Speeding up the High Throughput Searches for Glycan Analysis

Ningombam Sanjib Meitei[1], Arun Apte[2], Rupanjan Goswami[1], Julian Saba[3]
[1]PREMIER Biosoft, Indore, India; [2]PREMIER Biosoft, Palo Alto, U.S; [3]Thermo Fisher Scientific, San Jose, CA
Corresponding author e-mail: sanjib@premierbiosoft.com

Introduction

A typical LC-MS glycomic data processing includes peak detection, deconvolution of peaks of co-eluting isomers, molecular feature extraction, and selection of correct precursor *m/z* values from the isotope cluster for MS/MS data analysis. The currently existing bioinformatics tools do not adequately address the challenges of such workflows pertaining to time and accuracy of the identified glycans. Therefore, we have developed a software tool to streamline this process.

Methods

Sample Preparation
N-Linked glycans were released from glycoproteins (Bovine Fetuin) with PNGase F enzyme (New England BioLabs). The released glycans were labeled with 2-aminobenzamide (2AB) with slight modification from the reported procedure of Bigge et. al., [1] Prior to analysis, samples were dissolved in 100 µL D.I. water in a 250 µL auto sampler vial.

Liquid Chromatography
All glycans were separated on a Thermo Scientific™ GlycanPac™ AXR-1 (1.9 µm, 2.1 × 150 mm) column [2] by a Thermo Scientific™ Dionex™ Ultimate™ 3000 UHPLC instrument with either a fluorescence or an MS detector.

Mass Spectrometry
MS analysis was performed using a Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer in negative ion mode. LC-MS^2 experiments were conducted for structural elucidation.

Data Analysis
SimGlycan® 5.42 software (PREMIER Biosoft) was used for LC-MS and MS/MS data analysis. Thermo Scientific™ Xcalibur™ software is also used to visualize raw LC-MS data. The time taken to complete LC-MS data pre-processing is compared between SimGlycan v.5.42 and MZmine v.2.2 [3].

Results and Discussion
Challenges in LC-MS experiments based glycomic data analysis (especially high resolution instruments such as Q Exactive and Orbitrap Fusion mass spectrometers)

1. Large volume of data with high number of MS/MS scans (typical size of a .raw file >= 500 MB)
2. Convoluted LC peaks: Many glycan isomers elute under single LC peak.
3. Incorrect precursor *m/z* for MS/MS scans: Unlike isotope cluster for peptides and other small molecule metabolites, glycan isotopic clusters do not necessarily have base peak at monoisotopic *m/z* of a compound. Thus, many of the MS/MS scans in data independent acquisition (DIA) settings are acquired for higher isotopic peaks.
4. Lack of databases to store curated glycans and retention times to support LC-MS/MS based glycan structural identification in high throughput manner.
5. Lack of standardized report formats: In proteomics, popular database search tools namely Sequest, Proteome Discoverer (Thermo Fisher Scientific) and Mascot (Matrix Science) exports results in .sqt, .msf, and .dat formats which can be used for further downstream data analysis.

The latest version of SimGlycan software v. 5.42 is designed to address the above mentioned challenges.

Improved LC-MS data preprocessing

Figure 1 shows the total ion chromatogram (TIC) of the LC-MS and MS/MS data. The .raw file has the following properties: Size of the RAW file: 865 Megabytes, # of MS and MS/MS scans: 12845, Retention Time Range: 0-102 minutes, MS1 Mass Range: 400-2000, MS2 Mass Range: 200-3200.

The raw data file is processed using both SimGlycan v. 5.42 and MZmine v.2.2 on a desktop computer (Dell) with the following properties:

Operating System: Windows 10; *Processor:* Intel(R) Core(TM) i3-4130 CPU @ 3.40 GHz 3.40 GHz;
Installed memory (RAM): 8.00 GB; *System Type:* 64-bit Operating System.

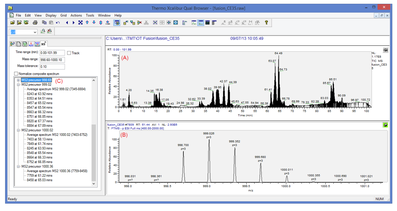

Figure 1: Typical Thermo Xcalibur software window showing (A) TIC, (B) Isotopic cluster of a glycan with base peak at M+1, and (C) MS/MS scans acquired at M+1, M+2 and higher isotopic peaks of a compound.

