Beilstein-Institut

Glyco-Bioinformatics – *Bits 'n' Bytes of Sugars*
October 4th – 8th, 2009, Potsdam, Germany

Glycoinformatic Platforms for Data Interpretation: An HPLC Perspective

Matthew P. Campbell*, Natalia V. Artemenko and Pauline M. Rudd

NIBRT Dublin-Oxford Glycobiology Laboratory, National Institute for Bioprocessing Research and Training, Conway Institute, University College Dublin, Ireland

E-Mail: *matthew.campbell@nibrt.ie

Received: 1st March 2010 / Published: 10th December 2010

Abstract

High-throughput and automated HPLC techniques allow for the rapid, detailed structural analysis of complex glycans. These advances have the potential for (i) validating a new generation of biomarkers that relate alterations in glycan processing to disease by mining the glycosylation patterns in a variety of disease types and (ii) for monitoring the production of therapeutic glycoproteins. The wealth of knowledge that can be generated justifies the requirement for databases, analytical tools and search facilities. Recent efforts by international consortia have increased the awareness of the need for glycoinformatics and several resources are now available including a suite of novel applications to assist the interpretation of HPLC data collections.

Introduction

Glycosylation is the most common and structurally diverse post-translational modification of proteins that has an impact on a wide range of biological functions [1]. It can have profound effects on the structure of and physico-chemical properties of proteins and their activity to half of all cellular and secretory proteins [2]. It is widely accepted that carbohydrates are important factors in many biological recognition processes and our understanding of their

functions is rapidly expanding with advances in high-throughput glycomic strategies, disease/biomarker profiling, improved understanding of the molecular glycosylation machinery and glycoinformatics.

The biosynthesis of glycans is a non-template-driven process in which saccharide donors, glycosyltransferases, and exoglycosidases play an interactive role. The resulting glycans can have complex structures with multiple branching points where each monosaccharide component can be one of approximately 15 residues present in nature [3]. Previous reports suggest that aberrant changes in protein glycosylation, specifically the attachment of specific monosaccharide residues or branch changes, have implications in various biological and pathological processes including infection, autoimmune disorders and cancer [4 – 7]. The characterization of disease-associated glycans from serum glycoproteins has helped our understanding of disease pathologies and provides the potential for identifying biomarkers for diagnosis and prognosis [8 – 11].

The inherent complexity of glycan structures and microheterogeneity (glycoforms in which a single protein is diversified by a heterogeneous array of glycans at each glycosylation site) makes analysis of glycoconjugates very challenging. Glycan analysis relies on the ability to detect small quantities of glycans on low abundant glycoproteins since there is a requirement to characterize glycans at the 1% level. A full and detailed characterization of glycan structure is a time consuming analytical process dependent on an array of sensitive, robust, high-resolution separation technologies that are needed to determine monosaccharide composition, linkages and branching sequences. Those currently in use include HPLC, Mass Spectrometry, tandem MS (MS/MS), LC-MS, capillary electrophoresis (CE), high-performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD), NMR and glycan arrays [12 – 17]

High performance liquid chromatography (HPLC) and mass spectrometry (MS) are the most widely used techniques to address the challenges as they offer high levels of sensitivity and the ability to handle complex mixtures of different glycan variations. There are advantages and disadvantages to each technique and many strategies incorporate an orthogonal approach due to the complex nature of glycans, where no single method can fully characterize a given glycan structure and/or function [17, 12].

Modern high-throughput HPLC-based methods are well-established and powerful methods used to obtain quantitative, reproducible and high-resolution separations of glycans at the femtomole level. The major benefits of this technology include, firstly, a well established technique using amide-based columns to generate highly reproducible and resolved separation profiles that is capable of separating structures with the same composition on the basis of sequence and linkage type; the ability to analyze both neutral and charged glycans in the same run; the 1:1 stoichiometry labelling of glycan with the fluorophore 2-aminobenzamide (2AB) that permits quantification; the normalization of glycans identified by using a stan-

dard, for example dextran, that enables the calibration of retention times to glucose unit values (GU) using Empower GPC software from Waters Ltd; finally, the use of specific exoglyosidases for the sequential digestion of glycan pools that allows complete characterization of complex glycans [16, 18].

