GlycoWorkbench: A Tool for the Computer Assisted Annotation of Mass Spectra of Glycans*

Alessio Ceroni[1], Kai Maass[2], Hildegard Geyer[2], Rudolf Geyer[2], Anne Dell[1], and Stuart M. Haslam[1]

[1] Division of Molecular Biosciences, Imperial College London, London, SW7 2AZ, UK

[2] Institute of Biochemistry, Faculty of Medicine, University of Giessen, Germany

* Dedicated to Dr. Claus-Wilhelm "Willi" von der Lieth

To whom correspondence should be addressed:

Dr. Stuart Haslam,

Division of Molecular Biosciences, Imperial College London, Exhibition Road, SW7 2AZ London, UK

Fax: +44 207 225 0458, E-mail: s.haslam@imperial.ac.uk

Abstract

Mass spectrometry is the main analytical technique currently used to address the challenges of glycomics as it offers unrivalled levels of sensitivity and the ability to handle complex mixtures of different glycan variations. Determination of glycan structures from analysis of MS data is a major bottleneck in high-throughput glycomics projects, and robust solutions to this problem are of critical importance. However, all the approaches currently available have inherent restrictions to the type of glycans they can identify and none of them has proved to be a definitive tool for glycomics.

GlycoWorkbench is a software tool developed by the EUROCarbDB initiative to assist the manual interpretation of MS data. The main task of *GlycoWorkbench* is to evaluate a set of structures proposed by the user by matching the corresponding theoretical list of fragment masses against the list of peaks derived from the spectrum. The tool provides an easy to use graphical interface, a comprehensive and increasing set of structural constituents, an exhaustive collection of fragmentation types, and a broad list of annotation options. The aim of *GlycoWorkbench* is to offer complete support for the routine interpretation of MS data. The software is available for download from: http://www.eurocarbdb.org/applications/ms-tools.

Keywords: (semi-) automated annotation / glycan structure analysis / mass spectrometry

1. Introduction

Carbohydrates are ubiquitous biological molecules and their roles in living organisms are varied and fundamental. Complex carbohydrates (also referred as *glycans*) are usually synthesized by sequential attachment of saccharide donors to a growing carbohydrate acceptor by specific enzymes. These monosaccharide units are covalently linked by glycosidic bonds, either in α or β configuration depending on the orientation of the anomeric centers. Glycans can have complex structures with multiple branching points, since each hydroxyl group of a monosaccharide constitutes a possible point of formation for a glycosidic bond. Further modifications of the basic monosaccharide unit at the various hydroxyl positions, such as substitution of the proton with other moieties or de-oxygenation, contribute to the structural complexity.

Glycans can be found as polymers made up exclusively of sugar residues but are usually observed in *glycoconjugates*, associated with other biomolecules such as lipids or proteins. Three main types of protein glycosylation exist: the carbohydrate can either be linked to the amide nitrogen atom of an asparagine residue (N-linked glycosylation), to the hydroxyl oxygen of a serine, threonine or hydroxyproline residue (O-linked glycosylation) or to the C-terminal amino-acid (GPI-linked).

Glycans can have structural and modulatory functions by themselves or can modulate the function of the molecules to which they are attached by the specific recognition of the glycan structure by carbohydrate-binding proteins. Glycans regulate both the folding and degradation of proteins. Moreover, since the outer cell membrane is covered by carbohydrates, they mediate interactions with other cells of the same organism or with pathogenic organisms such as viruses, bacteria and multi-cellular parasites. Glycans are increasingly implicated in playing a critical role in human diseases and their potential utility as biomarkers for pathological conditions is a major

driver for characterization of the glycome, the collection of all glycoconjugates synthesized by an organism.

1.1. Mass Spectrometry of Glycans

Mass spectrometry is the main analytical technique currently used to address the challenges of glycomics as it offers unrivalled levels of sensitivity and the ability to handle the complex mixtures of different glycan variations[1]. Modern MS techniques are capable of producing mass spectra of both the whole glycan (molecular ion) and the fragmented glycan (fragment ions). The high level of sequence information contained in the fragment ion spectra can be exploited to resolve the structure of a glycan molecule. Fragmentation by post-source decay (PSD)[2], high energy collision induced dissociation (CID)[3], infra-red lasers (IRMPD)[4] etc, involves the cleavage of one or more bonds in the glycan molecule. A popular nomenclature for identifying the various types of cleavages has been devised by Domon and Costello[5] and it is shown in Figure 1. The most common type of fragment produced in MS instrumentation[6] involves the cleavage of a glycosidic bond with the production of an ion that can maintain (Y and Z fragments) or not (B and C) the reducing-end of the original glycan. High energy collisions in the CID chamber of the MS instrument can also induce the breakage of the saccharide ring provoked by the cleavage of two bonds. The fragments resulting from these cross-ring cleavages can either maintain the reducing-end (X) or not (A).

The correct interpretation of glycan fragment ion mass spectra is fundamental to the determination of the glycan structure, just as the interpretation of peptide fragment ion mass spectra is fundamental to protein identification in proteomic experiments. However, the additional complexity of glycan structures compared to protein sequences poses greater difficulty during the analytical process. The monosaccharide units often have the same chemical

constitution, differing only in the stereochemistry of the hydroxyl groups, and cannot be distinguished by their mass. Moreover, detection of the linkage positions between monomers is dependent on the presence of specific cross-ring fragments which are not always produced. Therefore, other types of information such as knowledge of glycan biosynthetic pathways are usually incorporated during a complete structure assignment.

1.2. Automated Interpretation of Mass Spectrometry Data

Determination of glycan structures from analysis of MS data is a major bottleneck in high-throughput glycomics projects, and robust solutions to this problem are of critical importance. Therefore, it is not surprising that various experimentally oriented groups have been developing software solutions and algorithms to bypass this bottleneck. However, the current status of tools to analyze glycan MS data shows that automated interpretation of mass spectrometric data is still an evolving field. Up to now only a few software tools have been available to support experimentalists during the annotation process, and the capability of these tools is somewhat varied.

Library-based sequencing tools identify the glycan sequence by matching the unassigned mass spectra with data derived from known glycan structures. Similarly to the SEQUEST[7] method used for protein sequencing, *GlycosidIQ*[8] generates a theoretical peak list for each structure in the database by computing all its theoretical fragments. The best match between the theoretical peak lists and the mass spectra is then derived using a suitable scoring function. A totally different approach to library-based sequencing is through matching the unassigned spectra against a library of experimentally determined fragment spectra[9]. Both approaches are severely limited by the availability of reliable data, since no comprehensive and well curated collection of

5

experimentally derived glycan sequences exists at the moment, and no public collection of assigned MS^n spectra from pure glycans is available.

