, this study presents a semiautomated workflow for reliable glycoproteomic data analysis by the combination of software packages for MS/MS- and MS1-based glycopeptide identification as well as the integration of analyte quality control and quantification.

Introduction

Protein glycosylation mainly occurs in the form of *N*- and *O*-glycosylation. *N*-Glycans are attached to Asn within an amino acid consensus sequence (Asn-Xxx-Ser/Thr, Xxx ≠ Pro) and *O*-glycans are attached to Ser or Thr. Glycan compositions can range from monosaccharides (e.g., Tn antigen for *O*-glycans [1]) to large polysaccharides (e.g., *N*-glycans of recombinant human erythropoietin [2]). The most common building blocks of human protein glycans are hexoses (glucose, galactose, and mannose, Hex/H, 162.0528 Da), *N*-Acetylhexosamines (*N*-acetylglucosamine or *N*-acetylgalactosamine, HexNAc/N, 203.0794 Da), fucose (Fuc/F, 145.0579 Da), and sialic acid (*N*-acetylneuraminic acid, NeuAc/S, 291.0954 Da). The combinatorial possibilities of these building blocks and the variety of structural features, such as the linkage position and anomeric configuration, make protein glycosylation a highly complex posttranslational modification (PTM).

Glycoproteomics has become important for many life science disciplines, in particular for biomedical and biopharmaceutical research [3-5]. Glycopeptide-centered glycoproteomics aims at the characterization of macroheterogeneity and microheterogeneity of protein glycosylation [6]. Reversed-phase liquid chromatography coupled to high-resolution tandem mass spectrometry (RPLC–MS/MS) is a standard analytical method in the field of glycoproteomics [7]. The separation of glycopeptides in RPLC is mainly driven by the peptide portions. Thus, information on different proteins and glycosylation sites appears in the form of glycopeptide clusters. Next to the peptide portion, glycosylation features, such as sialic acids, can strongly influence the retention time [8]. Advances in MS technologies tremendously enhanced the detection and informative fragmentation of glycopeptides in the past years [9]. The large amount of highly complex data acquired using these technologies shifted the major bottleneck in glycopeptide analysis to the data processing steps. Next to the high complexity of glycosylation itself, data analysis is further complicated by interfering background signals from biological matrices and isomeric and near-isobaric ambiguities resulting from combinations of monosaccharides, adducts, amino acids, and amino acid modifications [10,11].

Efforts have been made in recent years in the development of bioinformatic tools to facilitate and automate data processing in glycopeptide-centered glycoproteomics [12]. Several reports have reviewed the functionalities and application areas of data analysis tools in the field of glycoproteomics [7,9,12,13]. MS/MS-based scoring software tools such as Byonic [14] are frequently used for glycopeptide identification [12]. Recently, software tools were developed that are based on the retention time (RT) characteristics and accurate mass differences of glycopeptide MS1 signals in RPLC–MS [10,15]. These tools detect inaccuracies of MS/MS assignments based on the RT and increase the number of identified glycopeptide compositions while keeping the false positive assignments low. Other reports performed glycopeptide identification using summed MS1 spectra of previously defined elution clusters [16]. This approach is applicable when the identity and elution behavior of the glycopeptides of interest is known and is aided by quality criteria such as mass accuracy and isotopic pattern matching. Furthermore, such approaches allow quantification in a high-throughput manner, which is advantageous e.g., in clinal cohort analysis [16-18].

Here, we present a workflow for the reliable and efficient analysis of glycopeptides from enriched glycoproteins. We performed a thorough evaluation of the software tools and workflows used in our laboratory for the identification and quantification of glycopeptides. For this, a sample containing immunoglobulins G and A (IgG and IgA), simultaneously captured from human plasma, was chosen. This sample showed a considerable level of complexity due to the presence of multiple glycoproteins of interest and cocaptured (glyco)proteins from the plasma. The tools included Byonic, GlycopeptideGraphMS, Skyline, and LaCyTools.

Results and Discussion

Glycoproteomics data analysis workflow

Affinity-copurified IgG/IgA from human plasma was chosen as a sample to demonstrate the integration of tools for the semiautomated glycoproteomic data analysis (Figure 1). The three main parts of this workflow cover glycopeptide identification (Byonic, GlycopeptideGraphMS), curation, and quantification (LaCyTools).

In the first step, Byonic was used for automated MS/MS-based (glyco)peptide identification. This initial step is crucial to validate the presence of glycopeptides and the assignment of the peptide portions. Next, the number of identified glycopeptides was maximized by performing an open search based on MS1 information (mass and RT) in GlycopeptideGraphMS. A preprocessing step in OpenMS was performed as described for the original GlycopeptideGraphMS workflow [15], including deisotoping and decharging of all features. The outcome of GlycopeptideGraphMS is a list with glycopeptide clusters (defined as LC–MS features (nodes) that are connected by Δmass and ΔRT within the provided limits for glycopeptides), for which at least one node should be confidently assigned by MS/MS to identify all glycopeptides in a cluster. The clusters are also presented in interactive graphs, which assist in the identification of false-positive connections (unlikely mass/RT shifts)

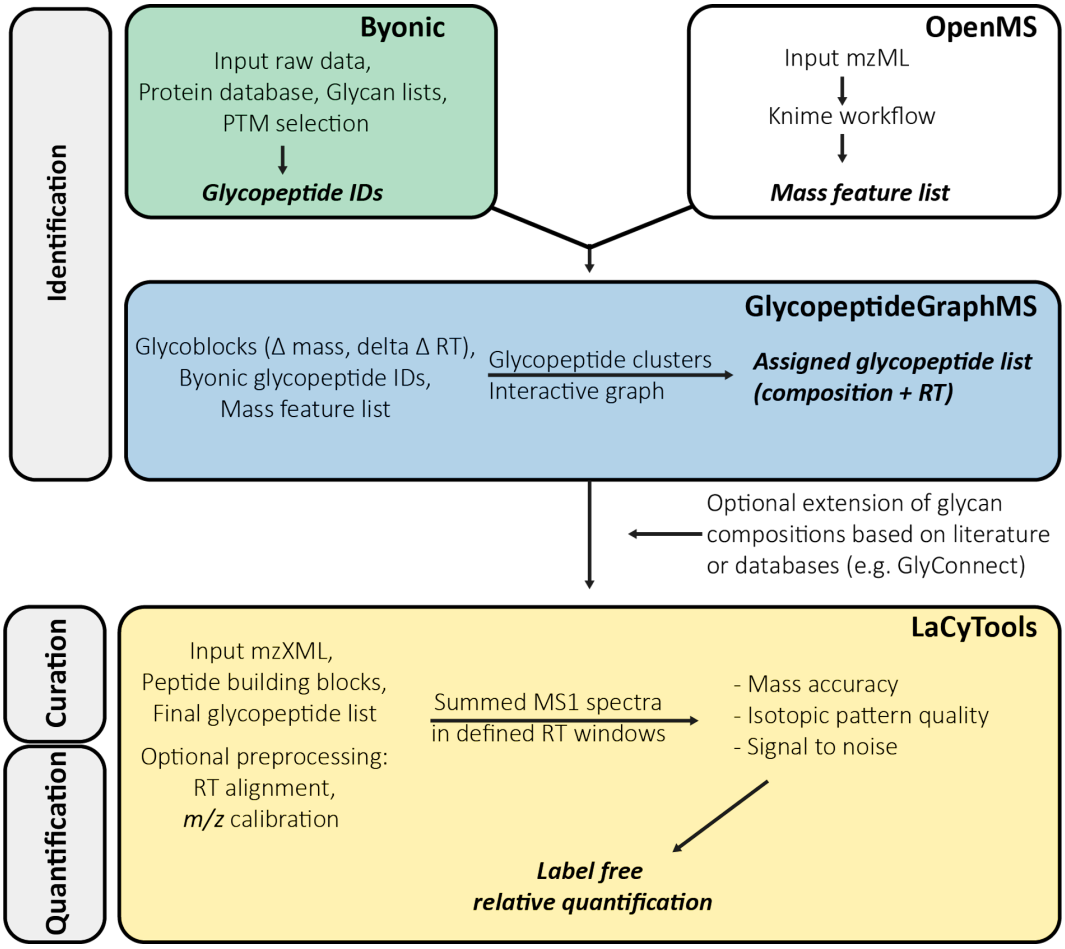


Figure 1: Integration of automated glycopeptide identification by Byonic and GlycopeptideGraphMS (aided by OpenMS) and subsequent analyte quality control and quantification by LaCyTools.

and unexpected glycopeptide clusters (e.g., missed cleaved products and peptide or glycan modifications). This information can be used in an iterative manner to adjust the search space for Byonic. Study-specific search criteria are listed in the Experimental section and a detailed manual for the use of GlycopeptideGraphMS can be found elsewhere [15]. Of note, separate LC–MS runs with exclusively MS1 information were acquired in order to maximize the MS1-based identification and to ensure the highest possible data quality for the quantification purposes.

Upon glycopeptide identification, the list of glycopeptides generated by GlycopeptideGraphMS was transformed to the input format required for targeted curation and quantification in LaCyTools [16]. A python script was developed to facilitate this step (Supporting Information File 3). LaCyTools was chosen because it is open-source, can be applied for a large number of samples (thousands of samples in one study have been reported [19]), and allows data curation and quantification. Importantly, LaCyTools requires RT clusters to be defined in which MS1 spectra can be summed and further processed, which is facilitated by the GlycopeptideGraphMS output. The analyte list may be extended by including glycan compositions (e.g., from the literature or databases such as GlyConnect [20]) within appropriate RT clusters (e.g., the same peptide portion and number of sialic acids). Furthermore, the user has the option to perform preprocessing steps, such as *m*/*z* calibration and RT alignment. For data curation, summed MS1 spectra were subjected to quality control based on user-defined cut-offs for mass accuracy, isotopic pattern matching, and the signal-to-noise ratio of an analyte. Finally, the integrated areas of all charge states passing the quality criteria were summed for each glycopeptide composition, the area was corrected for missing isotopes, and total area normalization was performed for label-free relative quantification. Study-specific parameters for the use of LaCyTools are provided in the Experimental section and further explanation on the use of this tool can be found elsewhere [21].

Glycopeptide identification

Automated MS/MS-based glycopeptide identification by Byonic

The automated and score-based MS/MS glycopeptide identification using Byonic resulted in the confident assignment of ten IgG/IgA *N*-glycopeptide clusters of interest (Table 1 and Figures S1–S10, Supporting Information File 2).

Assigned glycopeptides from copurified human plasma proteins other than IgG and IgA were not considered for further data processing (e.g., fibrinogen, alpha-1-antitrypsin, or clusterin, see Table S1A–E, Supporting Information File 1). Missed-cleavage variants were assigned for IgG1, IgG2/3, and IgA1/2 (Asn263) but not further considered because of their low abundance. For the IgA joining chain (JC), the elongated peptide with a missed cleavage was included for further data processing as the cleavage efficiency was previously determined to be glycoform dependent [18,22]. For the assignment of tryptic *N*-glycopeptides to specific proteins, ambiguities exist for one peptide moiety that could be assigned to either IgG2 or IgG3 and three moieties that were shared between IgA1 and IgA2 (Table 1) [3]. These ambiguities were not resolved using the proposed workflow. However, the presence of protein-specific (non)glycopeptides may indicate differences in the abundance of the individual proteins. For addressing these ambiguities, a more selective sample preparation is required, for example, using different enrichment strategies or proteases [23]. Interestingly, an additional allotype of the main IgG3 glycosylation site (EEQYNSTFR) was assigned in four out of five technical replicates by Byonic. This IgG3 glycopeptide is an isomer of the tryptic IgG4 glycopeptide (EEQFNSTYR). However, upon manual inspection of the data, only one scan of the assigned MS/MS spectra within all five technical replicates covered the relevant amino acids (position of Phe and Tyr), allowing an unambiguous discrimination between IgG3 or IgG4 (score 281, Figure S11, Supporting Information File 2). The IgA2 HYT glycopeptides had the lowest scores (max. 194) compared to the other glycosylation sites. It was detected in four out of five technical replicates and only with a maximum of one glycan composition. The low intensity of these glycopeptide signals resulted in a decreased likelihood for MS/MS selection. Of note, the IgA2 HYT glycopeptide covers a sequence stretch homologous to the hinge region of IgA1, carrying *O*-glycans. In a previous study the IgA1 peptide has been referred to as the HYT glycopeptide cluster as well [17]. The C-terminal IgA1/2 glycopeptides (LAGC/Y) were found mainly with methionine oxidation. Unoxidized peptide moieties were also assigned but with low scores (below 50). The manual check of the data revealed that in some cases, the selection of the wrong monoisotopic mass in Byonic led to misassignments of near-isobaric compositions, e.g., TPL H5N5F3 (3+, *m*/*z* 1074.8020, false) instead of TPL H5N5F1S1 (3+, *m*/*z* 1074.4619, correct). Other theoretical possible, but less common, tryptic IgG3, IgA1, and IgA2 glycopeptides were not detected [3,17]. One of the reported common miscleaved IgA2 *N*-glycopeptides (SESGQNVTAR) is likely to elute prior to MS acquisition as described previously for the applied gradient [17]. For the expected IgA1 *O*-glycopeptide cluster, the Byonic search failed to score any hits when performed as described previously [17]. Of note, the tryptic *O*-glycopeptide cluster could be detected upon manual inspection, albeit with low intensity (Figures S12 and S13, Supporting Information File 2). The reason for this was further investigated based on the GlycopeptideGraphMS results

Table 1: Automated MS/MS-based identification of IgG/IgA glycosylation sites by Byonic. For each glycopeptide moiety, a representative glycoform is shown (see Figures S1–S10, Supporting Information File 2 for the corresponding MS/MS spectra).

Protein	Glycopeptide	Glycosylation site[a]	Cluster	Mass error (ppm)	Score	Scan time (min)
IgG1	R.EEQYN[+H5N4F1]STYR.V	Asn297	IgG1	0.7	589	14.4
IgG2/3	R.EEQFN[+H3N4F1]STFR.V	Asn297	IgG2/3	0	693	18.5
IgG4	R.EEQFN[+H3N4F1]STYR.V	Asn297	IgG4	1.1	401	15.8
IgA1/2	R.LSLHRPALEDLLLGSEAN[+H5N4S1]LTC[+57]TLTGLR.D	Asn263	LSL	0.9	839	40.2
	R.LAGKPTHVN[+H5N5F1S2]VSVVM[+16]AEVDGTC[+57]Y.-[b]	Asn459	LAGY	0.4	601	25.5
	R.LAGKPTHVN[+H5N5F1S2]VSVVM[+16]AEVDGTC[+57].-[b]	Asn459	LAGC	2.9	649	25.9
IgA2	K.TPLTAN[+H5N4F1S1]ITK.S	Asn337	TPL	−1.2	728	19.1
	K.HYTN[+H5N5F1S1]SSQDVTVPC[+57]R.V	Asn211	HYT	1.3	194	15.6
JC	R.EN[+H5N4S2]ISDPTSPLR.T	Asn49	ENI	0.1	565	22.2
	R.IIVPLNNREN[+H5N4F1S1]ISDPTSPLR.T	Asn49	IIV	1.2	271	28.0

[a]Numbering according to [18]. [b]C-terminal peptide of the heavy chain, no C-terminal tryptic cleavage.

and is discussed in the section on automated MS1- and RT-based glycopeptide identification by GlycopeptideGraphMS.

In MS/MS scoring approaches such as Byonic, the definition of a threshold for the automated assignment of glycopeptides is generally a challenge as the scores depend largely on the fragmentation method, the peptide characteristics (e.g., peptide length or additional modifications), the glycome, and the sample matrix [11]. A recent study by us applied a threshold score of 200 for the IgG/IgA glycopeptides from human serum, aiming to find a balance between the exclusion of false positives while preventing false negatives [17]. Sensitive glycopeptide assignments relying only on oxonium ions and precursor mass, using a score above 30 were also described recently [15]. A suitable cut-off score should always be carefully evaluated for each (glyco)peptide moiety with respect to the glycoform coverage and accuracy [11].

Byonic identified the relevant *N*-glycosylation sites of IgG/IgA in all five technical replicates with the exception of the low-abundant IgA2 HYT glycopeptide. Further results and discussion of the accuracy and coverage of the investigated glycopeptides of interest are presented in the following section. Software tools for automated MS/MS-based assignments such as Byonic are highly useful in glycoproteomic data processing workflows. Other, noncommercial, automated MS/MS-based software tools for glycopeptide identification were recently reviewed [12] and have the potential to substitute Byonic in similar workflows as described here. However, these tools were not evaluated in the current study.

Automated MS1- and RT-based glycopeptide identification by GlycopeptideGraphMS

The glycopeptide identification was further extended by an open MS1 search based on mass and RT differences using GlycopeptideGraphMS [15]. RT clusters for all MS/MS assigned IgG/IgA glycopeptides were found using this tool (Figure 2). Of note, GlycopeptideGraphMS relies on the MS/MS assignment of at least one glycopeptide per RT cluster (be it automated or manual). The GlycopeptideGraphMS cluster with the highest number of connections contained the expected masses of the IgA1 *O*-glycopeptides, which were not assigned in the Byonic search (Figure S12, Supporting Information File 2). In line with the Byonic search, several other RT clusters of missed cleaved products or glycopeptides from other plasma proteins were present (data not shown).

Additional clusters with a +27.9949 Da (formylation) mass shift and an increased RT were observed for most of the IgG and IgA glycopeptides (see Figure S14, Supporting Information File 2 for representative IgG glycopeptide examples). The formylation was conveniently assigned to the glycan part (Figure S15, Supporting Information File 2) but may occur at the peptide portion as well [24-26]. Formylation is likely introduced by the exposure of the tryptic peptides to formic acid during the acid precipitation of sodium deoxycholate in the final step of the sample preparation and during subsequent storage [24]. Within the glycopeptide clusters of interest, Cys oxidation (+15.9949 Da) was assigned as an unexpected modification in all Cys-containing glycopeptides (five out of 11) at a high relative abundance (65.4–77.2%) and confirmed upon manual inspection of the MS/MS data (Figures S13 and S16–S18, Sup-

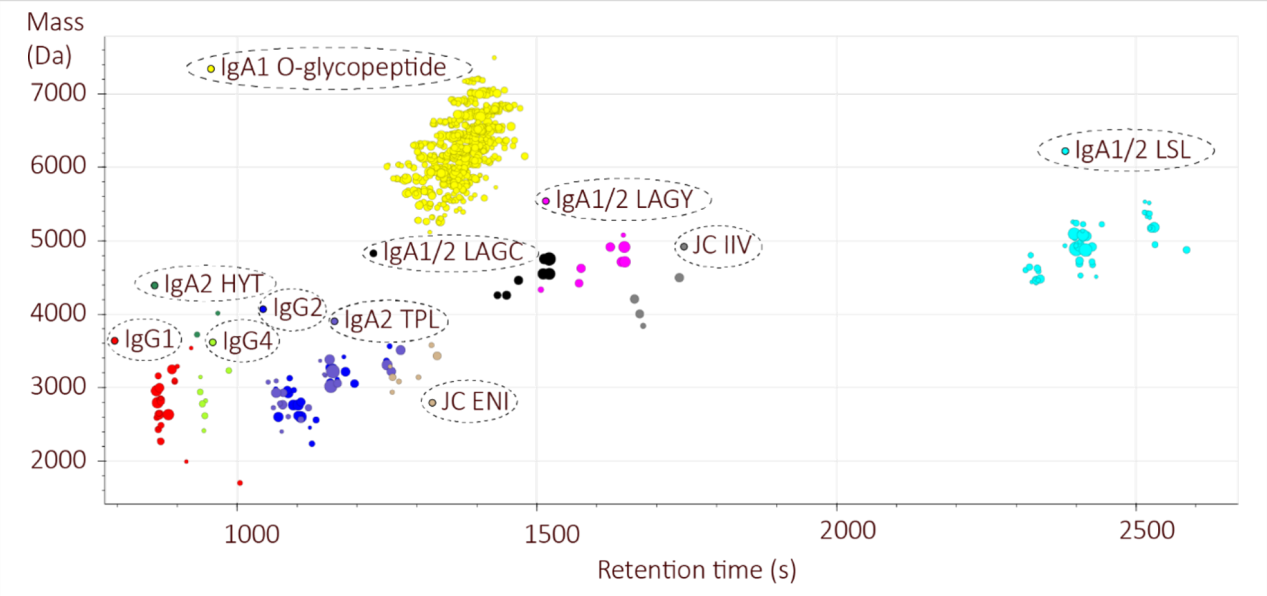


Figure 2: Representative IgG and IgA glycopeptide clusters detected by GlycopeptideGraphMS.

porting Information File 2). The y- and Y-fragment ions of (glyco)peptides with Cys oxidation showed a characteristic neutral loss of 107.0041 Da ($C_2H_5O_2NS$), as reported for singly oxidized carbamidomethylated Cys through an elimination reaction in the gas phase (Figures S13 and S16–S18, Supporting Information File 2) [27]. Peptides with Cys oxidation had a similar elution behavior as the unoxidized isomeric counterparts (with an additional hexose instead of a fucose unit), leading to a high degree of ambiguous, albeit often illogical compositions (e.g.. for the LSL cluster, Figure S16, Supporting Information File 2 and Table S2A–E, Supporting Information File 1) and false-positive assignments (e.g., the LAGC cluster, Figure S17, Supporting Information File 2) in GlycopeptideGraphMS. In line with these findings, the high number of illogical compositions and false-positive assignments of the IgA1 *O*-glycopeptide (three Cys residues) were due to modification variants on the Cys residue (Figures S12 and S13, Supporting Information File 2). In general, the assignment based on RT differences and MS1 information (manual or automated) had a highly increased uncertainty for the glycopeptides with partial Cys oxidation, and MS/MS was essential for confident identification in these cases. Of note, false-positive assignments related to Cys oxidation were also observed in the automated Byonic search upon manually reevaluation. For example, the LAGC glycopeptide composition H6N5S2 had a maximum score of 282, with no coverage of y-ions (Figure S17, Supporting Information File 2). This was due to the presence of the oxidized Cys residue at the C-terminus for which characteristic y- and Y-ions could be manually assigned in this scan. These findings substantiate that the scores in automated MS/MS searches may be still relatively high for false-positive assignments. Defining the appropriate search space with prior knowledge on relevant modifications and neutral losses is crucial to increase the identification accuracy for (glyco)peptides with unexpected modifications, such as Cys oxidation. The oxidation of Cys can appear biologically in the sample or artificially during/upon sample preparation [27,28]. In general, Met modifications are known for causing ambiguities in glycoproteomics due to partial oxidation, particularly in combination with carbamidomethylation [11,29]. To our knowledge, no study has previously reported on partial Cys oxidation as a confounder in glycoproteomics. As peptides containing the Cys oxidation had a higher abundance than the unoxidized counterparts, it is stressed that this modification should be carefully checked in Cys-containing glycopeptides as in the investigated sample, it had major implications on the IgA glycoprofiling accuracy. Further elaboration of the Cys-containing peptides, including modifications and correct glycan composition identifications, were considered beyond the scope of this study due to the largely increased complexity. Hence, the applicability of the proposed glycoproteomic data analysis workflow was demonstrated on a subset of six *N*-glycopeptide clusters, namely IgG1, IgG2/3, IgG4, JC (ENI, IIV), and IgA2 (TPL).

