Beilstein-Institut

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

SysBioWare: Structure Assignment Tool for Automated Glycomics

Sergey Y. Vakhrushev[1,*], Denis Dadimov[2], Jasna Peter-Katalinić[3,4,#]

[1]EMBL, Genome Biology Unit, Quantitative Proteomics, Meyerhofstrasse 1, 69117 Heidelberg, Germany

[2]MechSystemProject U.E., Bioinformatics division, 160 – 56 Mayakovsky str, Minsk 220028, Belarus

[3]University of Rijeka, Department of Biotechnology, S. Krautzeka bb, 51000 Rijeka, Croatia

[4]University of Münster, Institute of Pharmaceutical Biology and Phytochemistry, Hittorfstr. 56, 48149 Münster, Germany

E-Mail: *sergey.vakhrushev@embl-heidelberg.de, #jkp@uni-muenster.de

Received: 26th April 2010 / Published: 10th December 2010

Abstract

Glycomics as a part of systems biology closely related to proteomics encompasses knowledge acquainted by comprehensive and systematic studies of entire complement of carbohydrates in a cell, organ or organism. The prerequisite for these studies is a detailed information on molecular structure of complex carbohydrates which play a crucial role in processes like signalling, cell-cell recognition and immune response, and which act as therapeutic agents, vaccine or drug targets. Although the modern methods of mass spectrometry are well fitted for integrative "omics" experiment design, the interpretation of carbohydrate mass spectral data is still strongly linked to the human expertise. In this chapter we present a computational approach for automatic interpretation of mass spectral data of complex carbohydrates. We contribute to the field by designing a software package which will significantly reduce a need for human expertise in mass spectrometric

data interpretation derived from glycoconjugates and enable discovery and improvement of high-throughput protocol for automated glycomics. The proposed structure assignment tool named SysBioWare was constructed for automated processing raw MS and MS/MS performing isotopic grouping of detected peaks after de-noising and wavelet analysis. Monoisotopic *m/z* values render peak list association with the raw MS spectrum and allow compositional assignment according to the tuned building block library. This platform has been applied to human urinome and glycolipidome as a potent tool for rapid assignment of already known or/and new carbohydrate structures.

Introduction

Oligosaccharides, glycolipids, glycoproteins, glycopeptides, and proteoglycanes (mucopolysaccharides) occur ubiquitously in nature. Most of the naturally occurring proteins are glycosylated and these glycoproteins are found together with glycolipids in viruses, microorganisms, plants, and animals. Whereas the cell surface of bacteria, yeast, plants and other lower forms of life is primarily composed of polysaccharides, thus forming rigid cell walls, the cell membrane of mammalian cells contains carbohydrate residues linked to proteins and lipids. The ceramide residue (i. e., a long chain sphingoid base substituted at the amino group by a fatty acid) is by far the most prominent lipid constituent in addition to minor quantities of glycerol ethers and esters. In glycoproteins, glycan chains can be linked either N-glycosidically to asparagine residues or O-glycosidically to serine or threonine.

The growing perception of the biological importance of this group of compounds has greatly stimulated efforts in developing new and more powerful methods of structural elucidation. However, compared to the progress achieved in the structural analysis of other biopolymers such as proteins or nucleic acids, progress has been rather slow in the field of complex carbohydrates. This is mainly due to the complexity of carbohydrate structures, a consequence of the polyfunctional sugar molecule that allows several sites of anomeric linkages.

A complete structural analysis of a complex carbohydrate molecule involves the determination of (a) molecular mass and number of individual sugar components, (b) sites and anomeric configuration of glycosidic linkages, (c) conformation of sugar rings, (d) sequence of sugar components and pattern of branching, (e) secondary structure and spatial orientation, and (f) structure of the aglycon. Among methods for glycoanalysis, mass spectrometry have been found to be especially powerful for determining (a) molecular masses (b) sequence and pattern of branching, (c) in special cases, sites of glycosidic linkages and (d) structure of the aglycon. From the molecular mass, the number of sugar constituents in terms of deoxy hexose, hexose, hexosamine, etc. can, in most cases, be calculated [1]. Presently, MS is a most popular method for structural analysis due to its high sensitivity and speed on

the one hand, and the ability to analyze complex mixtures on the other. In particular, qualitative data interpretation of MS spectra in high-throughput projects appears of primary importance for rapid identification of biological routes [2].

Glycomics encompasses comprehensive and systematic genetic, physiological, pathological and structural studies of entire complement of carbohydrates in a cell, organ or organism. Although the modern MS instruments are fitted for integrative "omics" experiment design, the interpretation of carbohydrate mass spectral data is still strongly linked to the human expertise, and due to the carbohydrate structure complexity, it represents a complex task for specialists. In order to extent the limits and options of systems biology, efficient tools for glycomics are required. In this chapter we present a computational approach for automatic interpretation of mass spectra of complex carbohydrates to be used in glycomics.

Availability of Tools for Interpretation and Deposition of Complex Carbohydrate Structure

Over the past 20 years several groups reported on their attempts to develop rational tools for computational mass spectrometry in glycan analysis. Various products focus on compositional analysis, structure drawing, data base development, *in silico* fragmentation and spectra assignment, but very limited number of them efforts to incorporate rational tools into a single package. In some recent reviews the state-of-the-art in glycoinformatics, particularly in the field of computation of MS data, has been summarized [2 – 5].

In the "Glycomod" tool all possible glycan compositions are calculated from the values of their respective molecular ions, which can be done for the native permethylated and peracetylated structures. For glycopeptides, "Glycomod" infers compositional candidates if the mass and/or the sequence of the parent peptide is known [6, 7]. Upon the automatic creation of the composition hit list, it must be further manually analysed to remove from the hit list implausible glycan portion structure proposals, since a step for filtering of biologically non-relevant carbohydrate compositions is not provided.

"GlycoPep DB" has been designed as a web-based tool for glycopeptide analysis using a "smart search" concept where the human expertise for filtering implausible structures should be largely reduced. Designed for N-glycopeptide compositional assignment, this program can compare experimentally determined masses against the database of glycopeptides with N-linked glycans, where only biologically relevant structures are stored [8]. In this way the number of implausible glycan compositions in "GlycoPep DB" in comparison to "Glycomod" is reduced, but this approach is functionally limited to those structures, which are already known, and therefore present in databases.