De novo sequencing tools are not restricted to previously characterized structures but, of the many approaches that have been proposed, no single one has demonstrated the capability to deliver the desired accuracy and flexibility. Composition analysis tools, such as *GlycoMod*[10] and Glyco-Peakfinder[11], use data from single mass spectrometry measurements to estimate the quantities and classes of monosaccharide components of the glycan structure. The number of compositions matching a certain mass value scales exponentially with the number of different monomers that can form the solution; therefore taxonomic and biosynthetic information must be used to restrict the number of results. An innovative step in this direction has been taken by the *Cartoonist* tool[12], which generates only the N-linked glycans possibly synthesized by mammalian cells using a set of archetypal structures and a set of rules for the modification of said structures. The archetypes and rules have been compiled by a group of experts, and represent the current knowledge about biosynthetic pathways in mammalian organisms. Eventually, the multiple possibilities resulting from a composition analysis need to be validated by tandem mass spectrometry experiments.

Several approaches to de-novo sequencing have been proposed using data from MS^n fragmentation experiments for deriving the complete structure. *STAT*[13] generates all the possible structural topologies from a composition selected by the user amongst those compatible with the precursor mass. The structures thus produced are evaluated against the given peak list, and ranked accordingly. Like *STAT*, *Oscar*[14] generates candidate structures from an estimated composition but uses the information contained in fragmentation pathways of permethylated oligosaccharides as a basis for restricting the set of possible results, which must contain the

common N-glycan mammalian core ($Man_3GlcNAc_2$). In *StrOligo*[15], the differences between fragment masses are used to estimate the loss of known moieties and to produce a candidate composition for the precursor ion. Given the estimated composition, a set of structures is generated by applying biosynthetic rules specific to mammalian N-glycans. *GLYCH*[16] is derived from *de novo* peptide sequencing programs by allowing branches in the polymer structure (only binary branching is considered). The *GLYCH* algorithm performs a maximization of the number of assigned peaks by generating a series of B-ions starting from the leaves of the glycan tree structure. The complete structure is generated from the top-level B-ion by re-ranking the top scoring results according also to double cleavages.

1.3. Computer Assisted Interpretation of MS Data

All the approaches described in the previous section have inherent restrictions to the type of glycans they can identify and none of them has proved to be a definitive tool for glycomics. Expert knowledge about glycan biosynthesis is fundamental for the correct interpretation of a spectrum in order to restrict the number of solutions matching experimental data and to obtain reasonable results. Unfortunately, this information is not yet available in the form of comprehensive data collections, which makes completely automated annotation of generic glycan mass spectra still an unfeasible task.

The EUROCarbDB design study[17] aims to close this gap by creating the foundations for databases and bioinformatic tools in the realm of glycobiology and glycomics. The importance of the EUROCarbDB initiative in the development of glycan structure databases has been widely recognized[18, 19]. The EUROCarbDB project is currently establishing the technical infrastructure for a glycan database where all interested research groups could feed in their primary data, and it is already providing tools to aid the interpretation of these data.

GlycoWorkbench is a software tool developed by the EUROCarbDB initiative to assist the manual interpretation of MS data. Manual annotation of fragment spectra comprises a series of tedious and repetitive steps whose automation is straightforward, and can result in a substantial decrease of the time needed for sequencing a structure. Like other semi-automatic sequencing tools[20, 21], the main task of *GlycoWorkbench* is to evaluate a set of structures proposed by the user by matching the corresponding theoretical list of fragment masses against the list of peaks derived from the spectrum. Unlike any other semi-automatic tool, *GlycoWorkbench* provides an integrated environment with an easy to use graphical interface, a comprehensive and increasing set of structural constituents, an exhaustive collection of fragmentation types, and a broad list of annotation options.

GlycoWorkbench incorporates an intuitive visual editor of glycan structures, the *GlycanBuilder*[22], that enables a rapid assembly of structure models using a comprehensive collection of building blocks, and their display in several popular symbolic notations. The *in-silico* fragmentation engine computes a complete list of theoretical fragments including multiple glycosidic cleavages and all the possible ring fragments for every available type of monosaccharide. The annotation engine automatically matches the theoretical list of fragment masses with the experimental peak-list by taking into account several types of experimental techniques, various types and quantities of ion adducts, and neutral exchanges. The proposed annotations are presented using comprehensive and easily understandable reports that allow the determination of the correct structure by comparing the quality and the coverage of the different annotations from the structure candidates. The aim of *GlycoWorkbench* is to provide a complete support to the routine interpretation of mass spectrometric data and to form the basis for the development of a completely automatic assignment tool. The software is publicly available for download from the

8

EUROCarbDB website[23]. The features of the tool will be explained in more detail in the following sections.

2. *Material and methods*

GlycoWorkbench features a user friendly graphical interface designed to simplify and accelerate the routine steps performed during interpretation of a mass spectra. The typical semi-automatic annotation workflow involves: definition of the candidate structures, specification of the peak list, computation of fragments and relative mass to charge values, and annotation of peaks. All data produced with the tool, such as structures, peak-lists and annotations, can be printed or saved to file for later consultation.

2.1. Input and Display of Structures

Each intact or fragmented molecule is modeled in *GlycoWorkbench* as a tree structure whose nodes represent: monosaccharides, monosaccharide modifications, glycosidic or cross-ring cleavages, and reducing-end specificators or markers. The linkages between the monosaccharides are represented by the edges of the tree. Reducing-end specificators are used to identify possible modifications at the reducing-end terminal (e.g. reduction, fluorescent markers or no modification). Each node has a connection to a distinct parent node except those who describe unspecified linkages at the non-reducing end(s) of a glycan structure. A special node with no parents is defined for collecting glycan terminals with unspecified linkages.

The branching of constituents of a glycan molecule does not allow the input of the structure as straightforwardly as writing the linear sequence of amino-acids of a peptide chain. Additionally, numerous alternative notations are commonly adopted to graphically represent glycan structures and fragments. A user friendly input/output tool for glycan structures should provide an intuitive interface to build structures with minimal user interaction and create conventional and

informative graphical representations of glycans. *GlycoWorkbench* uses the *GlycanBuilder* tool[22] for visualizing and editing the candidate structures in the main drawing canvas (Figure 2), and for displaying the fragments in the annotation panels.

The *GlycanBuilder* tool is based on an automatic rendering algorithm that generates the monosaccharide symbolic or textual representations and determines their arrangement in the drawing panel. The most commonly used symbolic representation for glycans from the Consortium for Functional Glycomics[24] is available together with other less favored variations such as that utilized by the Oxford Glycobiology Institute[25]. The aspect and placement of residues and linkages is decided by a configurable set of rules specific for each notation. The flexibility of the rendering algorithm enables *GlycanBuilder* to be employed as an easy-to-use editor for defining structures as well as a component for the generation of pictorial representations of glycans and fragments.

