Biochimica et Biophysica Acta 1844 (2014) 108–116

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Validation of the curation pipeline of UniCarb-DB: Building a global glycan reference MS/MS repository ☆

Matthew P. Campbell [a], Terry Nguyen-Khuong [a], Catherine A. Hayes [b], Sarah A. Flowers [b], Kathirvel Alagesan [c], Daniel Kolarich [c], Nicolle H. Packer [a,*], Niclas G. Karlsson [b]

[a] Biomolecular Frontiers Research Centre, Macquarie University, Sydney, New South Wales, Australia
[b] Department of Medical Biochemistry, University of Gothenburg, Gothenburg, Sweden
[c] Department of Biomolecular Systems, Max Planck Institute of Colloids and Interfaces, Potsdam, Germany

ARTICLE INFO

Article history:
Received 31 October 2012
Received in revised form 1 April 2013
Accepted 16 April 2013
Available online 25 April 2013

Keywords:
Glycomics
Mass spectrometry
Database
Glycan
Standards
Glycobiology

ABSTRACT

The UniCarb-DB database is an emerging public glycomics data repository, containing over 500 tandem mass spectra (as of March 2013) of glycans released from glycoproteins. A major challenge in glycomics research is to provide and maintain high-quality datasets that will offer the necessary diversity to support the development of accurate bioinformatics tools for data deposition and analysis. The role of UniCarb-DB, as an archival database, is to provide the glycomics community with open-access to a comprehensive LC MS/MS library of *N*- and *O*- linked glycans released from glycoproteins that have been annotated with glycosidic and cross-ring fragmentation ions, retention times, and associated experimental metadata descriptions. Here, we introduce the UniCarb-DB data submission pipeline and its practical application to construct a library of LC–MS/MS glycan standards that forms part of this database. In this context, an independent consortium of three laboratories was established to analyze the same 23 commercially available oligosaccharide standards, all by using graphitized carbon-liquid chromatography (LC) electrospray ionization (ESI) ion trap mass spectrometry in the negative ion mode. A dot product score was calculated for each spectrum in the three sets of data as a measure of the comparability that is necessary for use of such a collection in library-based spectral matching and glycan structural identification. The effects of charge state, de-isotoping and threshold levels on the quality of the input data are shown. The provision of well-characterized oligosaccharide fragmentation data provides the opportunity to identify determinants of specific glycan structures, and will contribute to the confidence level of algorithms that assign glycan structures to experimental MS/MS spectra. This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era. Guest Editors: Martin Eisenacher and Christian Stephan.

1. Introduction

Protein glycosylation, the attachment of glycans or carbohydrates to proteins, is probably the most important and extensive post-translational modification that enhances the functional diversity of proteins and influences their biological activity. Glycans, largely located at the interface between the cell surface and extracellular matrix, are involved in a wide range of functions from affecting the structural properties of their conjugated partner to positively or negatively modulating cell development and growth, host-pathogen interactions, immune response, and cell–cell interactions, amongst other processes [1–7]. Understanding such functions is challenging, due to the nature of the sequential and competitive biosynthetic machinery, and the subsequent synthesis of many variants of glycosylated proteins (glycoforms).

Knowledge of glycan structural macro- and microheterogeneity is required to establish connections to their physiological and pathophysiological function. By definition, an integrated or systems approach is necessary to realize these goals, complemented by the critical requirement for precise, robust and sensitive analytical methodologies supported by bioinformatics strategies [8–13]. Motivated by the need to address such challenges, substantial progress has been made in the key areas of glycan structure determination, biochemical analysis of glycan–protein interactions, array platforms, functional genetic approaches including transgenic mice models, chemical synthesis methods, and the growth in bioinformatics and computational tools to obtain glycomics-specific information [14–20]. These continuing advancements in deciphering glycan structures and defining their physiological function have contributed to the maturation of glycomics as an analytical science.

Important components of functional glycomics are the in-depth and fine characterisation of the repertoire of glycan structures found

☆ This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era. Guest Editors: Martin Eisenacher and Christian Stephan.

* Corresponding author at: Macquarie University, Sydney, 2109, Australia. Tel.: +61 298508176.
E-mail address: nicki.packer@mq.edu.au (N.H. Packer).

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.04.018

on proteins in cells and tissues and the heterogeneity of these glycans that results from their non-template-driven biosynthesis. Several analytical and biochemical methodologies have been developed including mass spectrometry, high-performance liquid chromatography, capillary electrophoresis, nuclear magnetic resonance spectroscopy, and lectin arrays. All provide the means to probe the multiple structural attributes of carbohydrates, however, irrespective of the method, a trade-off exists between complete structure/sequence determination and high-throughput analysis. High-throughput analysis provides a snapshot of the most likely structures, while rigorous and detailed analysis is necessary to explicitly assign monosaccharide residues, branching and linkages.

Mass spectrometry (MS) has emerged as the leading tool for the structural characterization of glycans and dramatic improvements in tandem mass spectrometry instrumentation and experimental approaches have allowed glycomics to advance from the generation of oligosaccharide compositional lists to targeted observations of relative quantitative and dynamic structural changes [20–22]. This is exemplified by the MS contribution to the glycoprofiling of human and mouse cells and tissues by the Consortium for Functional Glycomics (CFG).

In terms of the wealth of information content associated with MS, and its ability to link structural determination with relative quantification, MS is likely to be a dominant technology for glycan analysis in the foreseeable future. However, despite the advances and the volume of spectra generated on a routine basis, and when compared to other omics fields, relatively few MS glycomics data repositories and analysis tools are available in the public domain. This is an unfortunate situation since the MS data can give biological information on glycans that is unique. For example, transcriptomic and microarray-based analyses cannot identify the changes in relative glycan levels or correlate specific changes in site-occupancy with any glycoprotein.

Despite the current absence of a journal requirement to make glycomics MS data publicly available, several repositories have been established to address the demand for storage and availability of glycomics data in the public domain. The US Consortium for Functional Glycomics established efforts towards the development of a bioinformatics platform capable of acquiring and disseminating diverse data collections that are focused on the provision of annotated MALDI-TOF MS glycan spectra. In addition, the European funded initiative, EUROCarbDB, and the Japanese KEGG Glycan contributed significant resources to the development of a bioinformatics framework comprising databases and analytical tools to address the challenges in glycomics. However few platforms provide access to manually annotated and validated MS/MS spectra of glycans, inclusive of fragment ion assignments that would facilitate structure elucidation, since few laboratories fully characterize glycan structures. As a consequence, mandatory data disclosure, such as is common in genomics and proteomics, in the glycomics field remains an important work in progress.

To address the central issue of providing mass fragmentation spectral data we present the synergistic approach of the UniCarb-DB initiative to collaborate with multiple institutions to collectively provide a database of MS/MS spectra. Here we present a comparison of data generated from known glycan standards collected in 3 separate laboratories. The ultimate goal is to facilitate data comparison, exchange and verification (similar to those objectives outlined by the Human Disease Glycomics/Proteome Initiative [23]) by establishing technical standards and demonstrating the capabilities of bioinformatics resources to freely disseminate glycomics data collections.