Previously we have reported a high-throughput 96 well platform method for N-glycan analysis [16]. Recent method development and advances to this technology have included interfacing fully automated liquid-handling robotic platforms, to optimize sample immobilization, enzymatic N-glycan release and fluorescent labelling; increasing sample productivity and reducing the time to complete sample preparation to eight hours. The availability of robotic-based instrumentation complements biomarker discovery and validation programs by accelerating the analysis of large sample cohorts where high-throughput methods are a prerequisite. Furthermore, automated platforms increase the application of glycomics by supporting alternative separation technologies, for example, specifically in the bioprocessing industry including quality by design, process analytical technology and critical feature analysis (glycosylation monitoring for therapeutic protein manufacturing) [20].

GLYCOINFORMATICS

The increasing amounts of data generated by mass spectrometry and high-throughput HPLC glycan analysis combined with high sensitivity, improved cycle speeds and robotic sample and handling platforms requires new ideas and methods for disseminating the conclusions drawn from such experiments. The sheer complexity and volume of data routinely generated necessitates bioinformatics solutions in the form of data repositories and analytical tools to facilitate data interpretation.

The traditional route of publishing articles is still the most practical method, however only a small proportion of the results and underlying data are readily accessible. When HPLC and mass spectrometry data is available it is generally only provided in specific formats. In addition the size of the data files result in raw data being archived and stored on local servers with only the final processed data being retained. To enable the community to effectively mine an increasingly rich data source, the experimental data needs to be collected in central public repositories.

In contrast to the genomic and proteomic areas, the glycosciences lack accessible, curated and comprehensive data collections that summarize the structure, characteristics, biological origin and potential function of glycans that have been experimentally verified and reported in the literature. The complexity of glycan structures and the variety of techniques available pose additional obstacles to the development of an accurate and user-friendly suite of tools and databases.

The current trend and population of databases is characterized by the existence of disconnected and incompatible collections of experimental data and proprietary applications. The lack of compatibility between these existing databases and there sparseness hinders the development of bioinformatics tools for the interpretation of data and implementing platforms for large-scale glycomics and glycoproteomic studies.

As glycan-related databases improve in both coverage and quality developers and experimentalists need to consider solutions and criteria requirements that maximize the value that can be extracted. The community has recognized the need for a more organized approach to accessing glycan related data. Over the last few years a number of international consortia have developed frameworks that support the growing requirement for storing and analyzing analytical data. Recent reviews highlight the approaches and difficulties the glycoinformatic community are facing and those steps being taken to create well-curated and annotated databases and tools [20, 21].

An Infrastructure Design Study (EUROCarbDB) was started under the sixth EU framework programme to establish the technical requirements for such a centralised and standardised architecture. The platform provides an introduction and/or recommendation of formats and nomenclatures; user-curated structural and experimental data; and open access to software programs and libraries to support continued development (http://www.eurocarbdb.org).

The partners of the EUROCarbDB initiative have developed tools and work flows to assist the interpretation of HPLC, MS and NMR experimental data, extending this infrastructure will utilize a strong background of resources to develop tools for data sharing and collection. This integration will meet the objectives set out by the ESF policy briefing (Structural Medicine – The Importance of Glycomics for Health and Disease, ESF Science Policy Briefing 27; 2006) and the NIH white paper [21]. A detailed overview of EUROCarbDB will be published elsewhere.

Recently, during a Consortium for Functional Glycomics (CFG) workshop on Analytical and Bioinformatic Glycomics a new Working Group on Glycomics Data(base) Standards (WGGDS) was established with the aim of defining standards and work flows for glycan data exchange. The WGGDS includes leading research partners from the CFG, GlycomeDB, RINGS and the EUROCarbDB consortia involved in glycomics and glycoinformatics.