For the six glycopeptide clusters of interest, the presence of 262 theoretical glycopeptides (based on the internal IgG/IgA glycan reference list [17] and Glyconnect entries for these peptides [20], Table S3, Supporting Information File 1) was manually evaluated in Skyline, and the presence of 83 glycopeptides in the used data was confirmed (Table S4, Supporting Information File 1). In total, 82 correct glycopeptide compositions were identified using GlycopeptideGraphMS with MS/MS validation, whereas the Byonic-only search resulted in 35 compositions (Table S4, Supporting Information File 1). Of note, four glycan compositions (H2N3F1, H2N4F1, H5N3F1S1, H5N5F2S1) were not included in the *N*-glycan search list of Byonic, and hence not included for the calculation of its glycopeptide coverage. Those glycans were only present in low abundance on the glycopeptides, and often no MS/MS spectrum was present (Figure 3). However, it highlights the importance of a complete glycan composition list for a database-based identification of glycopeptides, something that is less critical in MS1-based RT and accurate-mass-difference searches.

In the GlycopeptideGraphMS search, nine compositions were detected that were not within the internal IgG/IgA glycan reference list [17] or had an entry in Glyconnect for these peptides (Table S4, Supporting Information File 1) [20]. These analytes were present at very low relative abundances (<1%). The IgA2 TPL peptide showed the highest number (five) of additional compositions (H6N5F1, H5N5, H5N4F2S1, H5N5F2S1, and H6N4F1S1). For all glycoforms identified by GlycopeptideGraphMS, only one composition (TPL, H4N4F2S1) was determined as false-positive as no MS1 signals could be found for this analyte in the raw data. Of note, one TPL glycoform (H5N5S1) was detected with a low abundance in three out of five technical replicates but was excluded from the identification and further processing due to the presence of isobaric MS signals in the raw data. On the other hand, only one glycopeptide (IgG2/3 H3N3F1) was assigned manually, without being identified by GlycopeptideGraphMS as the correct mass and RT combination was not in the deconvoluted mass list. These false-positive and false-negative results are artifacts of the feature recognition, deconvolution, and deisotoping in OpenMS [15] prior to the GlycopeptideGraphMS analysis. Furthermore, the preprocessing steps caused some glycopeptides to be detected at multiple RT values (Figure 3), whereas the raw data showed only a single chromatographic peak. The applied OpenMS workflow has a reported accuracy of 91% for detecting the correct monoisotopic peak of a feature, and this workflow was not further optimized in the current study [15].

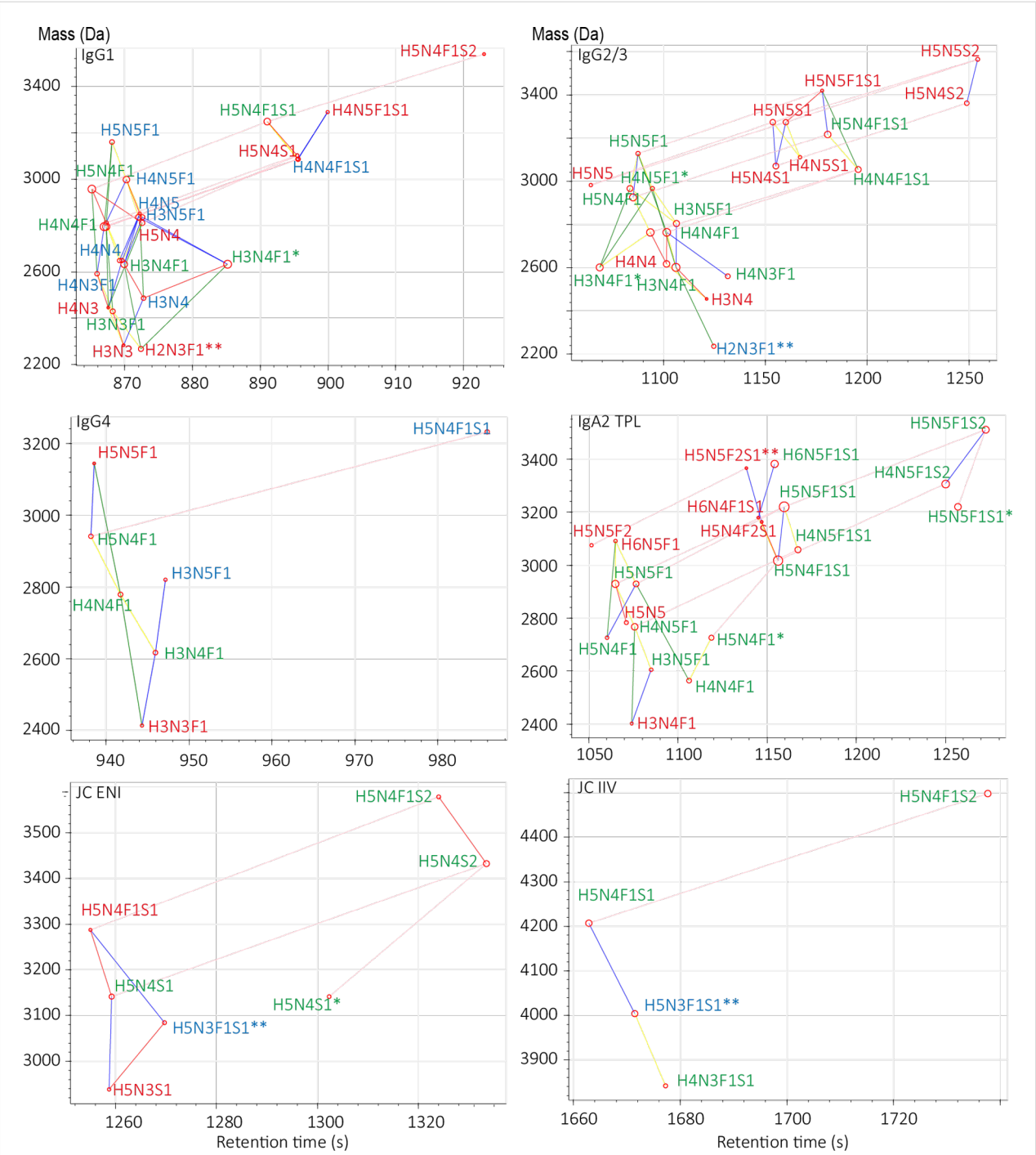


Figure 3: Representative GlycopeptideGraphMS output for peptides of interest. Assigned compositions were identified using MS/MS data via Byonic (green) or manual assignment (blue) or by MS1 only (red, GlycopeptideGraphMS with additional accurate-mass and isotopic pattern check of the raw data). The assignment of the compositions is based on information from all replicates. Lines between compositions indicate the mass difference for Hex (yellow), HexNAc (blue), HexHexNAc (green), Fuc (red), and NeuAc (purple). * Indicates potential deconvolution errors and ** indicates data not included in the Byonic search list.

With respect to the consistency of the glycopeptide identification in technical replicates, the MS1-based identification (GlycopeptideGraphMS) supported by MS/MS data showed a better performance than MS/MS identification alone (47 vs 16 glycopeptides detected in all replicates, respectively, Figure S19, Supporting Information File 2). Both automated identification approaches showed variations within the data of the technical replicates, and the glycopeptide coverage was maximized

by combining all measurements. For the MS1-based assignment, the variation between replicates was found in the minor glycan species, which were on the borderline of the limit of detection. Further, the stochastic nature of MS/MS selection is a known factor, which may cause variability in MS/MS-based assignments [15].

Overall, the GlycopeptideGraphMS workflow showed a high identification accuracy (82/83, 99%) and coverage (82/83, 99%, Table S4, Supporting Information File 1). In comparison, the accuracy of the Byonic search for the glycopeptides of interest was comparably high (35/37, 95%), whereas the glycopeptide coverage was moderate (35/77, 45%). This is in line with the reported near-perfect accuracy and limited coverage of the glycopeptide identification by Byonic [11]. Of note, the glycopeptide coverage of Byonic depends highly on the search parameters, fragmentation settings, and the presence and quality of MS/MS spectra. The latter is often compromised due to dynamic range limitations, especially in complex matrices [11,30]. The accuracy of both approaches (MS1 and MS/MS) may be impaired by unexpected peptide modifications, as exemplified for Cys oxidation. Thus, careful inspection of the result outputs (RT graphs in GlycopeptideGraphMS, automatically annotated MS/MS spectra in Byonic) is important. Indications of additional peptide modifications can then be considered for manual MS/MS verification and be included in the search space of automated MS/MS assignments in an iterative manner. Alternatively, a prior open search aimed at the identification of peptide modifications may be applied by software tools such as Preview [31]. Overall, this data shows that, while MS/MS-based assignment tools are essential for the confident identification of glycopeptide clusters, MS1-based approaches show a highly complementary performance by identifying glycopeptides for which no MS/MS data is present. For the latter, GlycopeptideGraphMS is a highly valuable tool as it is easy to use, fast, and open source.

Glycopeptide curation and quantification in LaCyTools

Upon glycopeptide identification, the analytes were curated and quantified by LaCyTools. The performance of LaCyTools was compared to that of Skyline (manual curation and quantification) and GlycopeptideGraphMS (quantification). The analytes and charge states passing the quality criteria (for LaCyTools: m/z accuracy <10 ppm, isotopic pattern quality value <0.2, signal-to-noise ratio >9; for Skyline: m/z accuracy <10 ppm, idotp >0.85) were highly similar between LaCyTools and Skyline (Table S5, Supporting Information File 1). Minor differences were observed for low-abundant glycopeptides. In GlycopeptideGraphMS, quality control is only based on mass accuracy and not included in this comparison.

The three software tools evaluated for targeted glycoform quantification resulted in comparable site-specific glycosylation profiles for human plasma IgG, JC, and IgA2 (Figure 4 and Table S5, Supporting Information File 1), which were in line with the literature (Table S6, Supporting Information File 1) [17,18]. Skyline and LaCyTools showed the highest similarity in the relative quantification results (Figure S20, Supporting Information File 2). Both tools had a median relative standard deviation (RSD) of 4% over all quantified glycopeptides. In contrast, GlycopeptideGraphMS integration resulted in a higher variability (median RSD: 15%, Figure S20, Supporting Information File 2) and slightly deviating glycosylation profiles, as compared to Skyline and LacyTools. As the data used for quantification were the same, the differences in the quantification precision are caused by the data processing performed by the different software tools. Of note, the automated quantification in GlycopeptideGraphMS required additional manual interference for analytes that had multiple RTs in the output file and only a single chromatographic peak in the raw data. Similar as for the glycopeptide identification, quantification with GlycopeptideGraphMS showed clearly that the preprocessing of the data is a crucial factor for the outcome. Further optimization of the OpenMS preprocessing steps to prevent double feature assignments may improve the quantification precision.

Within the investigated quantification tools, Skyline allows the highest control of the feature selection for quantification as the integrated EICs can be manually inspected for interferences, correct peak integration, and quality criteria (mass accuracy and isotopic pattern). LaCyTools provides information on the mass accuracy and isotopic pattern and integrates the isotopes of selected features in summed MS spectra within user-defined RT windows. Here, it is crucial to select appropriate RT windows and isotopes of interest before starting the analysis to prevent the inclusion of closely eluting isomeric and isobaric interferences. Of note, isomeric glycopeptide compositions were summed and not processed individually. This approach makes RT alignment a crucial step for a robust quantification. With the optimized parameters in place, LaCyTools allows highly automated data handling, making it an excellent tool for, e.g., clinical cohort analysis. In the current work, a python script was developed to streamline the connection between GlycopeptideGraphMS identification and LaCyTools quantification (Supporting Information File 3). All tools provided absolute values for glycopeptide quantification, which were subsequently total-area-normalized per glycosylation site, as commonly done in label-free relative quantification in glycoproteomics [16,17,30,32].

In the current study, all quantitative analyses were performed on the MS1-only runs to obtain the highest possible data

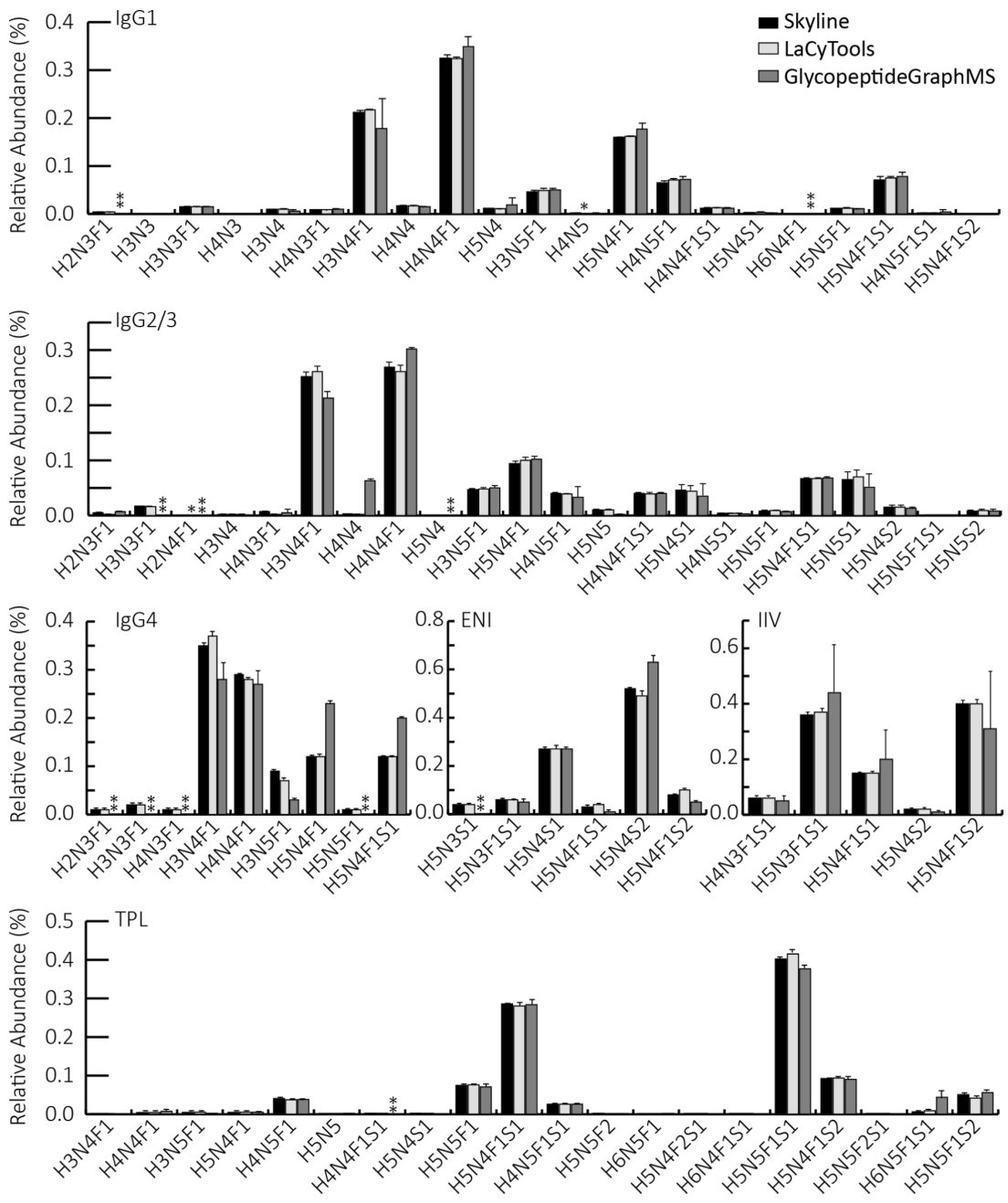


Figure 4: Comparison of quantification results obtained by manual integration of EICs in Skyline (black), automated integration of summed MS spectra in LaCyTools (light gray), and GlycopeptideGraphMS (dark gray). Error bars represent standard deviation of MS1-only measurements (n = 4 for LacyTools and Skyline; n = 3/4 for GlycopeptideGraphMS; in all detected replicates, n was at least 3. The first injection was excluded for all tools due to RT shifts and increased standard deviations). *: Did not pass the analyte curation (LaCyTools). **: Was not identified in at least 3 technical replicates (GlycopeptideGraphMS).

quality. However, runs including fragmentation scans are also suitable for quantification, albeit introducing a slightly higher variability in some cases due to a lower number of data points per chromatographic feature (in particular obvious for the IgG1 and IgG2/3 data in the current study, see Figure S21, Supporting Information File 2). The difference in the quantification accuracy between MS1-only and MS/MS data is highly dependent on the frequency of the MS1 scans, and thus the time spent on fragmentation scans. In most situations, it is likely that a compromise must be made to allow both robust quantification and data-rich MS/MS identification in the same LC–MS run. The introduction of MS1-based identification reduces the time needed for fragmentation.

Conclusion

Here, we demonstrated a semiautomated glycoproteomics data analysis workflow for enriched glycoproteins by integrating different tools for glycopeptide identification, curation, and quantification after RPLC separation and MS(/MS) detection. For this, a mix of the human plasma-enriched antibodies IgG and IgA was used as a representative glycoproteomics sample of moderate complexity. A similar approach can be applied to a more complex sample when targeting only a select set of glycoproteins. However, to capture the full complexity of, e.g., the human glycoproteome, improvements should be made in the automated integration between the described tools. In line with previous reports on single glycoproteins, the number of identified glycoforms was significantly maximized by combining MS1-based identification (using GlycopeptideGraphMS) in combination with MS/MS-based identification (using Byonic) as compared to fragmentation-based analysis alone. Moreover, the graphical approach allowed by GlycopeptideGraphMS is very powerful for identifying unexpected glycoforms as well as modifications of the glycopeptides and aids the optimization of the search space for MS/MS annotation in an iterative manner. Although an MS1-based approach alone allows the identification of more unique glycopeptides as compared to an MS/MS-based approach, a combined workflow is essential to prevent wrongly assigned glycopeptides as well as to identify the nature of specific modifications. The combination of Byonic and GlycopeptideGraphMS identification with LaCyTools-based curation and quantification of glycopeptides from enriched glycoproteins as presented in the current work provides a powerful workflow towards high-throughput glycopeptide analysis.

Experimental

Sample, chemicals, and enzymes

Human plasma Visucon-F was obtained from Affinity Biologicals (Ancaster, ON, Canada). Affinity matrix beads for IgG (CaptureSelect FcXL, capacity 25–35 g/L) and IgA (CaptureSelect IgA, capacity 8 g/L) were obtained from ThermoFisher Scientific (Leiden, Netherlands). All used chemicals were from Sigma-Aldrich (Zwijndrecht, Netherlands) except for trifluoroacetic acid (Merck, Darmstadt, Germany) and acetonitrile (Biosolve, Valkenswaard, Netherlands). Purified water was used from a Purelab Ultra system (Veolia Water Technologies Netherlands B.V., Ede, Netherlands). Sequencing-grade trypsin was obtained from Promega (Madison, WI).

Sample preparation

A detailed description of the methods for the immunoaffinity enrichment of the immunoglobulins and the glycopeptide preparation can be found elsewhere [17]. In brief, 5 µL of Visucon F plasma standard were diluted in PBS, and the immunoglobulins were enriched using a mix of CaptureSelect FcXL Affinity matrix beads for IgG and CaptureSelect IgA affinity matrix beads for IgA. Upon incubating the serum and the beads for 1 h at room temperature with agitation, the beads were washed three times with PBS and three times with water. The immunoglobulins were released by acid elution (100 mM formic acid) and collected into a 96-well PCR plate (Greiner Bio-One, Kremsmünster, Austria). Finally, the eluates were dried for 2.5 h at 60 °C by centrifugation under vacuum.

For tryptic digestion, the dried sample was reconstituted in 10 µL of reduction–alkylation buffer containing 100 mM Tris buffer, 1% w/v SDC, 10 mM tris(2-carboxyethyl)phosphine (TCEP), and 40 mM chloroacetamide (CAA). Upon mixing for 5 min, the samples were incubated for 5 min at 95 °C and cooled to room temperature. Tryptic digestion was started by the addition of 50 µL digestion buffer containing 50 mM ammonium bicarbonate pH 8.5 and 200 ng sequencing-grade trypsin. Upon mixing for 5 min, the sample was incubated at 37 °C overnight. Acid precipitation using 1.2 µL formic acid was performed on the following day. The precipitate was removed by centrifugation, and 40 µL of the supernatant was transferred to a V-bottom 96-well plate (Greiner). The sample was stored at −20 °C.

LC–MS/MS analysis

A 0.5 µL aliquot of the sample was analyzed five times with MS1 only (for MS1-based identification in GlycopeptideGraphMS and quantification in LaCyTools, Skyline, and GlycopeptideGraphMS) and five times with additional MS/MS (for fragmentation-based identification using Byonic and quantification using LaCyTools) in an alternating order. For the separation of the (glyco)peptides, the sample was injected into an Easy nLC 1200 system (Thermo Fisher Scientific) equipped with an in-house prepared precolumn (15 mm × 100 µm; Reprosil-Pur C18-AQ 3 µm, Dr. Maisch, Ammerbuch, Germany) and an analytical nanoLC column (15 cm × 75 µm; Reprosil-Pur C18-AQ 3 µm). As mobile phases 0.1% formic acid in water (A) and 20% water/80% acetonitrile + 0.1% formic acid (B) were used. A gradient from 10–40% of the mobile phase B was applied within 20 min. The LC was hyphenated to an Orbitrap Fusion Lumos MS (Thermo Fisher Scientific). For MS1 analysis, scans were acquired in a mass range of m/z 400–3,500 in positive mode. The resolution was set to 120,000. The target for automatic gain control (AGC) was set to 400,000. The maximum injection time was 50 ms. An intensity threshold of 20,000 was applied. For MS/MS analysis, charge states 2–7 were included for stepped higher-energy C-trap dissociation (HCD) with a normalized collision energy (NCE) of 35% ± 5% (30%, 35%, and 40% combined in one spectrum), a maximum injection time of 60 ms, and a AGC target of 50,000. Additionally, MS/MS fragmentation was trig-

gered for a HexNAc loss (204.087). For the triggered MS/MS analysis, a stepped HCD with an NCE of 35% ± 15% (20%, 35%, and 50% combined in one spectrum) was applied, and the AGC target was increased to 500,000 while the maximum injection time was increased to 200 ms. For all MS/MS scans, a precursor isolation width of *m*/*z* 1.2 was used. The MS/MS scan resolution was 30,000 and the *m*/*z* range was 110–3,500.