1.1. Why a reference database of glycan standards is important

An important factor in broadening the application of glycomics is the necessity to develop databases, as well as computational tools, to acquire and facilitate the interpretation of analytical data collections [24]. This is particularly important for meeting the technical demands of mass spectrometry, where the interpretation of a glycan MS/MS spectrum remains the critical step of a discovery-based glycomics experiment. The assignment of glycan structures is made challenging due to their complex branching and isomeric nature, whereby different structures can have identical molecular masses; for example, all hexose monosaccharide residues have the same molecular mass, as do the corresponding HexNAc derivatives. In the case of extended structures, the sequences are often identical but the topology, including branching type differs, which is further compounded by the anomeric and linkage configurations of the glycosidic bonds.

While high-throughput approaches (at the compositional level) can only offer an insight into potential glycan structures present, fine structural characterization involves the selection of a mass peak followed by fragmentation (MSn), and subsequent interpretation of signature ions to accurately assign glycan structures. The emergence of computational methods to assist the assignment of MS fragmentation patterns has significantly improved, but many are based upon theoretical fragmentation-pattern predictions, which in some cases can prove to be erroneous.

Some approaches to annotate the glycan compositions corresponding to MALDI-MS spectra have been developed, including GlycoMod [25] and Cartoonist [26]. The latter also assigns potential structures to a resultant monosaccharide composition based on pathway knowledge, the definition of biosynthetic rules and a library of several hundred archetypal glycans. Several tools, including GlycoFragment, GlycoSearchMS [27], GlycoWorkbench [28] and GlycoSidIQ [29] have focused on interpretation of mass spectrometry fragmentation patterns (MSn) through comparison of the experimental data to reference datasets using in silico fragmentation engines to compute theoretical fragments (glycosidic and cross-ring) to deduce the most likely glycan structure. Library-based sequencing tools are limited by the lack of comprehensive and well-curated collections of glycan structure fragments. A series of experimental fragmentation profile datasets of well-characterized glycan structures was adopted by STAT [30], OSCAR [31] and StrOligo [32], although efforts to improve their software functionality have been discontinued.

Access to well-characterized oligosaccharide standards provides the opportunity to elucidate and understand the unique fragmentation properties of individual glycans and corresponding isomers and to apply this knowledge to spectral matching algorithms. A comprehensive analysis of a wide range of oligosaccharide standards will thus provide invaluable information for the bioinformatics community to build programs similar to the high-throughput software, such as the MASCOT and SeQuest programs, now used to identify peptide sequences from proteomics data. The availability of such pure standard oligosaccharides of known structure is extremely limited at present and needs to be addressed by the glycomics community and industry.

To this end, a workflow for the creation of a library of chromatographic descriptions and curated MS/MS spectra of underivatized oligosaccharide alditol standards with known composition, linkage and branching information has been developed in partnership with UniCarb-DB [33]. The initial library consists of 23 biologically relevant (*N*-linked) glycan standards listed in Table 1 that have been fully determined at the MS/MS level including the identification of structure signature ions; this approach will facilitate the absolute and accurate glycan assignment to MS/MS spectra by filtering false positives and the noise associated with experimental artifacts. This small independent study, consisting of three laboratories analyzing the same oligosaccharides using graphitized carbon-liquid chromatography (LC)/electrospray ionization (ESI) mass spectrometry on three different ion trap instruments, aims to establish the technical standards of MS/MS fragmentation and to demonstrate the capabilities of bioinformatics resources to process data collections. This pilot study is focused on a comparative analysis of fragmentation patterns acquired from different ion trap MS/MS acquisition instruments using similar protocols and experimental-run conditions.

Table 1
Glycan standards used to develop a reference MS/MS fragmentation library with calculated mass to charge ratios (m/z) in negative ion mode. The sugar symbols are those employed by the Consortium for Functional Glycomics and linkage representation depicted by the Oxford Glycobiology Institute (UOXF) system [34].

Oligosaccharide standard	Negative ions		Composition
High mannose	$[M-2H]^{2-}$	$[M-H]^{-}$	
M1	293.1116	587.2232	$HexNAc_2Hex_1$
FM1	366.1406	733.2812	$HexNAc_2Hex_1dHex_1$
M2	374.138	749.276	$HexNAc_2Hex_2$
M3	455.1644	911.3288	$HexNAc_2Hex_3$
FM3	528.1934	1057.387	$HexNAc_2Hex_3dHex_1$
M5	617.2217	1235.443	$HexNAc_2Hex_5$
M6	698.2437	1397.487	$HexNAc_2Hex_6$
M9Glc1	1022.349	2045.699	$HexNAc_2Hex_{10}$
Complex type			
A3	759.7835	1520.567	$HexNAc_5Hex_3$
FA2	731.2728	1463.546	$HexNAc_4Hex_3dHex_1$
A2G1	739.2702	1479.54	$HexNAc_4Hex_4$
A2G2	820.2966	1641.593	$HexNAc_4Hex_5$
A4	861.3232	1723.646	$HexNAc_6Hex_3$
A3B	861.3232	1723.646	$HexNAc_6Hex_3$
FA2B	832.8125	1666.625	$HexNAc_5Hex_3dHex_1$
FA2G2	893.3256	1787.651	$HexNAc_4Hex_5dHex_1$
FA2BG1	913.8389	1828.678	$HexNAc_5Hex_4dHex_1$
A3G3	1002.863	2006.725	$HexNAc_5Hex_6$
A5B	1064.403	2129.805	$HexNAc_8Hex_3$
FA2BG2	994.8653	1990.731	$HexNAc_5Hex_5dHex_1$
A4G4	1185.429	2371.858	$HexNAc_6Hex_7$
A2G2S(6)1	965.8443	1932.689	$HexNAc_4Hex_5NeuAc_1$
A2G2S(6)2	1111.392	2223.784	$HexNAc_4Hex_5NeuAc_2$

GlcNac, GalNac, Gal, Man, Fuc, NeuAc; Linkage 6, 4, 3, 2

2. Material and methods

Dextra (Reading, UK) oligosaccharide standards were supplied in kind by Professor Jeremy Turnbull (University of Liverpool).

The detailed glycan structure and MS fragmentation profile of reduced standards were based on the protocol described by Jensen PH et al., 2012 [10], with minor modifications.

2.1. Reduction and purification

All *N*-linked glycans were reduced using $NaBH_4$ according to methods modified from previously described protocol [10]. In summary, samples were resuspended in LC–MS grade water to a concentration of 0.1 mg/mL–0.01 mg/mL before reduction. The oligosaccharides were reduced with 20 μL of 0.5–1 M $NaBH_4$/20–50 mM KOH at 50 °C for 2–16 h. Glacial acetic acid (1 μL) was added and the samples were desalted using cation exchange columns prepared in ZipTips (Millipore) or empty TopTips (Glygen) with 25–30 μL bed volume of exchange resin (AG 50W X8; Bio-Rad, Hercules, CA). Columns were washed with methanol, HCl and dH_2O before use. The reduced glycans were collected with two washes of dH_2O. Samples were dried using a vacuum centrifuge (Savant) and washed 5 times with MeOH to remove excess borate.