The fundamental aim of EUROCarbDB and WGGDS is to make it possible for labs to run experiments and combine results, to distribute workloads, provide access to a selection of tools, and to derive new methodologies by comparing results to help realise the potential of high-throughput glycomics. Each consortium and partner member can provide an extremely useful data resource contributing data acquired by the group or retrieved from external sources.

In addition to the EUROCarbDB and CFG resources the other most prominent and publicly available databases include the Kyoto Encyclopedia of Genes and Genomes glycome portal (KEGG GLYCAN) [22], the Japanese Consortium for Glycobioloy and Glycotechnology Databases (JCGGDB) and Glycoscience.de [23]. Other efforts in recent years are: the Russian Bacterial Carbohydrate Structure Database (BCSDB) [24] and GlycomeDB [25], an initiative to retrieve all structure entries, taxonomic annotations, and references from major databases. Most recently GlycoSuite [26], originally a commercial database, is now hosted by the Swiss Bioinformatics Institute.

However, further developments are required to establish a comprehensive infrastructure required for next generation analytical and informatic tools. Especially, collaborative efforts to standardize formats for structural, analytical, and reference data including controlled vocabularies. For example, a glycomics experiment can include many different terms and descriptions, therefore, many frameworks including EUROCarbDB have develop internal vocabularies or reference well established and controlled vocabularies or ontologies, for example the NCBI taxonomy and MeSH terminologies. The availability of maintained controlled vocabularies has provided the stability required for database and software development.

HPLC Data Analysis: Developing Tools and Databases

The complexity of glycoconjugates and techniques used to elucidate their structures does present significant bottlenecks to the development of integrated software and database packages. There are an increasing number of tools and database to support glycomics investigations especially for HPLC [27, 28], mass spectrometry [29 – 31] and NMR [32].

Recently, several new HPLC-based tools have been developed in conjunction with EUROCarbDB to help populate databases and to assist the process of data analysis and annotation. These tools and work flows have been designed primarily to support the group's high-throughput strategy with an open-source philosophy that is intended to progress the application of glycoinformatics in glycomics and glycoproteomics. This is achieved by the provision of semi-automated capabilities inclusive of a robust database framework, formats for data exchange, and support for storing and retrieving glycan structural and experimental data.

An innovative suite of integrated analytical tools, databases and standards are being developed to support the growing demand for HPLC data interpretation and access to well curated data collections such as GlycoBase, autoGU [28] and GlycoExtractor [27] (http://glycobase.nibrt.ie).

GlycoBase and autoGU

GlycoBase is a novel experimental database developed to support our HPLC methodology. The database contains the HPLC elution positions for 2AB-labelled N- and O-linked glycan structures (expressed in the form of glucose unit values), the predicted products of exoglycosidase digestions; supporting literature information; and a listing of subgroups in which the glycan has been identified. All structures were determined by a combination of Normal Phase-HPLC with exoglycosidase sequencing and mass spectrometry (MALDI-MS, ESI-MS, ESI-MS/MS, LC-MS, LC-ESI-MS/MS). Each carbohydrate structure is stored in the database using the GlycoCT [33] format that can be used to dynamically convert the pictorial representation of structures to a series of support nomenclatures including Oxford University [34] and CFG black and white and colour formats and textual representation. This is achieved by using the application programming interface created by EUROCarbDB; a detailed discussion of the EUROCarbDB features, frameworks and database structure will be published elsewhere.