MS/MS data evaluation

A manual inspection of the raw data was performed in Xcalibur (v. 2.2, Thermo Fisher Scientific). PMI-Byonic (v. 3.7.13 Protein Metrics) was used for the MS/MS-based protein and glycosylation site identifications [14]. Protein identification was based on a canonical *Homo sapiens* UniProt database including 71,591 protein sequences (20,205 from Swiss-Prot and 51,386 from TrEMBL). The C-terminal cleavage of lysine and arginine and a maximum of two missed cleavages was allowed. A tolerance of 10 ppm was applied for the precursors and 20 ppm for fragment ions. A carbamidomethylation was set as a fixed modification for cysteine residues. Methionine oxidation was enabled as a variable modification. The search for *N*- and *O*-glycopeptides was separately performed. For this purpose, either the database "*N*-glycan 309 mammalian no sodium" (Supporting Information File 7) or "*O*-glycan 78 mammalian" (Supporting Information File 6) was applied as a custom modification. For manual MS/MS assignments, the web tool ProteinProspector v. 6.2.1 was used (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct). All glycopeptide compositions that were not identified by Byonic were subjected to a manual check of the MS/MS raw data in Xcalibur. This check included verifying the presence of the characteristic MS/MS ions (Table S4, Supporting Information File 1). In addition, allotypes of IgG3 and IgA2, which can be present in a human plasma pool [3], were manually checked. For this, the peptide sequences TKPWEEQYNSTFR, GFYPSDIAVEWESSGQPENNYNTTPPMLDSDGSFFLYSK (IgG3 *N*-glycopeptides), and MAGKPTHINVSVVMAEADGTC(Y) (IgA2 *N*-glycopeptide) were checked for the presence of the Y1 (peptide + HexNAc) ion in the MS/MS data. In addition, the expected glycoforms H1N1, H1N1S1, and H1N1S2 of the IgG3 *O*-glycopeptide SCDTPPPCPR were checked.

GlycopeptideGraphMS analysis

MS1-based glycopeptide identification in all five MS1-only measurements and visualization was performed using GlycopeptideGraphMS (v. 2.06) according to the user manual [15]. In short, the raw data were first transformed to the mzML format using msconvert (ProteoWizard 3.0 suite). The data preprocessing included the deconvolution of all MS1 signals using an OpenMS workflow (KNIME_OPENMS_GraphMS_Preprocessing_120318) in KNIME [15,33,34]. This workflow was used with OpenMS 2.3. Adaptions in the parameters were made in the *m*/*z* range of 400–3500 and the charge states 2–7. For the glycopeptide identification in GlycopeptideGraphMS, the intensity threshold was set to 1,000,000, the allowed mass deviation of the glycan building blocks to 0.02 Da, and the maximum subgroup degree was set to 1. As composition searching blocks (see the example provided in Supporting Information File 5), hexose (Hex, 162.0528 Da, max. 30 s RT difference). *N*-Acetylhexosamine (HexNAc, 203.0794 Da, max. 30 s RT difference), hexose, and *N*-acetylhexosamine (HexHexNAc, 365.1322 Da, max. 30 s RT difference), deoxyhexose (Fuc, 146.0579 Da, max. 20 s RT difference), and *N*-acetylneuraminic acid (NeuAc, 291.0954 Da, max. 120 s RT difference) were enabled. For each glycopeptide cluster of interest, one data point was assigned to a composition that was verified by the Byonic search. For the visualization in GlycopeptideGraphMS, the diameter of the data points and the relative abundance of the glycopeptides were represented upon logarithmic scaling between intensities from 1×10^6 to 1×10^{12}. False-positive assignments containing negative values in the compositions (illogical compositions) based on the assigned reference data points of all glycopeptides were removed. Analytes (with logical compositions) connected solely to analytes with illogical compositions (i.e., negative features) were excluded as well. For quantitative comparisons, only analytes were considered which were identified in at least three technical replicates. Intensities of analytes present at more than one RT were summed in case of a close RT proximity (likely isomers) or manually checked in the raw data for multiple peaks and included or excluded, dependent on the presence of multiple peaks in the raw data.

Skyline analysis

In addition to the automated glycopeptide identification, a MS1 assignment and peak integration was performed in Skyline (v19.1.0.193). The correct peak integration was manually checked. A reference glycopeptide composition list was inserted into Skyline. This list contained the merged information from the automatically assigned compositions (Byonic and GlycopeptideGraphMS), compositions listed on GlyConnect [20] for IgG and IgA, and an in-house analyte list that was recently used for an IgG/IgA analysis (based on literature information and manual peak assignment in MS1) [17]. The transition settings were set to product ions, the charge states were set to 2–7, and the time window was adjusted for each different glycopeptide cluster. MS1 data of the glycopeptide compositions were manually inspected, and charge states with an isotope dot product (idotp) >0.85 and a mass accuracy <10 ppm were included. "Normalized Area" was used for quantification.

LaCyTools analysis

For automated quantification in LaCyTools (v 1.0.1) [16], the raw data were converted to the mzXML format by MSConvert. The generation of the LaCyTools analyte list was supported by an in-house Python (v 3.7.6) script (Supporting Information File 3), which converted a representative GlycopeptideGraphMS output to the required input format for LaCyTools. Glycopeptide compositions that were not assigned in the representative data set in GlycopeptideGraphMS were added to the list to an appropriate retention time cluster. Potentially false-positive results (no MS1 isotope pattern matching or no MS/MS verification) were manually removed. The applied analyte list is provided in Supporting Information File 5. Next, an alignment list was created by selecting the most abundant glycopeptide compositions for each RT cluster. The width of the retention time cluster was set to 15 s and adjusted to 7 s for analytes with closely eluting interference signals. The RT alignment of the technical replicates was performed within a time window of 30 s and an *m*/*z* window of 0.1. For analyte curation and quantification, an *m*/*z* window of 0.025 was used. Upon processing in LaCyTools, all charge states of analytes with an isotopic pattern quality value higher than 0.2, mass accuracies of >10 ppm, and a signal-to-noise ratio <9 were excluded. The peak areas of the remaining charge states were summed and corrected by being divided by the isotopic pattern fraction. Of note, for the comparison of the relative quantification of GlycopeptideGraphMS, Skyline, and LaCyTools, the relative abundance was not renormalized to the intersection of the analytes.

Supporting Information

Raw data were made available in MassIVE: https://doi.org/doi:10.25345/C5JJ00.

Supporting Information File 1

Supporting tables.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-253-S1.xlsx]

Supporting Information File 2

Supporting figures.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-253-S2.pdf]

Supporting Information File 3

Python script connecting the GlycopeptideGraphMS output and LaCyTools input.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-253-S3.ipynb]

Supporting Information File 4

LaCyTools analyte list.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-253-S4.ref]

Supporting Information File 5

Analyte search list.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-253-S5.csv]

Supporting Information File 6

Byonic *N*-glycan list.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-253-S6.txt]

Supporting Information File 7

Byonic *O*-glycan list.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-253-S7.txt]

Acknowledgements

The authors would like to thank Paul J. Hensbergen for critically reading the manuscript.

Funding

This research was funded by the European Commission H2020 (Analytics for Biologics project, Grant No. 765502).

ORCID® iDs

Steffen Lippold - https://orcid.org/0000-0002-1032-5808
Peter A. van Veelen - https://orcid.org/0000-0002-7898-9408
Magnus Palmblad - https://orcid.org/0000-0002-5865-8994
Manfred Wuhrer - https://orcid.org/0000-0002-0814-4995
Noortje de Haan - https://orcid.org/0000-0001-7026-6750

References

1. Fu, C.; Zhao, H.; Wang, Y.; Cai, H.; Xiao, Y.; Zeng, Y.; Chen, H. *HLA* **2016,** *88,* 275–286. doi:10.1111/tan.12900
2. Szabo, Z.; Thayer, J. R.; Reusch, D.; Agroskin, Y.; Viner, R.; Rohrer, J.; Patil, S. P.; Krawitzky, M.; Huhmer, A.; Avdalovic, N.; Khan, S. H.; Liu, Y.; Pohl, C. *J. Proteome Res.* **2018,** *17,* 1559–1574. doi:10.1021/acs.jproteome.7b00862
3. de Haan, N.; Falck, D.; Wuhrer, M. *Glycobiology* **2020,** *30,* 226–240. doi:10.1093/glycob/cwz048
4. Yang, Y.; Franc, V.; Heck, A. J. R. *Trends Biotechnol.* **2017,** *35,* 598–609. doi:10.1016/j.tibtech.2017.04.010
5. Pan, S.; Chen, R.; Aebersold, R.; Brentnall, T. A. *Mol. Cell. Proteomics* **2011,** *10,* R110.003251. doi:10.1074/mcp.r110.003251

6. Stavenhagen, K.; Hinneburg, H.; Thaysen-Andersen, M.; Hartmann, L.; Silva, D. V.; Fuchser, J.; Kaspar, S.; Rapp, E.; Seeberger, P. H.; Kolarich, D. *J. Mass Spectrom.* **2013,** *48,* 627–639. doi:10.1002/jms.3210
7. Thaysen-Andersen, M.; Packer, N. H. *Biochim. Biophys. Acta, Proteins Proteomics* **2014,** *1844,* 1437–1452. doi:10.1016/j.bbapap.2014.05.002
8. Ozohanics, O.; Turiák, L.; Puerta, A.; Vékey, K.; Drahos, L. *J. Chromatogr. A* **2012,** *1259,* 200–212. doi:10.1016/j.chroma.2012.05.031
9. Ruhaak, L. R.; Xu, G.; Li, Q.; Goonatilleke, E.; Lebrilla, C. B. *Chem. Rev.* **2018,** *118,* 7886–7930. doi:10.1021/acs.chemrev.7b00732
10. Klein, J.; Zaia, J. *J. Proteome Res.*, in press.
11. Lee, L. Y.; Moh, E. S. X.; Parker, B. L.; Bern, M.; Packer, N. H.; Thaysen-Andersen, M. *J. Proteome Res.* **2016,** *15,* 3904–3915. doi:10.1021/acs.jproteome.6b00438
12. Abrahams, J. L.; Taherzadeh, G.; Jarvas, G.; Guttman, A.; Zhou, Y.; Campbell, M. P. *Curr. Opin. Struct. Biol.* **2020,** *62,* 56–69. doi:10.1016/j.sbi.2019.11.009
13. Hu, H.; Khatri, K.; Zaia, J. *Mass Spectrom. Rev.* **2017,** *36,* 475–498. doi:10.1002/mas.21487
14. Bern, M.; Kil, Y. J.; Becker, C. *Curr. Protoc. Bioinf.* **2012,** *40,* 13.20.1–13.20.14. doi:10.1002/0471250953.bi1320s40
15. Choo, M. S.; Wan, C.; Rudd, P. M.; Nguyen-Khuong, T. *Anal. Chem. (Washington, DC, U. S.)* **2019,** *91,* 7236–7244. doi:10.1021/acs.analchem.9b00594
16. Jansen, B. C.; Falck, D.; de Haan, N.; Hipgrave Ederveen, A. L.; Razdorov, G.; Lauc, G.; Wuhrer, M. *J. Proteome Res.* **2016,** *15,* 2198–2210. doi:10.1021/acs.jproteome.6b00171
17. Momčilović, A.; de Haan, N.; Hipgrave Ederveen, A. L.; Bondt, A.; Koeleman, C. A. M.; Falck, D.; de Neef, L. A.; Mesker, W. E.; Tollenaar, R.; de Ru, A.; van Veelen, P.; Wuhrer, M.; Dotz, V. *Anal. Chem. (Washington, DC, U. S.)* **2020,** *92,* 4518–4526. doi:10.1021/acs.analchem.9b05722
18. Plomp, R.; de Haan, N.; Bondt, A.; Murli, J.; Dotz, V.; Wuhrer, M. *Front. Immunol.* **2018,** *9,* No. 2436. doi:10.3389/fimmu.2018.02436
19. Šimurina, M.; de Haan, N.; Vučković, F.; Kennedy, N. A.; Štambuk, J.; Falck, D.; Trbojević-Akmačić, I.; Clerc, F.; Razdorov, G.; Khon, A.; Latiano, A.; D'Incà, R.; Danese, S.; Targan, S.; Landers, C.; Dubinsky, M.; Campbell, H.; Zoldoš, V.; Permberton, I. K.; Kolarich, D.; Fernandes, D. L.; Theorodorou, E.; Merrick, V.; Spencer, D. I.; Gardner, R. A.; Doran, R.; Shubhakar, A.; Boyapati, R.; Rudan, I.; Lionetti, P.; Krištić, J.; Novokmet, M.; Pučić-Baković, M.; Gornik, O.; Andriulli, A.; Cantoro, L.; Sturniolo, G.; Fiorino, G.; Manetti, N.; Arnott, I. D.; Noble, C. L.; Lees, C. W.; Shand, A. G.; Ho, G.-T.; Dunlop, M. G.; Murphy, L.; Gibson, J.; Evenden, L.; Wrobel, N.; Gilchrist, T.; Fawkes, A.; Kammeijer, G. S. M.; Vojta, A.; Samaržija, I.; Markulin, D.; Klasić, M.; Dobrinić, P.; Aulchenko, Y.; van den Heuve, T.; Jonkers, D.; Pierik, M.; McGovern, D. P. B.; Annese, V.; Wuhrer, M.; Lauc, G. *Gastroenterology* **2018,** *154,* 1320–1333.e10. doi:10.1053/j.gastro.2018.01.002
20. Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.; Packer, N. H.; Lisacek, F. *J. Proteome Res.* **2019,** *18,* 664–677. doi:10.1021/acs.jproteome.8b00766
21. Falck, D.; Jansen, B. C.; de Haan, N.; Wuhrer, M. High-Throughput Analysis of IgG Fc Glycopeptides by LC-MS. In *High-Throughput Glycomics and Glycoproteomics: Methods and Protocols;* Lauc, G.; Wuhrer, M., Eds.; Methods in Molecular Biology; Humana Press: New York, NY, USA, 2017; pp 31–47. doi:10.1007/978-1-4939-6493-2_4
22. Deshpande, N.; Jensen, P. H.; Packer, N. H.; Kolarich, D. *J. Proteome Res.* **2010,** *9,* 1063–1075. doi:10.1021/pr900956x
23. Chandler, K. B.; Mehta, N.; Leon, D. R.; Suscovich, T. J.; Alter, G.; Costello, C. E. *Mol. Cell. Proteomics* **2019,** *18,* 686–703. doi:10.1074/mcp.ra118.001185
24. Zheng, S.; Doucette, A. A. *Proteomics* **2016,** *16,* 1059–1068. doi:10.1002/pmic.201500366
25. Nilsson, J.; Larson, G. In *Mass Spectrometry of Glycoproteins: Methods and Protocols;* Kohler, J. J.; Patrie, S. M., Eds.; Humana Press: Totowa, NJ, USA, 2013; pp 79–100. doi:10.1007/978-1-62703-146-2
26. Joenväärä, S.; Ritamo, I.; Peltoniemi, H.; Renkonen, R. *Glycobiology* **2008,** *18,* 339–349. doi:10.1093/glycob/cwn013
27. Steen, H.; Mann, M. *J. Am. Soc. Mass Spectrom.* **2001,** *12,* 228–232. doi:10.1016/s1044-0305(00)00219-1
28. Gupta, V.; Carroll, K. S. *Biochim. Biophys. Acta, Gen. Subj.* **2014,** *1840,* 847–875. doi:10.1016/j.bbagen.2013.05.040
29. Darula, Z.; Medzihradszky, K. F. *Anal. Chem. (Washington, DC, U. S.)* **2015,** *87,* 6297–6302. doi:10.1021/acs.analchem.5b01121
30. Delafield, D. G.; Li, L. *Mol. Cell. Proteomics*, in press. doi:10.1074/mcp.r120.002095
31. Kil, Y. J.; Becker, C.; Sandoval, W.; Goldberg, D.; Bern, M. *Anal. Chem. (Washington, DC, U. S.)* **2011,** *83,* 5259–5267. doi:10.1021/ac200609a
32. Rebecchi, K. R.; Wenke, J. L.; Go, E. P.; Desaire, H. *J. Am. Soc. Mass Spectrom.* **2009,** *20,* 1048–1059. doi:10.1016/j.jasms.2009.01.013
33. Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. *BMC Bioinf.* **2008,** *9,* No. 163. doi:10.1186/1471-2105-9-163
34. Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. *Nat. Methods* **2016,** *13,* 741–748. doi:10.1038/nmeth.3959

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Simulating the enzymes of ganglioside biosynthesis with Glycologue

Andrew G. McDonald* and Gavin P. Davey*

Full Research Paper

Open Access

Address:
School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland

Email:
Andrew G. McDonald* - amcdonld@tcd.ie;
Gavin P. Davey* - gdavey@tcd.ie

* Corresponding author

Keywords:
gangliosides; Glycologue; glycosyltransferases; neuropathy; Svennerholm nomenclature

Beilstein J. Org. Chem. **2021,** *17,* 739–748.
https://doi.org/10.3762/bjoc.17.64

Received: 21 December 2020
Accepted: 12 March 2021
Published: 23 March 2021

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: N. H. Packer

Abstract

Gangliosides are an important class of sialylated glycosphingolipids linked to ceramide that are a component of the mammalian cell surface, especially those of the central nervous system, where they function in intercellular recognition and communication. We describe an in silico method for determining the metabolic pathways leading to the most common gangliosides, based on the known enzymes of their biosynthesis. A network of 41 glycolipids is produced by the actions of the 10 enzymes included in the model. The different ganglioside nomenclature systems in common use are compared and a systematic variant of the widely used Svennerholm nomenclature is described. Knockouts of specific enzyme activities are used to simulate congenital defects in ganglioside biosynthesis, and altered ganglioside status in cancer, and the effects on network structure are predicted. The simulator is available at the Glycologue website, https://glycologue.org/.

Introduction

Gangliosides are glycosphingolipids that contain a sialylated carbohydrate linked to ceramide. Typically located in the plasma membranes of many tissues, gangliosides are most concentrated in the brain, where they are the dominant feature of the neuronal glycocalyx [1-3]. The oligosaccharide is based on a linear chain comprising of up to four monosaccharide units, containing glucose, galactose and *N*-acetylgalactosamine, to which are attached a variable number of sialic acid (*N*-acetylneuraminic acid) residues. The sialic acid content of the oligosaccharide, being anionic at pH 7, results in an overall negative charge. Figure 1 shows the structure of the monosialylated ganglioside GM1a.

The biosynthesis of gangliosides occurs in the endoplasmic reticulum and Golgi, where specific glycosyltransferases act, in stepwise fashion, by adding monosaccharides from sugar nucleotide donors, first to ceramide, and then to subsequent ceramide-linked glycoconjugate acceptors, before transport and

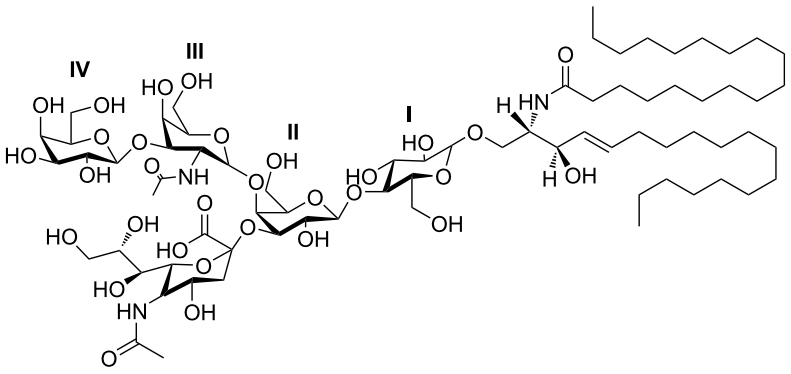


Figure 1: Chemical structure of ganglioside GM1a (a β-D-galactosyl-(1→3)-*N*-acetyl-β-D-galactosaminyl-(1→4)-[α-*N*-acetylneuraminyl-(2→3)]-β-D-galactosyl-(1→4)-β-D-glucosyl-(1↔1)-ceramide). Substituents of the core are labelled with Roman numerals **I–IV** (**I**, Glc; **II**, Gal; **III**, GalNAc; **IV**, Gal). IUPAC name: II^{3}Neu5Ac-Gg$_4$Cer.

eventual incorporation into the plasma membrane via vesicular fusion. Gangliosides, which function as antigenic determinants [4], may play a role in membrane organization [5], cell signaling [6], apoptosis [7], and in memory formation through neuromodulation of synaptic transmission [8]. Gangliosides are recycled in the lysosome through the action of glycohydrolases. The inhibition of membrane recycling has been shown to lead to an accumulation of lysosomal gangliosides resulting in neuronal death [9]. Congenital disorders of ganglioside biosynthesis can lead to a number of neuropathies, including motor deficits, microcephaly, sensory loss, and autistic features [10,11]. Certain gangliosides, such as GM2, have been identified as tumor markers for breast cancer stem cells [12], while members of the alpha-series gangliosides, such as GD1α, promote tumor-cell adhesion during metastasis [13]. The cholinergic neuron-specific gangliosides GQ1bα and GT1aα may contribute to the pathogenesis of Alzheimer's disease [14].

Previously, we described a deductive apparatus of a formal system for modelling the enzymes of mucin-type O-linked glycosylation, with a web-based application, O-Glycologue, that allows knockouts of enzymes of O-linked glycosylation and the assignment of custom "wild type" sets of enzyme activities to study the effects of differential knockouts on the resultant networks [15]. In this article, we describe an extension of this method to gangliosides, and to the enzyme reactions associated with their biosynthesis. The formalism and the associated web application, now renamed Glycologue, provide a way to explore the effects of mutations that result in a loss of functionality, or promotion of disease.