Using *GlycanBuilder*, a glycan can be rapidly specified starting from the reducing end by sequentially adding monosaccharides, modifications or reducing-end markers to the already drawn structure. Each addition is performed by selecting the point of attachment and the type of the new residue. The list of structural constituents contains a comprehensive collection of monosaccharides, substituents, reducing-end markers and monosaccharide modifications (see Table 1, Table 2 and Table 3 for the complete list). Additionally, a library of biologically relevant structural motifs (comprising both cores and terminals) is included to facilitate the input of structures. All stereo-chemical information about a monosaccharide (anomeric conformation, chirality, ring size, and linkage position) can be subsequently specified. Finally, the usual editing functions (cut & copy, undo/redo and drag & drop) are provided.

Structural data can be imported into the drawing panel from various encoding formats in use by existing databases initiatives, such as: LINUCS[26] used by the *Glycosciences.de*[27] portal, the format devised by GlycoMinds Ltd[28] used by the Consortium for Functional Glycomics[29] and Glyco-CT[30] developed by the EUROCarbDB initiative[17]. In this way structures that are already defined and stored in a database can be easily tested against the acquired spectra.

2.2. Computation of Masses

Each candidate structure defined in *GlycoWorkbench* is associated with a set of parameters that specifies the type of per-substitution (either none or one of per-methylation, per-acetylation, per-deuteromethylation and per-deuteroacetylation), the identities and quantities of ion adducts (H^+, Na^+, K^+, Li^+ are currently available) and the neutral exchanges. Modifications at the reducing-end (such as fluorescence labels) and single position substitutions (such as sulphates) are considered as constituents of the structure. A configuration file stores the value of masses and number of positions available to methyl and acetyl substituents for each possible structural constituent. The mass of the intact or fragment molecule is computed by traversing the structure, counting the mass of each component incremented by the possible per-substitutions and accounting for the mass loss given by the formation of glycosidic bonds. The mass to charge ratio is finally computed from the mass value by taking into account ion adducts and neutral exchanges.

2.3. Specification of a peak-list

The "PeakList" panel (Figure 2) allows the user to visualize and modify the list of labeled peaks (simply referred to as *peak-list* in the text and the software) that will be used during annotation. The peak-list can be loaded from a tab-separated text file, thus allowing for import from external applications such as peak-picking software, or can be created by typing mass and intensity values directly in the spreadsheet-like view. Alternatively, the raw spectrum can be loaded from file,

using several standard XML or vendor specific data formats (supported through the use of the ProteomeCommons IO library[31]). The data is displayed in the "Spectra" panel (Figure 2) and can be panned or zoomed to highlight specific regions. The user can then select the mass-to-charge values directly from the spectrum and add them to the peak-list. In *GlycoWorkbench* there are no functions for processing the spectra, like de-convolution or centroid discovery, since these features are already found in the software provided with the MS instrumentation from which the peak-list can be exported.

2.4. *In-silico* Fragmentation

The computation of fragments and their masses from the intact structure is a central step for the annotation of MS^n spectra. The fragmentation of glycans is very specific to the experimental conditions which can be extremely varied. Therefore, the strategy implemented in *GlycoWorkbench* is to generate all topologically possible fragmentations of the precursor molecular ion, applying both multiple glycosidic cleavages and cross-ring fragmentations, in order to cover the broadest possible range of conditions. The type and number of cleavages that are generated can be specified by the user. The list of cross-ring fragments is derived from a configuration file listing all possible cross-ring cleavages for each available monosaccharide type. For each entry in the file, the mass and the hydroxyl positions inherited from the intact monosaccharide ring are specified.

Fragments are computed by traversing the tree structure of the glycan and applying all the applicable cleavages at each single node. A fragment is allowed only if it contains at least an intact monosaccharide residue. In case of glycosidic cleavages two different sub-trees are created from the original structure: one corresponding to the sub-tree rooted at the current node (B or C cleavages) and the other being its complementary set of nodes (for Y and Z). In case of cross-ring

cleavages, the current node is first substituted with the corresponding cleaved ring. The algorithm then checks which hydroxyl positions of the monosaccharide ring are conserved by the cleavage, and leaves all the corresponding linkages intact while removing the other residues. Internal cross-ring fragments are not allowed (having both the reducing and non-reducing end sides) since they are rarely observed in practice. The fragmentation algorithm is recursively applied to fragmented structures in order to produce multiple cleavages.

The set of all generated fragments can be displayed in the "Fragments/List" panel by a tabular form that contains in each row: the fragment structure represented in the current symbolic notation, the type of fragment specified as a the list of cleavage types in the Domon and Costello notation, the mass to charge ratio given the ion adducts (inherited from the parent structure), the identities and quantities of ion adducts, the neutral exchanges (if any), and the mass of the fragment without adducts.

A visual editor of glycan fragments is also available (Figure 3a), where the user can specify in which positions the cleavages are occurring on the displayed structure in order to reproduce an already known fragment molecule. A single click on a glycosidic bond of a structure model generates the two resulting fragments. Similarly, the cross-ring fragments are generated by clicking on a monosaccharide residue. Multiple cleavages are produced by selecting the cleavage position on the already fragmented molecule. All the fragments are displayed with their mass and mass to charge value and can be copied to the structure editor for exporting to other software tools.

2.5. Automatic Annotation

The list of fragments generated by *in-silico* fragmentation of each candidate structure is finally tested for matches against the list of labeled peaks. Each fragment is tested against each peak to

check if the computed *m/z* value matches the experimentally derived one given the desired accuracy. For each fragment all possible combinations of ions adduct are generated. This feature allows the annotation of mass spectra derived from all sort of instrumentation by generating singly or multiply charged ions. The user can specify the maximum number and types of ions that can be associated with the glycan together with the possible number of neutral exchanges of charges (same choices available for the computation of masses). The maximum number of exchanges is determined by counting the charges available on the structure (given by the carboxylic, phosphate and sulphate groups) and can be further limited by specifying which ions are exchanged with protons.