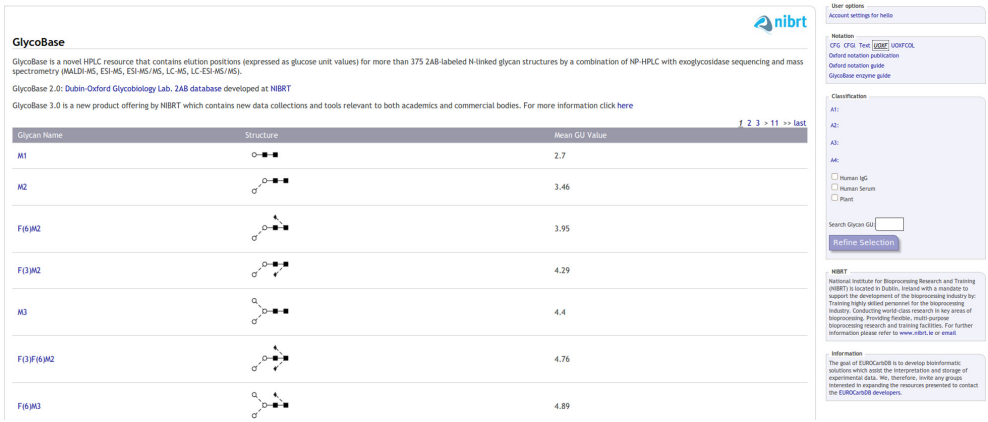

Figure 1. An overview of the front page to GlycoBase using the EUROCarbDB template. The page lists all structures stored in the database including a short name description; a pictorial representation in the Oxford nomenclature system which can be converted to other supported formats using the GlycoCT and EUROCarbDB API; and average reported glucose unit values. The refinement options on the right panel allow the user to display only those structures with specific structural features have been classified into specific subsets and/or fall within a user defined glucose unit range.

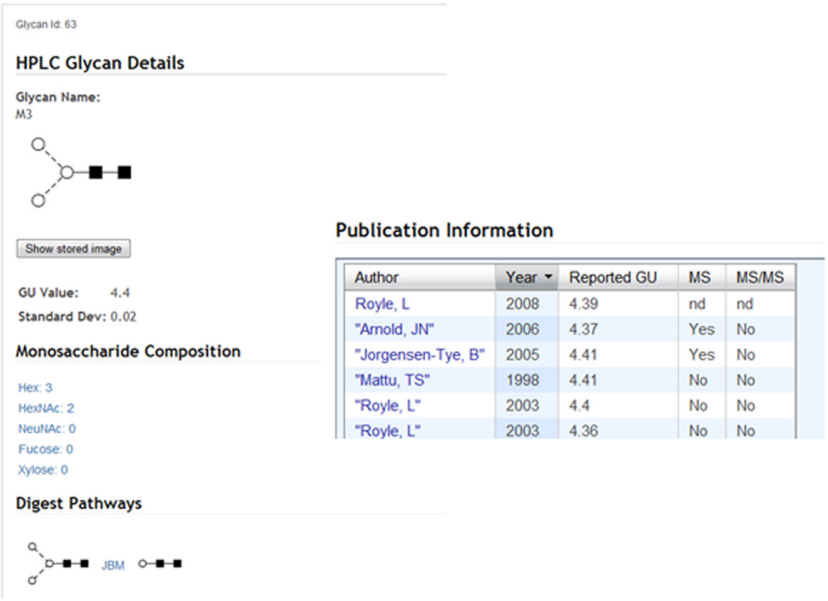

Figure 2. A data entry page listing for the Man3 glycan. For each glycan entry a description of the mean and standard deviation values is listed for all published references. In addition the monosaccharide composition is displayed followed by verified exoglycosidase digestions pathways. The publication of glucose unit values and products of digestions are essential aids to the manual interpretation of HPLC data collections.

The interpretation of HPLC data including exoglycosidase digestions can be time consuming and database-matching software (autoGU) is available to assist the assignment of possible glycan structures to each HPLC peak. When used in combination with data from a series of exoglycosidases, autoGU will create a refined list of structures based on the digest footprint i. e. shifts in GU values due to cleavage of terminal monosaccharides dependent on enzyme specificity.

GlycoBase and autoGU provide the backbone for the development of next generation glycoinformatics tools that are an invaluable aid to data interpretation. They provide the support for manual data interpretation whilst an active development program includes a number of new user-friendly features to meet the requirements of the end-user.

GlycoExtractor

A recent application, GlycoExtractor, is a novel approach for extracting large volumes of processed HPLC data from proprietary chromatography data software that stores locally acquired sample runs (Waters Empower). It is a web-based application that facilitates user demands to extract a series of sample sets generated by high-throughput methods. The tool

was developed to improve existing methods for processing and exporting large volumes of information for use with autoGU. Current methods are cumbersome, partly due to the lack of embedded workflows for exporting multiple sample sets. GlycoExtractor automates routine tasks by querying specific data attributes across all active projects to a single file rather than a set of disconnected output files, improving data interpretation [27].