2.2. Mass spectrometric analysis

2.2.1. Laboratory 1: HPLC-LTQ-CID ion trap MS analysis

Porous graphitized carbon (5 μm particle size) columns were used with a width of 250 μm and length of 10 cm packed in-house. Mobile phases consisted of 10 mM NH_4HCO_3 for solvent A and 10 mM NH_4HCO_3 with 80% acetonitrile for solvent B. The gradient, after 5 min of 100% solvent A increased solvent B to 45% in 41 min. The column was washed with 100% solvent B for 8 min and equilibrated at 100% solvent A for 25 min. Columns were attached to an Agilent 1100 series HPLC with a flow rate after passive splitting of 7–10 μL/min. An LTQ linear ion trap (Thermo Fisher Scientific) in negative ion mode was used for MS and MS/MS analysis using CID. A top three data dependent method was used with normalized collision energy of 35 for all samples. A standard containing porcine gastric mucin and fetuin released glycans was run either before or after each sample to monitor reproducibility of retention time and column sensitivity.

2.2.2. Laboratory 2: HPLC-XCT-ultra ion trap MS analysis

N-glycan alditols were separated using an Agilent 1100 capillary LC (Agilent Technologies, Santa Clara, CA) and analyzed using an Agilent MSD, three-dimensional ion-trap XCT mass spectrometer coupled to the LC. Separation was performed on a Hypercarb porous graphitized carbon column (5 μm particle size, 100 × 0.18 mm, Thermo Hypercarb, Thermo Scientific) across an 85 min gradient with a constant flow rate of 2 μL/min using a linear gradient up to 40% (v/v) CH_3CN in

10 mM NH_4HCO_3. ESI-MS was performed in negative mode with two scan events; MS full scan with mass range *m/z* 100–2000 and data dependant MS/MS scan after collision-induced dissociation (CID) of the top two most intense precursor ions.

2.2.3. Laboratory 3: HPLC-amazon ETD speed-CID ion trap MS analysis

Reduced glycans were resuspended in 10 mM NH_4HCO_3 prior to porous graphitized carbon LC–ESI MS/MS analysis on a UltiMate 3000 RSLCnano System (Dionex, part of Thermo Fisher Scientific) coupled to an Amazon ETD speed ion trap MS (Bruker Daltonics, Bremen, Germany). A calculated equivalent of around 75 pmol of glycan was loaded onto a Hypercarb™ column (100 × 0.32 mm, particle size: 5 μm; Thermo Fisher Scientific). The column was equilibrated in solvent A (10 mM NH_4HCO_3) and samples were loaded onto the column and desalted for 6 min before a linear gradient up to 20% solvent B (90% acetonitrile in 10 mM NH_4HCO_3) was developed over 60 mins. The column was further washed by increasing to 98% B for 8 mins before starting conditions were re-established. The flow rate was set to 6 μL/min throughout the whole analysis, the column oven temperature was set to 30 °C. Ions were detected in negative ion mode over a *m/z* range from 380 to 1800 Da. Data dependent MS/MS CID-fragmentation was performed for the three most intense ion within a precursor scan and a *m/z* range from 100 to 2500 Da was scanned for MS2 scans. The instrument was controlled using Hystar 3.2 software.

2.3. Data analysis

LC–MS/MS spectra were analyzed using Compass Data Analysis v4.0/4.1 (Bruker Daltonics, Bremen). The extracted ion chromatogram (EIC) for each glycan was derived based on the glycan $[M-H]^{1-}$/$[M-2H]^{2-}$ masses.

3. Results

The spectral mass lists of the combined MS/MS scan for each glycan standard with corresponding intensities were exported to GlycoWorkbench v1.2 for manual fragment mass annotation (http://download.glycoworkbench.org/). All monoisotopic peaks were annotated according to whether or not they are glycosidic or cross-ring fragmentation, and submitted to the UniCarb-DB database (http://www.unicarb-db.org).

3.1. The UniCarb-DB data submission pipeline

UniCarb-DB aims to provide a common interface to high-quality glycan MS/MS data, through a global collaborative effort, that will serve as a base for exchanging data and building improved analytical tools. UniCarb-DB was established in 2009 [33], with a mission to continue and further develop those tools, created by the EUROCarbDB initiative, as a public mass spectrometric data repository that provides access to high-quality, annotated MS-related glycomics experimental data.

In practical terms, the public sharing policy can only succeed if reliable and user-friendly software tools exist to streamline the data submission task and standardize the data entry. From a data management view, one of the difficulties that glycomics repositories encounter is handling the variety of workflows and data outputs commonly used, which are more diverse than genomic, transcriptomic and proteomic methodologies.

UniCarb-DB is initially centered on a top–down approach where the detected analytes are intact *N*- and *O*-linked glycans released from glycoproteins in reduced form. There might be a separation step, for example gel electrophoresis, blotting or the whole extract is digested with PNGase F and/or subjected to reductive beta-elimination followed by separation of the reduced alditols by porous graphitized carbon liquid chromatography interfaced to negative ESI-MS/MS. Other LC and MS techniques are of course available, with different experimental approaches serving different goals and UniCarb-DB will expand to accept these data. For example, some researchers use MALDI approaches to identify as many glycans as possible in any given sample at a compositional level, while others may fully characterize a small subset of structures. This plethora of approaches is relevant for data resources, since the experimental approach, the data processing workflow, and the data generated influence the type of data formats that must be accepted by glycomics repositories. Ultimately challenging is the design of a single unified resource for the presentation of the variety of glycomics MS data in a structured database. The UniCarb-DB submission workflow is summarized in Fig. 1.

An inherent problem with MS is the heterogeneity of data formats to export raw and processed data collections. These situations are improving with the development of standard formats (for MS data) by the Human Proteome Organization Proteomics Standards Initiative (PSI), and checklists for the standardization of experimental glycomics data and meta information by the MIRAGE (Minimum Information Required for A Glycomics Experiment) project.

UniCarb-DB has started to address those guidelines proposed by MIRAGE, which state the desirable minimal information that should be reported, by building a sequential workflow that captures high-level metadata descriptions during the submission process. The workflow intends to document (i) sample preparation encompassing release technique and/or methods that alter a glycan structure including exoglycosidase treatment and derivatization, (ii) complementary validation methods including lectin studies and monosaccharide analysis, and (iii) instrument description and parameters including ionization mode and LC-separation conditions. A full description of the terms and classification can be found in the data submission instructions on UniCarb-DB (http://www.unicarb-db.org).