The resulting annotated peak-list can be viewed using various panels that show different types of information. Each panel is based around a spreadsheet-like tabular form, whose cell values can be sorted by each column, and can be copied into other spreadsheet software. The "Annotation/Details" view (see Figure 4a) shows a detailed list of fragment-peak matches for each candidate structure. For each entry in the list, the peak intensity and *m/z* value are displayed together with the associated fragment structure, its mass and *m/z* value, the type of cleavages, the annotation accuracy (as the difference between the *m/z* values), the number and types of ion adducts and neutral exchanges. In *GlycoWorkbench* the type of the cleavage does not specify the position of the cleaved bond(s) (as in the Domon and Costello notation), since fragments with identical chemical structures are shown only once but can arise from cleavages in different parts of the glycan. This view can be used to refine the assignments by removing the matches that are not satisfactory given the user knowledge of the fragmentation pathway. The "Annotation/Summary" view (Figure 5b) lets the user compare the annotations for the different structures back-to-back in the same table. The matching fragments from the different candidates are shown in adjacent columns, with each row corresponding to a single peak. In this way, signals

that could distinguish the correct annotation from the other hypothetical models can be easily identified. The "Annotation/Statistics" view (Figure 5a) lets the user perform a quantitative comparison between the annotations, by showing a few aggregated indicators of the quality of the annotations. The coverage of the annotation is computed as the sum of the intensity values of all matched peaks divided by the sum of the intensities of all peaks. The average deviation between the acquired and the calculated mass to charge values is displayed in absolute and in *ppm* scale. The number of annotated peaks is displayed at three different thresholds of the relative intensities: for all the peaks, for peaks with intensity greater than 10% that of the highest peak and greater than 5%. The latter values focus on the major peaks to verify if the main signals in the spectrum are explained. Finally, the "Annotation/Calibration" view shows a scatter plot where each annotation has X coordinate corresponding to the real *m/z* value and Y coordinate corresponding to the accuracy of the annotation. For each peak, the best annotation giving the lowest deviation from the measured *m/z* value is highlighted in red. This view allows the user to verify the correct calibration of the mass spectra by highlighting trends in the annotation accuracy.

3. *Results and Discussion*

The use of *GlycoWorkbench* can greatly simplify the routine work conducted during interpretation of mass spectrometric data. The efficacy of the features offered by the tool can be best demonstrated using examples of practical annotation of mass spectra. In the following paragraphs several common use cases are shown, which include: detection of ion pairs from single bond cleavages to enhance manual interpretation of a mass spectrum, semi-automatic annotation of an MS/MS spectrum, differentiation between various structure candidates, location of an undetermined fucose, and detection of cross-ring fragments from a permethylated glycan.

The first four examples use data collected from the glycan structures present in a sample of batroxobin toxin from the *Bothrops moojeni* venom[32]. The investigated MALDI spectra of the pyridylaminated (-PA) N-glycans were recorded on an Ultraflex I (Bruker Daltonik, Bremen DE) in positive ion LID mode[33]. The last example uses data collected from a sample of Lacto-N-fucopentaose (Dextra, Reading UK). The glycan was permethylated using the procedure described by Dell[34] and the spectrum was obtained with an MALDI-ToF/ToF 4800 (Applied Biosystem, Foster City CA) in positive ion reflectron mode.

The figures showing annotated spectra have been produced by copying the fragments and structures drawn with *GlycoWorkbench* into a graphic editor.

3.1. Using the fragment editor for manual annotation

Manual interpretation of mass spectra of glycans is often a search for ion pairs which arise from the cleavage of single glycosidic bonds. The "Fragments Editor" uses the *in-silico* fragmentation tool to generate these fragments and allows a fast detection of such pairs from their *m/z* value. Figure 3A shows examples of such ion pairs generated from a bi-antennary N-glycan $Hex_3HexNAc_6$-PA. The ion pairs at *m/z* 204 and *m/z* 1598, *m/z* 407 and *m/z* 1395, *m/z* 569 and *m/z* 1233, describe the step by step degradation of one of the antennae of the N-glycan. Each pair has one peak representing a B-ion in the lower mass region and a corresponding peak representing a Y-ion in the higher mass region (Figure 3B). The "Fragments Editor" can also be useful to check the mass values of an already manually annotated spectrum. The completely assigned spectrum is given in the supplementary material (Figure S1).

3.2. Complete annotation of a spectrum

An almost complete annotation of the major peaks of a spectrum is necessary for the determination of a glycan structure by mass spectrometry. The automatic annotation tool from

16

GlycoWorkbench can be used to match the *in silico*-generated list of fragments of the given structure candidates with the list of peaks labeled in the spectrum. Figure 4A displays the automatic annotation of the peak list of the mass spectrum of a sodiated N-glycan $Hex_3HexNAc_6Fuc_1$-PA sorted by intensity of the mass signals. Only the most significant matches are shown to increase the clarity of the figure. The "Annotation Details" panel gives a detailed overview of the annotated peak list and allows a review of the annotation results. All assigned fragments are represented in the spectrum in Figure 4B. The flexibility of *GlycoWorkbench* allows parallel annotation of fragments with different ion adducts, such as sodiated and protonated fragments. The completely assigned spectrum is given in the supplementary material (Figure S2).

3.3. Discrimination between different structure candidates

The third example demonstrates how *GlycoWorkbench* can be effectively used when comparing more than one structure candidate with the acquired spectrum. After a composition analysis of the precursor mass of a fragment spectrum and a composition search in databases (e. g. using the Glyco-Peakfinder webservice[11]) candidates with more than one structure can be possible. As described in the previous example, the matching of the peak-list with the *in-silico* generated fragments can be done as a parallel calculation for more than one structure candidate. Figure 5A displays the "Stats" view of the matching of three candidates with the spectrum of a protonated N-glycan $Hex_5HexNAc_4Fuc_1$-PA.

In our example, the structure candidates either carrying fucose at an antenna or being of the complex-type N-glycan have noticeably worse coverage than the hybrid-type structure model. However, the choice between the candidates can only be made by rigorously comparing the annotations for each peak. Figure 5B gives a more detailed view of the matches between *in-silico*

fragmentation of all candidates and the mass list using the "Summary" panel. The final structure determination from the mass spectrum (for complete assignment see supplementary material) is based on the annotation of two major peaks in the spectrum: the signal at m/z 446 (FucHexNAc-PA) definitely shows a core fucosylation and the peak at m/z 407 (HexNAc$_2$) clearly proves the existence of only one complex-type antenna, since the complete structure comprises only 4 HexNAc in total.

3.4. Automatic positioning of residues with unknown attachment sites

The next example demonstrates a more advanced determination of structural details. Often the location of a fucose, as seen in the previous example, is one of the key questions for glyco-biologists. *GlycoWorkbench* incorporates in the annotation tool a feature that allows the automatic comparison of structure candidates arising from the placement of uncertain antennae in all possible positions within the structure. Figure 6 displays the positioning of a fucose in the bi-antennary N-glycan Hex$_3$HexNAc$_6$Fuc$_1$-PA. The decision where to locate the fucose residue correctly could already be given by looking at the "Stats" view. The coverage of the given intensity of the structure model with the fucose at the inner GlcNAc of the core is significantly superior to all the other possibilities. The complete annotation is then confirming this choice.