The method involves a set of regular-expression-based rules acting on strings of characters that representing the monosaccharide units, x, model the actions of transferases in the general form, Ax + B = A + xB, where Ax is a nucleotide sugar and B is the carbohydrate moiety of the acceptor, be it a glycolipid or some other oligosaccharide, the nucleotide A is the product of the donor and xB is the acceptor product. The strings can be seen as a compression of the familiar condensed linear IUPAC notation, using a single-letter notation to represent sugars, with upper-case denoting D, and lowercase, the L, sugars, and are read from right to left starting with the base (reducing-end) sugar. The letters a and b are reserved for α and β-anomers, respectively, while brackets are used to delimit branches, and the letter T is used to denote the connection point to ceramide, or to another conjugate depending on the context. In this work, we consider only four monosaccharides and their corresponding letters: Glc (G), Gal (L), Neu5Ac (S) and GalNAc (V). In IUPAC form (see Table 1), we can write the carbohydrate portion of the ganglioside GM1a (Figure 1) as any of the following:

- Galb1-3GalNAcb1-4[Neu5Aca2-3]Galb1-4GlcCer (IUPAC)
- Lb3Vb4[Sa3]Lb4GbT (Glycologue, full)
- L3Vb4[S3]L4GT (Glycologue, abbreviated)

Table 1: Single-letter codes used and their IUPAC equivalents.

Glycologue single-letter code	IUPAC symbol	IUPAC definition
G	Glc	β-D-glucose
L	Gal	β-D-galactose
S	Neu5Ac	*N*-acetylneuraminate
T	Cer	ceramide (*N*-acylsphingosine)
V	GalNAc	*N*-acetyl-α-D-galactosamine
a, b	α, β	anomeric configuration

Glycologue makes the further assumption that each sugar as it appears in the acceptor has a default anomer, with Glc and Gal being β, and Neu5Ac and GalNAc being α, which allows the a and b notation to be dropped in most instances. However, since GalNAc appears as the β anomer in gangliosides, the b is retained in the abbreviated Glycologue notation for any substrate in which it appears. The resulting string is referred to a *structure identifier* [15], since it also contains the instructions for drawing a 2-dimensional image of the oligosaccharide, in the manner of turtle graphics.

Results and Discussion

Model description

The simulator acts iteratively on an initial acceptor substrate, passing it to each enzyme in turn, and accumulating a set of acceptor products. The pool of novel acceptor products become the substrates at the next iteration, until either no new products are formed, or a user-determined maximum number of iterations has been reached. Table 2 lists the enzymes of ganglioside biosynthesis included in the current model, with an index number, **1**–**10**, the EC number, where available, a short name, a longer accepted name and a reaction pattern. The reactions in Table 2 are based on activities of enzymes already classified within the IUBMB Enzyme List, or from the cited references, wherever an EC number is not available. Glycosyltransferases can act on a variety of substrates, and in cases where substrate recognition follows a less specific rule, the reaction pattern uses an asterisk to denote parts of the acceptor that are of indeterminate length. The glossary in Table 2, footnote b, shows some of the assumptions implicit to the model, such as the configuration of the donors. From this information it is possible to infer that all of the enzymes of the model are all configuration-inverting, rather than configuration-retaining. This inversion of configuration refers to the stereochemistry of the anomeric carbon in the acceptor product, which has the opposite configuration to that of the donor substrate.

The formal language on which the method is based is a modelling language for glycosyltransferases, and can be used to classify the types of reaction catalyzed according to simple rules. We identify *extension* of a linear oligosaccharide as the default mode of action (Equation 1),

$$\mathrm{Ax} + \mathrm{yB} = \mathrm{xyB} + \mathrm{A} \quad (1)$$

Table 2: Enzymes of ganglioside biosynthesis and Glycologue reaction patterns.

enzyme no.	EC number	short name	accepted name[a]	reaction pattern[b]
1	EC 2.4.1.80	UGCG	ceramide glucosyltransferase	UDP-G + T = UDP + GbT
2	EC 2.4.1.47	β1Gal-T3	*N*-acylsphingosine galactosyltransferase	UDP-L + T = UDP + LbT
3	EC 2.4.1.274	β4Gal-T6	glucosylceramide β-1,4-galactosyltransferase	UDP-L + GbT = UDP + Lb4GbT
4	EC 2.4.99.9	ST3Gal-V	lactosylceramide α-2,3-sialyltransferase	CMP-S + Lb4GbT = CMP + [Sa3]Lb4GbT CMP-S + LbT = CMP + [Sa3]LbT
5	EC 2.4.99.8	ST8Sia-I	α-*N*-acetylneuraminate α-2,8-sialyltransferase	CMP-S + [Sa3]Lb4*T = CMP + [Sa8Sa3]Lb4*T
6	EC 2.4.99.-[16]	ST8Sia-V	(α-2,8-*N*-acetylneuraminate α-2,8-sialyltransferase)	CMP-S + [Sa8Sa3]*T = CMP + [Sa8Sa8Sa3]*T CMP-S + Sa8Sa3*T = CMP + Sa8Sa8Sa3*T
7	EC 2.4.1.92	β4GalNAc-T1	(*N*-acetylneuraminyl)-galactosylglucosylceramide *N*-acetylgalactosaminyltransferase	UDP-V + Lb4*T = UDP + Vb4Lb4*T UDP-V + [Sa3]Lb4*T = UDP + Vb4[Sa3]Lb4*T
8	EC 2.4.1.68	β3Gal-T4	ganglioside galactosyltransferase	UDP-L + Vb4*T = UDP + Lb3Vb4*T
9	EC 2.4.99.-[17,18]	ST3Gal-II	(β-1,3-galactosyl-ceramide α-2,3-sialyltransferase)	CMP-S + Lb3Vb4*T = CMP + Sa3Lb3Vb4*T
10	EC 2.4.99.-[19]	ST6GalNAc-V	(α1,3-Sia-β1,3-Gal-β1,3-GalNAc α-2,6-sialyltransferase)	CMP-S + Sa3Lb3Vb4*T = CMP + Sa3Lb3[Sa6]Vb4*T

[a]For enzymes without an EC number, a suggested name is given in parentheses. Literature references supporting the unclassified activities are provided in the EC number column, after the EC sub-subclass. [b]Asterisks act as a wildcard character, to denote an unspecified portion of the oligosaccharide. Symbols and abbreviations used in reaction patterns are those of Table 1 with the following additions: UDP, uridine 5′-diphosphate; CMP, cytidine 5′-phosphate; CMP-S, CMP-*N*-acetyl-β-neuraminate; UDP-G, UDP-α-D-glucose; UDP-L, UDP-α-D-galactose; UDP-V, UDP-*N*-acetyl-α-D-galactosamine.

where x and y are monosaccharides, Ax is the nucleotide-sugar donor, and yB the acceptor substrate. The formation of a single branch along a linear chain is described as *decoration*, where the pattern is (Equation 2).

$$\mathrm{Ax} + \mathrm{yB} = [\mathrm{x}]\mathrm{yB} + \mathrm{A} \quad (2)$$

Here we have assumed that [x]y is a substring of the parent acceptor and is a shorthand for *[x]y*B, the asterisks acting as a wildcard character. Double branches are used to form symmetric core structures, such as the trimannosyl core of N-glycans, or O-linked glycan cores based on GalNAc (Equation 3):

$$\mathrm{Ax} + \mathrm{yB} = [\mathrm{x}][\mathrm{y}]\mathrm{B} + \mathrm{A} \quad (3)$$

Capping of branches and linearly extended chains is achieved through *termination*, of which sialylation is a typical example. However, some enzymes can continue to act on such terminal elements, which can be called *termination with extension*, or *decoration with extension*. As an example of the latter, LacCer (L4GT) can be decorated on the galactose by enzyme **4** to give [S3]L4GT, and subsequently extended by enzymes **5** and **6** to give [S8S3]L4GT and [S8S8S3]L4GT. Termination with extension in this model occurs at the initial sialylation of L3Vb4L4GT by enzyme **9**, to yield S3L3Vb4L4GT, which can be further extended by two iterations of **6**, to produce S8S8S3L3Vb4L4GT. A separate category not considered in this model is *modification* of monosaccharides, for example through sulfation, acetylation or phosphorylation, which follow the same pattern as decoration. Glycologue structure identifiers order branches by linkage position, writing the branch with the lowest linkage first, reading from right to left. Modifiers are written before sugars units, and multiple modifiers on the same monosaccharide are again ordered by linkage position, from lowest to highest, reading right to left.

Nomenclature of gangliosides

Gangliosides are commonly labelled according to the abbreviated Svennerholm [20] nomenclature, or else by the expanded form recommended by IUPAC/IUBMB Joint Commission on Biochemical Nomenclature [21]. The original Svennerholm notation was a semi-systematic system, and its formation rules have not always consistently applied by those using it. We introduce here a more systematic Svennerholm nomenclature that reproduces, as far as possible, the traditional system, but which is capable of automatic assignment by Glycologue from the structural identifier, and then translation to the IUPAC form. A description of the method will be given, together with examples.

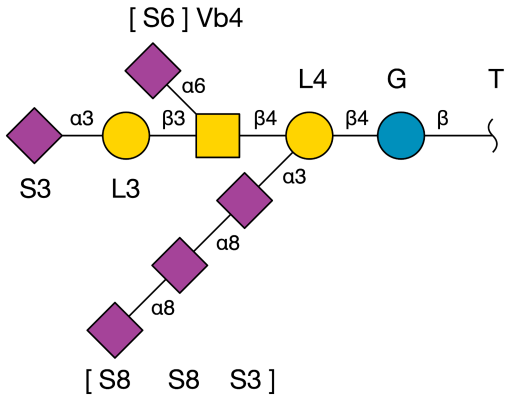
In the IUPAC system, ganglio-series of glycosphingolipids are given the core abbreviation, Gg, followed by the number of monosaccharides (-oses) in the linear core, as a subscript. Thus, the core descriptor $\mathrm{Gg}_n\mathrm{Cer}$ represents a core of length n attached to ceramide. (Formerly, the core descriptor was given as "$\mathrm{GgOse}_n\mathrm{Cer}$" [22], but this recommendation has since been rescinded [21].) To this base string are added a list of the sialic acids attached to each monosaccharide in the core, counting using the Roman numeral system, starting from the base glucose (cf. Figure 1). From the non-reducing end, write the position on the core where a sialic acid (or sialic acid chain) appears, as the uppercase Roman numeral, superscripting the linkage position after as the Arabic numeral, followed by "Neu5Ac"; if a chain of sialic acids is present, place the "Neu5Ac" in parentheses and subscript the number of residues after, e.g., $\mathrm{IV}^3(\mathrm{Neu5Ac})_2$. Repeat this procedure for as many units of the core as are sialylated, separating with commas, then append a hyphen, followed by the core descriptor.

The systematic Svennerholm nomenclature system used by Glycologue counts the total number of sialic acids in the carbohydrate, and appends this as a single letter (A: Asialo = 0, M: Monosialo = 1, D: Disialo = 2, T: Tri = 3, etc.) to the "G" of ganglioside. The core is assigned a number based on its length, where 1 denotes fully extended, Galb1-3GalNAcb1-4Galb1-4GlcCer, 2 is GalNAcb1-4Galb1-4GlcCer and 3 is Galb1-4GlcCer. When the core is fully extended, the series letter, a, b or c, is appended, which denotes the presence of either one, two or three sialic acids on position II of the core. The presence of sialylation on the root galactose of LacCer thus determines the series into which the ganglioside is categorized. A common practice, although not recommended by IUPAC, is to add α at the end of the code, when an α-2,6-linked Neu5Ac is present on position III, which is the GalNAc β4-linked to Gal. The composition of the systematic Svennerholm name (SSN) is then

"G" + (the total sialic-acid count, as a capital letter) + $(4 - n + 1)$ + (series letter a–c, where $n = 4$) + (α),

where n is the core length as defined above. GalCer is denoted by the core number 4, and hence is synonymous with GA4, while its monosialylated form is denoted GM4.

To illustrate the method with an example (see Figure 2), the simulator predicts the existence of the ganglioside carbohydrate with structure identifier S3L3[S6]Vb4[S8S8S3]L4GT. The total sialic acid count of this structure is 5, which is assigned the letter P (penta-sialylated). It is a c-series ganglioside, with 3 sialic acid residues on position II. There is 1 sialic acid attached to the GalNAc, which means that there must be one (5–3–1) sialic acid on the terminal non-reducing Gal (position



Structure identifier	SSN	Procedure
S3L3[S6]Vb4[S8S8S3]L4**GT**	**G**	1. Determine the base unit (GlcCer, LacCer or G)
S3L3[**S6**]Vb4[**S8S8S3**]L4GT	G**P**	2. Count the total number of sialic acids present
S3**L3**[S6]**Vb4**[S8S8S3]**L4**GT	GP**1**	3. Determine the core extension level (3, 2 or 1)
S3L3[S6]Vb4**[S8S8S3]L4**GT	GP1**c**	4. Determine the series letter, a, b, c
S3L3**[S6]Vb**4[S8S8S3]L4GT	GP1c**α**	5. Examine sialylation on position III (α-series)

Figure 2: Construction of the Svennerholm name GP1cα from its Glycologue structure identifier. At each step of the procedure, the parts of the structure identifier that determine the corresponding part of the systematic Svennerholm name (SSN) are shown in bold face.

IV). The core is fully extended, which gives $n = 4$, therefore the SSN is GP1cα.

The IUBMB name follows from the systematic Svennerholm code. The fact that there are five (P) sialic acids, three of which are on position II (c), and one on position III (α), implies that the core must be fully extended, with the remaining sialic acid on position IV. The IUPAC name of this structure is therefore IV^{3}Neu5Ac,III^{6}Neu5Ac,II^3(Neu5Ac)$_3$-Gg$_4$Cer.

The SSN is identical to that which is generally used, with the exceptions noted in Table 3. It should be noted that the name "GM1" becomes ambiguous if the letter a–c is not consistently applied when the core is fully extended ($n = 4$), since there are two galactose residues (positions II and IV) at which sialylation can occur. The name "GM1b", which has been used to refer to GM1 (IV^{3}Neu5Ac-Gg$_4$Cer) is formally incorrect, since it is not a b-series ganglioside, with two Neu5Ac residues on position I, but in the 0-series gangliosides that derive from GA2 (Figure 3).

Enzymes of the model

Using the preceding classification, we see that the enzymes of Table 2 fall into five categories: extension (activities **1**–**3**, **5**, **7**, and **8**), decoration with extension (**4**), decoration (**10**), termination with extension (**9**), and termination (**6**), all of which follow the reaction patterns of Equation 1 and Equation 2, and none follow the double-branching pattern of Equation 3. The first iteration, starting from ceramide (T), produces GlcCer, catalyzed by ceramide glucosyltransferase (EC 2.4.1.80). Also included is the activity of *N*-acylsphingosine galactosyltransferase (EC 2.4.1.47), which produces GalCer, the starting member of the Gala series of galactocerebrosides. GlcCer, but not GalCer, can at the next iteration be extended with a galac-

Table 3: Commonly used Svennerholm names that differ from their systematic counterparts.

Svennerholm name (incorrect)	systematic Svennerholm name (SSN)	IUPAC name
GM1b	GM1	IV^{3}Neu5Ac-Gg$_4$Cer
GM1	GM1a	II^{3}Neu5Ac-Gg$_4$Cer
GD1c	GD1	IV^3(Neu5Ac)$_2$-Gg$_4$Cer
GD1aα	GD1α	IV^{3}Neu5Ac,III^{6}Neu5Ac-Gg$_4$Cer

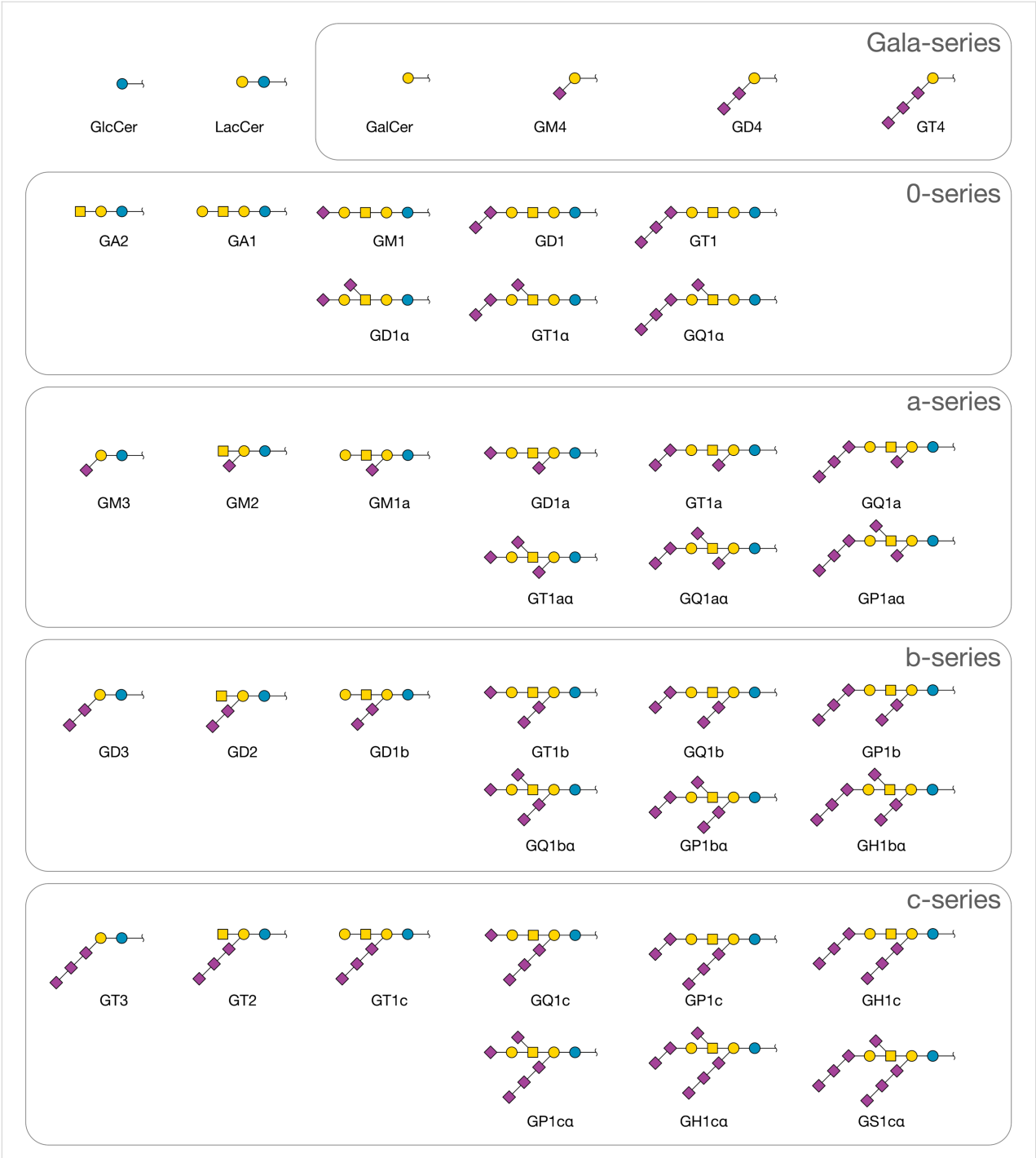


Figure 3: Ganglioside carbohydrates predicted by the model. All structures are linked to ceramide at the base glucose or galactose residue.

tose, β4-linked to the glucose, to form lactosylceramide, LacCer, catalyzed by β4Gal-T6 (EC 2.4.1.274). This is the first asialo-ganglioside, with core level 3. Two further extensions are possible, to yield core levels 2 and 1, by adding a GalNAc residue β4-linked to the preceding galactose (EC 2.4.1.92), followed by a further galactose in a β3-linkage to GalNAc. The maximally extended core oligosaccharide is thus L3Vb4L4GT in the abbreviated Glycologue notation. Sialylation can occur in the model through decoration of the base galactose, or termination of the β3-linked galactose, by the α-2,3-sialyltransferase enzymes ST3Gal-V (**4**) and ST3Gal-II (**9**), also known as GM3 synthase and GM1 synthase [23,24], respectively. It is assumed that the sialyltransferase activity **5** occurs before core extension with GalNAc by EC 2.4.1.92 (**7**). Up to two further sialylation

steps can occur, on each of the sialylated galactose positions, with α-2,8-linkage (activities **5** and **6**). Ganglioside GM1 can act as the substrate of ST6GalNAc-V (**10**), which adds a α-2,6-linked Neu5Ac to the central GalNAc residue of the core. As it is known that certain isoforms of ST3Gal-V can act on GalCer [25,26], this alternative activity of the enzyme has been included in the model (Table 2). After 11 iterations of the method, the simulator produces 41 unique structures shown in Figure 3, in 49 reactions, with the network shown in Figure 4.

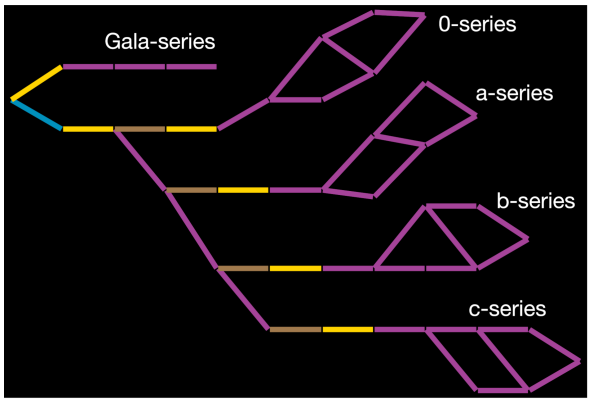


Figure 4: Ganglioside biosynthetic reaction network predicted by the Glycologue enzyme simulator. Starting from ceramide, which is the root (leftmost) node, 41 carbohydrate structures are predicted using 10 enzymes. The edges of the graph represent enzyme reactions, colored according to the type of sugar transferred: yellow (galactosyltransferases); blue (glucosyltransferases); brown (*N*-acetylgalactosaminyltransferases); magenta (sialyltransferases).