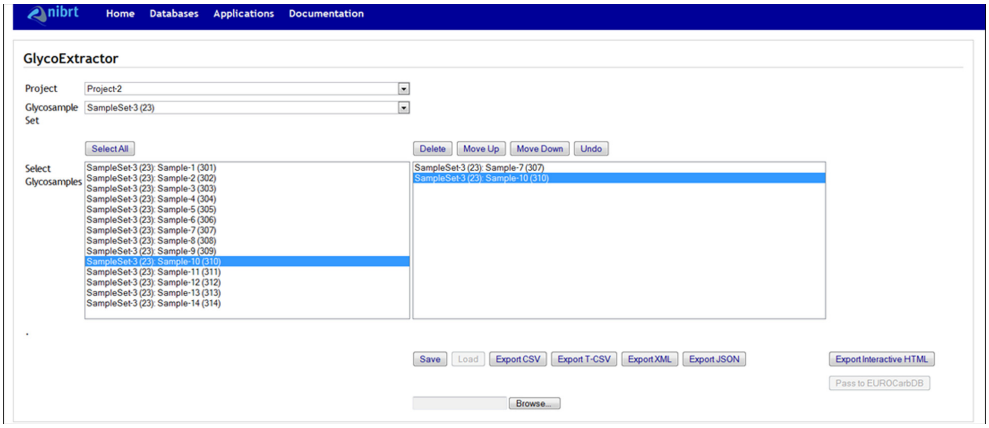

Figure 3. GlycoExtractor interfacing with an example database of HPLC data. The corresponding glucose unit values, peaks areas and project summary can be exported to a single file in XML, JSON or CSV format.

The development of these novel tools was achieved by EU efforts to develop centralized resources for the glycobiology community. It is anticipated that the next release will be a fully integrated solution of tools interfacing with the EUROCarbDB framework that will further improve both data processing and interpretation.

Conclusions

The availability and development of glycoinformatics databases and tools have increased considerably in recent years as a result of the efforts of international collaborations and consortia. However, further development is required to establish a comprehensive infrastructure required for next generation analytical and bioinformatics tools that rivals those platforms available for genomics and proteomics.

The formation of partnerships between leading consortia, such as the recent CFG sponsored WGGDS program is opening opportunities for developers and experimentalists to develop an infrastructure for cross-referencing data collections; the integration of resources that allows users to search data collections in a unified and user-friendly way. It is essential that the work achieved to date continues and the availability of organized and annotated databases of analytical data will enable the development of new technologies and supporting tools for automated, high-throughput identification of glycan structures.

Acknowledgements

The authors would like to thank the collaboration with the EUROCarbDB initiative (http://www.eurocarbdb.org) RIDS contract number: 011952.

References

[1] Dwek, R.A. (1996) Glycobiology: Toward Understanding the Function of Sugars. *Chem. Rev.* **96**: 683 – 720.
doi: http://dx.doi.org/10.1021/cr940283b.

[2] Apweiler, R., Hermjakob, H., Sharon, N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta* **1473**:4 – 8.

[3] Helenius, A., Aebi, M. (2004) Roles of N-linked glycans in the endoplasmic reticulum. *Annu. Rev. Biochem.* **73**:1019 – 1049.
doi: http://dx.doi.org/10.1146/annurev.biochem.73.011303.073752.

[4] Fuster, M.M., Esko, J.D. (2005) The sweet and sour of cancer: glycans as novel therapeutic targets. *Nat. Rev. Cancer* **5**:526 – 542.
doi: http://dx.doi.org/10.1038/nrc1649.

[5] Green, R.S., Stone, E.L., Tenno, M., Lehtonen, E.*, et al.* (2007) Mammalian N-glycan branching protects against innate immune self-recognition and inflammation in autoimmune disease pathogenesis. *Immunity* **27**:308 – 320.
doi: http://dx.doi.org/10.1016/j.immuni.2007.06.008.