Using "GlycoSuite DB" the glycan composition can be assigned to the experimentally determined selected precursor ion, which is present in the glycan structure database either on the basis of the mass search through glycan structure database or is deposited according to its composition. According to the last update in May 2009, in this database 9436 entries from 864 published references were collected. Additionally, the query for the particular glycan can be performed either by the sequence, or the type of glycosidic linkage, or its biological source, accession, and taxonomy, by the type of disease, the attached protein and is associated with the appropriate reference (http://glycosuitedb.expasy.org/glycosuite/glycodb). For interpretation of oligosaccharide fragmentation patterns the "GlycosidIQ^{TM}" software, based on the principle of matching experimental MS/MS data with those generated *in silico* from the GlycoSuite DB, is available [9]. The output is ranked according to proprietary algorithm which takes into account the fragmentation properties of branched glycans as well as their biological probabilities.

Based on a similar approach "GlycoFragment" is another tool by which the generation of all theoretically possible A-, B-, C-, X-, Y- and Z-fragments of oligosaccharides can be performed and each peak of a measured mass spectrum compared with the calculated fragments of all structures contained in the SweetDB database [10, 11].

In "STAT" a list of all possible saccharide compositions based on its mass alone will be generated. To find all possible N-linked oligosaccharide structures for this particular composition, the fragments imputed manually can be analysed by rejecting all structures not containing a trimannosyl core. However, other biosynthetic pathway rules are not included. Fucosylated glycans and glycans containing a bisecting GlcNAc residue are not considered [12].

The web-based application "Glyco-Peakfinder" was developed for a rapid assignment of glycan compositions. To provide entirely *de novo* compositional assignment, this platform does not operate with prior information from glycan databases or pre-calculated known archetypes. In the computational stage it can accept a number of glycan derivatizations such as permethylation, peracetylation, perdeuteromethylation, and acetylation [13]. This tool is complemented by "GlycoWorkbench", which is a suite of software tools designed for rapid drawing of glycan structures and for assisting the process of structure determination from mass spectrometry data. The graphical interface of GlycoWorkbench provides an environment in which structure models can be assembled. Mass values from the structure candidate list can be computed, their fragments automatically matched with MS^n data and the results compared to assess the best candidate followed by MS/MS data annotation.

Another development based on matching of *in silico* generated fragments against measured MS/MS data is a commercial product "SymGlycan". Candidate structures for *in silico* fragmentation are retrieved from the proprietary database of biologically plausible structures.

Each candidate structure is scored to indicate how closely it matches the experimental data. Additionally, this program provides elements of project management, for associating results with input profile and managing search parameters [14].

As long as updates of carbohydrate databases are still far behind developments in proteomics and because *in silico* approach imposes functional limitations on application to glycomics unlike in proteomics, computational mass spectrometry for glycomics is highly focused on the development of tools for *de novo* glycan MS/MS interpretation.

"StrOligo" is a tool for automated interpretation of tandem MS spectra of complex N-linked glycans. Tandem mass spectra are simplified leaving only monoisotopic peaks and based on assumption that all glycans contain a trimannosyl core. The remaining number of GlcNAc in the composition will determine the number of antennae, where the possibility of bisecting GlcNAc is eliminated if the number of remaining Gal is the same as the number of GlcNAc. If the number of Gal or NeuAc or NeuGc is lower than the number of antennae, the different structural isomers are assessed. The introduction of polylactosamine repeats and branch fucosylation is allowed and experimental mass spectra will be evaluated by matching them with those generated *in silico* [15, 16].

Another program for automated annotation of glycan MS data is "Cartoonist", which was developed for the automated annotation of N-glycan MALDI TOF mass spectra. "Cartoonist" is constructed on the principle of labelling MALDI peaks with cartoons representing the most plausible glycan assemblies biosynthesized in mammals using 300 manually determined archetypes [17]. By this approach the numbers of implausible structural candidates for a certain precursor ion is largely reduced, but on the other hand due to the fixed library of created archetypes, its application area is limited by the size of the library itself. Automated identification of N-glycopeptides was realized by extending Cartoonist to another program called "Peptoonist" using a combination of MS and MS/MS data [18].

All previously mentioned tools are mostly focused on single functional task like compositional proposal, data base search and spectra annotation. Moreover, laborious spectrum processing tasks such as peak recognition and charge deconvolution still require manual user intervention in most cases.

The SysBioWare suite presents first platform which pipelines all crucial steps of computational mass spectrometry into one process, in particular:

- raw spectrum processing (baseline adjustment and noise removal),
- peak shape detection using continuous wavelet analysis,
- charge deconvolution and isotopic clustering,
- proposing candidate compositions,

- filtering biologically implausible candidates,
- generating candidate structures using *de novo* approach or open database search,
- assessing candidate structures using *in silico* fragmentation,
- providing reporting facilities for comparing composition lists, etc.

SysBioWare is accessible in a web-based interface or as a desktop application. Desktop version additionally provides data storage and management modules which allows user to organize experiments in tree-like structures with annotation and searching facilities. Also, user can tune analysis parameters to adjust to specific hardware and sample and save parameter sets as analysis profiles. Calculation-intensive tasks such as shape detection and isotopic distribution calculation are implemented using Intel Performance Primitives library which fully exploits processor hardware optimizations [19].

Here is a typical analysis scenario in SysBioWare:

1. User loads MS spectrum data. mzXML or plain ASCII formats are currently supported.
2. User chooses peak detection profile that best suits hardware used and experiment settings.
3. Next, user can run peak detection as a single step. In our experiments this step typically completed within one minute providing true positive rate in the range of 85 – 95%. Additionally, user can step through individual processing stages and adjust individual parameters. This is useful for developing and debugging analysis profiles.
4. User chooses compositional analysis profile that best describes sample substance used in the experiment. This profile determines which building blocks (monosaccharides, adducts and modifications) will be tried and describes plausible biological structures as a set of rules. Every profile element can be adjusted if necessary.
5. Candidate composition list is automatically produced. Since it is based on profile selected at the previous step, only feasible compositions are retained. In our experiments, we achieved one or two compositions per assigned peak.
6. If the goal of this experiment is structure identification, user selects precursor ion and most likely composition and proceeds to MS/MS analysis phase.
7. Here, as in MS-analysis step, user loads MS/MS data and runs peak detection module first.

8. Then, user selects MS/MS analysis profile. This profile sets *de novo* synthesis and *in silico* fragmentation parameters. Alternatively, user can construct molecule structure interactively or block by block using IUPAC notation
9. *De novo* structure generation algorithm then generates candidate structure list which is ranked to bring up the structures which have *in silico* fragmentation pattern closest to observed MS/MS data.
10. The top structure is selected and passed on to *in silico* fragmentation and annotation phase. Here structure and spectrum can be annotated with individual fragments.