Networks and knockouts

The structures shown in Figure 3 are divided into five subsets: the Gala series, 0-series, a-series, b-series, and c-series gangliosides. Starting from GalCer (GA4), the model predicts three downstream products, GM4, GD4, and GT4, through the sequential action of ST3Gal-V (**4**) and two applications of the enzyme ST8Sia-I (**5**). These structures have previously been observed in bovine milk [27]. The 0-series gangliosides produced, in addition to asialo-gangliosides, the sialylated forms GM1, GD1, and GT1, along with their α2-,6-sialylated counterparts, GD1α, GT1α, and GQ1α. Also in Figure 3 are the a-series gangliosides GM3, GM2, and GM1a, and the derivatives of the GM1a with terminal sialylation, GD1a, GT1a [28], and GQ1a. The b-series gangliosides, GD1b, GT1b (both downregulated in Alzheimer's disease [14,29]), GQ1b and GP1b, and their three α-variants are predicted. There is no GT1bα, since a terminal sialic acid on position IV is required by enzyme rule **10**. The c-series gangliosides follow the same pattern as those of the b series. The members of this series with fully extended core (GT1c, GQ1c, GP1c and GH1c) appear in stellate neurons of adult human brain [30], but are also found in extraneural tissues in species such as rat [31]. The model predicts the structures of GP1cα, GH1cα and GS1cα, which may correspond to the penta-, hexa-, and septa-sialylated gangliosides observed in embryonic chicken brain [32].

The effects of knocking out each enzyme of Table 2 individually are shown in Figure 5. Comparing the pattern of glucosylation, sialylation, galactosylation, and GlcNAc-ylation events among the different knockouts, and with that of the full network in Figure 4, reveals that the most pronounced effects on ganglioside complexity occur with enzyme activities **1**, **3**, and **7**, which result in fewer than 10 reactions each. That any structures are formed in the absence of UGCG (**1**) is because of the Gala structures formed by *N*-acylsphingosine galactosyltransferase (**2**). Mutations in the gene coding for enzyme **7** (β4GalNAc-T1; also known as GM2/GD2 synthase [33]) are responsible for spastic paraplegia [10]. The knockouts affecting

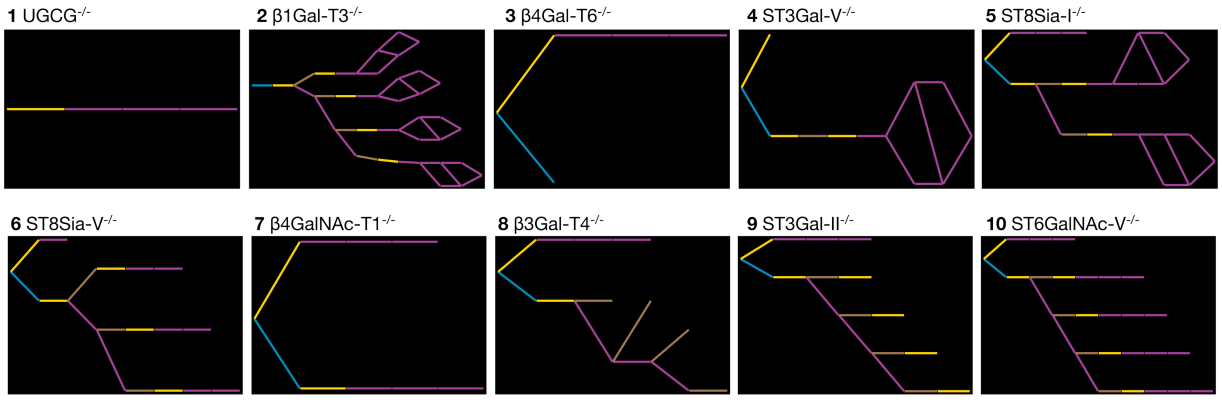


Figure 5: Predicted effects on the pathways of ganglioside biosynthesis when individual enzyme activities are completely inhibited or knocked-out. Panels 1–10 correspond to the enzymes of Table 2. Enzyme reactions are shown as lines colored according to the type of sugar transferred: yellow (galactosyltransferases); blue (glucosyltransferases); brown (*N*-acetylgalactosaminyltransferases).

entire series, discounting those of Gala, are the enzymes **4**, **5**, and **6**, which produce only the 0-series (**4**), 0- and a-series (**5**) and 0-, a-, and b-series gangliosides (**6**). The loss of GM3 through ST3Gal-V (**4**) deficiency is associated with auditory impairment in mouse and human [34]. The loss of complex polysialylated structures is evident in the knockouts of enzyme activities **8** and **9**, which are unable to form terminal sialic acid or α-type structures. Knockout of ST3Gal-II (**9**) reduces GD1and GT1b levels in the brain by 50%, whereas brain-protein sialylation is unchanged [35]. A loss of ST3Gal-II also leads to late-onset obesity and insulin resistance [36].

Glycologue web application

The Glycologue ganglioside simulator is available at https://glycologue.org/g/, along with the source code of the simulator in the Python programming language. Glycologue exports networks as SBML, for import into Copasi [37], CellDesigner [38], Tellurium [39], or other modelling software supporting this format. Glycan structures can be imported or exported as GlycoCT [40], and exported as Linear Code [41] or IUPAC condensed linear formats. Sets of structures can be downloaded as CSV or GlycoCT. A key function of Glycologue is the ability to predict the enzymes required for the biosynthesis of a given glycan; the subset of the enzyme activities can then be used to generate all of the ganglioside carbohydrates, starting from ceramide, or any other structure. By setting a baseline knockout of enzyme activities, the effects of further knockouts on the number and type of glycans formed can be predicted. In the web application version of Table 2, reactants and reactions are linked to ChEBI [42] and Rhea [43] by their identifiers, where these are available.

Future development

In addition to biosynthesis, the biochemistry of gangliosides includes cellular transport and recycling. A limitation of the model is that it considers only absolute changes to enzyme activity, in which an activity is either off or on, which a kinetic model based on differential-equation-based rate laws [44-48] or stochastic kinetics [49,50] would improve upon. Nevertheless, we have shown that the knockouts are able to reproduce the distinct species-specific features and disease states arising from congenital defects of ganglioside biosynthesis. Kinetic models based on the networks described here can be generated in modelling software, using the SBML output provided. In such models, we suppose that, owing to the multi-branched structure of the networks (cf. Figure 4), multi-substrate competition effects would need to be taken into account [44], since multiple substrates compete for the same enzyme. Competing fluxes downstream of branch points will also influence the kinetics [48,51]. Future extensions to this work will consider the effects of acetylation of sialic acid residues, since this modification reduces the negative charge of the carbohydrate, thus altering binding affinity, while an increased incidence of 9-*O*-acetylated GD3 is associated with melanoma [52]. The activities of glycosidases might be added to the simulators as a way to model lysosomal storage diseases (LSDs) such as Tay-Sachs, in which ganglioside GM2 accumulates as a result of a deficiency in β-*N*-acetylhexosaminidase activity (EC 3.2.1.52) [53]. Since Glycologue structure identifiers can be exported as Linear Code, future support for the recently introduced LiCoRR (Linear Code for Reaction Rules) formalism [54] is also possible. Glycologue can incorporate the Neu5Gc and KDN variants of sialic acid, and predict structures containing these residues. However, they have not been considered here, from a de novo standpoint, owing to the combinatorial complexity that would arise from equal participation of the donors CMP-Neu5Ac, CMP-Neu5Gc, and CMP-KDN.

Acknowledgements

Prof. Keith Tipton (Trinity College Dublin) is thanked for a critical reading of the manuscript.

ORCID® iDs

Andrew G. McDonald - https://orcid.org/0000-0003-2727-176X
Gavin P. Davey - https://orcid.org/0000-0002-8667-8781

References

1. Schengrund, C.-L. *Trends Biochem. Sci.* **2015,** *40,* 397–406. doi:10.1016/j.tibs.2015.03.007
2. Schnaar, R. L. *J. Mol. Biol.* **2016,** *428,* 3325–3336. doi:10.1016/j.jmb.2016.05.020
3. Schnaar, R. L. The Biology of Gangliosides. *Advances in Carbohydrate Chemistry and Biochemistry;* Academic Press: New York, London, 2019; Vol. 76, pp 113–148. doi:10.1016/bs.accb.2018.09.002
4. Gregson, N. A.; Kennedy, M.; Leibowitz, S. *Nature* **1977,** *266,* 461–463. doi:10.1038/266461a0
5. Ohmi, Y.; Ohkawa, Y.; Yamauchi, Y.; Tajima, O.; Furukawa, K.; Furukawa, K. *Neurochem. Res.* **2012,** *37,* 1185–1191. doi:10.1007/s11064-012-0764-7
6. Lopez, P. H.; Schnaar, R. L. *Curr. Opin. Struct. Biol.* **2009,** *19,* 549–557. doi:10.1016/j.sbi.2009.06.001
7. Bieberich, E.; MacKinnon, S.; Silva, J.; Yu, R. K. *J. Biol. Chem.* **2001,** *276,* 44396–44404. doi:10.1074/jbc.m107239200
8. Rahmann, H. *Behav. Brain Res.* **1995,** *66,* 105–116. doi:10.1016/0166-4328(94)00131-x
9. Boutry, M.; Branchu, J.; Lustremant, C.; Pujol, C.; Pernelle, J.; Matusiak, R.; Seyer, A.; Poirel, M.; Chu-Van, E.; Pierga, A.; Dobrenis, K.; Puech, J.-P.; Caillaud, C.; Durr, A.; Brice, A.; Colsch, B.; Mochel, F.; El Hachimi, K. H.; Stevanin, G.; Darios, F. *Cell Rep.* **2018,** *23,* 3813–3826. doi:10.1016/j.celrep.2018.05.098
10. Li, T. A.; Schnaar, R. L. Congenital Disorders of Ganglioside Biosynthesis. *Progress in Molecular Biology and Translational Science;* Academic Press: New York, London, 2018; Vol. 156, pp 63–82. doi:10.1016/bs.pmbts.2018.01.001

11. Trinchera, M.; Parini, R.; Indellicato, R.; Domenighini, R.; dall'Olio, F. *Mol. Genet. Metab.* **2018,** *124,* 230–237. doi:10.1016/j.ymgme.2018.06.014
12. Battula, V. L.; Shi, Y.; Evans, K. W.; Wang, R.-Y.; Spaeth, E. L.; Jacamo, R. O.; Guerra, R.; Sahin, A. A.; Marini, F. C.; Hortobagyi, G.; Mani, S. A.; Andreeff, M. *J. Clin. Invest.* **2012,** *122,* 2066–2078. doi:10.1172/jci59735
13. Taki, T.; Ishikawa, D.; Ogura, M.; Nakajima, M.; Handa, S. *Cancer Res.* **1997,** *57,* 1882–1888.
14. Ariga, T. *Mol. Neurobiol.* **2017,** *54,* 623–638. doi:10.1007/s12035-015-9641-0
15. McDonald, A. G.; Tipton, K. F.; Davey, G. P. *PLoS Comput. Biol.* **2016,** *12,* e1004844. doi:10.1371/journal.pcbi.1004844
16. Kim, Y.-J.; Kim, K.-S.; Do, S.-i.; Kim, C.-H.; Kim, S.-K.; Lee, Y.-C. *Biochem. Biophys. Res. Commun.* **1997,** *235,* 327–330. doi:10.1006/bbrc.1997.6725
17. Kojima, N.; Lee, Y.-C.; Hamamoto, T.; Kurosawa, N.; Tsuji, S. *Biochemistry* **1994,** *33,* 5772–5776. doi:10.1021/bi00185a014
18. Giordanengo, V.; Bannwarth, S.; Laffont, C.; Miecem, V.; Harduin-Lepers, A.; Delannoy, P.; Lefebvre, J.-C. *Eur. J. Biochem.* **1997,** *247,* 558–566. doi:10.1111/j.1432-1033.1997.00558.x
19. Okajima, T.; Fukumoto, S.; Ito, H.; Kiso, M.; Hirabayashi, Y.; Urano, T.; Furukawa, K.; Furukawa, K. *J. Biol. Chem.* **1999,** *274,* 30557–30562. doi:10.1074/jbc.274.43.30557
20. Svennerholm, L. Ganglioside Designation. In *Structure and Function of Gangliosides;* Svennerholm, L.; Mandel, P.; Dreyfus, H.; Urban, P.-F., Eds.; Advances in Experimental Medicine and Biology, Vol. 125; Springer US: Boston, MA, USA, 1980; p 11. doi:10.1007/978-1-4684-7844-0_2
21. Chester, M. A. *Pure Appl. Chem.* **1997,** *69,* 2475–2488. doi:10.1351/pac199769122475
22. IUPAC-IUBCommission on Biochemical Nomenclature. *Eur. J. Biochem.* **1977,** *79,* 11–21. doi:10.1111/j.1432-1033.1977.tb11778.x
23. Groux-Degroote, S.; Guérardel, Y.; Julien, S.; Delannoy, P. *Biochemistry (Moscow)* **2015,** *80,* 808–819. doi:10.1134/s0006297915070020
24. Yu, R. K.; Tsai, Y.-T.; Ariga, T.; Yanagisawa, M. *J. Oleo Sci.* **2011,** *60,* 537–544. doi:10.5650/jos.60.537
25. Berselli, P.; Zava, S.; Sottocornola, E.; Milani, S.; Berra, B.; Colombo, I. *Biochim. Biophys. Acta, Gene Struct. Expression* **2006,** *1759,* 348–358. doi:10.1016/j.bbaexp.2006.07.001
26. Chisada, S.-i.; Yoshimura, Y.; Sakaguchi, K.; Uemura, S.; Go, S.; Ikeda, K.; Uchima, H.; Matsunaga, N.; Ogura, K.; Tai, T.; Okino, N.; Taguchi, R.; Inokuchi, J.; Ito, M. *J. Biol. Chem.* **2009,** *284,* 30534–30546. doi:10.1074/jbc.m109.016188
27. Rivas-Serna, I. M.; Polakowski, R.; Shoemaker, G. K.; Mazurak, V. C.; Clandinin, M. T. *J. Food Compos. Anal.* **2015,** *44,* 45–55. doi:10.1016/j.jfca.2015.06.006
28. Ando, S.; Hirabayashi, Y.; Kon, K.; Inagaki, F.; Kojima, K.; Tate, S.-i.; Whittaker, V. P. *J. Biochem.* **1992,** *111,* 287–290. doi:10.1093/oxfordjournals.jbchem.a123751
29. Crino, P. B.; Ullman, M. D.; Vogt, B. A.; Bird, E. D.; Volicer, L. *Arch. Neurol. (Chicago)* **1989,** *46,* 398–401. doi:10.1001/archneur.1989.00520400054019
30. Heffer-Lauc, M.; Cacic, M.; Serman, D. *Glycoconjugate J.* **1998,** *15,* 423–426. doi:10.1023/a:1006938221704
31. Saito, M.; Sugiyama, K. *Biochim. Biophys. Acta, Gen. Subj.* **2000,** *1474,* 88–92. doi:10.1016/s0304-4165(99)00222-6
32. Rösner, H. *J. Neurochem.* **1981,** *37,* 993–997. doi:10.1111/j.1471-4159.1981.tb04486.x
33. Furukawa, K.; Takamiya, K.; Furukawa, K. *Biochim. Biophys. Acta, Gen. Subj.* **2002,** *1573,* 356–362. doi:10.1016/s0304-4165(02)00403-8
34. Inokuchi, J.-i.; Go, S.; Yoshikawa, M.; Strauss, K. *Biochim. Biophys. Acta, Gen. Subj.* **2017,** *1861,* 2485–2493. doi:10.1016/j.bbagen.2017.05.025
35. Sturgill, E. R.; Aoki, K.; Lopez, P. H.; Colacurcio, D.; Vajn, K.; Lorenzini, I.; Majić, S.; Yang, W. H.; Heffer, M.; Tiemeyer, M.; Marth, J. D.; Schnaar, R. L. *Glycobiology* **2012,** *22,* 1289–1301. doi:10.1093/glycob/cws103
36. Lopez, P. H.; Aja, S.; Aoki, K.; Seldin, M. M.; Lei, X.; Ronnett, G. V.; Wong, G. W.; Schnaar, R. L. *Glycobiology* **2017,** *27,* 129–139. doi:10.1093/glycob/cww098
37. Hoops, S.; Sahle, S.; Gauges, R.; Lee, C.; Pahle, J.; Simus, N.; Singhal, M.; Xu, L.; Mendes, P.; Kummer, U. *Bioinformatics* **2006,** *22,* 3067–3074. doi:10.1093/bioinformatics/btl485
38. Matsuoka, Y.; Funahashi, A.; Ghosh, S.; Kitano, H. Modeling and Simulation Using CellDesigner. In *Transcription Factor Regulatory Networks;* Miyamoto-Sato, E.; Ohashi, H.; Sasaki, H.; Nishikawa, J.; Yanagawa, H., Eds.; Methods in Molecular Biology, Vol. 1164; Humana Press: New York, NY, USA, 2014; pp 121–145. doi:10.1007/978-1-4939-0805-9_11
39. Medley, J. K.; Choi, K.; König, M.; Smith, L.; Gu, S.; Hellerstein, J.; Sealfon, S. C.; Sauro, H. M. *PLoS Comput. Biol.* **2018,** *14,* e1006220. doi:10.1371/journal.pcbi.1006220
40. Herget, S.; Ranzinger, R.; Maass, K.; Lieth, C.-W. v. d. *Carbohydr. Res.* **2008,** *343,* 2162–2171. doi:10.1016/j.carres.2008.03.011
41. Banin, E.; Neuberger, Y.; Altshuler, Y.; Halevi, A.; Inbar, O.; Nir, D.; Dukler, A. *Trends Glycosci. Glycotechnol.* **2002,** *14,* 127–137. doi:10.4052/tigg.14.127
42. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. *Nucleic Acids Res.* **2016,** *44,* D1214–D1219. doi:10.1093/nar/gkv1031
43. Lombardot, T.; Morgat, A.; Axelsen, K. B.; Aimo, L.; Hyka-Nouspikel, N.; Niknejad, A.; Ignatchenko, A.; Xenarios, I.; Coudert, E.; Redaschi, N.; Bridge, A. *Nucleic Acids Res.* **2019,** *47,* D596–D600. doi:10.1093/nar/gky876
44. Iber, H.; Zacharias, C.; Sandhoff, K. *Glycobiology* **1992,** *2,* 137–142. doi:10.1093/glycob/2.2.137
45. Umaña, P.; Bailey, J. E. *Biotechnol. Bioeng.* **1997,** *55,* 890–908. doi:10.1002/(sici)1097-0290(19970920)55:6<890::aid-bit7>3.0.co;2-b
46. Krambeck, F. J.; Bennun, S. V.; Andersen, M. R.; Betenbaugh, M. J. *PLoS One* **2017,** *12,* e0175376. doi:10.1371/journal.pone.0175376
47. Jimenez del Val, I.; Nagy, J. M.; Kontoravdi, C. *Biotechnol. Prog.* **2011,** *27,* 1730–1743. doi:10.1002/btpr.688
48. McDonald, A. G.; Hayes, J. M.; Bezak, T.; Głuchowska, S. A.; Cosgrave, E. F. J.; Struwe, W. B.; Stroop, C. J. M.; Kok, H.; van de Laar, T.; Rudd, P. M.; Tipton, K. F.; Davey, G. P. *J. Cell Sci.* **2014,** *127,* 5014–5026. doi:10.1242/jcs.151878
49. Fisher, P.; Spencer, H.; Thomas-Oates, J.; Wood, A. J.; Ungar, D. *Cell Rep.* **2019,** *27,* 1231–1243.e6. doi:10.1016/j.celrep.2019.03.107
50. Liang, C.; Chiang, A. W. T.; Hansen, A. H.; Arnsdorf, J.; Schoffelen, S.; Sorrentino, J. T.; Kellman, B. P.; Bao, B.; Voldborg, B. G.; Lewis, N. E. *Curr. Res. Biotechnol.* **2020,** *2,* 22–36. doi:10.1016/j.crbiot.2020.01.001
51. McDonald, A. G.; Hayes, J. M.; Davey, G. P. *Curr. Opin. Struct. Biol.* **2016,** *40,* 97–103. doi:10.1016/j.sbi.2016.08.007

52. Manzi, A. E.; Sjoberg, E. R.; Diaz, S.; Varki, A. *J. Biol. Chem.* **1990,** *265,* 13091–13103. doi:10.1016/s0021-9258(19)38271-7
53. Fernandes Filho, J. A.; Shapiro, B. E. *Arch. Neurol. (Chicago)* **2004,** *61,* 1466–1468. doi:10.1001/archneur.61.9.1466
54. Kellman, B. P.; Zhang, Y.; Logomasini, E.; Meinhardt, E.; Godinez-Macias, K. P.; Chiang, A. W. T.; Sorrentino, J. T.; Liang, C.; Bao, B.; Zhou, Y.; Akase, S.; Sogabe, I.; Kouka, T.; Winzeler, E. A.; Wilson, I. B. H.; Campbell, M. P.; Neelamegham, S.; Krambeck, F. J.; Aoki-Kinoshita, K. F.; Lewis, N. E. *Beilstein J. Org. Chem.* **2020,** *16,* 2645–2662. doi:10.3762/bjoc.16.215

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

A systems-based framework to computationally describe putative transcription factors and signaling pathways regulating glycan biosynthesis

Theodore Groth[1], Rudiyanto Gunawan[1] and Sriram Neelamegham[*1,2,3,§]

Full Research Paper

Open Access

Address:
[1]Chemical and Biological Engineering, University at Buffalo, State University of New York, Buffalo, NY 14260, USA, [2]Biomedical Engineering, University at Buffalo, State University of New York, Buffalo, NY 14260, USA and [3]Medicine, University at Buffalo, State University of New York, Buffalo, NY 14260, USA

Email:
Sriram Neelamegham[*] - neel@buffalo.edu

* Corresponding author
§ Address for correspondence: 906 Furnas Hall, Buffalo, NY 14260, USA; phone: 716-645-1200; fax: 716-645-3822

Keywords:
ChIP-Seq; glycoinformatics; glycosylation; TCGA transcription factor

Beilstein J. Org. Chem. **2021,** *17,* 1712–1724.
https://doi.org/10.3762/bjoc.17.119

Received: 20 August 2020
Accepted: 12 July 2021
Published: 22 July 2021

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: N. H. Packer

Abstract

Glycosylation is a common posttranslational modification, and glycan biosynthesis is regulated by a set of glycogenes. The role of transcription factors (TFs) in regulating the glycogenes and related glycosylation pathways is largely unknown. In this work, we performed data mining of TF–glycogene relationships from the Cistrome Cancer database (DB), which integrates chromatin immunoprecipitation sequencing (ChIP-Seq) and RNA-Seq data to constitute regulatory relationships. In total, we observed 22,654 potentially significant TF–glycogene relationships, which include interactions involving 526 unique TFs and 341 glycogenes that span 29 the Cancer Genome Atlas (TCGA) cancer types. Here, TF–glycogene interactions appeared in clusters or so-called communities, suggesting that changes in single TF expression during both health and disease may affect multiple carbohydrate structures. Upon applying the Fisher's exact test along with glycogene pathway classification, we identified TFs that may specifically regulate the biosynthesis of individual glycan types. Integration with Reactome DB knowledge provided an avenue to relate cell-signaling pathways to TFs and cellular glycosylation state. Whereas analysis results are presented for all 29 cancer types, specific focus is placed on human luminal and basal breast cancer disease progression. Overall, the article presents a computational approach to describe TF–glycogene relationships, the starting point for experimental system-wide validation.