To streamline data submission the UniCarb-DB workflow supports annotated GlycoWorkbench files. GlycoWorkbench is an open source, platform independent tool that assists the visualization and annotation of MS/MS data in an interactive manner. It provides functionality to generate an annotated list of peaks derived from a variety of MS

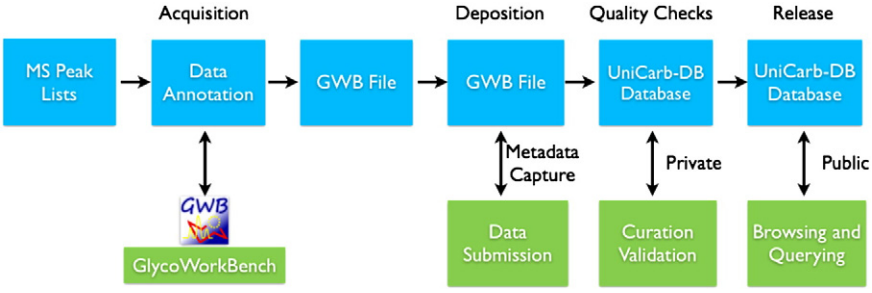

Fig. 1. The UniCarb-DB MS data submission workflow.

formats or selected peak lists (exported from vendor data acquisition software). The software supports a number of structural constituents and annotation options, in addition to an extensive library of fragmentation types. Data providers can use this standalone tool to transform a variety of MS formats or selected peak lists into the GlycoWorkbench GWP format. It allows the user to carry out an initial assessment on data quality and reliability, and can be used to interpret and annotate data prior to submission or validation with the UniCarb-DB requirements. A list of supported data formats, requirements and processing conditions is documented at www.glycoworkbench.org.

The UniCarb-DB web-interface reinforces and adopts key functionality previously introduced by EUROCarbDB and here we demonstrate how new extensions and improvements have been used to create an application for storing and providing MS/MS data on glycan standards. The data submission process is split into four sections (i) sample preparation by the user using a template GlycoWorkbench model (Table 2), (ii) description of the experimental evidence (experimental conditions) and its deposition, (iii) automated and curator validation and quality checks, and (iv) assignment of contributor identifiers and public release. The initial launch of an intuitive web-interface and workflow is designed to encourage data deposition by providing access to minimalistic entry points, which capture the core metadata data descriptions and provide efficient access to the MS collections.

Once the data has been successfully submitted, a unique identifier is provided to the contributor. Initially, the data is not accessible to the public domain as the intention is to allow the contributor to interact with the administration team to rectify any significant errors that may have occurred during the data deposition process, and to provide extra information that may not be supported either by GlycoWorkbench or the metadata terms.

The availability of MS/MS data in a public repository is only the first step. The ability to inspect, spectral match and score experimental data and validate reported and published results is the important functionality of such a database. UniCarb-DB intends to provide such a mechanism that enables the interrogation of spectra and data quality and will, as the amount of data increases, allow improved structural analysis of glycans released from glycoproteins.

3.2. Spectral processing and similarity scoring

As a validation of the usability of such a MS/MS repository for spectral matching by diverse laboratories employing different methodologies, the reproducibility of LC–ESI-MS/MS obtained from three separate laboratories on the same set of 23 standard oligosaccharides was assessed. The samples used were commercially available oligosaccharides of good quality and representative of biologically relevant glycans comprising a range of *N*-linked structures including pauci-mannose, high-mannose and complex type with structural features that are bisecting, neutral fucosylated, and a degree of branching from bi-antennary to penta-antennary.

A principal objective of this multi-institutional study was to compare the correspondence of the acquired MS fragmentation data and to identify the data parameters that influence the similarity scores of the compared spectra. The intention, in the first instance, was to establish a central lookup engine documenting the rules of fragmentation of glycans in the negative ion mode. Although differences can be observed in each tandem MS, there was the expected high degree of convergence and similarity of most structures.

The dot-product function available in the R package OrgMassSpecR (http://orgmassspecr.r-forge.r-project.org), a general package for mass spectrometry analysis, was used to view all sets of spectra and to calculate the similarity scores of the three sets of data (Table 3). Prior to statistical evaluation the spectra were exported in the NIST MSP format, a text-based format for storing centroid *m/z* values and corresponding intensities. For each set of spectra acquired from the three ion trap instruments, a similarity score was calculated to determine spectral similarity, using a head-to-tail plot of two mass spectra with the query spectrum (bottom) and a reference spectra (top). An example alignment for the asialo, agalacto and tetraantennary glycans is shown in Fig. 2. The laboratory 1 spectra were used as the reference spectra, and the *m/z* values of peaks in the partner spectra within t of a reference *m/z* value are paired with the reference peaks. Ideally, a single peak from the top spectrum should be paired with a single peak from the reference spectrum. Peaks without a match are paired with an intensity of zero. The score is calculated as below where $(u\% * \%v)$ is the dot product that measures the cosine of the angle between spectra represented as vectors. Previous work on the MS analysis of small molecules has demonstrated that the spectral dot product is the most effective scoring function [35].

$$\cos\theta = (u\% * \%v) \Big/ \left(\mathrm{sqrt}\left(\mathrm{sum}\left(u^{\wedge}2\right)\right) * \mathrm{sqrt}\left(\mathrm{sum}\left(v^{\wedge}2\right)\right)\right)$$

u and v are aligned intensity vectors of the two spectra, and t is used to align the intensities.

Note that all of these parameters are easily adjustable including the definition of baseline noise. The parameters (0.25 *m/z* threshold and 5% peak intensity threshold) described above, though not necessarily optimal for all possible situations, were found to be adequate for the majority of the standards and achieved the expected similarity

Table 2
A selection of metadata terminology used in UniCarb-DB.

Release method	Protein preparation	Glycosidase treatment	Complementary strategies
Reductive beta elimination	AG-PAGE composite gel	α-Mannosidase	Monosaccharide analysis
PNGase F treatment	Coomassie blue staining	β-Galactosidase	Sialic acid analysis
Hydrazine treatment	PVDF blotting	Sialidase	Smith degradation

Table 3
Similarity scores for each glycan standard. The ion trap mass spectrometers used in this study were Thermo Finnigan LTQ linear ion trap (lab 1), Agilent XCT-ultra ion trap (lab 2) and Bruker Amazon ETD speed ion trap (lab 3).

Standard name (see Table 1)	Similarity relative to lab 1	
	Lab 2	Lab 3
A3	0.56	0.97
FA2	0.84	0.53
A2G1*	0.69	0.18
A2G2*	0.91	0.01
A4	0.96	0.92
A3B	0.92	0.95
FA2B	0.68	0.96
FA2G2	0.67	0.68
FA2BG1	0.88	0.5
A3G3	0.81	0.87
A5B	0.68	0.8
FA2BG2	0.87	0.52
A4G4	0.5	0.76
M1	0.67	0.98
FM1	0.96	0.98
M2	0.92	0.98
M3	0.94	0.93
FM3	0.94	0.97
M5	0.77	0.94
M6	0.97	0.75
M9Glc1	0.85	0.91
A2G2S1	0.85	0.59
A2G2S2	0.99	0.99

* Low score due to the different charge states of the selected precursors chosen for fragmentation; $M–2H^{2-}$ laboratory 1 and $M–H^{1-}$ laboratory 3 (see below for discussion on this observation).