The SysBioWare web site www.sysbioware.com provides access to all functionality described. It also provides video introduction and demo analysis sessions for selected carbohydrates structures.

Architecture and General Features of SysBioWare Demonstrated on N-Glycan Mixtures

Computational Mass Spectrometry for Glycomics

The software utilities are constructed of distinct modules dealing with mass spectrometric data processing and their interpretation (Figure 1). *"Composition Blocks"* constructor (Figure 1) allows users to create a library of potential building block components, adducts and modifications, which can be used at further analysis steps. Component increment masses are calculated automatically from elemental composition formula. *"Molecule Classes"* library defines a list of molecule classes to be observed (Figure 1, inset). It is defined by: a) the list of all possible building blocks, which provides the foundation for compositional assignment; b) biological feasibility rules (Bio-filter). These rules, written as a line code, specify conditions or the ratio between building blocks; c) an average elemental composition model is used for more accurate isotopic peak grouping. At the current stage two types of model biomolecules are implemented: peptides and glycans. *"Laboratory Module"* provides user interface to the database of experiments.

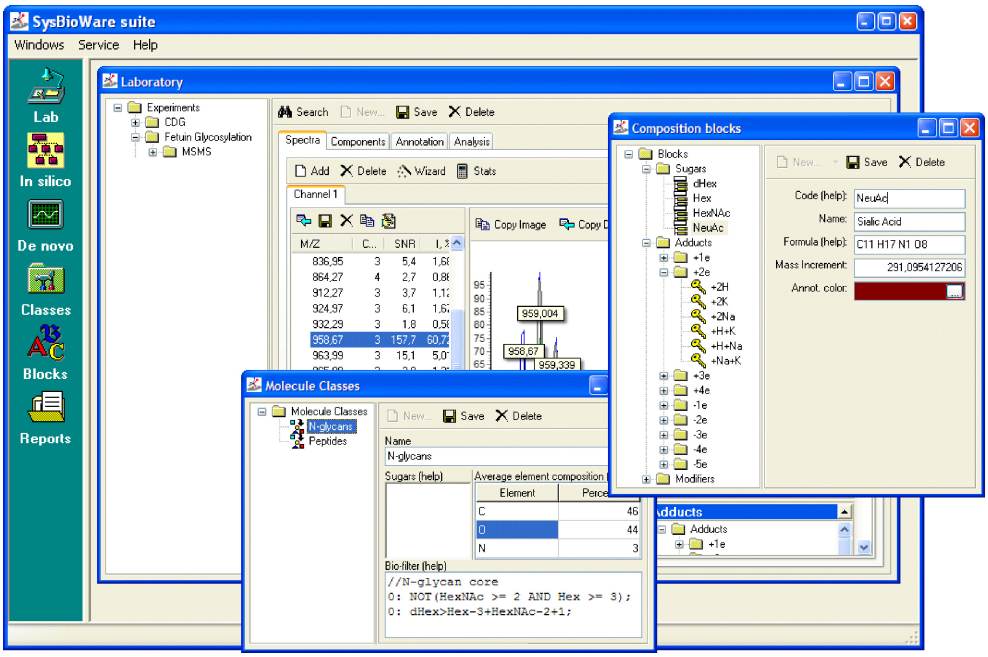

Figure 1. The main window of the SysBioWare 2.0.0 software with activated ‘‘Library’’, Compositions blocks’’ and ‘‘Molecule Classes’’ modules.

The user starts by importing raw mass spectrum at the ‘‘Spectra’’ tab, where Peak Detection Wizard generates initial monoisotopic peak list. The ‘‘Components’’ tab requires users to provide the lists of building blocks, adducts and modifications they are expecting to see as well as the list of bio-filter rules which can be imported from the molecule classes library. ‘‘Annotation’’ tab allows the user to describe various aspects of the experiment such as materials, methods, instrument settings, biological sources, etc. These annotation elements can be later searched to retrieve past experiments and make cross-references. Finally, ‘‘Analysis’’ tab is the place where compositional assignment is performed. This procedure is based on the principle of modelling of the respective glycoconjugate ions using different combinations of potential building blocks defined by the user in the ‘‘Components’’. Afterwards, assignment results can be compared between different experiments and exported to Excel.

Peak Detection Wizard

The wizard provides mass spectrometric data processing (base line correction and noise level determination), peak shape determination (using continuous wavelet transform), automated monoisotopic *m/z* values recognition and charge state determination (Figure 2).

Data import and resampling. The raw spectrum needs to be cleaned from duplicates *m/z* values (spectrum points with the equal *m/z* values and non-equal intensity values) and re-sampled to have uniform *m/z* values. A number of optimization steps, such as the speed of wavelet analysis calculation, are enhanced (Figure 2A-B).

Baseline Correction. We have introduced two modes of baseline correction, which can be optionally selected by the user. The first mode detects the electronic noise, which is assumed to have normal distribution and then uses this noise as a lowest signal level. The second mode is based on the removal of specified proportion of lowest measurements. Afterwards, the baseline can be subtracted from the spectrum (Figure 2C).

Smoothing and Peak Detection. At this step, noise spikes and coarse quantization effects are removed from the spectrum (Figure 2D).