Introduction

The glycan signatures of cells and tissue are controlled by the expression pattern of 300–350 glycosylating-related genes that are together termed glycogenes [1,2]. These glycogenes include the glycosyltransferases, glycosidases, sulfotransferases, transporters, etc. The expression of these glycogenes is in turn driven by the action of a class of proteins called transcription factors

(TFs). These TFs regulate gene expression by binding proximal to the promoter regions of genes, facilitating the binding of RNA polymerases. They may homotropically or heterotropically associate with additional TFs in order to directly or indirectly control messenger RNA (mRNA) expression. Among the TFs, some "pioneer factors" can pervasively regulate gene regulatory circuits and access chromatin despite it being in a condensed state [3]. These TFs act as "master regulators", promoting the expression of several genes across many signaling pathways, such as differentiation, apoptosis, and cell proliferation. The precise targets of the TFs are controlled by their tissue-specific expression, DNA binding domains, and nucleosome interaction sequences [3]. Additional factors regulating transcriptional activity include: i) cofactors and small molecules that enable TF-DNA recognition and RNA polymerase recruitment [3]; ii) chromatin modifications, such as acetylation, methylation, and phosphorylation, which alter TF access; and iii) methylation of CpG islands in promoter regions that inhibit gene expression [4,5].

There are currently several isolated studies of TF–glycogene interactions, but a systematic "systems-level analysis" is absent. Many of these previous studies are based on discrete glycogene promoter region analysis and reporter assays. These studies have established some notable TF–glycogene relationships, though they are limited to distinct cell types. Examples include the regulation of MGAT5 by ETS2 in NIH3T3 fibroblasts [6], control of the α2-6 sialyltransferases ST6Gal-I/II by hypoxic nuclear factor 1-α (HNF1-α) in HepG2 cells [7], c-JUN-B3GNT8 regulatory relationships in gastric carcinoma cell lines [8], and SP1-B4GALT1 relations in lung cancer A549 cells [9]. A recent study also used computational predictions and wet-lab experiments to determine that ZNF263 is a potential heparin sulfate master regulator [10]. This TF regulates two sulfotransferases, HS3ST1 and HS3ST3A1. The above approaches have limitations: i) they do not consider the cellular epigenetic state that could impact TF binding; ii) proximal regulators are studied, but enhancers present several kilobases away from the transcription state site (TSS) are neglected; and iii) most of these reported TF–glycogene relationships only have partial support in established bioinformatics databases (DBs, see Supporting Information File 1). Thus, these are limited hypothesis-based investigations that do not describe the breadth of the regulatory landscape, based on current knowledge.

In the current article, we propose that more global and higher-throughput TF–glycogene relationships under biologically relevant conditions may be discovered using multiomics data mining. To this end, we sought to utilize multiomics experimental datasets and curated pathway DBs to relate cell-specific signaling processes to TFs, TFs to glycogenes, and glycogenes to glycosylation pathways (Figure 1A). These connections were made using data available from Cistrome Cancer DB [11], Reactome DB [12], and by the manual curation of various human glycogenes into pathways at GlycoEnzDB (https://virtu-

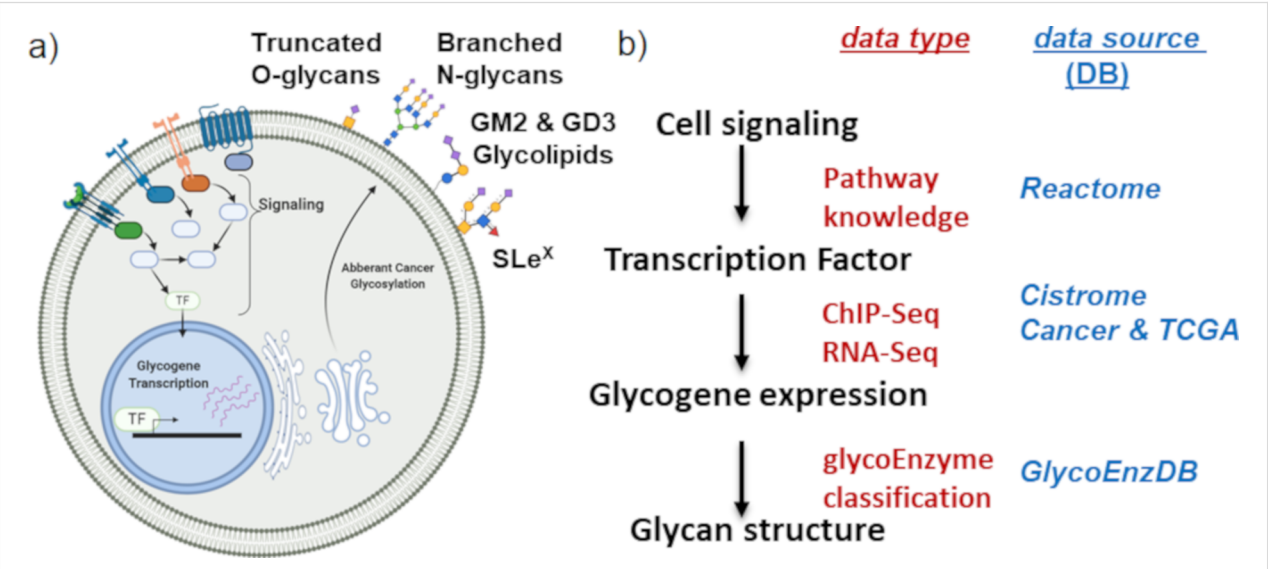


Figure 1: A systems glycobiology framework to link multi-OMICs data. a) Cell signaling proceeds to trigger TF activity. The binding of TFs to sites proximal to the TSS triggers glycogene expression. A complex set of reaction pathways then results in the synthesis of various carbohydrate types, many of which are either secreted or expressed on the cell surface. b) Data available at various resources can establish the link between cell signaling and glycan biosynthesis. The Reactome DB contains cell signaling knowledge. Chip-Seq and RNA-Seq data available at the Cistrome Cancer DB describe the link between the TFs and glycogenes. Pathway curation at GlycoEnzDB establishes the link between glycogenes and glycan structures. Cell illustration created using BioRender (https://biorender.com/).

alglycome.org/GlycoEnzDB, Figure 1B). Here, the Cistrome Cancer DB uses TF–gene binding data from previously published chromatin immunoprecipitation sequencing (ChIP-Seq) studies for various cell systems and cancer tissue RNA-Seq data from the Cancer Genome Atlas (TCGA) [13]. It provides putative TF–gene relationships for 29 TCGA cancer types provided they satisfy three inclusive criteria: i) TFs should be expressed at a high level in a given tissue; ii) changes in TF gene expression should correlate with RNA changes in target genes; and iii) ChIP-Seq data must support the TF–gene binding proximal to the TSS. Next, knowledge curated in the Reactome DB [12] was used to establish links between TFs and signaling pathways. In the final step, manually curated glycogene classifications were utilized to determine TFs that disproportionately regulate individual glycosylation pathways. It is important to note that the findings from this study represent computational inferences that are yet to be validated in the wet lab. Nevertheless, it provides a systems-based framework for the design and analysis of studies that link TFs to glycosylation pathways and glycan structures.

Results

TF–glycogene interaction map and relation to cell signaling pathways

The article follows a workflow shown in Figure 2. It mines TF–glycosylation pathway relationships from the Cistrome Cancer DB [14], which involves curating TF–gene relationships by integrating ChIP-Seq data from Cistrome DB and RNA-Seq data from TCGA. The Cistrome Cancer DB uses three filtering criteria to determine putative TF–gene relationships: i) The TF should be active in a cancer type, i.e., the reads per kilobase million (RPKM) value in a cancer type must be greater than the median RPKM expression of the TF across all 29 different cancer types; ii) the RNA expression of the TF and target gene should be correlated. To determine this, Cistrome first compares the selected TF–gene correlation with a null distribution computed by randomly selecting 1 million TF–gene pairs. Linear regression and statistical analysis are then performed on the top 5% hits (positive and negative coefficients) to establish TF–gene correlations. This analysis accounts for target gene copy number, tumor purity, and promoter methylation extent; and iii) TF–gene relationships must be supported by ChIP-Seq evidence. Here, a nonlinear weighted sum called regulatory potential (RP) quantifies the strength of TF–gene interactions based on the proximity of TF binding site to the gene TSS and also the number of TF–gene binding interactions based experimentally detected ChIP peaks [15,16].

In the current article, we passed the TF–gene relationships established in Cistrome Cancer DB to identify TFs potentially interacting with 341 glycogenes (Supporting Information File 3, Table S1). The two metrics for this selection were RP ≥ 0.5 and TF–glycogene expression correlation coefficient ρ ≥ 0.4. Such analysis was performed for 29 cancer types listed in Supporting Information File 3 (Table S2). Based on our selected thresholding, the analysis revealed 22,654 potential TF–glycogene interactions. The above data were used for two types of analysis described below. Here, the number of putative TF–glycogene relationships can be tuned by modifying the RP and ρ values.

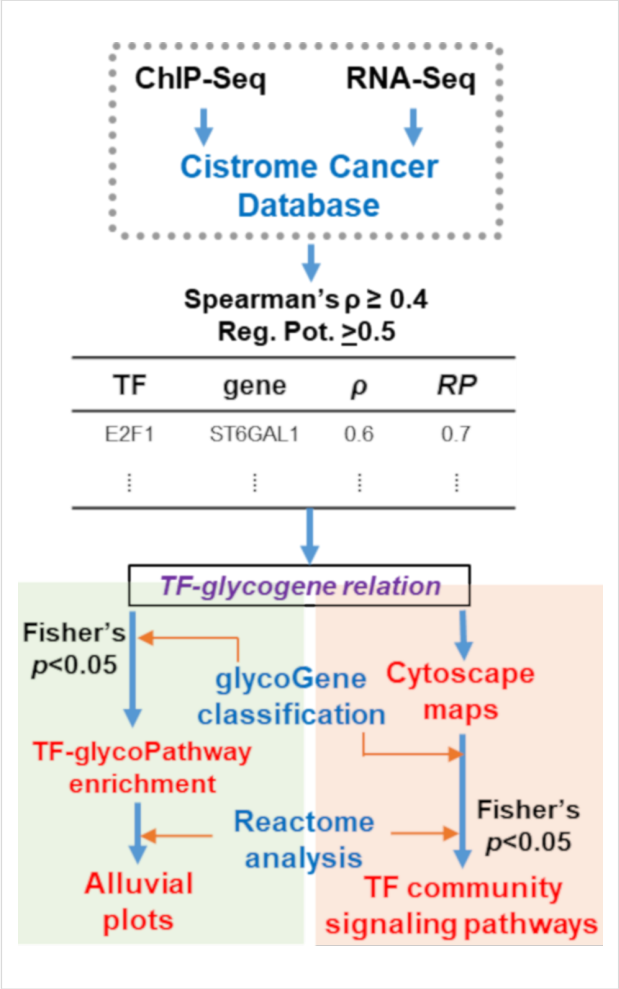


Figure 2: Analysis workflow: ChiP-Seq provides evidence of TF binding to promoter regions with 0 ≤ RP ≤ 1, quantifying the likelihood that this is functionally important. RNA-Seq quantifies Spearman's correlation (ρ) between TF and gene expression. Filtering these data establishes potential TF–glycogene interactions in specific cancer types. TFs disproportionately regulating specific glycosylation pathways were identified using the above TF–glycogene relationships as well as biochemical knowledge available at GlycoEnzDB (green region). Reactome DB analysis helped to establish cell signaling-TF–glycosylation pathway connectivity that are visualized using alluvial plots. Independently, Cytoscape maps enabled visualization of TF–glycogene relationships in different cancer types (orange region). Clusters in the resulting interactomes were related to pathway maps and signaling processes, and thus developing TF–community signaling pathway relationships.

First, the Fisher's exact test was used to infer TF–glycogene interactions that may regulate individual glycosylation pathways. This analysis was based on pathway classifications from GlycoEnzDB (Supporting Information File 3, Table S3) that grouped 208 glycogenes into 20 glycosylation pathways/groups. TFs having a disproportionately larger number of relationships with individual glycosylation pathways were determined with respect to all TF–glycogene relationships. Reactome DB was then used to associate these TFs to potential signaling pathways. This resulted in a relationship between cell signaling, TF activity regulation, and glycan structure changes (Supporting Information File 3, Tables S4 and S5). The data are presented as alluvial plots for the 29 cancer types (Supporting Information File 1). Here, the TFs were linked to glycosylation pathways by colored bands if they were found to regulate a disproportionately high fraction of glycogenes belonging to that pathway. Likewise, biological pathways were linked with TFs if that TF was found to be enriched in the biological pathway. Reading these alluvial plots from the left to the right, one can deduce which biological pathways may be potentially involved in regulating TFs, and how these TFs could regulate glycosylation.

Second, we visualized TF–glycogene interactions using Cytoscape maps for each of the cancer types individually (Supporting Information File 2). Regulatory modules were identified with graph clustering methods to identify groups of TFs that regulate common groups of glycogenes. Using our glycosylation pathway definitions, we used Fisher's exact test to describe what kinds of glycosylation pathways were disproportionately over-represented in each cluster. This analysis revealed 335 glycopathway enrichments in the TF–glycogene communities across the 29 cancer types (Supporting Information File 3, Table S6). Next, we determined, using the Reactome DB overrepresentation API, if the TFs identified in these clusters could be related to specific cell signaling pathways. Here, we noted 901 pathway enrichments across the different cancer types (Supporting Information File 3, Table S7). Common TFs that we observed across all TF–glycogene communities include the TCF and LEF families, FOXO and FOXP, the RUNX family, and IRF family TFs, which were found to regulate diverse glycosylation pathways, such as sialylation pathways, complex N-linked glycan synthesis, as well as chondroitin and dermatan sulfate synthesis.

Overall, the above analysis revealed the existence of communities of TF–glycogene relationships that could be linked to both cell signaling processes and specific glycosylation pathways.

TF–pathway relationships in breast cancer

We provide a more detailed description of our findings in breast cancer as an example. This disorder appears in 5 unique molecular subtypes based on the PAM50 classification [17]. These include the following: i) normal-like; ii) and iii) luminal A and luminal B, respectively, which overexpress estrogen receptor ESR1; iv) Her2+ tumors, which overexpress the epidermal growth factor receptor (ERBB); and v) basal (triple negative), which express neither ESR1 nor ERBB. Each of these subtypes has unique signaling mechanisms that may contribute to different glycan signatures.

In our analysis, TF–glycogene relationships for breast cancer derived by filtering Cistrome Cancer DB were enriched for the glycosylation pathways. Figure 3 summarizes these cancer-related TF–glycosylation pathway relationships for luminal (type A and B together) and basal breast cancer. Here, glycans potentially affected by the enriched TFs are shown in SNFG format [18,19]. The analysis suggests that TF transformations accompanying cancer progression may impact all four major classes of glycans: O- and N-glycans found on glycoproteins, glycosaminoglycans, and glycolipids. Thus, multiple glycan changes may accompany oncological transformation.

TF–glycogene communities in luminal and basal breast cancer

Cytoscape plots were generated for luminal breast cancer (Figure 4a). Here, using the bipartite graph community detection methods [20], we identified three large communities of TF–glycogene interactions. The largest community detected in this analysis had TFs enriched for RUNX3 signaling, IL-21 signaling, MECP2, and PTEN regulation. Overrepresentation glycosylation pathway analysis performed on the TFs in this community suggests that these TFs may regulate pathways related to sialylation, hyaluronan synthesis, as well as chondroitin and dermatan sulfate elongation. Here, STAT1, 4, and 5 proteins were enriched in the IL-21 signaling pathway. Luminal breast cancer types are known to express STAT1 and 3 as well as STATs 2 and 4. STAT5 is known to be constitutively active in luminal breast cancer and confers antiapoptotic characteristics to cells [21]. The other two communities detected consisted primarily of chromatin-modifying enzymes. Complex N-linked glycan synthesis and the dolichol pathway were significantly enriched in the second community. In the third community, O-linked mannose and LacdiNAc synthesis were disproportionately regulated. Overall, the pathway maps suggest that chromatin remodeling enzymes could potentially play roles in regulating glycan synthesis in luminal breast cancer.

Like luminal, basal breast cancer TF–glycogene relationships were also clustered into three communities. Here, the first community was enriched for chromatin-modifying enzymes, with complex N-linked glycan synthesis being the primary glycosylation pathway being affected (Figure 5a). The second

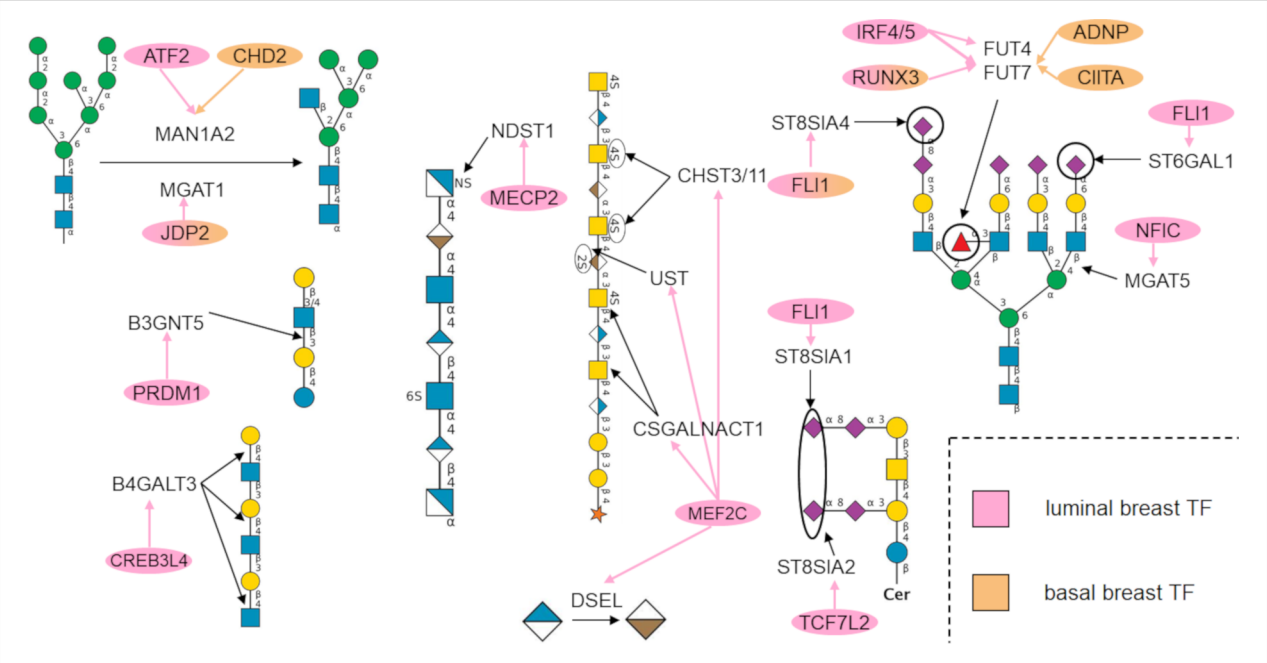


Figure 3: Summary of TFs enriched to glycosylation pathways for luminal and basal breast cancer: The TFs found to be enriched to glycosylation pathways and the glycogenes they regulate are shown in pink for luminal and orange for basal breast cancer. Note that some of the TFs shown above do not appear in the alluvial plots in the subsequent figures because they were not enriched to a signaling pathway in Reactome. The glycans synthesized by the enriched glycogenes are shown in SNFG format [18]. All figures were generated using DrawGlycan-SNFG [19].

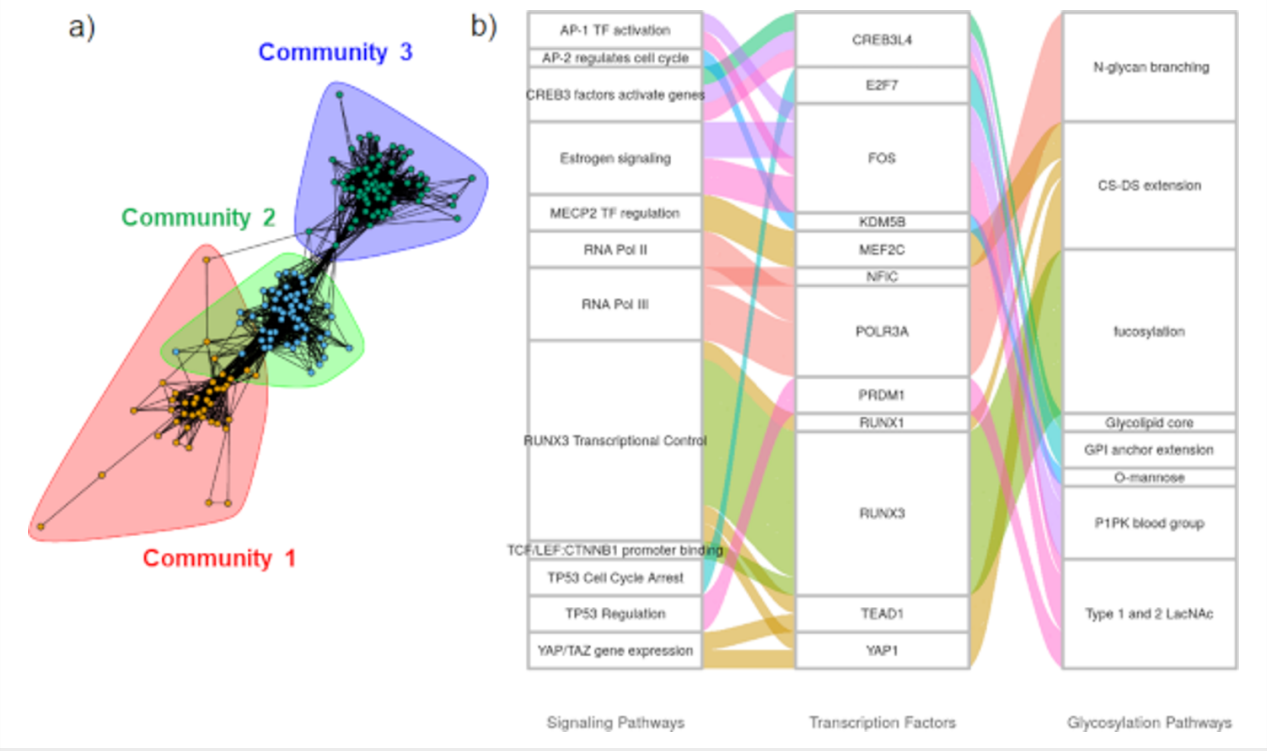


Figure 4: Luminal breast cancer signaling pathway enrichment and glycogene connections. a) TF-to-glycogene communities in luminal breast cancer: Three large TF-to-glycogene communities were discovered in the luminal breast subnetwork. Community 1 was enriched for pathways involving RUNX3, RUNX1, IL-21, and PTEN. Communities 2 and 3 consist primarily of chromatin-modifying enzymes. b) Signaling pathway enrichment analysis for luminal breast cancer: Connections between signaling pathways and TFs found to be statistically significant for luminal breast cancer. Some pathways enriched to TFs were condensed to conserve space. More TF-to-glycogene relationships exist in luminal breast cancer and these can be viewed in the Cytoscape figures (Supporting Information File 1).

