Beilstein-Institut

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

Glycoinformatics for Structural Glycomics

Stuart M. Haslam[1] **and David Goldberg**[2]

[1]Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, U.K.

[2]Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304, U.S.A.

E-Mail: [1]s.haslam@imperial.ac.uk, [2]goldberg@parc.com

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

Abstract

Ultra-high sensitivity mass spectrometric strategies incorporating MALDI-MS/MS and nano-electrospray(ES)-MS/MS enable very complex mixtures of glycoproteins and glycolipids from biological extracts of cells and tissues to be screened thereby revealing the types of glycans present and, importantly, providing clues to structures that are likely to be functionally important. However, in contrast to the genomic and proteomic areas, the glycosciences lack accessible, curated and comprehensive data collections that summarize the structure, characteristics, biological origin and potential function of glycans that have been experimentally verified and reported in the literature. This lack of glycan databases has been identified by glycobiologists as the single biggest hindrance to their research. Additionally, the sparseness of glycan databases hampers the realization of bioinformatics tools for the interpretation of experimental data and the automatic determination of the glycan structure, therefore limiting the possibility of large scale glycomics studies. The current status of the field and possible future developments are outlined.

Introduction

Glycans, both in the form of polysaccharides or glycoconjugates (bound to proteins and lipids), are the most abundant class of biomolecules and are increasingly being implicated in human health. Glycosylation is by far the most important post-translational modification in terms of the number of proteins modified and the diversity generated. Since glycoproteins, glycolipids and glycan-binding proteins (GBPs, also called lectins, which specifically recognize particular glycan epitopes) are frequently located on the cell's primary interface with the external environment, the cell surface, many biologically significant events can be attributed to glycan recognition. For this reason the rapidly expanding glycoscience field is being increasingly recognized as an important component of life science research. In contrast to the genomic and proteomic areas, the glycosciences lack accessible, curated and comprehensive data collections that summarize the structure, characteristics, biological origin and potential function of glycans that have been experimentally verified and reported in the literature. This lack of glycan databases has been identified by glycobiologists as the single biggest hindrance to their research. Additionally, the sparseness of glycan databases hampers the realization of bioinformatics tools for the interpretation of experimental data and the automatic determination of the glycan structure, therefore limiting the possibility of large scale glycomics studies. The complexity of the glycan structures and the variety of techniques that are used for their study, pose additional obstacles to the development of a single automated tool that could have the same impact on glycomics as Mascot and SEQUEST have had for proteomics.

Mass Spectrometry, Glycomics and Glycoproteomics

Rapidly-increasing developments in the field of mass spectrometry, particularly over the past twenty years, have led to the achievements of new milestones regarding sensitivity and determination of molecular weight, with mass accuracies of 0.01% of the total molecular weight of the sample now routinely being attained. These features have extended the capabilities of mass spectrometers to the study of large biopolymers such as glycoproteins, which can be several hundreds of thousands of Daltons in mass. The ability to analyse minute quantities of sample within a complex mixture, together with enhanced sensitivity and accurate mass analysis has led to mass spectrometry becoming a method of choice for the analysis of carbohydrates. The objectives of a glycomics experiment are to define the complete complement of carbohydrate structures in a system. Depending on the experimental set up the system could be purified glycoproteins, SDS-PAGE gel bands, cells, tissues or organs biopsies or a complete organism such as *Caenorhabditis elegans*. The best strategy for such experiments is MALDI-TOF mass spectrometry analysis of permethylated derivatised glycan samples [reviewed in 1]. Such methodologies have been utilized by The Analytical Glycotechnology Core C of The Consortium for Functional Glycomics (CFG; http://www.functionalglycomics.org), which was established in 2001 with 'glue' grant funding from the National Institute of General Medical Sciences. The overarching goal of the

CFG is to define the paradigms by which protein–carbohydrate interactions mediate cell communication. The objectives of a glycoproteomics experiment are to define glycan populations at individual glycosylation sites in an individual glycoprotein. This is more complex and resource intensive, both in terms of equipment and man hours, than a glycomics experiment and is achieved by MALDI-TOF and/or ES-mass spectrometry analysis of glycopeptides. It can be greatly facilitated by prior glycomic analysis [reviewed in 2].