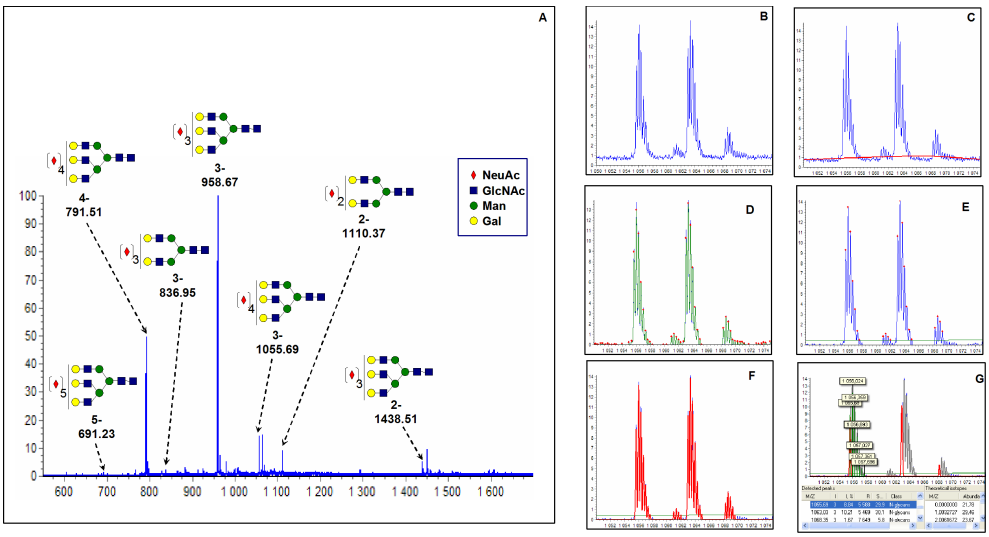

Figure 2. (A) The (-) nanoESI Q-TOF MS data of the mixture of N-glycan digested from the bovine fetuin by PNGase F. Different stages of the working process of Peak Detection Wizard are shown on the example from zoomed area of the raw spectrum (Figure 2A) in the *m/z* range from 1050 to 1075. (B) Data import and resampling stage; (C) Baseline Correction stage; (D) Smoothing and peak detection stage (shown in green); (E) Noise level determination. Peaks detected above the noise level are shown by red circles; (F) Peak Shape Detection stage based on the continuous wavelet analysis. Signals matched by Gaussians are shown in red; (G) Isotopes grouping stage: grouping of different peaks into a single isotopic envelope and charge state determination. The raw spectrum after baseline subtraction and smoothing is shown in blue. Signals matched by Gaussians are shown in grey. The recognized monoisotopic peaks are shown in red.

Noise level determination. The intensity of noisy peaks is assumed to follow normal distribution with varying local standard deviation. The mean of the noise is estimated as the median of peak intensities, and local standard deviation is estimated from the local median absolute deviation. The peaks which intensity exceeds specified percentile are considered to be useful signals (Figure 2E).

Peak detection. Peaks that have passed noise filtering are matched against one of the expected peak shapes (Figure 2F). Currently, Gaussian and Lorentzian shapes are used. Shape matching is implemented using continuous wavelet transform.

Isotope Grouping. After detecting potential peaks by shape matching, SysBioWare attempts to group them as isotopes and determines their charge state (Figure 2G). Starting with a maximum possible charge state, the program tries to find peaks at designated locations that form the packet of the same shape as a corresponding theoretical isotopic distribution. Isotopic distributions are probed by SysBioWare for different molecule classes, which can be selected by the user from the ''Molecule Classes library'' module, *e.g.* glycans and peptides (Figure 1). Finally, interpreted data can be referenced from subsequent experiment records and organized into table as a user report [19].

Isotopic distribution modelling

To show accuracy of approximation in oligosaccharide molecule modelling a comparison between the isotopic distributions of the real composition and computed from the model molecule has been performed (Figure 3). After baseline subtraction peak shape determination, a number of isotopic envelopes corresponding to tri-sialylated triantennary N-glycan (Figure 3A), di-sialylated biantennary N-glycan (Figure 3B) and penta-sialylated triantennary N-glycan (Figure 3C) have been compared with theoretical isotopic distribution normalised to the monoisotopic peak computed from hypothetical glycan model molecule (green dashed line). For these classes of molecules very good consistency was observed. By this, indication of the specific calculation for model molecules at masses higher than 2500 Da is demonstrated. Moreover, preliminary information about sample nature allows the user to tune the Peak Detection Wizard more precisely and correct the oligosaccharide model molecule according to any specific case.

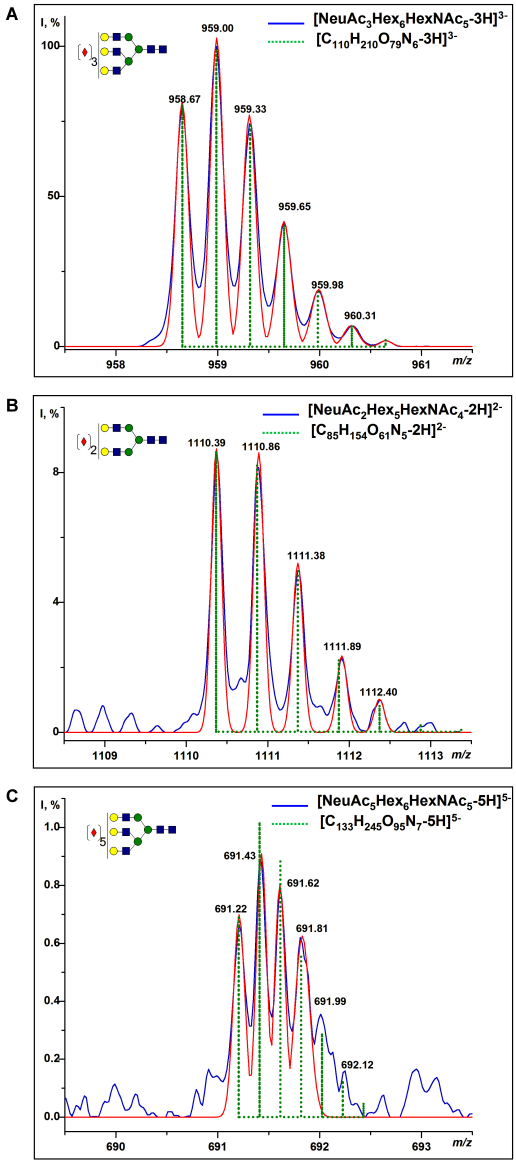

Figure 3. A comparison between baseline corrected raw spectrum from Figure 2A (blue line), peak recognised signal (red line) and isotopic distributions theoretically simulated from the hypothetical analogue calculated from the mass values based on glycan model (green dashed line). As examples the following ions have been considered: trisialylated triantennary N-glycans $NeuAc_3Hex_6HexNAc_5$ (A), pentasialylated triantennary N-glycan $NeuAc_5Hex_6HexNAc_5$ (B) and di-sialylated biantennary N-glycan $NeuAc_2Hex_5HexNAc_4$ (C). 2D glycan pictogram nomenclature is the same as in Figure 2.