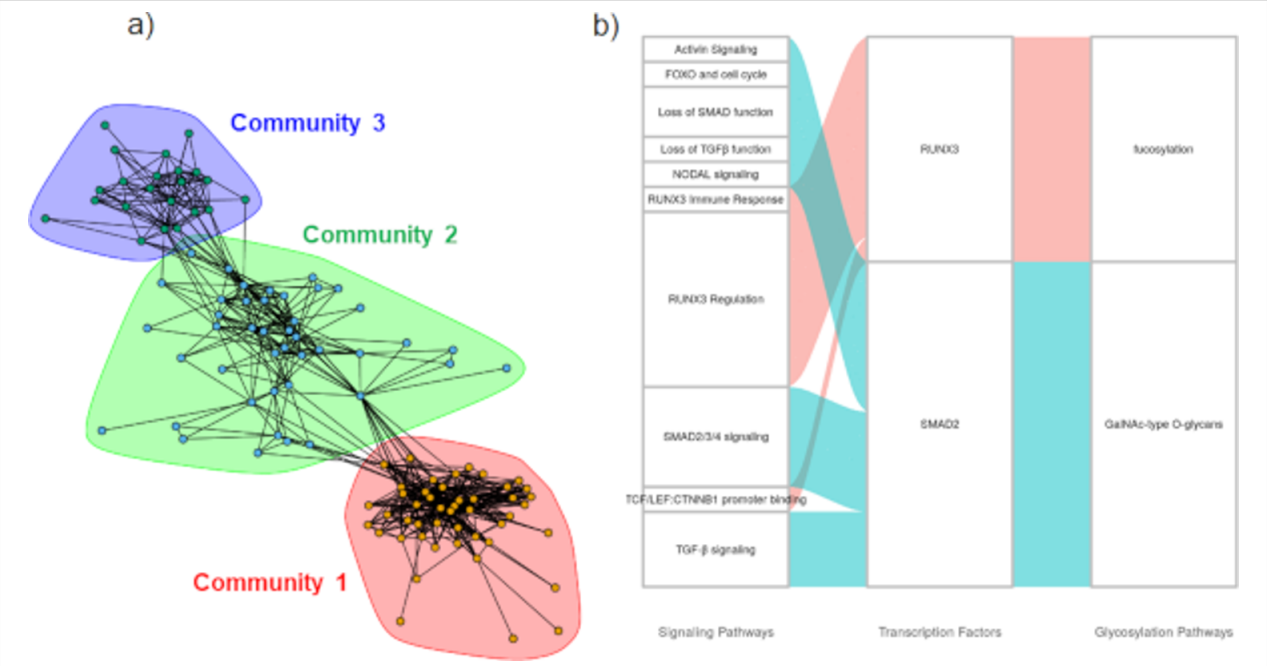


Figure 5: Basal breast cancer signaling pathway enrichments and glycogene connections. a) TF-to-glycogene communities in basal breast cancer: Three large TF-to-glycogene communities were discovered in the basal breast subnetwork. Community 1 has TFs enriched to chromatin-modifying enzymes, and community 2 has TFs enriched to interferon α/β/γ signaling. Community 3 did not have any signaling pathways enriched. b) Signaling pathway enrichment analysis for basal breast cancer: Connections between signaling pathways and TFs found to be statistically significant for basal breast cancer. TFs displayed have been enriched to the displayed glycosylation pathways using Fisher's exact test.

community was enriched for interferon α/β/γ signaling pathways, with interferon regulatory factor (IRF) TFs being enriched. In this regard, the TFs IRF-1 and IRF-5 have been shown to act as tumor suppressors in breast cancer [22,23]. Their loss of function in breast cancer could potentially downregulate O-linked fucosylation. The third community did not exhibit any specific TF pathway enrichments.

Linking cell signaling to TF and glycogenes for luminal breast cancer

The links between biological signaling pathways, TFs, and glycosylation pathways are shown in alluvial plots for luminal (Figure 4b) and basal breast cancer (Figure 5b), with additional plots provided for additional cancer types in Supporting Information File 1 for luminal breast cancer.

CREB3L4 and PRDM1 disproportionately affect the type I and II LacNAc pathway in luminal breast cancer: Our analysis suggests that CREB3L4 (enrichment p-value = 0.036) and PRDM1 (enrichment p-value = 0.039) may regulate the type 1 and 2 LacNAc pathways. CREB3L4 is known to primarily be expressed in the prostate and some breast cancer cell lines and has been linked to diverse roles involving chromatin organization in spermiogenesis, adipocyte regulation, and dysregulation in prostate cancer [24,25]. It has been found to be upregulated in breast cancer with respect to normal-like. PRDM1, also known as Blimp-1, is a transcriptional repressor, and its upregulation in cancer is known to dysregulate other proteins [26]. The increase poly-LacNAc structures have been shown to play roles in cancer metastasis [27]. CREB3L4 was found to regulate B4GALT3 glycogene (ρ = 0.56, RP = 0.94), which adds galactose in a β1-4 linkage. PRDM1 was found to regulate B3GNT5, which is critical for lacto/neolacto series of glycolipids (ρ = 0.60, RP = 0.84).

MEF2C disproportionately regulates glycosaminoglycan synthesis pathways: MEF2C was found to regulate several genes in the chondroitin and dermatan sulfate synthesis pathways (p = 0.008). This TF plays roles in development, particularly in the development of neurons and hematopoietic cell differentiation towards myeloid lineages. It is known that MEF2C is directly impacted by TGF-β signaling, and thus increasing the metastatic potential of cancer [28]. MEF2C was found to be inhibited by MECP2 based on Reactome pathway enrichment. Since the glycosaminoglycan elongation pathways positively correlate to MEFC2 expression and MEFC2 is amplified in cancer, it is possible that MECP2 may not be sufficiently expressed to repress MEFC2 in call cancer cells. MEF2C was found to regulate CSGALNACT1 (ρ = 0.66, RP = 0.71), CHST3 (ρ = 0.50, RP = 0.74), CHST11 (ρ = 0.47, RP = 0.84), DSEL (ρ = 0.40, RP = 0.81), and UST (ρ = 0.42, RP = 0.95). Here, CSGALNACT1 is responsible for the addition of GalNAc

to glucuronic acid to increase chondroitin polymer length, CHST3, CHST11, and UST are involved in the sulfation of GalNAc and iduronic acid, and DSEL is the epimerase which converts glucuronic acid to iduronic acid in CS/DS chains.

MECP2 disproportionately regulates heparan sulfate chain elongation: The MECP2 (enrichment p-value = 0.037) was found to positively regulate heparan sulfate elongation. MECP2 regulates gene expression by binding to methylated promoters and then by recruiting chromatin remodeling proteins to condense DNA and repress gene expression [29,30]. MECP2 was found to regulate sulfotransferase NDST1 (ρ = 0.41, RP = 0.67).

Linking cell signaling to TF and glycogenes for basal breast cancer

Fewer TFs were found to be enriched to signaling pathways in basal breast cancer compared to luminal cancer (Figure 4b). Despite this, there are many other TF–glycosylation pathway enrichments for basal breast cancer available for analysis in Supporting Information File 1. The roles of two enriched TFs and their relation to glycogenes and cancer is elaborated below.

RUNX3 and fucosylation: The terminal fucosyltransferase FUT7 (ρ = 0.49, RP = 0.89) was found to be positively regulated by the RUNX3 TF (enrichment p-value = 0.033). The RUNX family of TFs (including RUNX1–3), are involved in several developmental processes, including hematopoiesis, immune cell activation, and skeletal development. It was discovered that RUNX3 acts as a tumor suppressor gene in breast cancer. Upon cancer development, the RUNX3 promoter is hypermethylated, leading to reduced TF activity and loss of tumor suppression activity [31]. Our data suggest that this may be associated with a reduction of FUT7 activity, and thus impacting the expression of the sialyl Lewis-X antigens in basal tumors. Sialyl Lewis-X is considered to be an important regulator of cancer metastasis as it binds the selectins on various vascular and blood cell types.

Regulation of GalNAc-type O-linked glycans by SMAD2: SMAD2 was found to significantly affect core 1 and 2 O-linked glycan structures (enrichment p-value = 0.035). SMAD proteins are activated by TGF-β signaling and bind to DNA to act as cofactors to recruit TFs. SMAD2 has been shown to act as a tumor metastasis suppressor in cell lines [32,33]. This TF was found to regulate GALNT1 (ρ = 0.54, RP = 1.00), which adds GalNAc to serine or threonine residues to being core 1 and 2 O-linked glycan synthesis. Thus, SMAD2 may play a key role in regulating Tn antigen expression in proteins such as MUC-1 that are associated with breast cancer progression.

Refinement of TF–glycopathway enrichments after false discovery correction

The number of enrichments above is high. In order to reduce the findings to a smaller set, we applied the Benjamini–Hochberg correction to our TF–glycopathway enrichments. While possibly reducing false positives, this may also reduce true positives. Nevertheless, after this correction, a total of 121 TF–glycopathway enrichments were found to be statistically significant across all cancer types (Figure 6 and Supporting Information File 3, Table S8). Here, basal breast cancer (BRCA_2), adrenocortical carcinoma (ACC), liver hepatocellular carcinoma (LIHC), lung squamous cell carcinoma (LUSC), and skin cutaneous melanoma (SKCM) did not have any TFs enriched to any glycopathway, and thus are not depicted.

Filtering our TF–glycopathway enrichments illuminates the fact that pancreatic cancer shows a high degree of enrichment to the GalNAc-type O-glycan pathways, which is consistent with our prior experiments [34]. FOXA1 (P_{adj} = 0.00096), KLF5 (P_{adj} = 0.0012), MECOM (P_{adj} = 0.029), and TCF7L2 (P_{adj} = 0.000087) were found to regulate several GalNAc transferases. FOXA1 is an important regulatory TF involved in the development of endoderm-derived organs. Upon pancreatic cancer development, FOXA1 expression is known to decrease, which drives the epithelial to mesenchymal transition [35]. Kruppel-like factor 5 (KLF5) is commonly upregulated in several cancer types and promotes pancreatic cancer proliferation by targeting the cell cycle [36]. MECOM (also known as PRDM3) is a nuclear TF known to ablate inflammatory responses and tumorigenesis in pancreatic cancer contexts [37]. Transcription factor 7-like 2 (TCF7L2) is regulated by Wnt β-catenin signaling. This TF is important in gluconeogenesis in the liver, adipogenesis, regulation of hormone synthesis, and pancreas homeostasis. TCF7L2 exhibits polymorphisms which results in loss of function and can promote metastatic phenotypes in colorectal cancer [38]. O-Linked glycosylation via GALNT3 and B3GNT3 has been shown to regulate differentiation of pancreatic cancer stem cells [39]. FOXA1 (RP = 0.97, ρ = 0.49) and KLF5 (RP = 0.71, ρ = 0.68) were found to regulate GALNT3, and KLF5 (RP = 0.98, ρ = 0.67) and TCF7L2 (RP = 0.95, ρ = 0.62) were found to regulate B3GNT3. Since KLF5 and TCF7L2 have been shown to be upregulated in pancreatic cancer stem cells, it would be interesting to validate if GALNT3 and B3GNT3 are driven by any of these TFs.

Discussion

In the current analysis, we mined public high-throughput ChIP-Seq and RNA-Seq data to identify putative TF–glycogene relationships across 29 different cancer types. Approximately three glycogenes were regulated by a given TF based on our filtering

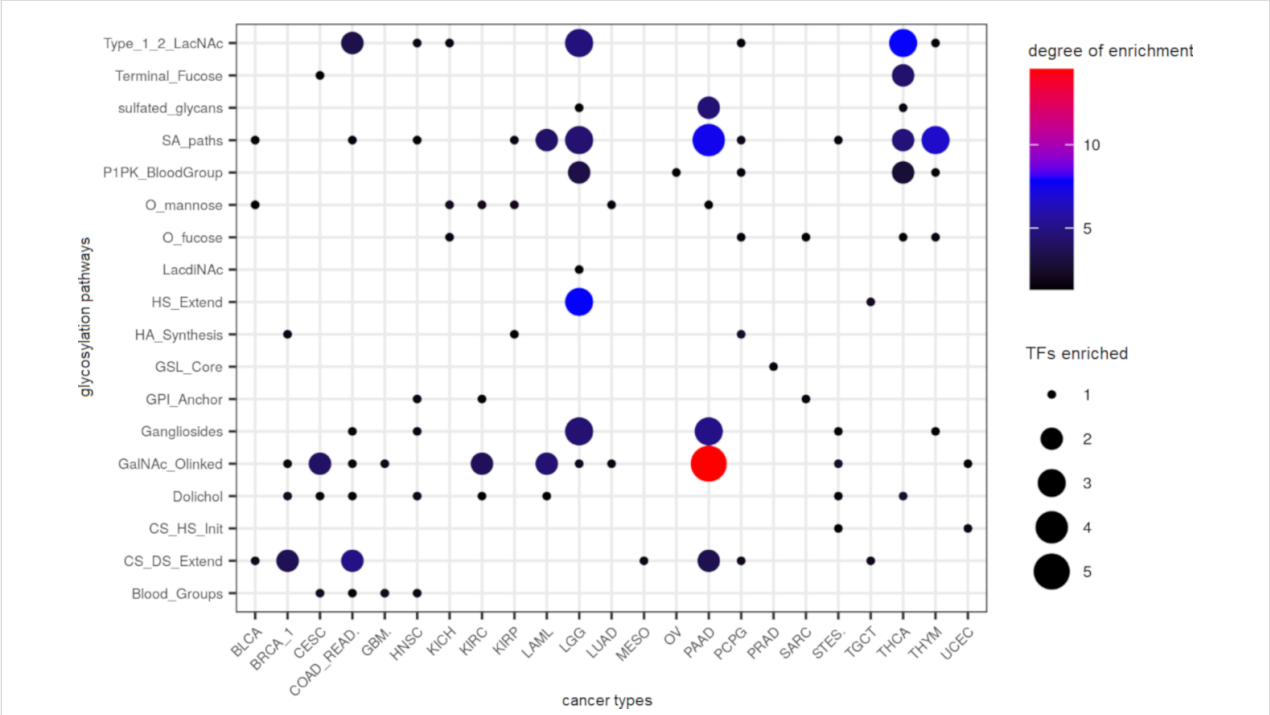


Figure 6: Summary of TF–glycopathway enrichments across all cancer types: TF enrichments to glycopathways across all cancer types are depicted as dots (Fisher's exact test adjusted $P < 0.05$ for overrepresentation). The dot size corresponds to the number of TFs that were found to regulate the pathway. The degree of regulation is defined as the sum of all $-\log_{10}$ (adjusted enrichment p-values) across all TFs for a given cancer–glycopathway pair.

criteria, with this number ranging from 1–10. These findings are tissue-specific, as TF and glycogene expression vary widely among the different cell types. The analysis also suggests putative TF–glycogene interactions that disproportionately impact specific glycosylation pathways. Knowing which TF regulates which glycogene and pathway in a context-dependent manner can provide insight as to how signaling pathways contribute to altered glycan structures in diseases such as diabetes and cancer. Thus, this work represents a rich starting point for wet-lab validation and glycoinformatics DB construction.

Visualizing TF–glycogene interaction networks revealed communities of glycogenes in each cancer type. The presence of chromatin-modifying enzymes in large regulatory communities in both luminal and basal breast cancer suggests a role of epigenetics in glycogene regulation. To date, a systems-level investigation evaluating the epigenetic states of cell systems on the resulting glycome has not been performed. Our results suggest that complex N-linked branching and glycosylation may be sensitive to these processes. The signaling pathways enriched in the largest community in luminal breast cancer were reflected in our pathway enrichment findings. RUNX3, interleukin signaling, and the involvement of MECP2 regulation were all found to disproportionately regulate sialic acid and GAG synthesis pathways.

Several of the TFs enriched to glycosylation pathways were either regulated by or involved in TGF-β signaling and Wnt β-catenin signaling. These TFs primarily affected glycosaminoglycan synthesis pathways, sialylation, and type-2 LacNAc synthesis. Cell cycle and metabolic regulatory TFs were shown to regulate some glycogenes involved in the dolichol pathway. The crosstalk between cell cycle and glycosylation is not well explored and may potentially be important for understanding N-linked glycosylation flux in cancer. Some TFs were found to interact with methyl CpG-binding TFs when regulating glycosaminoglycan proteins, implicating methylation as a possible modulator of glycosylation in cancer.

Our TF–glycogene relationships, mined from Cistrome Cancer DB, represent a starting point for experimentally discovering the TFs regulating glycosylation. The findings would likely vary between cell types, and thus additional efforts are necessary before a wet-lab-validated framework emerges. Orthogonal datasets containing other ChIP-Seq and omics data may also enhance in silico validation. Some examples include: i) data from the Gene Transcription Regulatory Database (GTRD) [40], which has analyzed publicly available ChIP-Seq data with multiple algorithms to systematically catalog TF–gene relationships across several organisms and cellular contexts; ii) the Regulatory Circuits DB [41], which relies on the activity of

promoter and enhancer regions through cap analysis of gene expression (CAGE), TF motif instances, and expression quantitative trait loci (eQTL) to evaluate weights (evidence scores) for TF–gene isoform relationships; and iii) integration of TF-binding motifs, protein–protein interactions, and coexpression networks using data from GTEx and a method called PANDAS [42]. Such analyses represent next steps in this project, as extensive data harmonization is required for cross-platform validation. Care should be taken when integrating these data, however, as the kind of omics data, degree of experimental evidence, and the statistical approaches taken by other investigators can influence the set of TF–gene relationships found. In addition to in silico validation, perturbational experiments, such as performing CRISRP-Cas9 knockouts with single-cell RNA-Seq, followed by glycomics/glycoproteomics-based mass spectrometry, would further support the proposed TF–glycogene relationships [43].

Some caveats in our analysis are important to note. First, we only used selected values of RP and ρ to filter TF–glycogene relationships from the Cistrome Cancer DB. Further studies are needed in order to determine how the selected thresholds affect the discovered relationships. A full list of TF–glycogene relationships found Cistrome Cancer DB are provided in Supporting Information File 3 (Table S9) for readers to test alternative thresholds. Second, the glycogenes in individual pathways in this article were classified using current knowledge of glycobiology. Different classification methods meant to address different glycosylation pathways may result in different TF–glycopathway enrichments [44]. Third, while Cistrome Cancer DB systematically filters TF–gene relationships based on ChIP-Seq and RNA-Seq evidence, the DB has some biases. In one aspect, only TFs that were considered to be sufficiently expressed were considered in this analysis. Lower expressed TFs that may also be functional are excluded. Additionally, while RNA-Seq relationships in Cistrome Cancer DB are selected based on the specific tissue type, supporting ChIP-Seq evidence is not cell-type-specific. Regardless of these limitations, the current study presents a framework for thinking in the glycosciences, so that knowledge of genes and transcripts can be linked to glycans and their function [2].

Conclusion

A majority of current studies in the Glycoscience field use experimental data and curations related to glycans only. Fewer investigations examine the links between the glycans, glycogenes and glycosylation pathways, and other nonglyco datasets. We set out to identify these relationships by mining publicly-available data. Using this, we describe putative regulatory relationships between TFs and glycogenes across 29 cancer types. Some TFs appear to regulate glycogenes in communities, indicating potential cross-talk across pathways in regulating glycosylation. The communities varied with cancer type, even in a single tissue, suggesting that these TF–glycogene interactions are dynamic in nature. Groups of TFs enriched to glycosylation pathways were also associated with signaling pathways. Thus, a connection between cell signaling, TF activity and glycosylation begins to emerge. Overall, the putative TF–glycosylation pathway enrichments found here represent the starting point for wet-lab and orthogonal dataset validation. Such studies could enhance our fundamental understanding of glycosylation pathway regulation, and lead to novel ways to control the glycogenes and glycan structures during health and disease.

Experimental

Glycogene-pathway classification

A list of 208 unique glycogenes involved in 20 different glycosylation pathways were used in this work (Supporting Information File 3, Table S3). These data were collated from GlycoEnzDB (https://virtualglycome.org/GlycoEnzDB), with original data coming from various sources in literature [45,46]. The following is a summary of the pathways studied and the enzymes involved:

1) Glycolipid core: The enzymes in this group are involved in the biosynthesis of the glucosylceramide (GlcCer) and galactosylceramide (GalCer) lipid core. Here, the GlcCer core is formed by the UDP-glucose:ceramide glucosyltransferase (UGCG), which transfers the first glucose. Following this, lactosylceramide is formed by the action of the β1-4GalT activity of B4GalT5 (and possibly also B4GalT3, 4, and 6). The GalCer core is typically structurally small and is made by UDP-Gal:ceramide galactosyltransferase (UGT8). These structures can be further sulfated by GAL3ST1 or sialylated by ST3GAL5.

2) P1-Pk blood group: The Pk, P1, and P antigens are synthesized on lactosylceramide glycolipid core. The activity of α1-4GalT (A4GALT) on this core results in the Pk antigen, followed by β1-3GalNAcT (B3GALNT1) to form the P antigen. The P1 antigen, on the other hand, is formed by the sequential action of β1-3GlcNAcT (B3GNT5), β1-4GalT (B4GALT1-6), and α1-4GalT (A4GALT) on the glycolipid core.