Glycoinformatics

The application of glycomic and glycoproteomic methodologies outlined above has lead to the generation of large volumes of carbohydrate structural data. The manual interpretation of such large data sets is time consuming and requires expert knowledge. This bottleneck in the process has causes a considerable slowing of progress. Compared to the field of proteomics the automated interpretation of MS spectra of glycans is still an evolving field. In the following sections the most powerful glycoinformatic tools for MS glycan data are described.

Cartoonist

Cartoonist is a family of programs that annotate mass spectra of glycopeptides or detached glycans. The annotation is done by labelling peaks of the spectrum with *Cartoons*, which are graphical representations of glycans. Cartoons are widely used in the glycobiology community, because they give a quick sense of a glycan's structure. They are especially useful when annotating a spectrum containing many different glycans – it is much easier to see trends and patterns in a page of graphical cartoons than in a page of chemical formulas.

The original version of Cartoonist was described in [3]. In this article we summarize some of the improvements and new applications since that original publication: specifically, automatic cartoon libraries, separate java program for displaying annotations, searching the CFG (Consortium for Functional Glycomics) database and principled setting of parameters. Although there are versions of Cartoonist for N-glycopeptides and O-glycans we will only discuss the version for detached N-glycans.

Automatic Cartoon Libraries

One component of Cartoonist is a large library of plausible cartoons. Since the publication of [3], we have automated library construction. The user only need specify a set of antennae. The default set of eleven is shown in Figure 1 on the left. The cartoon library is generated in two steps. In the first step, the antennae are placed into "slots" of a template N-glycan, to generate a large set of base cartoons (we also add a small number of cartoons to the base set which do not fit this pattern). In the second step, rules are applied to expand the base set to a much larger library of cartoons. Our current library-builder uses three rules and generates

145,010 cartoons from 15,256 base cartoons (with the default set of antennae). The rules are illustrated on the right hand side of Figure 1. The first rule takes a cartoon with a single fucose and generates a cartoon without the fucose. The second rule adds a bisecting GlcNAc to a cartoon. The third rule systematically substitutes one sialic acid for another. For example, if the base cartoon has two NeuAc monosaccharides, this rule will generate three additional cartoons: one with the first NeuAc substituted with a NeuGc, one with the second NeuAc substituted, and one with both NeuAcs substituted.

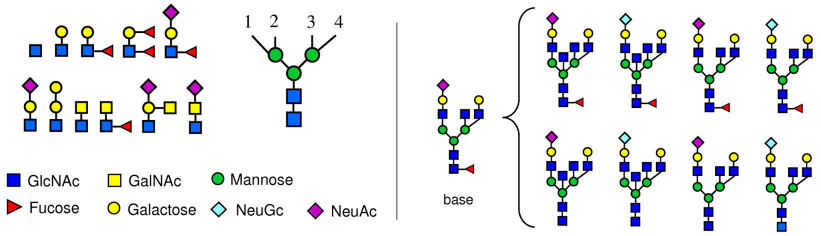

Figure 1. The standard Cartoon library is automatically generated using the 11 antennae shown on the left. Placing antennae into the slots numbered 1 – 4 gives the set of base cartoons. This set is further expanded using rules illustrated on the right.

Java Browser

The original version of Cartoonist was a monolithic program that computed annotations and produced a postscript file of the annotated spectrum which could be printed or displayed. But it was a static view of the spectrum. The current version has been modularized: the front end of Cartoonist produces a '.msa' file describing the annotations in human readable form. The back end is a Java browser, which reads the .msa file and displays it as an annotated spectrum. In addition to modularization, the advantage of this design is that the back end is a program that can pan and zoom through the spectrum and annotations. Figure 2 shows three different views of the browser on a spectrum of Human Monocytes from the CFG website. The first view shows the entire spectrum (Fig. 2A). There is not enough space to show all the annotations. The second view is zoomed in around Man-9 (2397 m/z), showing additional annotations (Fig. 2B). The third view illustrates one of the browser options: the cartoons can be magnified (or shrunk) (Fig. 2C).