3.5. Annotation of spectra of persubstituted glycans showing evidence of ring fragmentation

In the previous examples all the structures were underivatized and only the fragments resulting from glycosidic bond cleavages were used to annotate the spectrum. In this further example the applicability of *GlycoWorkbench* to different types of experimental setups is demonstrated. Figure 7 shows the detailed annotation of a list of peaks selected from a spectrum of the permethylated oligosaccharide Lacto-N-fucopentaose. Cross-ring fragments can be extremely

useful in identifying the linkage positions of monosaccharides by MS without additional linkage analysis. The *in-silico* fragmentation tool is able to compute cross-ring fragments for all available monosaccharides and to use them for annotation of the mass spectrum as shown in the figure.

4. Concluding remarks

Determination of glycan structures from analysis of MS data is a major bottleneck in high-throughput glycomics projects, and robust solutions to this problem are of critical importance. However, the current status of tools to analyze glycan MS data shows that completely automated interpretation of generic mass spectrometric data is still unfeasible. *GlycoWorkbench* is a semi-automatic annotation tool developed by the EUROCarbDB initiative to assist the manual interpretation of MS data. *GlycoWorkbench* provides an integrated environment with an easy to use graphical interface that allows a sensible simplification of the determination of glycan sequences from mass spectrometric data.

The visual editor of glycan structures based on *GlycanBuilder*[22] enables a rapid assembly of structure models and their display in various symbolic notations. The annotation process allows the assignment of experimental peaks with a complete list of theoretical fragments by taking into account several types of experimental techniques. The annotation reports assist the determination of the correct structure by allowing the comparison of quality and coverage of the different assignments. The examples shown in section 3 demonstrate how the tool can provide a complete support to the routine interpretation of mass spectrometric data.

The possibility of importing structure candidates into *GlycoWorkbench* using several sequence encoding formats allow the user to integrate the tool with existing structure databases and with composition analysis tools such as *Glyco-Peakfinder*[11] to assist during the selection of potential candidates . Tight integration of the upcoming structure database from EUROCarbDB and of the

Glyco-Peakfinder tool into the *GlycoWorkbench* interface will enhance the tool with the capability of profiling glycan structures by mass value and will provide a complete workflow from raw data to completely annotated spectra.

GlycoWorkbench has been developed to offer a complete set of features that cover a broad spectrum of experimental MS techniques. The tool has been publicly available[23] from the very beginning as to fulfill the open access philosophy of EUROCarbDB. The sum of these factors has resulted in several laboratories already employing the *GlycoWorkbench* to assist their research. The experiences and feedback obtained from the users are of great importance for the constant development of the tool to further enhance its usability and flexibility. The tool is continuously updated and is designed to enable the addition of new features as pluggable components.

GlycoWorkbench has been developed for EUROCarbDB and as part of this initiative its components are being used to develop this database. With the progression of the database development and the collection of valuable data into it, the *GlycoWorkbench* will be connected to a precious source of expert knowledge that will be used to increase the level of automation in the annotation process. Information such as experimentally derived structures and previously assigned spectra could be directly applied to the annotation of new data, while other information such as biosynthetic and fragmentation pathways could be extracted from the data and used to build more intelligent features into the tool. With the addition of new components and the continuous development the tool is undergoing, *GlycoWorkbench* is projected to become a complete platform for analysis of glycomic MS data.

5. Acknowledgements

6. Availability

The software is freely available and can be downloaded from http://www.eurocarbdb.org/ms-tools. The use of the tool requires the installation of Java 5.0. Further information is provided in the download page.

7. Supporting information

Supporting Information Available: This material is available free at http://pubs.acs.org.

Figure S1: see Figure 3, fully assigned spectrum. For peaks with multiple possible assignments only one is displayed.

Figure S2: see Figure 4, fully assigned spectrum. For peaks with multiple possible assignments only one is displayed.

Figure S3: see Figure 7, fully assigned spectrum, enlarged version.

8. Tables

Table 1: List of available monosaccharides

Type	Symbol	Description	Symbol	Description
Deoxypentose	dPen	Deoxypentose	dRib	Deoxyribose
Pentose	Pen	Pentose	Ara	Arabinose
	Rib	Ribose	Xyl	Xylose
Deoxyhexose	dHex	Deoxyhexose		
	Rha	Rhamnose	Fuc	Fucose
	dTal	6-Deoxytalose	Qui	Quivonose
Hexose	Hex	Hexose	MeH	3-Methyl-hexose
	Glc	Glucose	Gal	Galactose
	Tal	Talose	Man	Mannose
	Fru	Fructose	All	Allose
Hexosamine	HexN	Hexosamine	GalN	Galactosamine
	GlcN	Glucosamine	ManN	Mannosamine
Acidic sugar	HexA	Hexuronic Acid		
	GlcA	Glucuronic Acid	GalA	Galacturonic Acid
	ManA	Mannuronic Acid	IdoA	Iduronic Acid
Unsaturated acidic sugar	4uHexA	4-unsaturated HexA		
	4uGlcA	4-unsaturated GlcA	4uGalA	4-unsaturated GalA
	4uManA	4-unsaturated ManA	4uIdoA	4-unsaturated IdoA
Deoxyheptose	dHept	Deoxyheptose	dHept	Deoxyheptose
Heptose	Hept	Heptose	Hept	Heptose
N-acetyl hexosamine	HexNAc	N-acetylhexosamine	GalNAc	N-acetylgalactosamine
	GlcNAc	N-acetylglucosamine	ManNAc	N-acetylmannosamine
Acidic sugar	MurNAc	Muramic acid	Neu	Neuraminic acid
	KDN	KDN	NeuAc	N-Acetyl Neuraminic acid
	KDO	KDO	NeuGc	N-glycolyl Neuraminic acid

Table 2: List of available reducing-end modifications

Symbol	Description
freeEnd	Free reducing end
redEnd	Reduced reducing end
PA	2-Aminopyridine
2AP	2-Aminopyridine
2AB	2-Aminobenzamide
AA	Anthranilic Acid
DAP	2,6-Diaminopyridine
4AB	4-Aminobenzamidine
DAPMAB	4-(N-[2,4-Diamino-6-pteridinylmethyl]amino)benzoic acid
AMC	7-Amino-4-methylcoumarin
6AQ	6-Aminoquinoline
2AAc	2-Aminoacridone
FMC	9-Fluorenylmethyl carbazate
DH	Dansylhydrazine