3) Gangliosides: This pathway encompasses all glycogenes responsible for synthesizing a/b/c gangliosides. UGCG is included to consider the addition of glucose to ceramide. ST3GAL5 and ST8SIA enzymes are added to take the core ganglioside structures to the a, b, and c levels. B4GALTs and B4GALNT1 are included to account for ganglioside elongation. Decoration of the gangliosides with sialic acid occurs using ST6GALNAC3-6 and also ST8SIA1/3/5.

4) Dolichol pathway: This results in the formation of the dolichol-linked 14-monosaccharide precursor oligosaccharide. This glycan is cotranslationally transferred en bloc onto Asn-X-Ser/Thr sites of the newly synthesized protein as it enters the endoplasmic reticulum. The enzymes involved is such synthesis include the ALG (asparagine-linked N-glycosylation) enzymes and additional proteins (part of OSTA and OSTB) involved in the transfer of the glycan to the nascent protein.

5) Complex N-glycans: This pathway includes glycogenes responsible for processing the N-linked precursor structure emerging from the dolichol pathway into complex structures. Enzymes involved include mannosidases, glucosidases, some enzymes facilitating protein folding, and also enzymes that direct acid hydrolases to the lysosome.

6) N-glycan branching: These glycogenes are responsible for the addition of GlcNAc to processed N-linked glycan structures. These include all the MGAT enzymes.

7) GalNAc-type O-glycans: O-linked glycans are attached to serine (Ser) or threonine (Thr) on peptides, where GalNAc is the root carbohydrate. This is mediated by a family of about 20 Golgi-resident polypeptide *N*-acetylgalactosaminyltransferases (ppGalNAcTs or GALNTs). Core 1 structures result from the attachment of β1-3 linked galactose to the core GalNAc using C1GALT1 and the corresponding chaperone C1GALT1C1. Core 2 structures then form upon addition of β1-6-linked GlcNAc by GCNT1. Modifications of core 3 and core 4 glycans can occur during disease, and thus this classification includes core 3-forming B3GNT6 and core 4-forming GCNT3. Other O-glycan core types are rare in nature.

8) Chondroitin sulfate and heparan sulfate initiation: Chondroitin and heparan sulfate glycosaminoglycans all have a common core carbohydrate sequence attaching them to the corresponding proteins. These are constructed by the activity of specific xylotransferases (XYLT1 and XYLT2), galactosyltransferses B4GALT7 and B3GALT6 that sequentially add two galactose residues to xylose, and the glucuronyltransferase B3GAT3 that adds glucuronic acid to the terminal galactose. Also involved in the formation of this core is FAM20B, a kinase that 2-*O*-phosphorylates xylose. At this point, the addition of GalNAc to GlcA by CSGALNACT1 and 2 results in the initiation of chondroitin sulfate chains. The attachment of GlcNAc by EXTL3 to the same GlcA results in heparan sulfates.

9) Chondroitin sulfate and dermatan sulfate extension: Chondroitin sulfates and dermatan sulfates are extended via the addition of GalNAc-GlcA repeat units. This is catalyzed by CSGALNACT1, which is better suited for the initial GalNAc attachment, followed by CSGALNACT2, which is preferred for synthesizing disaccharide repeats. CHSY1, CHSY3, CHPF, and CHPF2 all exhibit dual β1-3GlcAT and β1-4GlcAT activity. Additional enzymes mediate sulfation. Epimerization of glucuronic acid to iduronic acid by DSE and DSEL results in the conversion of chondroitin sulfates to dermatan sulfates.

10) Heparan sulfate extension: EXT1 and EXT2 both have GlcUA and GlcNAc transferase activities and are together responsible for HS chain polymerization. EXTL1–3 are additional enzymes with GlcNAc transferase activity that facilitate heparin sulfate biosynthesis. Additional enzymes that are critical for heparin sulfate function include the HS2/3/6ST sulfotransferases, the GlcA epimerase GLCE, and additional enzymes mediating N-sulfation (i.e., NDSTs).

11) Hyaluronan synthesis: This pathway consists of the three hyaluronan synthases, HAS1–3.

12) Glycophosphatidylinositol (GPI) anchor extension: This pathway includes glycogenes responsible for the synthesis of GPI-anchored proteins in the ER. This involves the synthesis of a glycan–lipid precursor that is en bloc transferred to proteins.

13) O-Mannose: This is initiated by the addition of mannose to Ser/Thr using POMT1 or POMT2. β1-2 or β1-4 GlcNAc linkages can then be made using POMGNT1 or POMGNT2 to yield M1 or M3 O-linked mannose structures, respectively. MGAT5B can facilitate β1-4 GlcNAc linkage onto the M1 structure to yield the M2 core. Additional carbohydrates typically found on complex N-linked glycan antennae can then be attached. In particular, such extensions may be initiated by members of the B4GALT family or B3GALNT2. Specific variants are noted on α-dystroglycans.

14) O-linked fucose: This pathway includes POFUT1, the enzyme responsible for the addition of fucose to Ser/Thr residues. MFNG, LFNG, and RFNG can attach β3GlcNAc to this fucose.

15) Type 1 and 2 LacNAc: These enzymes help construct either Galβ1-3GlcNAc (type 1) or Galβ1-4GlcNAc (type 2) lactosamine chains on antennae of N-linked glycan, O-linked glycans, and glycolipids. Also included are GCNT1–3 that can facilitate formation of I-branches on N-glycans.

16) Sialylation: This group encompasses all kinds of sialyltransferases: ST6GAL, ST3GAL, ST8SIA, and ST6GALNACs. Enrichments to this pathway capture overall increase in sialylation regardless of context.

17) Fucosylation: these include α1-2 (FUT1, 2) and α1-3 (FUT3–7, 9) fucosyltransferases that can act on N-glycans, O-glycans and glycolipids.

18) ABO blood group synthesis: these are enzymes involved in the biosynthesis of ABO antigens.

19) LacDiNAc: glycogenes involved in the synthesis of LacDiNac structures.

20) Sulfated glycan epitopes: this includes the enzymes attaching sulfate to different types of carbohydrates.

Mining TF–glycogene relationships in Cistrome Cancer DB

Regulatory potential and gene correlation data were downloaded from the Cistrome Cancer DB in tab-delimited form (http://cistrome.org/CistromeCancer/CancerTarget/) [14]. TF–gene relationships were filtered for the 341 glycogenes in this article (Supporting Information File 3, Table S1). In total, the full dataset contained 45,238 TF-to-glycogene relationships, including relational data for 570 unique TFs found in the 29 cancer systems across all the glycogenes. Positive regulatory relationships between TFs and glycogenes were selected based on RP ≥ 0.5 and ρ ≥ 0.4 (Figure 2). This filtering resulted in 22,654 TF–glycogene relationships including 526 unique TFs across 29 cancer types.

Cytoscape was used to visualize TF–glycogene regulatory relationships [47]. To achieve this, all TF–glycogene relationship data were loaded into Cytoscape as a network. These data were filtered based on RP and ρ thresholds defined previously. A binding potential (BP) score was computed by taking the product of RP and ρ for each TF–glycogene relationship. TF–glycogene relationships for each cancer type were separated into subnetworks. The Prefuse Force Directed Layout algorithm in Cytoscape was used to arrange nodes in each cancer subnetwork. The closeness of nodes to one another is weighted by 1-BP. Thus, nodes with high BPs will be placed closer together, whereas smaller BPs will be placed further away. Since there are two classes of nodes (TFs and glycogenes), we treated TF–glycogene networks as bipartite and applied the corresponding procedure for community detection [20]. Firstly, the bipartite TF–glycogene graphs are projected into two different unipartite graphs, where TFs and glycogenes are placed into separate graphs. The edge weights connecting TFs is computed as the number of shared glycogenes they regulate. The TF unipartite graph was then subjected to a greedy modularity optimization-based approach implemented in the igraph R package [48]. TF–glycogene interactions in each community were subjected to overrepresentation analyses to identify enriched signaling and glycosylation pathways.

Relating TF–glycogene interactions to glycosylation and signaling pathways

A one-sided Fisher's exact test was applied to determine if a particular TF disproportionately regulates one of the 20 glycosylation pathways described in Supporting Information File 3 (Table S3). Input data to the test consisted of all TF–gene interactions that passed the RP and ρ thresholds for the cancer type being analyzed. TFs were considered to be disproportionately regulating a glycosylation pathway if Fisher's exact test resulted in a p-value ≤ 0.05. These p-values were then adjusted using the Benjamini–Hochberg method to identify the strongest enrichments across all cancer types.

TFs enriched to glycosylation pathways were associated with putative regulatory pathways using the Reactome DB overrepresentation analysis API, which also uses Fisher's exact test, to associate the TFs with signaling pathways [12]. Signaling pathway enrichments with adjusted p(FDR) < 0.1 were kept. A high p-value cutoff was chosen to allow users to gain a high-level perspective as to what potential pathways may be regulating enriched TFs. The connection between cell signaling pathways and TFs and that between the TFs and glycosylation pathways were visualized using alluvial plots generated using the R package ggalluvial. Only signaling pathways with <30 members are presented for brevity. A comprehensive listing of enriched signaling pathways is available in Supporting Information File 3 (Table S5).

Supporting Information

Supporting Information File 1

Comparison of wet-lab studies and entries in DBs as well as Alluvial plots for all cancer types.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-17-119-S1.pdf]

Supporting Information File 2

Cistrome Cancer TF-to-glycogene subnetworks.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-17-119-S2.cys]

Supporting Information File 3

Supplementary tables.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-17-119-S3.zip]

Funding

This work was supported by US National Institutes of Health grants HL103411, GM133195, and GM126537.

ORCID® iDs

Sriram Neelamegham - https://orcid.org/0000-0002-1371-8500

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://biorxiv.org/cgi/content/short/2020.08.19.257956v1

References

1. Zhu, Y.; Groth, T.; Kelkar, A.; Zhou, Y.; Neelamegham, S. *Glycobiology* **2021,** *31,* 173–180. doi:10.1093/glycob/cwaa074
2. Neelamegham, S.; Mahal, L. K. *Curr. Opin. Struct. Biol.* **2016,** *40,* 145–152. doi:10.1016/j.sbi.2016.09.013
3. Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R.; Weirauch, M. T. *Cell* **2018,** *172,* 650–665. doi:10.1016/j.cell.2018.01.029
4. Namba, S.; Sato, K.; Kojima, S.; Ueno, T.; Yamamoto, Y.; Tanaka, Y.; Inoue, S.; Nagae, G.; Iinuma, H.; Hazama, S.; Ishihara, S.; Aburatani, H.; Mano, H.; Kawazu, M. *Cancer Sci.* **2019,** *110,* 1096–1104. doi:10.1111/cas.13937
5. Malta, T. M.; de Souza, C. F.; Sabedot, T. S.; Silva, T. C.; Mosella, M. S.; Kalkanis, S. N.; Snyder, J.; Castro, A. V. B.; Noushmehr, H. *Neuro-Oncology (Cary, NC, U. S.)* **2018,** *20,* 608–620. doi:10.1093/neuonc/nox183
6. Chen, L.; Zhang, W.; Fregien, N.; Pierce, M. *Oncogene* **1998,** *17,* 2087–2093. doi:10.1038/sj.onc.1202124
7. Svensson, E. C.; Conley, P. B.; Paulson, J. C. *J. Biol. Chem.* **1992,** *267,* 3466–3472. doi:10.1016/s0021-9258(19)50754-2
8. Jiang, Z.; Liu, Z.; Zou, S.; Ni, J.; Shen, L.; Zhou, Y.; Hua, D.; Wu, S. *Oncol. Rep.* **2016,** *36,* 1353–1360. doi:10.3892/or.2016.4959
9. Muramoto, K.; Tange, R.; Ishii, T.; Miyauchi, K.; Sato, T. *Biol. Pharm. Bull.* **2017,** *40,* 1282–1288. doi:10.1248/bpb.b17-00212
10. Weiss, R. J.; Spahn, P. N.; Toledo, A. G.; Chiang, A. W. T.; Kellman, B. P.; Li, J.; Benner, C.; Glass, C. K.; Gordts, P. L. S. M.; Lewis, N. E.; Esko, J. D. *Proc. Natl. Acad. Sci. U. S. A.* **2020,** *117,* 9311–9317. doi:10.1073/pnas.1920880117
11. Mei, S.; Qin, Q.; Wu, Q.; Sun, H.; Zheng, R.; Zang, C.; Zhu, M.; Wu, J.; Shi, X.; Taing, L.; Liu, T.; Brown, M.; Meyer, C. A.; Liu, X. S. *Nucleic Acids Res.* **2017,** *45,* D658–D662. doi:10.1093/nar/gkw983
12. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; Loney, F.; May, B.; Milacic, M.; Rothfels, K.; Sevilla, C.; Shamovsky, V.; Shorser, S.; Varusai, T.; Weiser, J.; Wu, G.; Stein, L.; Hermjakob, H.; D'Eustachio, P. *Nucleic Acids Res.* **2020,** *48,* D498–D503. doi:10.1093/nar/gkz1031
13. Grossman, R. L.; Heath, A. P.; Ferretti, V.; Varmus, H. E.; Lowy, D. R.; Kibbe, W. A.; Staudt, L. M. *N. Engl. J. Med.* **2016,** *375,* 1109–1112. doi:10.1056/nejmp1607591
14. Mei, S.; Meyer, C. A.; Zheng, R.; Qin, Q.; Wu, Q.; Jiang, P.; Li, B.; Shi, X.; Wang, B.; Fan, J.; Shih, C.; Brown, M.; Zang, C.; Liu, X. S. *Cancer Res.* **2017,** *77,* e19–e22. doi:10.1158/0008-5472.can-17-0327
15. Tang, Q.; Chen, Y.; Meyer, C.; Geistlinger, T.; Lupien, M.; Wang, Q.; Liu, T.; Zhang, Y.; Brown, M.; Liu, X. S. *Cancer Res.* **2011,** *71,* 6940–6947. doi:10.1158/0008-5472.can-11-2091
16. Wang, S.; Sun, H.; Ma, J.; Zang, C.; Wang, C.; Wang, J.; Tang, Q.; Meyer, C. A.; Zhang, Y.; Liu, X. S. *Nat. Protoc.* **2013,** *8,* 2502–2515. doi:10.1038/nprot.2013.150
17. Parker, J. S.; Mullins, M.; Cheang, M. C. U.; Leung, S.; Voduc, D.; Vickery, T.; Davies, S.; Fauron, C.; He, X.; Hu, Z.; Quackenbush, J. F.; Stijleman, I. J.; Palazzo, J.; Marron, J. S.; Nobel, A. B.; Mardis, E.; Nielsen, T. O.; Ellis, M. J.; Perou, C. M.; Bernard, P. S. *J. Clin. Oncol.* **2009,** *27,* 1160–1167. doi:10.1200/jco.2008.18.1370
18. Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; Varki, A.; Woods, R. J.; The SNFG Discussion Group. *Glycobiology* **2019,** *29,* 620–624. doi:10.1093/glycob/cwz045
19. Cheng, K.; Zhou, Y.; Neelamegham, S. *Glycobiology* **2017,** *27,* 200–205. doi:10.1093/glycob/cww115
20. Zhou, T.; Ren, J.; Medo, M.; Zhang, Y.-C. *Phys. Rev. E* **2007,** *76,* 046115. doi:10.1103/physreve.76.046115
21. Miklossy, G.; Hilliard, T. S.; Turkson, J. *Nat. Rev. Drug Discovery* **2013,** *12,* 611–629. doi:10.1038/nrd4088
22. Bi, X.; Hameed, M.; Mirani, N.; Pimenta, E. M.; Anari, J.; Barnes, B. J. *Breast Cancer Res.* **2011,** *13,* No. R111. doi:10.1186/bcr3053
23. Yanai, H.; Negishi, H.; Taniguchi, T. *OncoImmunology* **2012,** *1,* 1376–1386. doi:10.4161/onci.22475
24. Khan, H. A.; Margulies, C. E. *Front. Genet.* **2019,** *10,* No. 591. doi:10.3389/fgene.2019.00591
25. Sampieri, L.; Di Giusto, P.; Alvarez, C. *Front. Cell Dev. Biol.* **2019,** *7,* No. 123. doi:10.3389/fcell.2019.00123
26. Romagnoli, M.; Belguise, K.; Yu, Z.; Wang, X.; Landesman-Bollag, E.; Seldin, D. C.; Chalbos, D.; Barillé-Nion, S.; Jézéquel, P.; Seldin, M. L.; Sonenshein, G. E. *Cancer Res.* **2012,** *72,* 6268–6278. doi:10.1158/0008-5472.can-12-2270
27. Dimitroff, C. J. *Proc. Natl. Acad. Sci. U. S. A.* **2019,** *116,* 13729–13737. doi:10.1073/pnas.1900268116
28. Yu, W.; Huang, C.; Wang, Q.; Huang, T.; Ding, Y.; Ma, C.; Ma, H.; Chen, W. *Tumor Biol.* **2014,** *35,* 10943–10951. doi:10.1007/s13277-014-2403-1
29. Nan, X.; Ng, H.-H.; Johnson, C. A.; Laherty, C. D.; Turner, B. M.; Eisenman, R. N.; Bird, A. *Nature* **1998,** *393,* 386–389. doi:10.1038/30764
30. Jones, P. L.; Veenstra, G. J. C.; Wade, P. A.; Vermaak, D.; Kass, S. U.; Landsberger, N.; Strouboulis, J.; Wolffe, A. P. *Nat. Genet.* **1998,** *19,* 187–191. doi:10.1038/561
31. Chen, F.; Liu, X.; Bai, J.; Pei, D.; Zheng, J. *Oncol. Rep.* **2016,** *35,* 1227–1236. doi:10.3892/or.2015.4515
32. Tian, F.; DaCosta Byfield, S.; Parks, W. T.; Yoo, S.; Felici, A.; Tang, B.; Piek, E.; Wakefield, L. M.; Roberts, A. B. *Cancer Res.* **2003,** *63,* 8284–8292.
33. Petersen, M.; Pardali, E.; van der Horst, G.; Cheung, H.; van den Hoogen, C.; van der Pluijm, G.; ten Dijke, P. *Oncogene* **2010,** *29,* 1351–1361. doi:10.1038/onc.2009.426
34. Chugh, S.; Barkeer, S.; Rachagani, S.; Nimmakayala, R. K.; Perumal, N.; Pothuraju, R.; Atri, P.; Mahapatra, S.; Thapa, I.; Talmon, G. A.; Smith, L. M.; Yu, X.; Neelamegham, S.; Fu, J.; Xia, L.; Ponnusamy, M. P.; Batra, S. K. *Gastroenterology* **2018,** *155,* 1608–1624. doi:10.1053/j.gastro.2018.08.007
35. Song, Y.; Washington, M. K.; Crawford, H. C. *Cancer Res.* **2010,** *70,* 2115–2125. doi:10.1158/0008-5472.can-09-2979
36. Li, Y.; Kong, R.; Chen, H.; Zhao, Z.; Li, L.; Li, J.; Hu, J.; Zhang, G.; Pan, S.; Wang, Y.; Wang, G.; Chen, H.; Sun, B. *Aging* **2019,** *11,* 5035–5057. doi:10.18632/aging.102096

37. Ye, J.; Huang, A.; Wang, H.; Zhang, A. M. Y.; Huang, X.; Lan, Q.; Sato, T.; Goyama, S.; Kurokawa, M.; Deng, C.; Sander, M.; Schaeffer, D. F.; Li, W.; Kopp, J. L.; Xie, R. *Cell Death Dis.* **2020,** *11,* No. 187. doi:10.1038/s41419-020-2371-x
38. Wenzel, J.; Rose, K.; Haghighi, E. B.; Lamprecht, C.; Rauen, G.; Freihen, V.; Kesselring, R.; Boerries, M.; Hecht, A. *Oncogene* **2020,** *39,* 3893–3909. doi:10.1038/s41388-020-1259-7
39. Barkeer, S.; Chugh, S.; Karmakar, S.; Kaushik, G.; Rauth, S.; Rachagani, S.; Batra, S. K.; Ponnusamy, M. P. *BMC Cancer* **2018,** *18,* No. 1157. doi:10.1186/s12885-018-5074-2
40. Yevshin, I.; Sharipov, R.; Kolmykov, S.; Kondrakhin, Y.; Kolpakov, F. *Nucleic Acids Res.* **2019,** *47,* D100–D105. doi:10.1093/nar/gky1128
41. Marbach, D.; Lamparter, D.; Quon, G.; Kellis, M.; Kutalik, Z.; Bergmann, S. *Nat. Methods* **2016,** *13,* 366–370. doi:10.1038/nmeth.3799
42. Sonawane, A. R.; Platig, J.; Fagny, M.; Chen, C.-Y.; Paulson, J. N.; Lopes-Ramos, C. M.; DeMeo, D. L.; Quackenbush, J.; Glass, K.; Kuijjer, M. L. *Cell Rep.* **2017,** *21,* 1077–1088. doi:10.1016/j.celrep.2017.10.001
43. Kelkar, A.; Zhu, Y.; Groth, T.; Stolfa, G.; Stablewski, A. B.; Singhi, N.; Nemeth, M.; Neelamegham, S. *Mol. Ther.* **2020,** *28,* 29–41. doi:10.1016/j.ymthe.2019.09.006
44. Narimatsu, Y.; Joshi, H. J.; Nason, R.; Van Coillie, J.; Karlsson, R.; Sun, L.; Ye, Z.; Chen, Y.-H.; Schjoldager, K. T.; Steentoft, C.; Furukawa, S.; Bensing, B. A.; Sullam, P. M.; Thompson, A. J.; Paulson, J. C.; Büll, C.; Adema, G. J.; Mandel, U.; Hansen, L.; Bennett, E. P.; Varki, A.; Vakhrushev, S. Y.; Yang, Z.; Clausen, H. *Mol. Cell* **2019,** *75,* 394–407.e5. doi:10.1016/j.molcel.2019.05.017
45. Taniguchi, N.; Honke, K.; Fukuda, M.; Narimatsu, H.; Yamaguchi, Y.; Angata, T., Eds. *Handbook of Glycosyltransferases and Related Genes,* 2nd ed.; Springer: Tokyo, Japan, 2014. doi:10.1007/978-4-431-54240-7
46. Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H., Eds. *Essentials of Glycobiology,* 3rd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2017.
47. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. *Genome Res.* **2003,** *13,* 2498–2504. doi:10.1101/gr.1239303
48. Csardi, G.; Nepusz, T. *InterJournal* **2006,** No. 1695.

License and Terms