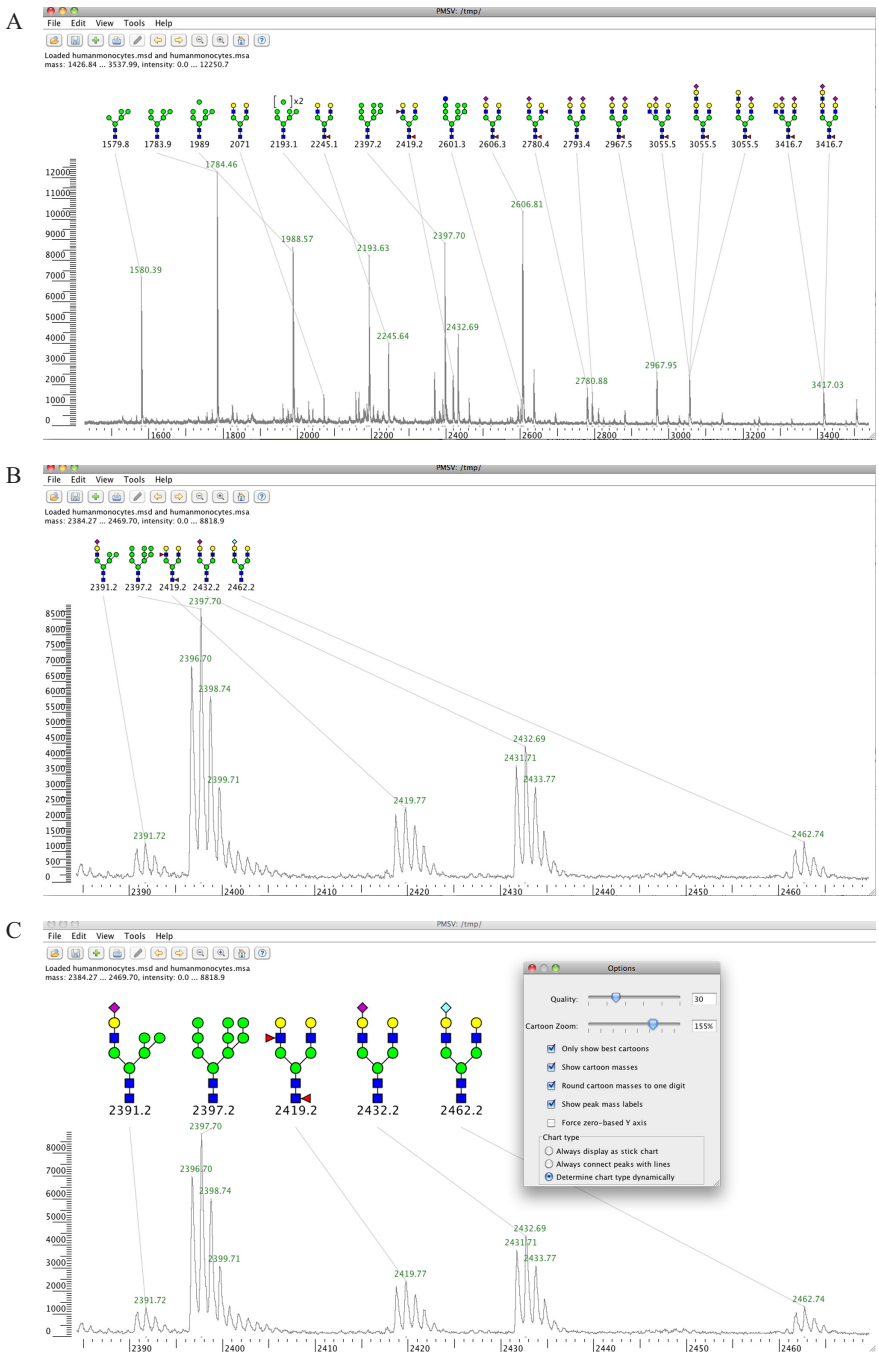

Figure 2. (A) The browser viewing a spectrum of N-glycans of Human Monocytes (from the CFG website). (B) The second view is zoomed in version of the first, the third view (C) has used the Options window to magnify the size of the cartoons.