[6] Hakomori, S. (2002) Glycosylation defining cancer malignancy: new wine in an old bottle. *Proc. Natl. Acad. Sci. U.S.A*. **99**:10231 – 10233.
doi: http://dx.doi.org/10.1073/pnas.172380699.

[7] Taniguchi, N., Ekuni, A., Ko, J.H., Miyoshi, E.*, et al.* (2001) A glycomic approach to the identification and characterization of glycoprotein function in cells transfected with glycosyltransferase genes. *Proteomics* **1**:239 – 247.
doi: http://dx.doi.org/10.1002/1615-9861(200102)1:2<239::AID-PROT239>3.0.CO;2-K.

[8] Abd Hamid, U.M., Royle, L., Saldova, R., Radcliffe, C.M.*, et al.* (2008) A strategy to reveal potential glycan markers from serum glycoproteins associated with breast cancer progression. *Glycobiology* **18**:1105 – 1118.
doi: http://dx.doi.org/10.1093/glycob/cwn095.

[9] Arnold, J.N., Saldova, R., Hamid, U.M.A., Rudd, P.M. (2008) Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation. *Proteomics* **8**:3284 – 3293.
doi: http://dx.doi.org/10.1002/pmic.200800163.

[10] Butler, M., Quelhas, D., Critchley, A.J., Carchon, H.*, et al.* (2003) Detailed glycan analysis of serum glycoproteins of patients with congenital disorders of glycosylation indicates the specific defective glycan processing step and provides an insight into pathogenesis. *Glycobiology* **13**:601 – 622.
doi: http://dx.doi.org/10.1093/glycob/cwg079.

[11] Wuhrer, M. (2007) Glycosylation profiling in clinical proteomics: heading for glycan biomarkers. *Expert. Rev. Proteomics* **4**:135 – 136.
doi: http://dx.doi.org/10.1586/14789450.4.2.135.

[12] Domann, P.J., Pardos-Pardos, A.C., Fernandes, D.L., Spencer, D.I.R.*, et al.* (2007) Separation-based glycoprofiling approaches using fluorescent labels. *Proteomics* **7**(Suppl. 1):70 – 76.
doi: http://dx.doi.org/10.1002/pmic.200700640.

[13] Estrella, R.P., Whitelock, J.M., Roubin, R.H., Packer, N.H., Karlsson, N.G. (2009) Small-scale enzymatic digestion of glycoproteins and proteoglycans for analysis of oligosaccharides by LC-MS and FACE gel electrophoresis. *Methods Mol. Biol.* **534**:171 – 192.

[14] Harvey, D.J. (2005) Proteomic analysis of glycosylation: structural determination of N- and O-linked glycans by mass spectrometry. *Expert. Rev. Proteomics* **2**:87 – 101.
doi: http://dx.doi.org/10.1586/14789450.2.1.87.

[15] Harvey, D.J. (2009) Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update for 2003 – 2004. *Mass Spectrom. Rev.* 28:273 – 361.
doi: http://dx.doi.org/10.1002/mas.20192.

[16] Royle, L., Campbell, M.P., Radcliffe, C.M., White, D.M.*, et al.* (2008) HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. *Ana.l Biochem.* **376**:1 – 12.
doi: http://dx.doi.org/10.1016/j.ab.2007.12.012.

[17] Wada, Y., Azadi, P., Costello, C.E., Dell, A.*, et al.* (2007) Comparison of the methods for profiling glycoprotein glycans – HUPO Human Disease Glycomics/Proteome Initiative multi-institutional study. *Glycobiology* **17**:411 – 422.
doi: http://dx.doi.org/10.1093/glycob/cwl086.

[18] Royle, L., Dwek, R.A., Rudd, P.M. (2006) Determining the structure of oligosaccharides N- and O-linked to glycoproteins. *Curr. Protoc. Protein Sci. Chapter 12* Unit 12.16.

