Published in final edited form as:

Methods Mol Biol. 2013 ; 951: . doi:10.1007/978-1-62703-146-2_18.

Software tools for glycan profiling

Chuan-Yih Yu, Anoop Mayampurath, and Haixu Tang*

School of Informatics & Computing, Indiana University, Bloomington, IN USA 4708

1. Introduction

Glycosylation is a common post-translational modification that affects the protein function through the attachment of glycans. Alterations of protein glycosylation are indicative of diseases [1–3] and may occur through changes of the glycans (alterations in monosaccharide composition, glycan structure or linkage), aberrant glycosylation and dynamics of microheterogeneities. The aim of mass spectrometry based glycomics is to detect these changes through glycan profiling by first characterizing the glycans and then looking for changes across conditions. Software tools in glycomics aim to first detect the glycans from MS platforms such as MALDI-TOF and then annotating them. Annotation can be done at three levels – composition level, sequence level (through glycan *cartoons*) and sequence plus linkage level as shown in Table 1. Tools for platforms such as MALDI typically allow annotation through composition and cartoons. Usually, a single spectrum that contains peaks indicating the presence of digested glycans from glycopeptide or glycoprotein is used for detection and annotation of all possible glycan candidates within the sample. The candidate glycans can be detected through two different ways. One is through database searching using mass from a curated glycan database and the other is through *de novo* sequencing algorithms. These two types of methods have their own advantages and disadvantages. Database searching is commonly used for MALDI-based glycomics and provides results through a fast and precise search of the spectra, but cannot find any novel glycans that are not collected in the database. *De novo* sequencing methods typically involve fragmentation spectra and although the algorithms are slow and prone to error, novel glycans can be discovered. There are many tools that utilize both approaches for glycan detection and annotation. In this chapter, we describe three software tools that are used in N-linked glycomics.

2. Software

In this section, we will briefly introduce the methods and usage for the software tools. The software is listed in alphabetic order.

I. Cartoonist

Cartoonist is an automated N-glycan profiling tool that can be used to annotate the spectrum. It begins with an archetype set of N-glycan, which it then expands using sets of per-defined rules based on synthetic glycosylation pathways. By using three rules and an initial set of 300 archetypes, a total of 2,800 N-glycan candidates are derived.

Cartoonist assigns potential N-glycan to a peak based on mass, and uses the top 15 intense peaks to calibrate the result. The calibration simply takes the mass difference between

*Corresponding author: Haixu Tang, Informatics 225, Informatics Building, 901 E. 10th St. Bloomington, IN 47408-3912, hatang@indiana.edu, Phone: (812)-856-1859, Fax: (812)-856-4764.

predicted and observed mass value, and uses those pairs to fit with a linear model. Then this model is used to reevaluate the glycan assignments within each spectrum.

Cartoonist is integrated into the PARC Mass Spectrum Viewer [5] wrapped as an executable jar file. Users can directly execute or use a java command. These two execution methods will exhibit different behaviors in memory consumption and running time. Using the java command line (java -jar MassSpecViewer.jar) will consume more memory than direct execution of the jar file, but the command line method will run faster. We ran a small sample data and list the performance both on running time and memory usage in Table 2.

Usage steps

1. File -> Open, select msd format
(Note: although PARC Mass Spectrum Viewer supports many formats of mass spectra, but only the msd format can be used for Cartoonist annotation)
2. Click “Yes” in “Download cartoons” pop-up window.
3. The result will show in the window

Website: http://bio.parc.com/mass_spec

Supported spectrum format: msd

II. GlycoWorkbench [2]

GlycoWorkbench is a suite of programs for glycan profiling and interpretation. It not only supports various data formats, but also provides basic spectrum processing tools. It also provides a user-friendly interface to draw glycan structures which can then be used to annotate the mass spectrum. Additionally, a simulated fragmentation mechanism also allows to view putative fragment peaks of the glycan in the spectrum.

In this manner, GlycoWorkbench can deal with both MS and tandem MS data. Users can load tandem MS data to obtain glycan structure and linkage information. Here, we focus on the glycan profiling functionality of the software from a single MS spectrum prospect. Users can either draw a specific structure or search certain glycans in public glycan database via GlycoWorkbench. It supports four different formats of public databases (CFG [4], Carbbank [6], GlycomeDB [7] and Glycosciences [8]). The software has excellent interpretation of glycan structures, but consumes a lot of memory for profiling a whole spectrum.