Searching the CFG database

Cartoonist can be used not only to annotate spectra as they roll off of a mass spectrometer, but also to analyze existing databases of spectra. Motivated by the controversy about the existence of NeuGc in human cells and tissues [4] we used Cartoonist to search human samples in the CFG for NeuGc. Those spectra were annotated by human experts (without any NeuGc), but in this experiment we do not use that information, only the annotations made independently by Cartoonist. A systematic search found several matches, the most compelling of which were in Human Monocytes. Figure 3 shows the browser zoomed in on two possible NeuGc's in human monocytes. Each example has an additional peak with the correct m/z to be a NeuAc/NeuGc substitution, which is more evidence for the validity of the assignment to NeuGc.

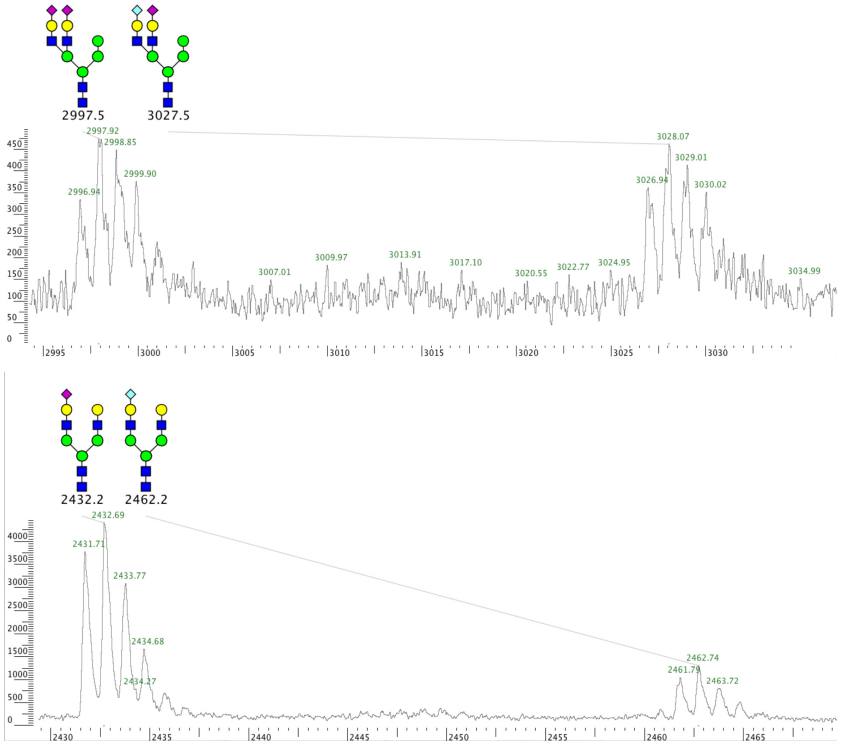

Figure 3. Peaks that may represent NeuGc in Human Monocytes. Principled Statistical Setting of Parameters

Cartoonist assigns a score to each annotation, which gives a rough sense of the probability that the annotation is correct. In the original version of Cartoonist, the score computation was rather *ad hoc*. The large set of human-expert annotations from the CFG web have been used to replace the original *ad hoc* setting of parameters in Cartoonist with statistically

sound settings. One of the factors that make up the score is how close the observed isotope envelope is to the theoretical envelope. Cartoonist computes a number Δ that measures the difference in shape between the ideal and observed isotope envelopes. Figure 4 shows a histogram of Δ values for high-confidence annotations in the CFG spectra. By fitting this to a curve, we can convert Δ to a probability, specifically the ''tail probability'', that is, the area under the curve to the right of Δ. The blue dots lie on the best fitting curve of the form e^{-ax}, the red dots on xe^{-ax}. Clearly the latter is a better fit: specifically, the red curve is $(4/\mu^2)\, x\, e^{-(2x/\mu)}$, a probability distribution with mean of μ. For this data, $\mu = 0.18$. Using this curve gives an explicit formula for converting the Δ value of an isotope envelope to a probability.