[19] Blow, N. (2009) Glycobiology: A spoonful of sugar. *Nature* **457**:617 – 620. doi: http://dx.doi.org/10.1038/457617a.

[20] Aoki-Kinoshita, K.F. (2008) An introduction to bioinformatics for glycomics research. *PLoS Comput. Biol.* **4**:e1000075. doi: http://dx.doi.org/10.1371/journal.pcbi.1000075.

[21] Packer, N.H., von der Lieth, C.-W., Aoki-Kinoshita, K.F., Lebrilla, C.B.*, et al.* (2008) Macquarie University, Sydney, New South Wales, NSW 2109, Australia. 2008, pp. 8 – 20.

[22] Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F.*, et al.* (2006) KEGG as a glycome informatics resource. *Glycobiology* **16**:63R-70R. doi: http://dx.doi.org/10.1093/glycob/cwj010.

[23] Lutteke, T., Bohne-Lang, A., Loss, A., Goetz, T.*, et al.* (2006) GLYCOSCIENCES.de: an internet portal to support glycomics and glycobiology research. *Glycobiology* **16**:71R-81R. doi: http://dx.doi.org/10.1093/glycob/cwj049.

[24] Toukach, P., Joshi, H. J., Ranzinger, R., Knirel, Y., von der Lieth, C.-W. (2007) Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de. *Nucleic Acids Res.* **35**:D 280 – 286. doi: http://dx.doi.org/10.1093/nar/gkl883.

[25] Ranzinger, R., Frank, M., von der Lieth, C.-W., Herget, S. (2009) Glycome-DB.org: a portal for querying across the digital world of carbohydrate sequences. *Glycobiology* **19**:1563 – 1567. doi: http://dx.doi.org/10.1093/glycob/cwp137.

[26] Cooper, C.A., Harrison, M.J., Wilkins, M.R., Packer, N.H. (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.* **29**:332 – 335. doi: http://dx.doi.org/10.1093/nar/29.1.332.

[27] Artemenko, N.V., Campbell, M.P., Rudd, P.M. (2010) GlycoExtractor – a web-based interface for high throughput processing of HPLC-glycan data. *J. Proteome Res*. **9**(4):2037 – 2041. doi: http://dx.doi.org/10.1021/pr901213u.

[28] Campbell, M.P., Royle, L., Radcliffe, C.M., Dwek, R.A., Rudd, P.M. (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* **24**:1214 – 1216.
doi: http://dx.doi.org/10.1093/bioinformatics/btn090.

[29] Goldberg, D., Sutton-Smith, M., Paulson, J., Dell, A. (2005) Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* **5**:865 – 875.
doi: http://dx.doi.org/10.1002/pmic.200401071.

[30] Ceroni, A., Maass, K., Geyer, H., Geyer, R.*, et al.* (2008) GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.* **7**:1650 – 1659.
doi: http://dx.doi.org/10.1021/pr7008252.

[31] Maass, K., Ranzinger, R., Geyer, H., von der Lieth, C.W., Geyer, R. (2007) "Glycopeakfinder" – *de novo* composition analysis of glycoconjugates. *Proteomics* **7**:4435 – 4444.
doi: http://dx.doi.org/10.1002/pmic.200700253.

[32] Loss, A., Stenutz, R., Schwarzer, E., von der Lieth, C.W. (2006) GlyNest and CASPER: two independent approaches to estimate 1 H and 13C NMR shifts of glycans available through a common web-interface. *Nucleic Acids Res.* **34**:W733 – 737.
doi: http://dx.doi.org/10.1093/nar/gkl265.

[33] Herget, S., Ranzinger, R., Maass, K., Lieth, C.W. (2008) GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr. Res.* **343**:2162 – 2171.
doi: http://dx.doi.org/10.1016/j.carres.2008.03.011.

[34] Harvey, D.J., Merry, A.H., Royle, L., Campbell, M.P.*, et al.* (2009) Proposal for a standard system for drawing structural diagrams of N- and O-linked carbohydrates and related compounds. *Proteomics* **9**:3796 – 3801.
doi: http://dx.doi.org/10.1002/pmic.200900096.