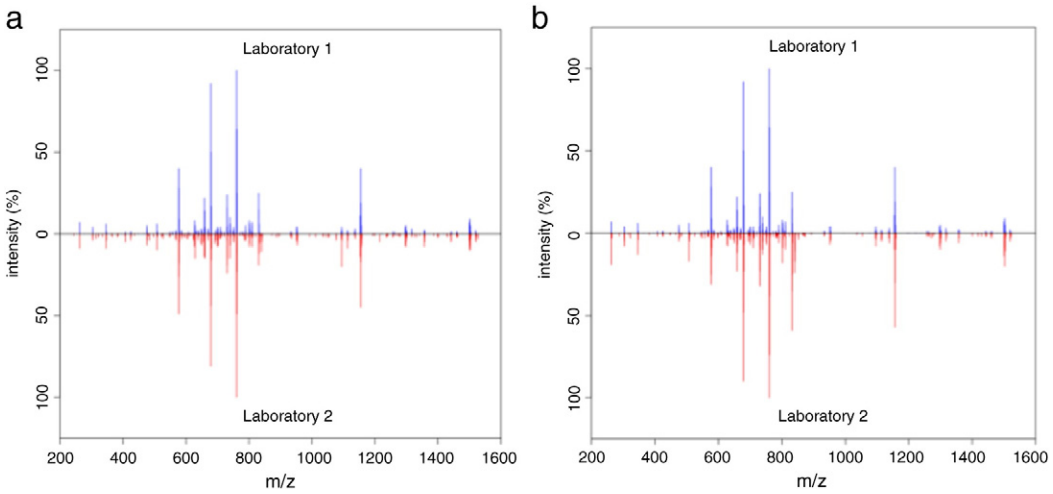

Fig. 2. Representative tandem MS acquired for the asialo, agalacto and tetraantennary (A4) glycans standard analyzed at laboratory 1 (top spectra in blue) compared to (a) laboratory 2 with a 0.96 similarity score, and (b) laboratory 3 with a 0.92 similarity score.

in the majority of cases. Large positive dot-product scores are obtained when peaks in the reference spectrum align with peaks in the target spectrum (within a defined mass tolerance). Conversely, a tandem pattern with few matching peaks receives a poor score. The output from the analysis is an array containing scan number, dot-product score, and match count — the number of matching peaks within the defined tolerance.

From this analysis it is evident that the similarity of fragmentation is high in most cases and the assigned statistical confidence consolidates the reproducibility between the institutes, despite minor differences in sample preparation and data acquisition methods. In spectral library matching the ideal situation is to obtain a perfect match of the unknown experimentally derived spectrum with a single library spectrum with a one-to-one correspondence between each peak intensity. A value of 1 indicates a perfect overlay of the two MS/MS spectra. In reality however, because of instrument variability this is not generally possible. In addition to the spectral similarity calculation, differences in score are also dependent on the data components such as instrument processing of the raw data consisting of such parameters as peak selection, peak intensity scaling and weighting of peak intensity by mass position. Here, we observed (Fig. 2) that tandem spectral comparisons between 0.92 and 0.96 have a degree of similarity between the most intense ions in relation to intensity and *m/z*. The scores in Table 3 ranged between 0.5 (A4G4) to 0.99 (A2G2S2) for laboratory 2 and 0.01 (A2G2) to 0.99 (A2G2S2) for laboratory 3 with the performance for the optimized dot-product approach on average 0.82 and 0.77 for glycans analyzed by laboratory 2 and laboratory 3 relative to the reference laboratory (laboratory 1).

Since there were no guidelines provided to the participating laboratories about preferred charge state of precursor ions for fragmentation analysis, the generated data also provided insights into the quality of spectra in relation to the charge state of the parent ions. As expected, singly charged fragment spectra of A2G2 (similarity score 0.01) and A2G1 (similarity score of 0.18) from laboratory 3 compared to doubly charged precursor fragment spectra from laboratory 1 provided different structural information about the oligosaccharides.

In order to investigate the types of fragmentation generated by different charge states, we applied the dot-product algorithm to show how fragmentation differed. Fig. 3 shows the MS/MS spectra of standard A2G1 from laboratories 1 and 3. The spectral similarity score for this alignment as given by OrgMassSpecR (using the default tolerance of 0.25) is 0.18, or 0.32 (with an increased tolerance of 0.5). While it is apparent that the overall spectra are different there are regions where the similarity is higher. This is the case in the region

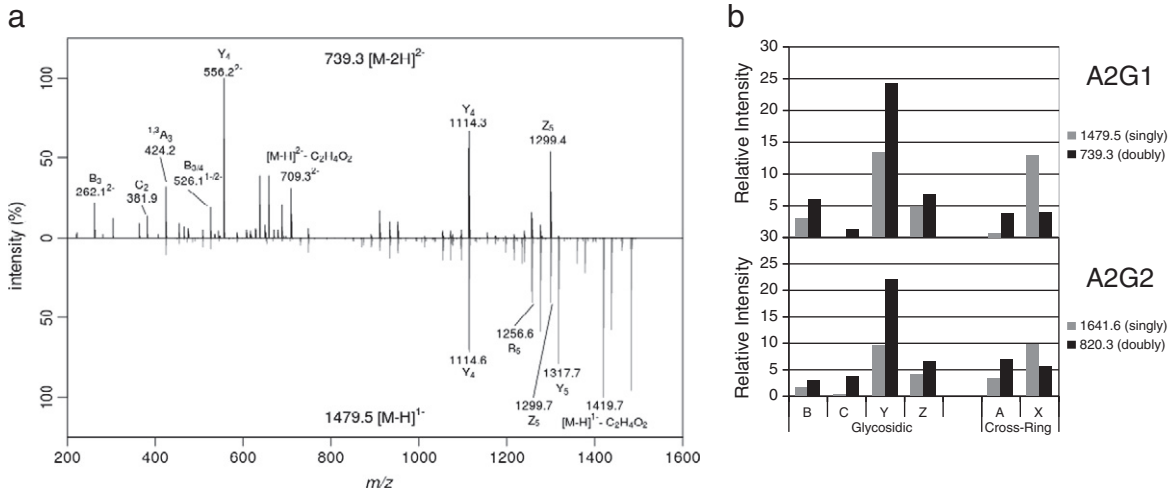

Fig. 3. (a) A comparison of the MS^2 fragmentation of the glycan standard A2G1 acquired in singly and doubly charged state from laboratories 1 and 3 respectively. Bar charts (b) show the relative intensities of the fragment ions observed for A2G1 (upper) and A2G2 (lower).

where only singly charged fragments are found. For instance, in the mass region 700 to 1300 (with tolerance of 0.5), the spectra have a similarity score of 0.78. The bar graphs (right of figure) show the percentage intensities of the annotated fragments found in A2G1 (top) and A2G2 (bottom) that had been fragmented as either singly or doubly charged precursors and emphasizes the importance of the charge state of the parent ion as a necessary parameter for entry into the curation pipeline of UniCarb-DB in the same way as is required in proteomics spectral matching software. From the bar graphs it is apparent that the doubly charged species are more prone to glycosidic cleavages, in particular Y fragments, whereas the singly charged species both have an X-type cross-ring cleavage (loss of $C_2H_4O_2$) as one of their most intense peaks (Fig. 3b). The X-type cross-ring fragmentation is responsible for the prevalence of satellite peaks below the Y and Z fragments of the singly charged species, especially pronounced in the 1300 to 1500 *m/z* region. This is in contrast to the low mass region of the doubly charged species where there are more A-type cross-ring fragments, providing quality sequence information as opposed to the inconclusive X-type fragmentation. These observations emphasize that the scoring of the similarity, between acquired MS/MS fragmentation and the reference database, only reflects the quality of the overlay without questioning if the reference spectrum is actually conclusively defining the structure. Further refinement of the scoring algorithms will need to consider both the spectral similarity and the confidence of the assigned sequence, based on the analysis of the fragments observed.