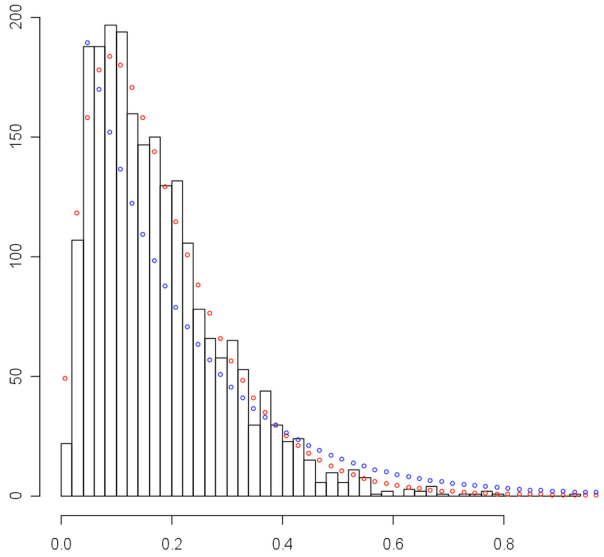

Figure 4. Histogram showing the distribution of Δ over a large set of isotope envelopes from the CFG spectra. The number Δ measures the discrepancy in shape between the observed and expected isotope envelopes. The red dots show the curve xe^{-ax} fitted to the histogram, the blue dots the curve e^{-ax}.

We also investigated whether a single Δ curve is sufficient, or whether we need to take into account how Δ varies with mass. To study the dependence on mass, we remove ''outliers'', the 10% of peaks with lowest intensity. For the remaining peaks the median mass is 2634. The mean of the Δ values is 0.17 both for peaks below the median and above the median. So it appears there is no gain to be had by having separate curves for different masses. However, as seen Figure 5 below, the high mass histogram is more ragged.

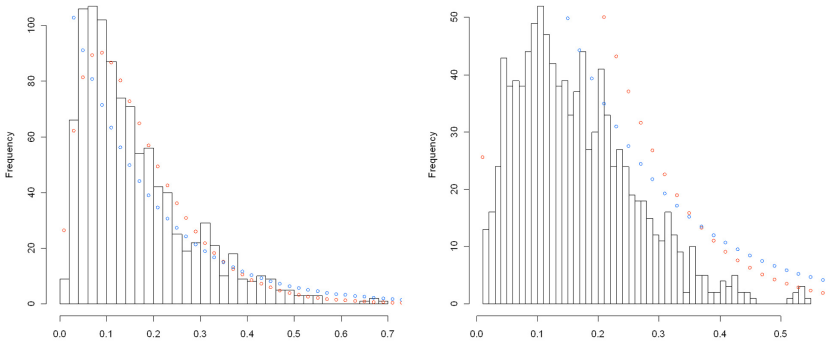

Figure 5. The histogram on left shows peaks with mass < 2600, on the right with mass ≥ 2600. For the left histogram, again the red curve is a better fit. Each histogram only considers peaks with rank < 557. In each case, $\mu = 0.18$. The red and blue curves are fit as in Figure 4.

Next we study deviation, which is the difference between the computed m/z and the observed (after recalibration) m/z. For judging the quality of a peak assignment, we would like to know the variance of the mass deviations, so that we can estimate how many "sigmas" the mass of an observed peak has strayed from the theoretical mass. That graph of this variance for the CFG spectra is shown in Figure 6.

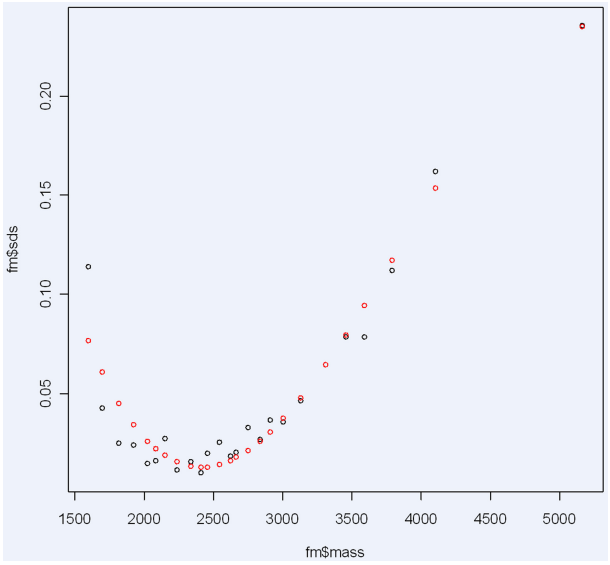

Figure 6. The histogram on left shows peaks with mass < 2600, on the right with mass ≥ 2600. For the left histogram, again the red curve is a better fit. Each histogram only considers peaks with rank < 557. In each case, $\mu = 0.18$. The red and blue curves are fit as in Figure 4.