System Validation. Automated Peak Detection and Assignment of N-glycan standard mixture

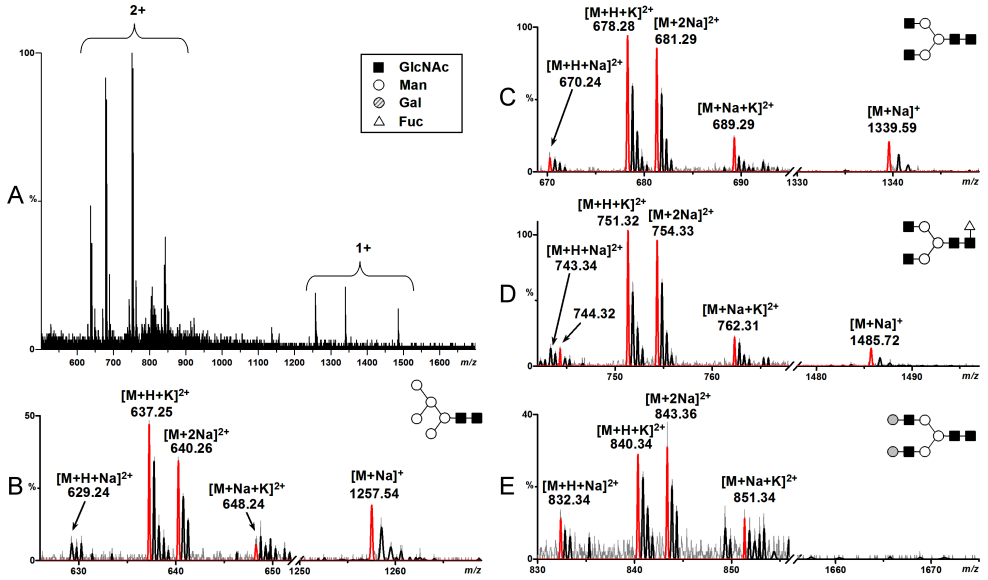

Figure 4. Example of the automated monoisotopic *m*/*z* peak recognition. (A) Positive ion mode nanoESI Q-Tof MS of the equimolar mixture of neutral N-glycan standards Man5, NGA2, NGA2F, and NA2. (B) Expansion of the singly (left) and doubly charged (right) detection area of the Man5 N-glycan. (C) Expansion of the singly (left) and doubly charged (right) detection area of the NGA2 N-glycan. (D) Expansion of the singly (left) and doubly charged (right) detection area of the NGA2F N-glycan. (E) Expansion of the singly (left) and doubly charged (right) detection area of the NA2 N-glycan. © 2008 American Chemical Society, reprinted with permission from [19].

To test the ability of Peak Detection Wizard in managing with high-throughput projects MS data of equimolar mixture of N-glycan standards at low mass concentration and obtained at short acquisition time (few seconds) has selected as submitted for processing. Four neutral N-glycan standards: Man5 (Catalogue No. M-00250S), NGA2 (Catalogue No. C-0720), NGA2F (Catalogue No. C-004301), NA2 (Catalogue No. C-0024300 M) purchased at (Oxford GlycoSciences, Abington, U.K.) have been dissolved each in MeOH/H_2O 1/1(v/v) at concentration of 0.5 pmol/µl and submitted to nanoESI Q-Tof MS analysis in the positive ion mode, simulating high-throughput procedure (Figure 4). At two scans, which correspond approximately to approximately 2 seconds of acquisition time, oligosaccharide standards have been detected as a group of singly and doubly charged ions at different signal abundances, showing the presence different forms of cations mostly formed by [M+Na], [M+K], $[M+H+Na]^{2+}$, $[M+H+K]^{2+}$, $[M+2Na]^{2+}$ and $[M+Na+K]^{2+}$ ionic species. Under the selected

conditions the monoisotopic *m/z* values and charge states for all ionic species corresponding to N-glycan standards have been correctly recognized except for the $[M+H+Na]^{2+}$ at *m/z* 629 corresponding to the Man5 glycan.

Application to Glycourinomics

Turning the focus to complex carbohydrate analysis of human urine (glycourinomics) upon non-invasive sampling has been indicated by investigations, in which a large-scale accumulation of carbohydrate metabolites has been described [20]. Abnormal urinary oligosaccharide and glycopeptide excretion has been observed in clinical cases of hereditary diseases including α-*N*-acetylgalactosaminidase deficiency, G_{M1} gangliosidosis and sialidoses [21].

"Congenital Disorders of Glycosylation" is a group of inherited metabolic diseases caused by low or missing activities of enzymes involved in biosynthesis of oligosaccharides in organism [22 – 24]. Abnormal processing of carbohydrates may lead to changes in glycome profile between patient and control. Human urine was selected as a potential source for monitoring glycoconjugate pattern for the further development of methodological protocol for high-throughput screening and biomarkers discovery [25 – 31]. The one of the most important element of this strategy is application of SysBioWare computational platform for glycomics data processing and interpretation.

The validation of this platform is demonstrated on an example of complex carbohydrate analysis in urine (Figure 5). The oligomeric carbohydrate portion from the patient's urine was considered to contain possible specific metabolic components, which might be indicative for the diagnostic identification of the clinical defect. To obtain the oligomeric carbohydrate portion from the patient's urine a fractionation using gel permeation chromatography steps followed by anion-exchange chromatography has been carried out. A general characteristics of the fractions obtained was their high heterogeneity, the complexity of the profiles obtained by mass spectrometry and a high dynamic range of single components concerning the quantification [32, 33]. Usually, less than 30% of components were in the range of relative abundance 5 – 100%, where more than 70% of all components were in the range from 1 to 5%. By manual evaluation of mass spectra only more abundant components could be safely recognized according to their monoisotopic *m/z* values. Due to the presence of chemical noise clusters, the correct manual recognition in the intensity range below 1% was not reliable.