Threshold settings defined in the dot-product model can also be adjusted to improve similarity scoring. For example, extending the *m/z* tolerance (0.5 ppm) and comparing peaks with intensities above 5% improve scoring for A2G2 and A2G1, from 0.01 to 0.2 and 0.18 to 0.4 respectively. Changes to these signal-to-noise parameters can significantly alter the quality of scoring, therefore visual inspection and understanding of the spectrum acquisition and data processing are recommended considerations in parallel to any statistical analysis.

Primarily, UniCarb-DB contains processed and de-isotoped MS^2 spectra with partial or full assignment of structure feature ions. De-isotoping significantly reduces the number of signals corresponding to a single glycan structure, thereby increasing signal intensity and improving signal-to-noise ratios. The mass list produced can be used directly for database searching, enabling users to compare acquired data against the growing library of validated spectra that are intended to help simplify interpretation of fragment ion spectra. To compare the accuracy and viability of using the de-isotoped MS/MS profile data to match experimental unprocessed spectra, two raw datasets obtained on FA2B and M3 from laboratories 1 and 3 were compared against the de-isotoped (glycan standard) data now stored in UniCarb-DB. By using the dot-product approach a similarity score of 0.6 was calculated for the unprocessed data on FA2B from both laboratories. The raw data on M3 scored 0.93 (laboratory 1) and 0.87 (laboratory 3) against the corresponding de-isotoped data (Fig. 4). These data indicate that a degree of confidence can be assigned to this method of matching

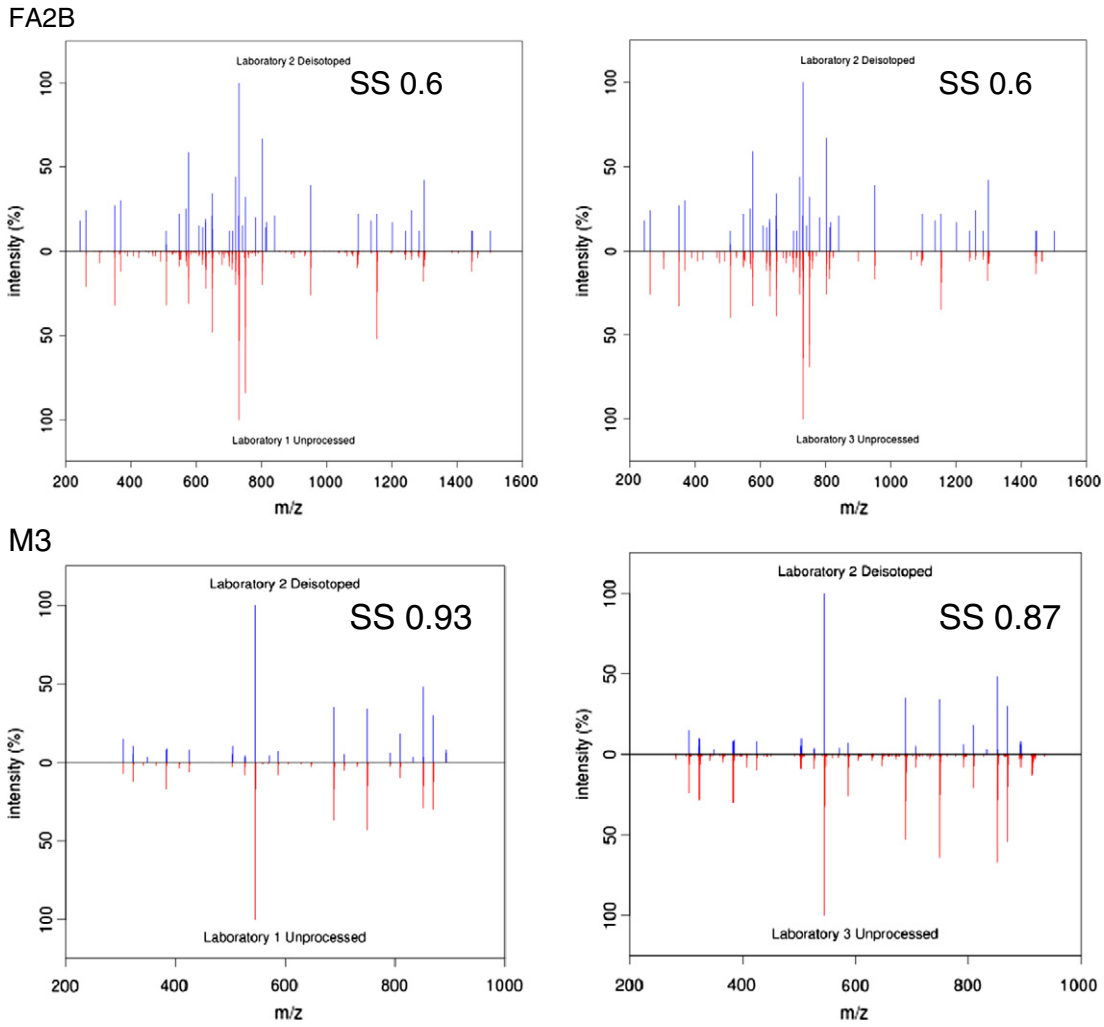

Fig. 4. Correlation between de-isotoped data in UniCarb-DB (laboratory 2) and unprocessed spectra collected from laboratories 1 and 3 for glycan standards FA2B and M3. The similarity score (SS) was calculated from the head-to-tail plots of the de-isotoped (top) and unprocessed (bottom) spectra.

unprocessed against de-isotoped data for preliminary structure assignment, but further emphasize the importance of comparing like-to-like data when interrogating UniCarb-DB to confidently assign MS^2 spectra.

Overall, the quality of data and fragmentation pattern similarities were high scoring between the three laboratories, and the informative annotated fragment ions provide the start of a knowledgebase for glycan MS/MS interpretations. The reproducibility and correlation of fragmentation data in this database imply that the identification of an unknown glycan through finding a best match in a database that measures agreement between the tandem mass spectra and the candidate structure is plausible. Differences in correlation in this study imply that such a comparison does carry caveats; importantly, data agreement is improved by comparison of presented spectra as the same (i) processed de-isotoped peak list, (ii) precursor charge state, and (iii) instrument sensitivity. The issue of sensitivity is more difficult to correct for, as it is dependent on the limit of detection of the individual mass spectrometer. The effects can range from incomplete dissociation to poor ion statistics for fragment ions that may make them indistinguishable from noise; and glycans present at the limit of detection can produce poor quality fragmentation. Furthermore, sensitivity and instrument resolution can impact the extent of glycosidic and cross-ring fragmentation ions identified, which necessitates the need to score the spectral features in order to validate the identification of structures.