The plot has grouped the peaks into 25 bins (of approximately 90 peaks each), and computed the standard deviation of the mass deviation for each bin. The black dots are the actual SD's, the red dots are the SD's fit to the cubic,

$$0.7443 - 0.0007138 \times mass + 0.0000002149 \times mass^2 - 0.00000000001855 \times mass^3$$

Cartoonist uses the data in Figure 4 and Figure 6 to compute a 'score' measuring the quality of an annotation. It makes the simplifying assumption that the deviation and isotope fit (Δ) are independent, and so their probabilities can be multiplied. The probability associated with Δ is the tail probability of the curve shown in Figure 4. The probability associated with the deviance is computed a bit differently. We assume the deviance has a Gaussian distribution, with a standard deviation σ given by Figure 6. So the probability associated with the deviation is again a tail probability, but this time the right hand tail of a Gaussian curve with standard deviation σ. The total probability is the product of these two numbers, and the final score is the minus logarithm of the probability.

EUROCarbDB

An Infrastructure Design Study (EUROCarbDB) was started in the European Union FP6 framework to lay out the foundations of a carbohydrate structural database and associated informatics tools. The partners of the EUROCarbDB design study have successfully established formats and standards for carbohydrate data exchange, developed tools to assist the interpretation of carbohydrate experimental data and designed a database architecture that can be used to store and retrieve glycan data. EUROCarbDB (http://www.ebi.ac.uk/eurocarb/) comprises an open-access relational database of glycan (carbohydrate) structures and primary research data, accessed via a web-based, browse/search/contribution interface. The database component is complemented by a suite of associated glyco-informatics tools, designed to aid in the elucidation and submission of glycan structures when used in conjunction with contemporary carbohydrate research workflows. Fundamentally, the database can be sub-divided into core and experimental components. The core database is comprised of four modules, which are interconnected. These four core modules are: sequence and structure, biological context, evidence, and reference, each of which is subsequently described. The relationship between each of these major modules is many-to-many; for example, a specific carbohydrate sequence may be found in (i.e. linked to) multiple species/tissues, each of which may have several contributions of experimental evidence, such as MS, HPLC or NMR data or a single mass spectrum may be itself evidence for multiple carbohydrate sequences.

Two software tools, ''Glyco-Peakfinder'' and ''GlycoWorkbench'', have also been developed to assist in MS-carbohydrate structural data assignments and both freely provided by the EUROCarbDB initiative. Both tools can be used independently as they focus on different stages of the interpretation process of MS data. However, they can interact in a way that

allows a complete and smooth workflow from raw data to a completely assigned spectrum and, furthermore, an easy to use way to collect all mandatory information for a new database entry. ''Glyco-Peakfinder'' is a web-application developed for *de novo* composition analysis of glycoconjugates. It is designed to ease the time intensive manual annotation of all kinds of MS spectra. Glycan profiles can be analysed as well as fragment spectra. ''Glyco-Peakfinder'' assigns all types of fragmentations including monosaccharide cross-ring cleavages (A-, B-, C-, X-, Y-, Z- fragments, gain and loss of small molecules). The tool provides full user control to handle modified glycans, including modifications at the reducing end, or of the whole structure. The option to calculate multiply-charged ions increases the range of application to mass spectrometric techniques other than just MALDI. Although the derived information for each entered m/z value is completely independent from the results of neighbouring peaks, a cross-linking of the results for several peaks provides additional sequence information. In addition the proposed glycan compositions can be used to search open access databases to assess if such a composition has previously been reported [5].