Table 3: List of available substituents

Symbol	Description
Me	Methyl
Ac	Acetate
NAc	N-Acetate
Pv	Pyruvate
P	Phosphate
S	Sulphate

9. References

1. Dell, A.; Morris, H. R., Glycoprotein structure determination by mass spectrometry. *Science* **2001,** 291, (5512), 2351-2356.
2. Spengler, B.; Kirsch, D.; Kaufmann, R.; Cotter, R. J., Metastable decay of peptides and proteins in matrix-assisted laser-desorption mass spectrometry. *Rapid Communications in Mass Spectrometry* **1991,** 5, (4), 198-202.
3. Weiskopf, A. S., Characterization of oligosaccharide composition and structure by quadrupole ion trap mass spectrometry. *Rapid communications in mass spectrometry* **1997,** 11, (14), 1493.
4. Hakansson, K.; Chalmers, M. J.; Quinn, J. P.; McFarland, M. A.; Hendrickson, C. L.; Marshall, A. G., Combined Electron Capture and Infrared Multiphoton Dissociation for Multistage MS/MS in a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer. *Analytical Chemistry* **2003,** 75, (13), 3256-3262.
5. Domon, B.; Costello, C. E., A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate journal* **1988,** 5, (4), 397.
6. Dell, A., F.A.B.-mass spectrometry of carbohydrates. *Advances in carbohydrate chemistry and biochemistry* **1987,** 45, 19-72.
7. Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of The American Society for Mass Spectrometry* **1994,** 5, (11), 976.
8. Joshi, H. J.; Harrison, M. J.; Schulz, B. L.; Cooper, C. A.; Packer, N. H.; Karlsson, N. G., Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* **2004,** 4, (6), 1650-1664.
9. Kameyama, A.; Kikuchi, N.; Nakaya, S.; Ito, H.; Sato, T.; Shikanai, T.; Takahashi, Y.; Takahashi, K.; Narimatsu, H., A Strategy for Identification of Oligosaccharide Structures Using Observational Multistage Mass Spectral Library. *Analytical Chemistry* **2005,** 77, (15), 4719-4725.
10. Cooper, C. A.; Gasteiger, E.; Packer, N. H., GlycoMod—A software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* **2001,** 1, 340–349.
11. Maass, K.; Ranzinger, R.; Geyer, H.; Von der Lieth, C. W.; Geyer, R., De Novo Composition Analysis of Glycoconjugates. *Proteomics* **2007,** in press.
12. Goldberg, D.; Sutton-Smith, M.; Paulson, J.; Dell, A., Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* **2005,** 5, (4), 865-875.
13. Gaucher, S. P.; Morrow, J.; Leary, J. A., STAT: A Saccharide Topology Analysis Tool Used in Combination with Tandem Mass Spectrometry. *Analytical Chemistry* **2000,** 72, (11), 2331-2336.
14. Lapadula, A. J.; Hatcher, P. J.; Hanneman, A. J.; Ashline, D. J.; Zhang, H.; Reinhold, V. N., Congruent Strategies for Carbohydrate Sequencing. 3. OSCAR: An Algorithm for Assigning Oligosaccharide Topology from MS^n Data. *Analytical Chemistry* **2005,** 77, (19), 6271-6279.
15. Ethier, M.; Saba, J. A.; Spearman, M.; Krokhin, O.; Butler, M.; Ens, W.; Standing, K. G.; Perreault, H., Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid communications in mass spectrometry* **2003,** 17, (24), 2713-20.
16. Tang, H.; Mechref, Y.; Novotny, M. V., Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* **2005,** 21, (suppl_1), i431-439.

17.   Design Studies Related to the Development of Distributed, Web-based European Carbohydrate Databases (EUROCarbDB). http://www.eurocarbdb.org/

18.   *Structural Medicine - The Importance of Glycomics for Health and Disease*; European Science Foundation: 2006.

19.   Aoki, K. F.; von der Lieth, C. W.; Raman, R.; York, W. S. *Urgent requirements for the development of informatics for glycomics and glycobiology*; National Institute of Health: 2007.

20.   Clerens, S.; Van den Ende, W.; Verhaert, P.; Geenen, L.; Arckens, L., Sweet Substitute: a software tool for in silico fragmentation of peptide-linked N-glycans. *Proteomics* **2004,** 4, (3), 629-32.

21.   Lohmann, K. K.; von der Lieth, C. W., GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Research* **2004,** 32, (Web Server issue), W261-266.

22.   Ceroni, A.; Dell, A.; Haslam, S. M., The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code for Biology and Medicine* **2007,** 2, (1), 3.

23.   EUROCarbDB   -   Tools   to   analyse   MS   spectra: GlycoWorkbench. http://www.eurocarbdb.org/applications/ms-tools

24.   The Consortium for Functional Glycomics nomenclature for representing glycan structures. http://glycomics.scripps.edu/CFGnomenclature.pdf

25.   Royle, L.; Dwek, R. A.; Rudd, P. M., Unit 12.6 Determining the structure of oligosaccharides N- and O-linked to glycoproteins. In *Current Protocols in Protein Science*, Coligan, J. E.; Dunn, B. M.; Speicher, D. W.; Wingfield, P. T., Eds. John Wiley and Sons: 2006.

26.   Bohne-Lang, A.; Lang, E.; Forster, T.; von der Lieth, C. W., LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydrate Research* **2001,** 336, (1), 1-11.

27.   Lutteke, T.; Bohne-Lang, A.; Loss, A.; Goetz, T.; Frank, M.; von der Lieth, C. W., GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* **2006,** 16, (5), 71R-81R.

28.   Ehud, B.; Yael, N.; Yaniv, A.; Asaf, H.; Ori, I.; Dotan, N.; Avinoam, D., A Novel Linear Code Nomenclature for Complex Carbohydrates. *Trends Glycoscience Glycotechnology* **2002,** 14, (77), 127-137.

29.   Functional Glycomics Gateway. http://www.functionalglycomics.org/fg/

30.   Herget, S.; Ranzinger, R.; von der Lieth, C. W. A sequence format and namespace for complex oligo- and polysaccharides. http://www.eurocarbdb.org/recommendations/encoding/

31.   ProteomeCommons   IO   Meta-Information. http://www.proteomecommons.org/current/531/

32.   Lochnit, G.; Geyer, R., Carbohydrate Structure Analysis of Batroxobin, a Thrombin-Like Serine Protease from Bothrops moojeni Venom. *European Journal of Biochemistry* **1995,** 228, 805-816.