4. Discussion

An important factor in strengthening the wider adoption of glycomics by the general scientific community is the urgent need to develop databases and computational tools to acquire, integrate, annotate and mine glycomics data sets such as analytical MS/MS collections. Over the last few years significant efforts by the EUROCarbDB, CFG and KEGG initiatives have developed solutions to help disseminate data collections. However, the generation and public availability of glycomics data are still overshadowed by the sheer volume of data accessible in the proteomics and genomics domains even though there is the same need. The role of UniCarb-DB to further extend and continue the efforts of the discontinued EUROCarbDB is starting to yield positive growth with the quality and quantity of data submitted to UniCarb-DB gaining traction.

Since the last UniCarb-DB publication, there have been significant developments in the UniCarb-DB framework. One of the most important improvements is the availability of a data submission workflow that aims to simplify the process of deposition by integrating the features of the standalone tool GlycoWorkbench to store tandem MS/MS data collections. The submission process has made data submission much easier and more straightforward, and the intuitive workflow ensures that both the experienced glycoscientist and graduate researcher familiar with GlycoWorkbench can easily contribute to this growing knowledgebase.

Similar to other 'omics' disciplines, the complexity of glycan structural and analytical data sets impacts strategies to automate deposition. It is clear that, in the immediate future, the role of data curation will continue to be necessary in the preparation and quality control of a UniCarb-DB deposition. The construction of workflows and resources that support the accumulation of data sets represents an opportunity to perform this quality control analysis on glycomics data, and, subsequently to learn from the fragmentation properties of glycans. Moreover, the creation of such a MS glycan knowledgebase will enable us to build more advanced tools for complex queries and data retrieval, dataset comparison and access to additional automated annotation of submitted data sets. It will be possible to identify common features of fragmentation, which will enable the identification of structural families of oligosaccharides that share similarities in their fragmentation in the negative ion mode in the first instance. To achieve this we have expanded the captured fragment annotation to include high intensity glycosidic cleavages (Y, Z, B and C type ions), cross-ring fragmentation (X and A ions), and internal fragments as well as neutral loss of small molecules (water or carbon dioxide). The assignment of fragmentation will allow us (and the community) to identify common fragmentation pathways and differences in the pattern between neutral, sialylated and sulfated oligosaccharide fragmentation. For more sophisticated data mining purposes, users can also download the corresponding XML file for each glycan standard from the UniCarb-DB site in compressed zipped files.

Editorial and discussion chapters in the Beilstein GlycoBioinformatics Workshop publications have recently addressed the current lack of data sharing policies. In this context, the value of UniCarbKB as an integrated data knowledgebase [36] was recognized and further strengthened by its inclusion in such initiatives as MIRAGE and the implementation of systems that will enable data sharing (Glyco-RDF and the CFG sponsored web services project) between members of the US, European, Japanese and Australian glycoscience centers, with UniCarb-DB as the mass spectrometry submission point. Such programs aim to alleviate the restrictions and reticence for data sharing by building a semantic web approach, which is capable of working with the inherent complexity of the data and the existence of different structure encoding formats. The UniCarb initiative has always been supportive of community data standards, therefore, in the next phase of development the infrastructure will support MIRAGE guidelines, the integration of RESTful services and the design of RDF data formats (Fig. 5).

At this stage in development, automated checking and validation of metadata descriptions are not yet implemented, but we are in the process of developing procedures that identify and indicate missing or erroneous elements of the deposited data set in order to increase the quality of the data.

5. Conclusions

A multi-dimensional approach is required to comprehensively characterize the glycosylation status of any glycoprotein. Fortunately, there is great potential of the highly sensitive technology of LC–MS/MS to advance our understanding of the heterogeneity of oligosaccharides in complex biological materials. Through the UniCarb-DB directive, we have initiated a fragmentation library of glycan reference standards with detailed fragmentation properties. This initiative is intended to support current strategies, and move towards the automated quantitation and high confidence assignment of glycan structures in negative mode ESI MS/MS glycomics experiments in the first instance.

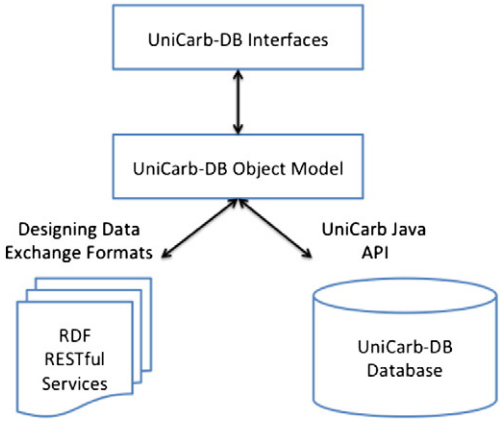

Fig. 5. UniCarb-DB model. The web interfaces allow access, querying and data submission to UniCarb-DB. The Java API is a collection of open-source libraries that support the object model representations of the database structure, which can be used to interact with the data content and rendering of web interfaces. RDF and RESTful are new services designed to support data sharing, developed in collaboration with international efforts.

The study described here validates the potential usefulness of this approach, based on the high degree of fragment spectra similarity that was observed for the majority of the structures tested in the three laboratories, despite the differences in the ion trap instrument used. Furthermore, access to a platform that enables glycomics data-sharing (adopting MIRAGE and MCP reporting guidelines) will support the realization of more advanced data analysis algorithms. This will be achieved through a collaborative approach to build a global MS/MS fragmentation library (UniCarb-DB) as part of a knowledgebase of experimental data on glycan structures (UniCarbKB), thus representing a milestone in the advancement of glycoanalysis. The potential for this MS/MS reference database to improve assignment of detailed glycan structures will be realized when applied to complex biological samples.

Funding

This study was funded by the National eResearch Collaboration Tools and Resources (NeCTAR); the Australian National Data Service (ANDS); Agilent Technologies; and the Swedish Foundation for International Cooperation in Research and Higher Education (STINT); and KA was supported by a PhD scholarship from the Beilstein Institut.

Acknowledgements

The authors thank Dextra Laboratories Ltd, UK, for the provision of the oligosaccharide standards. Many of the features described in this manuscript is built on the success of EUROCarbDB and we acknowledge all those involved for their on-going support. We also would like to thank the Max Planck Society, Bruker Daltonics and Agilent Technologies for their support.