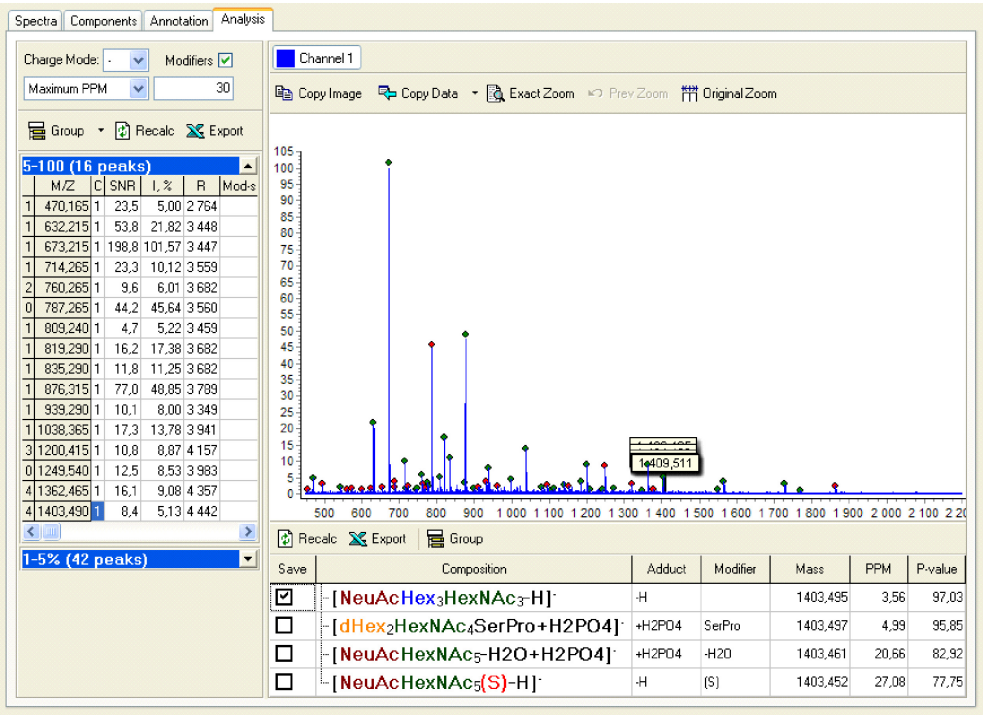

Figure 5. Example of a high-throughput screening analysis: complex glycoconjugate mixture KLM 3 from the CDG patient acquired by the (-) nanoESI Q-Tof MS. Channel 1 contains data related to the KLM 3 fraction. © 2008 American Chemical Society, adapted with permission from [19].

Using SysBioWare function Peak Detection Wizard the parameters were optimized to keep the detection of false positive signals as low as possible. Concerning the ions in the relative intensity range from 5 to 100% all manually determined monoisotopic peaks have been recognized correctly without any true negative and false positive signals. From peaks determined manually in the range of 1 – 5% of the relative abundance none of false positive signals have been recognized and more than 60% monoisotopic *m/z* values have been identified correctly by the Peak Detection Wizard. After the activation of the "Recalc" function within the mass deviation window less than 30 ppm, the list of all possible compositions related to the selected *m/z* values was computed. In the urine sample KLM 3 from 42 molecular ions detected within the relative intensity range 1 to 5% 12 have been uniquely assigned to free oligosaccharides and 1 to a Thr-linked glycan. For seven ionic species, at *m/z* 1112.45, 1143.38, 1208.47, 1271.47, 1313.45, 1565.55 and 1727.58, respectively, two different compositions each were proposed by the platform. With respect to ionic species at *m/z* 1143.38, 1565.55, and 1727.58, the composition with the highest biological relevance was correlated with the higher mass accuracy calculation. For molecular ions at *m/z* 1271.47 and 1313.45, under the current instrumental conditions, both proposed composi-

tions could be acceptable. For ions at *m/z* 1112.45 and 1208.47 both proposed compositions could not be accepted taking in account biosynthetic rules for the glycan assembly. This fact reflects a high necessity for future improvements of a biofilter library regarding the non-relevant glycan moieties. For more accurate compositional evaluation for these ions, high resolution MS and/or fragmentation analysis are necessary. For the first structural evaluation of MS data which can cover around 70% of compositions the SysBioWare platform offers a time scale of less than 1 min, in which the analysis of the raw mass spectrum along with the compositional assignment of glycoconjugate species can be accomplished.

Application to Glycosphingolipidomics

Glycosphingolipids (GSLs) are major components of the outer leaflet of the cell membrane, playing pivotal roles in a variety of biological processes, including the development of cancer [34, 35]. In course of distinct cell surface events during infection or tumour development they show disease-related expression changes. Thus, they could serve as useful targets for biomarker discovery. The molecular structure of GSLs is composed of a hydrophilic, highly variable carbohydrate chain and a lipophilic ceramide anchor. Both portions exhibit numerous structural variations, the combination of which results in a large diversity of GSL structures that can potentially exist [36]. Several hundred GSLs have already been characterized that differ only with regard to the glycan structures (http://sphingolab.biology.gatech.edu) and (http://www.glycosciences.de). The ceramide moiety exhibits variation in the number of carbon atoms, double bonds incorporated into the aliphatic chains, and hydroxyl groups present. Such variation of potential GSL structures, along with the biosynthetic rules, has to be taken into account upon data interpretation.

Applying SysBioWare to glycolipid analysis [37] the program was first adapted according to the structural specificities of GSLs, considering the glycan and the ceramide moiety as separate moieties, accordingly that the ceramide moiety can be introduced as a new building block. In humans, the ceramide is composed of a long chain base, the sphingosine, and a fatty acid bound to the amino group. The GSL model includes all possible structures with fatty acid substitutions from C 14 to C 26, where up to two double bonds can be assigned. The N-acyl linked fatty acids can be hydroxylated or non-hydroxylated and may have an odd number of C atoms. The so-called lyso-forms of GSLs, composed only of the glycan and the long chain base, i. e. without fatty acid substitution, are also supported by this model. The restrictive rules of the BioFilter were derived from the set of GSL structure found in humans (http://sphingolab.biology.gatech.edu).

MS data acquired from a complex GSL mixture purified from human serum were evaluated by the SysBioWare program, in which the Peak Detection Wizard was optimized for GSL analysis, achieving up to 94% of the true positive monoisotopic *m/z* values. By applying the BioFilter to the compositional analysis, it was possible to reduce the number of candidates by a factor of 2 to 4 for 75% of the cases, and by a factor of 5 to 9 for 18% of the cases. 14

peaks were unambiguously assigned to a single composition, whereas in 14 other cases, the correct compositions were placed at the highest rank when scored according to the mass accuracy.

Distinguishing between isomeric structures for correct assignment is a challenging task to be performed from MS data by the BioFilter toolbox. A fucose containing GSL with a hydroxy fatty acid substitution has the same elemental composition as the congruent GSL with a hexose instead of the fucose and a non-hydroxy fatty acid. These aspects were shown to be accurately solved by fragmentation analysis in MS/MS. The interpretation of MS/MS data was supported by *in silico* fragmentation module, calculating a comprehensive set of predicted fragment ions. This function was considered to be a powerful tool for the validation of proposed GSL structures.