33.   Lewandrowski, U.; Resemann, A.; Sickmann, A., Laser-Induced Dissociation/High-Energy Collision-Induced Dissociation Fragmentation Using MALDI-TOF/TOF-MS Instrumentation for the Analysis of Neutral and Acidic Oligosaccharides. *Analytical Chemistry* **2005,** 77, (10), 3274-3283.

34.   Dell, A., Mass spectrometry of carbohydrate-containing biopolymers. *Methods in enzymology* **1994,** 230, 108.

10. Figures
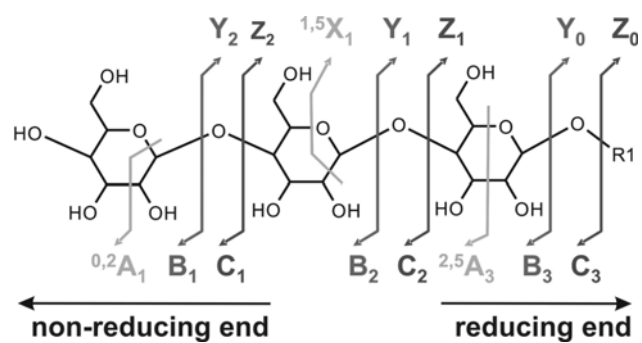
Figure 1: Nomenclature of fragments of carbohydrates as defined by Domon and Costello[5].
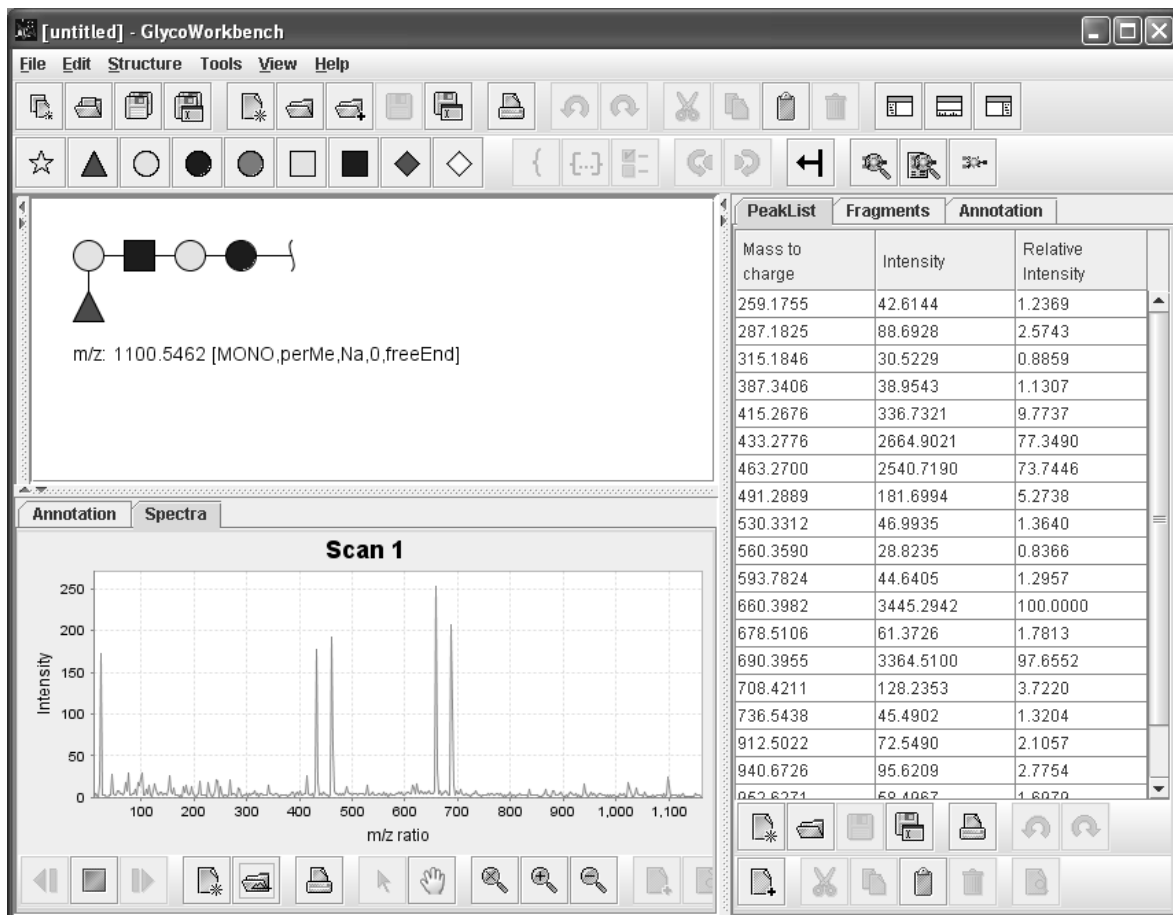
Figure 2: Graphical interface of the GlycoWorkbench tool. In this figure the main drawing canvas, the spectra panel and the peaklist panel are shown.
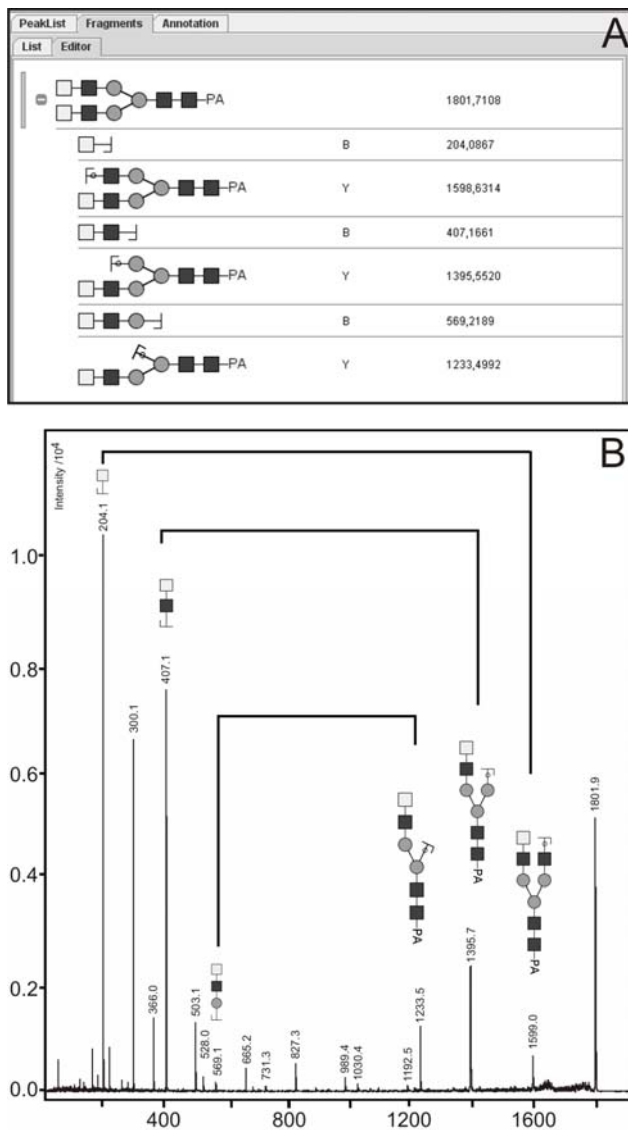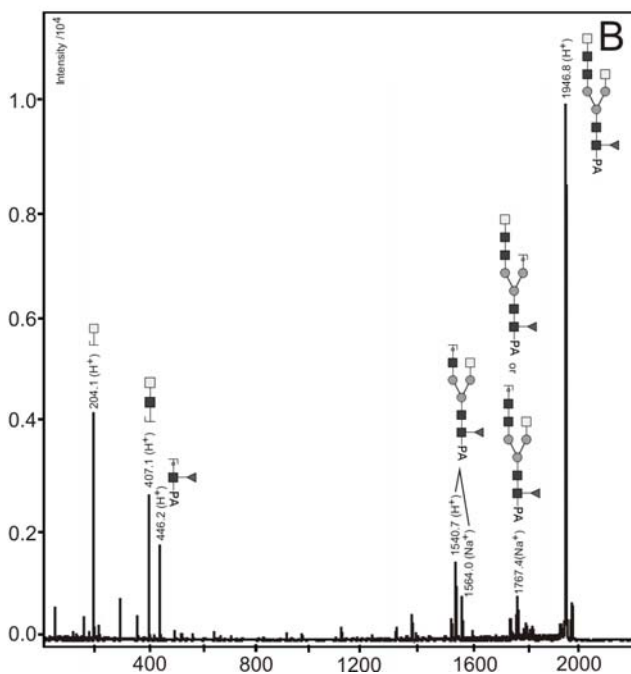
Figure 3: Example of manual interpretation of mass spectra. The fragment editor is used to find ion pairs resulting from single glycosidic bond cleavages (A). The ion pairs at *m/z* 204 and *m/z* 1598, *m/z* 407 and *m/z* 1395, *m/z* 569 and *m/z* 1233, describe the step by step degradation of one of the antennae of the N-glycan. Each pair has one peak representing a B-ion in the lower mass region and a corresponding peak representing a Y-ion in the higher mass region (B).