Table 1 shows the steps and corresponding parameters used for peak detection and deconvolution:

Sl. no	Feature	Instruction	Parameters	Values	Time
1	Mass Detection	Raw data methods> Peak detection> Mass detection	Raw data files	Fusion_CE35.raw	1 minute
			Scans> Set Filters		
			Scan number	1 – 12,845	
			Retention time	0.00 – 102.00	
			MS level	1	
			Scan Definition	--	
			Polarity	Negative	
			Spectrum type	Any	
			Mass detector	Exact Mass	
			Mass list name	masses	
2	Chromatogram Builder	Raw data methods> Peak detection> Chromatogram builder	Raw data files	fusion_CE35	35 minutes
			Scans	Retrieves selection from Mass Detection	
			Mass list	masses	
			Min time span (min)	0.30	
			Min height	10000	
			m/z tolerance	0.05 (*m/z*), 20 (ppm)	
			Suffix	chromatograms	
3	Chromatogram Smoothing	Peak list methods> Peak Detection> Smoothing	Peak lists	fusion_CE35.raw chromatogram	10 seconds
			Filename suffix	smoothed	
			Filter width	15	
			Remove original peak list	Not checked	
4	Chromatogram Deconvolution	Peak list methods> Peak Detection> Chromatogram deconvolution	Peak lists	fusion_CE35.raw chromatogram smoothed	12 seconds
			Suffix	deconvoluted	
			Algorithm	Local minimum search	
			Chromatographic threshold	0.1%	
			Search minimum in RT range (min)	0.3	
			Minimum relative height	1%	
			Minimum absolute height	10000	
			Min ratio of peak top/edge	0.1	
			Peak duration range (min)	0.00 – 10 mins	
			Remove original peak list	Not checked	

Table 1: The four steps of MZmine, corresponding parameters settings, and time taken to complete each step. The total time taken to complete all the four steps is 36 minutes and 22 seconds.

Figure 2 shows the typical graphical user interfaces (GUI) wherein users can enter the peak detection and picking parameters.

SimGlycan software completed the data processing task in 2 minutes. The significant improvement in time is tested and verified using multiple files acquired using both Q Exactive and Orbitrap Fusion mass spectrometers (Data not shown).

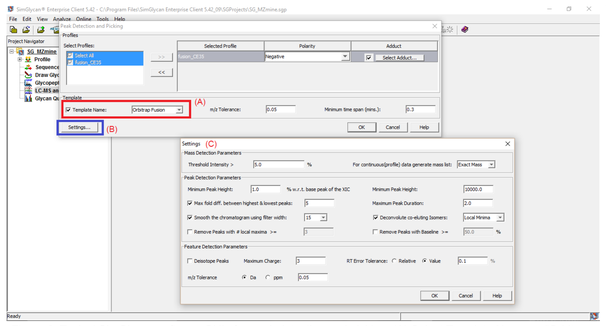

Figure 2: Typical SimGlycan software GUIs for peak detection and picking. (A) Select Template Name as “Orbitrap Fusion” and Click “OK” to process the data, (B) Users can click “Settings” to edit the peak detection and picking parameters, and (C) A typical GUI wherein users can change the peak detection and picking parameters.

Table 3 shows comparative statistics of the results from both the tools. Figure 3 shows the number of unique and common LC-peaks detected by SimGlycan and MZmine tools.

Feature	SimGlycan v. 5.42	MZmine v. 2.20
Time to complete Peak Detection	2 minutes	36 minutes & 22 seconds
# of LC-Peaks detected	2861	13428
# of unique compounds after molecular feature finding	1909	NA
Cluster MS/MS scans into LC-detected compounds	Yes	NA
Pick M+0 peak *m/z* as the precursor *m/z* of the MS/MS scans acquired at M+1 and higher isotopes	Yes	NA

Table 3: Comparison between SimGlycan v. 5.42 vs MZmine v. 2.2; NA: Functionality Not Available

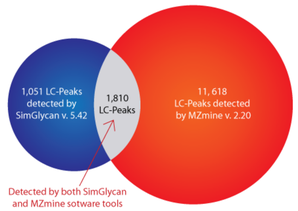

Figure 3: Venn diagram depiction of unique vs common LC peaks detected by both the software tools.