References

[1] S.M. Haslam, S. Julien, J.M. Burchell, C.R. Monk, A. Ceroni, O.A. Garden, A. Dell, Characterizing the glycome of the mammalian immune system, Immunol. Cell Biol. 86 (2008) 564–573.
[2] Y. van Kooyk, G.A. Rabinovich, Protein–glycan interactions in the control of innate and adaptive immune responses, Nat. Immunol. 9 (2008) 593–601.
[3] P.M. Rudd, T. Elliott, P. Cresswell, I.A. Wilson, R.A. Dwek, Glycosylation and the immune system, Science 291 (2001) 2370–2376.
[4] A. Varki, Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins, Nature 446 (2007) 1023–1029.
[5] C. Bonomelli, K.J. Doores, D.C. Dunlop, V. Thaney, R.A. Dwek, D.R. Burton, M. Crispin, C.N. Scanlan, The glycan shield of HIV is predominantly oligomannose independently of production system or viral clade, PLoS One 6 (2011) e23521.
[6] A. Imberty, A. Varrot, Microbial recognition of human cell surface glycoconjugates, Curr. Opin. Struct. Biol. 18 (2008) 567–576.
[7] P.C. Pang, P.C. Chiu, C.L. Lee, L.Y. Chang, M. Panico, H.R. Morris, S.M. Haslam, K.H. Khoo, G.F. Clark, W.S. Yeung, A. Dell, Human sperm binding is mediated by the sialyl-Lewis(x) oligosaccharide on the zona pellucida, Science 333 (2011) 1761–1764.
[8] S.J. North, P.G. Hitchen, S.M. Haslam, A. Dell, Mass spectrometry in the analysis of N-linked and O-linked glycans, Curr. Opin. Struct. Biol. 19 (2009) 498–506.
[9] M. Pabst, F. Altmann, Glycan analysis by modern instrumental methods, Proteomics 11 (2011) 631–643.
[10] D. Kolarich, P.H. Jensen, F. Altmann, N.H. Packer, Determination of site-specific glycan heterogeneity on glycoproteins, Nat. Protoc. 7 (2012) 1285–1298.
[11] H. Geyer, R. Geyer, Strategies for analysis of glycoprotein glycosylation, Biochim. Biophys. Acta 1764 (2006) 1853–1869.
[12] C.D. Rillahan, J.C. Paulson, Glycan microarrays for decoding the glycome, Annu. Rev. Biochem. 80 (2011) 797–823.
[13] D.J. Harvey, C.A. Scarff, M. Crispin, C.N. Scanlan, C. Bonomelli, J.H. Scrivens, MALDI-MS/MS with traveling wave ion mobility for the structural analysis of N-linked glycans, J. Am. Soc. Mass Spectrom. 23 (2012) 1955–1966.
[14] L.N. Robinson, C. Artpradit, R. Raman, Z.H. Shriver, M. Ruchirawat, R. Sasisekharan, Harnessing glycomics technologies: integrating structure with function for glycan characterization, Electrophoresis 33 (2012) 797–814.
[15] J.C. Paulson, O. Blixt, B.E. Collins, Sweet spots in functional glycomics, Nat. Chem. Biol. 2 (2006) 238–248.
[16] B. Lepenies, J. Yin, P.H. Seeberger, Applications of synthetic carbohydrates to chemical biology, Curr. Opin. Chem. Biol. 14 (2010) 404–411.
[17] C. Rademacher, J.C. Paulson, Glycan fingerprints: calculating diversity in glycan libraries, ACS Chem. Biol. 7 (2012) 829–834.
[18] S.J. North, J. Jang-Lee, R. Harrison, K. Canis, M.N. Ismail, A. Trollope, A. Antonopoulos, P.C. Pang, P. Grassi, S. Al-Chalabi, A.T. Etienne, A. Dell, S.M. Haslam, Mass spectrometric analysis of mutant mice, Methods Enzymol. 478 (2010) 27–77.
[19] C.W. von der Lieth, A.A. Freire, D. Blank, M.P. Campbell, A. Ceroni, D.R. Damerell, A. Dell, R.A. Dwek, B. Ernst, R. Fogh, M. Frank, H. Geyer, R. Geyer, M.J. Harrison, K. Henrick, S. Herget, W.E. Hull, J. Ionides, H.J. Joshi, J.P. Kamerling, B.R. Leeflang, T. Lutteke, M. Lundborg, K. Maass, A. Merry, R. Ranzinger, J. Rosen, L. Royle, P.M. Rudd, S. Schloissnig, R. Stenutz, W.F. Vranken, G. Widmalm, S.M. Haslam, EUROCarbDB: An open-access platform for glycoinformatics, Glycobiology 21 (2011) 493–502.
[20] R. Raman, S. Raguram, G. Venkataraman, J.C. Paulson, R. Sasisekharan, Glycomics: An integrated systems approach to structure–function relationships of glycans, Nat. Methods 2 (2005) 817–824.
[21] J. Zaia, Mass spectrometry and glycomics, OMICS 14 (2010) 401–418.
[22] D.J. Harvey, L. Royle, C.M. Radcliffe, P.M. Rudd, R.A. Dwek, Structural and quantitative analysis of N-linked glycans by matrix-assisted laser desorption ionization and negative ion nanospray mass spectrometry, Anal. Biochem. 376 (2008) 44–60.
[23] N. Taniguchi, Human disease glycomics/proteome initiative (HGPI), Mol. Cell Proteomics 7 (2008) 626–627.
[24] N.H. Packer, C.W. von der Lieth, K.F. Aoki-Kinoshita, C.B. Lebrilla, J.C. Paulson, R. Raman, P. Rudd, R. Sasisekharan, N. Taniguchi, W.S. York, Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11–13, 2006), Proteomics 8 (2008) 8–20.
[25] C.A. Cooper, E. Gasteiger, N.H. Packer, GlycoMod — A software tool for determining glycosylation compositions from mass spectrometric data, Proteomics 1 (2001) 340–349.
[26] D. Goldberg, M. Sutton-Smith, J. Paulson, A. Dell, Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra, Proteomics 5 (2005) 865–875.
[27] K.K. Lohmann, C.W. von der Lieth, GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates, Nucleic Acids Res. 32 (2004) W261–W266.
[28] A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell, S.M. Haslam, GlycoWorkbench: A tool for the computer-assisted annotation of mass spectra of glycans, J. Proteome Res. 7 (2008) 1650–1659.
[29] H.J. Joshi, M.J. Harrison, B.L. Schulz, C.A. Cooper, N.H. Packer, N.G. Karlsson, Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data, Proteomics 4 (2004) 1650–1664.
[30] S.P. Gaucher, J. Morrow, J.A. Leary, STAT: A saccharide topology analysis tool used in combination with tandem mass spectrometry, Anal. Chem. 72 (2000) 2331–2336.
[31] A.J. Lapadula, P.J. Hatcher, A.J. Hanneman, D.J. Ashline, H. Zhang, V.N. Reinhold, Congruent strategies for carbohydrate sequencing. 3. OSCAR: An algorithm for assigning oligosaccharide topology from MSn data, Anal. Chem. 77 (2005) 6271–6279.
[32] M. Ethier, J.A. Saba, M. Spearman, O. Krokhin, M. Butler, W. Ens, K.G. Standing, H. Perreault, Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry, Rapid Commun. Mass Spectrom. 17 (2003) 2713–2720.
[33] C.A. Hayes, N.G. Karlsson, W.B. Struwe, F. Lisacek, P.M. Rudd, N.H. Packer, M.P. Campbell, UniCarb-DB: A database resource for glycomic discovery, Bioinformatics 27 (2011) 1343–1344.
[34] D.J. Harvey, A.H. Merry, L. Royle, M.P. Campbell, P.M. Rudd, Symbol nomenclature for representing glycan structures: extension to cover different carbohydrate types, Proteomics 11 (2011) 4291–4295.
[35] S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, J. Am. Soc. Spectrom. 6 (1994) 1044-0305.
[36] M.P. Campbell, C.A. Hayes, W.B. Struwe, M.R. Wilkins, K.F. Aoki-Kinoshita, D.J. Harvey, P.M. Rudd, D. Kolarich, F. Lisacek, N.G. Karlsson, N.H. Packer, UniCarbKB: putting the pieces together for glycomics research, Proteomics 11 (2011) 4117–4121.