Figure 4: Automatic annotation of the peak list of a LID spectrum of a sodiated N-glycan $Hex_3HexNAc_6Fuc_1$-PA sorted by intensity of the mass signals. Only the most significant matches are shown to increase the clarity of the figure. The "Annotation Details" panel (A) gives a detailed overview of the annotated peaklist and allows a review of the annotation results. All assigned fragments are represented on the spectra (B).

Figure 5: Parallel annotation of the same peaklist with multiple structure candidates. A) "Stats" view of the matching of three candidates with the spectrum of a protonated N-glycan $Hex_5HexNAc_4Fuc_1$-PA. The structure candidates with the fucose at the antennae and the complex N-glycan have noticeable worse coverage than the hybrid structure model; B) more detailed view of the matches between *in-silico* fragmentation of all candidates and the mass list using the "Summary" panel. The signal at *m/z* 446 (FucHexNAc-PA) definitely shows a core fucosylation and the peak at *m/z* 407 (HexNAc_2) clearly proves the existence of only one complex antenna, since the complete structure comprises only 4 HexNAc in total.

Figure 6: Automatic positioning of a fucose in the biantennary N-glycan $Hex_3HexNAc_6Fuc_1$-PA. The decision where to locate the fucose residue correctly can be again directly judged from the "Stats" view. The coverage of the given intensity of the structure model with the fucose at the reducing end GlcNAc of the core is significantly superior to all the other possibilities.
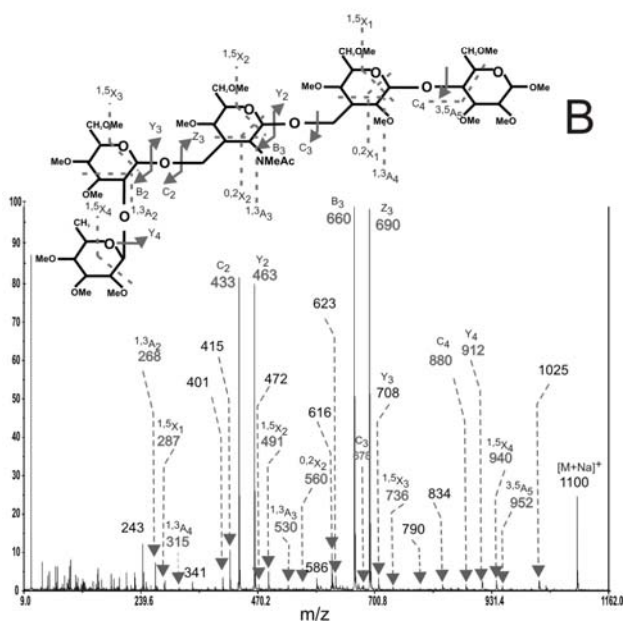
Figure 7: Detailed annotation of a list of peaks selected from a spectrum of permethylated Lacto-N-fucopentaose. Cross-ring fragments can be extremely useful in identifying the linkage positions of monosaccharides by MS without additional linkage analysis. The *in-silico* fragmentation tool is able to compute cross-ring fragments for all available monosaccharides and use them to annotate the mass spectrum as shown here.

11. Table of contents
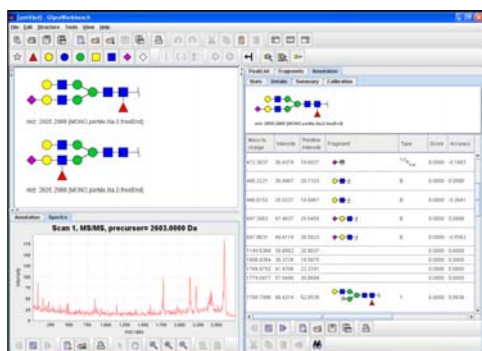
GlycoWorkbench is a software tool developed to assist the interpretation of MS data of glycans. The main task of *GlycoWorkbench* is to evaluate a set of structures proposed by the user by matching the corresponding list of fragment masses against the list of peaks derived from the spectrum. The tool provides an easy to use graphical interface and a broad set of features. The software can be downloaded from http://www.eurocarbdb.org/applications/ms-tools.