Custom LC-MS/MS Glycan Templates

SimGlycan software provides web-based modules that facilitate users to store glycans along with retention time. Users can add glycans into database by importing corresponding KEGG Chemical Function (KCF) format file in batch mode (Figure 4). In case KCF file is not readily available, users can draw the structure in SimGlycan software using “Draw” module and save the corresponding KCF data (Figure 5).

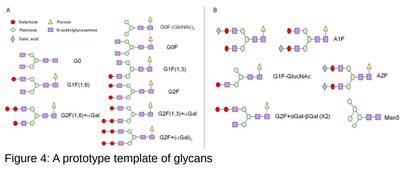

Figure 4: A prototype template of glycans

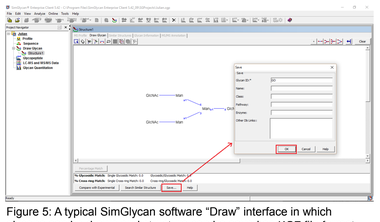

Figure 5: A typical SimGlycan software “Draw” interface in which glycans can be drawn and structure can be saved as KCF file format.

Additional functionality such as browse glycans by searching from taxonomy or structure search, move/copy data from a database to another database, edit information of existing glycans etc. are allowed. Figure 6 shows a typical SimGlycan web browser showing a list of searched glycans. In order to store retention times corresponding to glycans, just click “Add” (Figure 7) and enter (multiple) detailed information from LC-experiment/s (Figure 8).

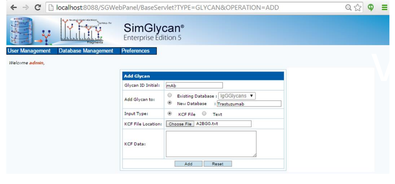

Figure 6: SimGlycan software interface to add glycans into an existing or new database

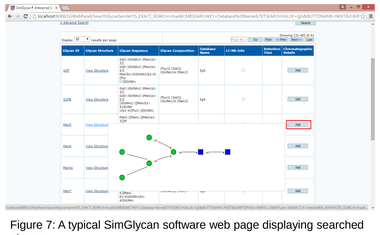

Figure 7: A typical SimGlycan software web page displaying searched glycans

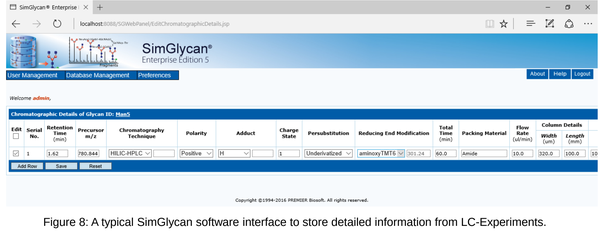

Figure 8: A typical SimGlycan software interface to store detailed information from LC-Experiments.

Once the database is constructed, users can restrict database search to custom database/s that contains curated data, thereby increasing the confidence of identification (Figure 9).

Glycan Identification using MS/MS Data

Figure 9 shows a typical SimGlycan software interface showing Search Parameter dialog with (A) the list of MS/MS scans with corrected precursor *m/z* values, (B) the filters such as Class, Sub-Class, Biological Source, etc., (C) the fragmentation patterns specific to the experimental settings, and (D) the curated database containing both glycan structures and retention times when users specify retention times as the initial database search predicate. After performing MS/MS database search, identified glycans are listed in the results pane. Identified glycan structures can be sorted based on retention time, intensity and other parameters. Structures, fragments and other information can be viewed in a single workspace (Figure 10, top left). The MS/MS spectra will be automatically annotated with the fragment ions of identified structures using three modes of annotation namely Domon and Costello fragmentation nomenclature, successive loss of monosaccharide residues and cartoons showing moveable fragment structures (Figure 10).

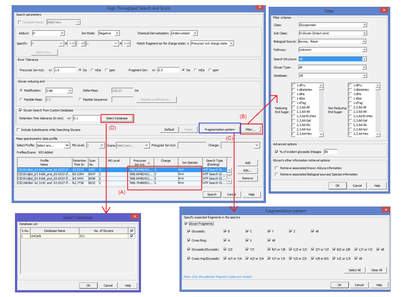

Figure 9: A typical SimGlycan software interface showing Search Parameters dialog. The list of MS/MS scans with corrected precursor *m/z* values is shown in the dialog.