According to the SysBioWare assignment proposal, several candidate structures for the signal at *m/z* 1249.74 would be possible, among them the isobaric structure variants with or without fucosylation and hydroxylated fatty acids. *In silico* fragmentation pattern of globotetraosylceramide (Gb4Cer) (d18:1, C 16:0) $[M+Na]^+$ was in good agreement with that obtained by CID experiments of the precursor ions at *m/z* 1249.74, providing a high sequence coverage. Besides, the presence of low amounts of an isomeric or isobaric structure that is likely to contain a terminal hexose in contrast to Gb4Cer with a terminal *N*-acetylhexosamine unit was indicated by low abundant ions at *m/z* 1069.72 and *m/z* 1087.70, which can possibly be assigned to the isobaric neolactotetraosylceramide (nLc4Cer) (d18:1, C 16:0) $[M+Na]^+$, a minor component of the human serum glycosphingolipidome [38]. Upon *in silico* fragmentation of the nLc4Cer component, 43 of the computed ions could be matched with the experimental data, many of which overlapping with fragment ions of its isobar, Gb4Cer.

Combined with *in silico* fragmentation matching, the potential of the SysBioWare platform was shown to be extendable also to GSL samples. Accordingly, it can be considered for rapid and accurate evaluation of complex data sets obtained from different patients' consortia. In protocols, where GSL profiles obtained from clinical samples by MS and MS/MS analysis SysBioWare may serve as a powerful tool for high-throughput glycolipidomics.

Conclusions and Outlook

Algorithms for monoisotopic *m/z* values recognition based the peak shape matching and isotopic grouping have been developed. Peak detection algorithm has been optimized for glycomics applications and successfully tested for mass spectra at different level of complexity. Tuned BioFilter toolbox, providing filtration of unreliable structures has been incorporated into Compositional Module. A novel software platform with an integrated Peak Detection Module and Compositions has been developed, providing the option for high-throughput data analysis within the time scale less than one minutes and their organization into the

local data base for further analysis and mutual correlation. The developed software has been tested for the analysis of complex glycoconjugate mixtures from CDG patients and validated as a potential tool for clinical applications. Combined with *in silico* fragmentation matching, the potential of the SysBioWare platform was shown to be extendable to different glycoconjugate species.

The development of technologies for mass spectrometric analysis during the past 15 years opened new horizons for glycomics. Using modern mass spectrometry very small amounts of complex biological material can be analyzed in a very short time very accurately. Following the progress in instrumentation for data acquisition, huge amount of data are presently generated which require efficient ways to be classified and deposited in databases for applications in life sciences, biotechnology and medicine. Therefore is it high time to get new computational tools available for the broad community of scientists, which are not necessarily always trained as glycochemists or glycobiologists, but who will be using these glycomics tools to make significant contributions to our understanding of the function and the functionally or developmentally induced changes of carbohydrate structures in cells or subcellular compartments. To reach new frontiers in systems biology, glycomics will make progress by providing new and spectacular information, but the task in coming years will be to integrate the work on the fundamentals of this approach and to provide sufficient intellectual and financial resources for the integrated platforms.

Abbreviations

CDG	Congenital Disorders of Glycosylation
CID	Collision induced dissociation
Da	Dalton
dHex	Deoxyhexose
ESI	Electrospray ionization
Gal	Galactose
GlcNAc	*N*-acetylglucosamine
GSL	Glycosphingolipid
Hex	Hexose
HexNAc	*N*-acetylhexosamine
NeuGc	*N*-glycolylneuraminic acid
MALDI	Matrix-assisted laser desorption/ionization
Man5	Oligomannose

MS	Mass spectrometry
MS/MS (MS^n)	Tandem mass spectrometry (n stages)
m/z	Mass to charge ratio
NA2	Asialo-, galactosylated biantennary N-glycan
NeuAc	*N*-acetylneuraminic acid
NGA2	Asialo-, agalacto-, biantennary N-glycan
NGA2F	Asialo-, agalacto-, biantennary N-glycan with core fucose
ppm	Parts per million
Q	Quadrupole
r.i.	Relative intensity
S	Sulphate
Ser	Serine
Thr	Threonine
TOF	Time-of-flight

References

[1] Egge, H., Peter-Katalinić (1987) *J. Mass Spectrom. Rev.* **6**:331 – 393. doi: http://dx.doi.org/10.1002/mas.1280060302.

[2] Packer, N.H., von der Lieth, C.W., Aoki-Kinoshita, K.F., Lebrilla, C.B., Paulson, J.C., Raman, R., Rudd, P., Sasisekharan, R., Taniguchi, N., York, W.S. (2008) *Proteomics* **8**:8 – 20. doi: http://dx.doi.org/10.1002/pmic.200700917.

[3] Aoki-Kinoshita, K.F. (2008) *PLoS Comput. Biol.* **4**. doi: http://dx.doi.org/10.1371/journal.pcbi.1000075.

[4] von der Lieth, C.W., Bohne-Lang, A., Lohmann, K.K., Frank, M. (2004) *Brief Bioinform.* **5***:*164 – 178. doi: http://dx.doi.org/10.1093/bib/5.2.164.

[5] von der Lieth, C.W., Lutteke, T., Frank, M. (2006) *Biochim. Biophys. Acta* **1760**:568 – 577. doi: http://dx.doi.org/10.1016/j.bbagen.2005.12.004.

[6] Cooper, C.A., Gasteiger, E., Packer, N.H. (2001) *Proteomics* **1**:340 – 349. doi: http://dx.doi.org/10.1002/1615-9861(200102)1:2<340::AID-PROT340>3.3.CO;2-2.

[7] Cooper, C.A., Joshi, H.J., Harrison, M.J., Wilkins, M.R., Packer, N.H. (2003) *Nucleic Acids Res*. **31**:511 – 513.
doi: http://dx.doi.org/10.1093/nar/gkg099.

[8] Go, E.P., Rebecchi, K.R., Dalpathado, D.S., Bandu, M.L., Zhang, Y., Desaire, H. (2007) *Anal. Chem.* **79**:1708 – 1713.
doi: http://dx.doi.org/10.1021/ac061548c.

[9] Joshi, H.J., Harrison, M.J., Schulz, B.L., Cooper, C.A., Packer, N.H., Karlsson, N.G. (2004) *Proteomics* **4**:1650 – 1664.
doi: http://dx.doi.org/10.1002/pmic.200300784.