The further annotation process of MS/MS and MSn spectra can be performed by using the ''GlycoWorkbench'' suite of software tools. The graphical interface of ''GlycoWorkbench'' provides an environment in which structure models can be rapidly assembled, automatically matched with MS/MS and MSn data, and compared to assess the best candidate assignment. The main component of ''GlycoWorkbench'' is ''GlycanBuilder'', a flexible visual editor especially designed for a user-friendly input of glycan structures [6]. Glycans often exhibit tree-like non-linear structures, and their constituents exhibit great diversity. Because of the tree-like structure, the input of a glycan sequence is not as straightforward as writing a sequence of characters, as for DNA, RNA and peptide sequences. The lack of a suitable user-friendly graphical interface to input complex carbohydrate structures in a computer readable format has long been a severe deficiency in the practical application of glyco-related databases. Additionally, numerous alternative notations are commonly adopted to graphically represent glycan structures. After the desired input structures have been defined with ''GlycanBuilder'', the remaining components of ''GlycoWorkbench'' can be used to derive their fragments, compute the fragment masses, build a peak-list and annotate it. The computation of fragments and their masses from the intact structure is a central step for the annotation of MS/MS and MSn spectra. The ''fragmentation tool'' creates all topologically possible fragmentations of the precursor molecular ion, applying both multiple glycosidic cleavages and cross-ring fragmentations. The fragments are computed by recursively traversing the tree structure of the glycan and applying all the possible cleavages at each position. Fragmented structures are then subjected to the same process to produce multiple cleavages. For a given glycan fragment, the m/z ratio can be calculated both for native and derivatised structures (per-methylated/per-acetylated) taking into account several types and quantities of ion adducts. A visual editor of glycan fragments is also available, 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. The next step in the annotation process is the assignment of possible fragments to each m/z-value in a given peak list.

In ''GlycoWorkbench'' a peak list can either be loaded from a tab-separated text file, thus allowing for import from peak-picking software, or it can be created by typing mass and intensity values directly into the application. Once the peak-list is ready, the fragment m/z values from the *in silico* fragmentation are matched with a given accuracy to each peak in the list. The annotated peak-list can be displayed using various panels that show its different aspects. Each panel is based around a spreadsheet-like table view, whose cell values can be sorted by each column, and can be copied into spreadsheet applications. The annotated peak list can be stored in a specific XML format for later consultation or export for example into the EUROCarbDB database [7].

Acknowledgments

Thanks to Shane Ahern for his work in implementing the Java browser, Alessio Ceroni, Kai Maass and René Ranzinger for development of Glyco-Peakfinder and GlycoWorkbench. Funding: NIGMS (NIH Grant R01GM074128 to D.G.); the glycan analyses were performed by the Analytical Glycotechnology Core of the Consortium for Functional Glycomics (NIGMS GM62116 and the NCRR); the Biotechnology and Biological Sciences Research Council (BBSRC) Grant Nos. BBF0083091 and B19088 and the sixth European Union Research Framework Programme (EUROCarbDB RIDS Contract No. 011952).

References

[1] North, S.J., Hitchen, P.G., Haslam, S.M., Dell, A. (2009) Mass spectrometry in the analysis of N-linked and O-linked glycans. *Curr. Opin. Struct. Biol*. **19**:498 – 506. doi: http://dx.doi.org/10.1016/j.sbi.2009.05.005.

[2] Tissot, B., North, S.J., Ceroni, A., Pang, P.C., Panico, M., Rosati, F., Capone, A., Haslam, S.M., Dell, A., Morris, H.R. (2009) Glycoproteomics: past, present and future. *FEBS Lett.* **583**:1728 – 1735. doi: http://dx.doi.org/10.1016/j.febslet.2009.03.049.

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

[4] Tangvoranuntakul, P., Gagneux, P., Diaz, S., Bardor, M., Varki, N., Varki, A., Muchmore, E. (2003) Human uptake and incorporation of an immunogenic nonhuman dietary sialic acid. *Proc. Natl. Acad. Sci. U.S.A.* **100**:12045 – 12050. doi: http://dx.doi.org/10.1073/pnas.2131556100.

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

[6] Ceroni, A., Dell, A., Haslam, S.M. (2007) The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol. Med*. **7**:2 – 3.
doi: http://dx.doi.org/10.1186/1751-0473-2-3.

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