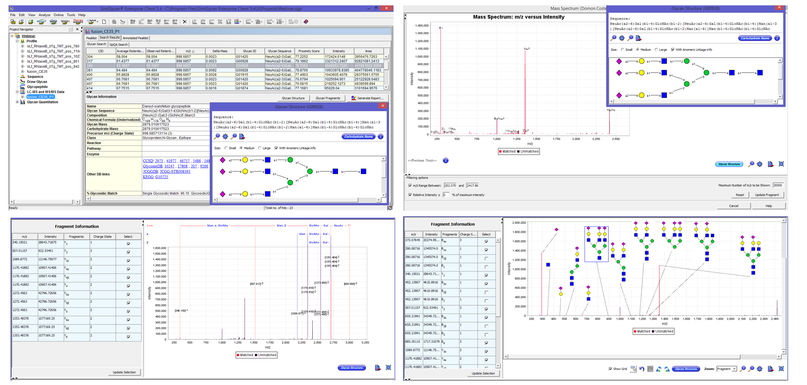

Figure 10: SimGlycan software search results pane (top left); MS/MS peak annotation with Domon and Costello fragment nomenclature (top right); MS/MS peak annotation showing successive loss of monosaccharide residues (bottom right); MS/MS peak annotation with cartoons (bottom right).

Portable Reports

Spreadsheet based reports facilitate easy reviewing of results for further verification, downstream analysis and dynamic information sharing. One major challenge with spreadsheet based report format is to save glycan structures into spreadsheet cells so that information such as retention time, precursor *m/z*, glycan ID etc. can be processed using spreadsheet operations to further organize the results. SimGlycan software generates report in MS excel file wherein glycan structures are also exported along with other structure specific information (Figure 11).

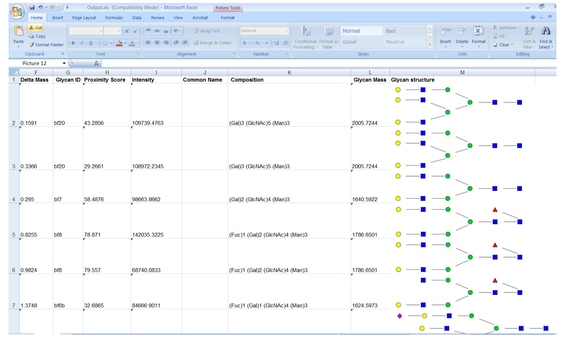

Figure 11: Typical view of MS excel file containing results outputted by SimGlycan software after performing MS/MS database search.

Conclusion

- The GlycanPac AXR-1 column separates glycans based on charge, size, polarity and isomeric structure, thereby providing a greater number of resolved peaks compared to commercial amide HILIC columns for 2AB-labeled bovine fetuin *N*-linked glycans.
- Faster Orbitrap enables higher scan rates at higher resolution. This translates to increased sensitivity and better quality MS/MS data for both abundant and low abundance glycans.
- SimGlycan® 5.42 software provides informatics support for LC-MS and MS/MS data analysis by enabling users to create LC-MS glycan templates.
- 10,000 MS/MS scans can be subjected to SimGlycan database search in batch mode for structural identification of glycans.
- Results including glycan structures can be exported into MS excel file, thereby facilitating easy review of results as well as dynamic sharing of information for further post identification data analysis.

Reference

1. Bigge, J. C. et al., Anal. Biochem. 1995, 230, 229-238.
2. Thermo Scientific GlycanPac AXR-1 Product Specification: http://www.dionex.com/en-us/webdocs/114170-PS-GlycanPac-AXH1-Column-PS20695_E.pdf
3. Pluskal, T, et al., BMC bioinformatics 11.1 (2010): 1.
4. Apte A., Meitei N. S. (2010) In Functional Glycomics, Humana Press; p. 269-81.
5. Chester, T.L., Anal. Chem. 2013, 85 (2), 579-589.
6. Ruhaak, L.R. et al., Anal. Bioanal. Chem. 2010, 397, 3457-3481.
7. Application Note 20786: Structural Analysis of Labeled N-Glycans from Proteins by LC-MS/MS Separated Using a Novel Mixed-Mode Stationary Phase.
8. Application Note 20827: Structural Analysis of Native N-Glycans Released from Proteins Using a Novel Mixed-Mode Column and a Hybrid Quadrupole-Orbitrap Mass Spectrometer.