[10] Lohmann, K.K., von der Lieth, C.W. (2003) *Proteomics* **3**:2028 – 2035.
doi: http://dx.doi.org/10.1002/pmic.200300505.

[11] Lohmann, K.K., von der Lieth, C.W. (2003) *Glycobiology* **13**:846 – 846.

[12] Gaucher, S.P., Morrow, J., Leary, J.A. (2000) *Anal. Chem.* **72**:2331 – 2336.
doi: http://dx.doi.org/10.1021/ac000096f.

[13] Maass, K., Ranzinger, R., Geyer, H., von der Lieth, C.W., Geyer, R. (2007) *Proteomics* **7**:4435 – 4444.
doi: http://dx.doi.org/10.1002/pmic.200700253.

[14] Apte, A., Meitei, S.M. (2010) *Methods Mol. Biol.* **600**:269.
doi: http://dx.doi.org/10.1007/978-1-60761-454-8_19.

[15] Ethier, M., Saba, J.A., Ens, W., Standing, K.G., Perreault, H. (2002) *Rapid Commun. Mass. Spectrom.* **16**:1743 – 1754.
doi: http://dx.doi.org/10.1002/rcm.779.

[16] Ethier, M., Saba, J.A., Spearman, M., Krokhin, O., Butler, M., Ens, W., Standing, K.G., Perreault, H. (2003) *Rapid Commun. Mass Spectrom.* **17**:2713 – 2720.
doi: http://dx.doi.org/10.1002/rcm.1252.

[17] Goldberg, D., Sutton-Smith, M., Paulson, J., Dell, A. (2005) *Proteomics* **5**:865 – 875.
doi: http://dx.doi.org/10.1002/pmic.200401071.

[18] Goldberg, D., Bern, M., Parry, S., Sutton-Smith, M., Panico, M., Morris, H. R., Dell, A. (2007) *J. Prot. Res.* **6**:3995 – 4005.
doi: http://dx.doi.org/10.1021/pr070239f.

[19] Vakhrushev, S.Y., Dadimov, D., Peter-Katalinić, J. (2009) *Anal. Chem.* **81**:3252 – 3260.
doi: http://dx.doi.org/10.1021/ac802408f.

[20] Linden, H.U., Klein, R.A., Egge, H., Peter-Katalinić, J., Dabrowski, J., Schindler, D. (1989) *Biol. Chem. Hoppe Seyler* **370**:661 – 672.

[21] Schindler, D., Kanzaki, T., Desnick, R.J. (1990) *Clin. Chim. Acta* **190**:81 – 91. doi: http://dx.doi.org/10.1016/0009-8981(90)90282-W.

[22] Jaeken, J. (2004) *J. Inherit. Metab. Dis.* **27**:423 – 426. doi: http://dx.doi.org/10.1023/B:BOLI.0000031221.44647.9e.

[23] Jaeken, J., Carchon, H. (2004) *Curr. Opin. Pediatr.* **16**:434 – 439. doi: http://dx.doi.org/10.1097/01.mop.0000133636.56790.4a.

[24] Jaeken, J., Matthijs, G. (2007) *Annu. Rev. Genomics Hum. Genet.* **8**:261 – 278. doi: http://dx.doi.org/10.1146/annurev.genom.8.080706.092327.

[25] Frösch, M., Bindila, L., Zamfir, A., Peter-Katalinić, J. (2003) *Rapid Commun. Mass Spectrom.* **17**:2822 – 2832. doi: http://dx.doi.org/10.1002/rcm.1273.

[26] Vakhrushev, S.Y., Langridge, J., Campuzano, I., Hughes, C., Peter-Katalinić, J. (2008) *J. Clin. Proteom.* **4**:47 – 57. doi: http://dx.doi.org/10.1007/s12014-008-9010-3.

[27] Vakhrushev, S.Y., Langridge, J., Campuzano, I., Hughes, C., Peter-Katalinić, J. (2008) *Anal. Chem.* **80**:2506 – 2513. doi: http://dx.doi.org/10.1021/ac7023443.

[28] Vakhrushev, S.Y., Mormann, M., Peter-Katalinić, J. (2006) *Proteomics* **6**:983 – 992. doi: http://dx.doi.org/10.1002/pmic.200500051.

[29] Vakhrushev, S.Y., Snel, M.F., Langridge, J., Peter-Katalinić, J. (2008) *Carbohydrate. Res.* **343**:2172 – 2183. doi: http://dx.doi.org/10.1016/j.carres.2007.11.014.

[30] Vakhrushev, S.Y., Zamfir, A., Peter-Katalinić, J. (2004) *J. Am. Soc. Mass Spectrom.* **15**:1863 – 1868. doi: http://dx.doi.org/10.1016/j.jasms.2004.09.008.

[31] Zamfir, A., Vakhrushev, S., Sterling, A., Niebel, H.J., Allen, M., Peter-Katalinić, J. (2004) *Anal. Chem.* **76**:2046 – 2054. doi: http://dx.doi.org/10.1021/ac035320q.

[32] Vakhrushev, S.Y., Mormann, M., Peter-Katalinić, J. (2006) *Proteomics* **6**:983 – 992. doi: http://dx.doi.org/10.1002/pmic.200500051.

[33] Vakhrushev, S.Y., Zamfir, A., Peter-Katalinić, J. (2004) *J. Am. Soc. Mass Spectrom.* **15**:1863 – 1868.
doi: http://dx.doi.org/10.1016/j.jasms.2004.09.008.

[34] Hakomori, S. (1996) *Cancer Res.* **56**:5309 – 5318.

[35] Kannagi, R., Yin, J., Miyazaki, K., Izawa, M. (2008) *Biochim. Biophys. Acta* **1780**:525 – 531.
doi: http://dx.doi.org/10.1016/j.bbagen.2007.10.007.

[36] Peter-Katalinić, J., Egge, H. (1990) *Methods Enzymol.* **193**:713 – 733.
doi: http://dx.doi.org/10.1016/0076-6879(90)93446-R.

[37] Souady, J., Dadimov, D., Kirsch, S., Bindila, L., Peter-Katalinić, J., Vakhrushev, S.Y. (2010) *Rapid Commun. Mass Spectrom.* **24**:1 – 10.
doi: http://dx.doi.org/10.1002/rcm.4479.

[38] Kundu, S.K., Diego, I., Osovitz, S., Marcus, D.M. (1985) *Arch. Biochem. Biophys.* **238**:388 – 400.
doi: http://dx.doi.org/10.1016/0003-9861(85)90179-1.