Usage steps

Search all peaks

1. Load mass spectra file(in the formats of plain text, xml, mzXML, mzData or t2d)
2. Click on the Profiler button in the Tools tab and click on “Annotate peaks with structures from database”
3. Choose one or multiple databases, derivatization and reducing end.
4. Choice the fragment options in the pop-up window and click OK.
5. The result will be shown in the Search panel

Website: <http://www.glycoworkbench.org/>

Supported mass spectrum data format: Plain text (Peak list), xml, mzData, mzXML, and t2d

III. Multi N-Glycan

Multi N-Glycan uses pre-derived N-glycans as candidates that are created based on known N-glycan synthetic pathways [9]. There are 328 N-glycans of different masses are used in the analysis; but users are allowed to specify their own N-glycans in this candidate list. Glycan annotation is then done at a composition level and thus, structure information is absent. It is worth mentioning that Multi N-Glycan not only directly uses mass as a feature for detection but also uses sophisticated mixture models to improve N-glycan annotation. Multi N-Glycan calculates the correlation between theoretical and experimental isotopic envelopes. It uses three different models for constructing theoretical isotopic envelopes for each glycan candidate. First, the glycan candidate mass is directly used to create an isotopic envelope. Second, a composite overlapping theoretical isotopic envelope comprising of two glycan masses with a mass difference within a mass tolerance window is constructed and matched to the observed isotopic envelope. Finally, a composite envelope is also created using a candidate glycan and an unknown compound. Using these three models, individual as well as *overlapping* glycan isotopic envelopes can be annotated, leading to an increasing number of identified glycans. Multi N-Glycan also provides utilities that detect profile abundance variations across two samples. The correlation based fit score is used to filter in confident glycan identifications. Then, similar to the gene-shaving technique [10], a glycan shaving algorithm based on principal component analysis (PCA) is used to identify the top 'n' (n is a user-specified number) glycan species that contribute most to the abundance variation [11]. This is particular useful for glycan biomarker discovery where abundance variations could be related to change of state between healthy and disease samples [1–3]. Multi N-Glycan can also be used for O-glycan profiling and biomarker discovery if the user can input a pre-defined list of O-glycan compositions.

Usage steps

1. Load spectrum file by click “Load” button in “Peak List Setting” panel.
2. Choose related options for the N-glycans profiling experiment, and load default or user-defined N-glycans composition file in “Glycan List Setting” Panel.
3. Click “Calculation”.
4. Result will be shown in the lower table.

Website: <http://mendel.informatics.indiana.edu/~chuyu/MultiNGlycan/>

Supported spectrum format Plain text (Peak list), mzXML, and RAW file (Thermo Scientific instruments)

3. Comparison of three software tools

We took two MALDI TOF/TOF data sets from previous publication [12]. One data set from hepatocellular carcinoma patients and the other is from healthy control. All software tools were run on default settings, and results were compared with each other. For ease, glycans that differ only in linkages are considered to be a single glycan. In Cartoonist result, we took generated msa file and count non-redundant peaks. In Multi N-Glycan, we only took glycans, which have the correlation score above 0.7 in one of three models. We note that before running GlycoWorkbench annotation, we loaded the spectrum (in peak list) and conducted peak centroid via default setting. Total number of identified glycans is listed in Table 3. We also illustrate the common glycans between software tools in Figure 4.

We can see that Multi N-Glycan and Cartoonist have a few common identifications. This is because the majority of Cartoonist annotated glycans mass are more than 3000, which not

pick up by the Multi N-Glycan. GlycoWorkbench has more than nearly 600 glycans in both sets, and this huge amount of glycans might be falsepositive.

4. Conclusion

Here, we introduced three software tools for high throughput glycan annotation and profiling in glycomics. To gather comprehensive glycan annotation at the sequence and linkage level, we need to acquire more information from the different avenues. A preliminary glycan profile only gives us the mass of glycans but since some monosaccharides have exactly the same mass, e.g. mannose and galactose, GalNAc and GlcNAc etc., the precise monosaccharide composition cannot be deciphered from mass spectra alone. However, we can utilize tandem MS (MSⁿ) in combination with other software tools to elucidate both sequence and linkage information [13]. Here we limited our discussion on the tools for N-glycan profiling. These methods can be extended to other types of glycosylations such as O-linked glycans. O-glycans have more diverse core structures, which mean the space of candidate of glycans is larger than that of N-glycans. As a result, it is harder to explore O-glycans using above software directly.

The lack of a comprehensive glycan database is a drawback since novel and rare structures are not identified using database search techniques. Discovery of novel glycan structures becomes more and more important in cases where diseases are related to rare glycosylation. We need to be more careful when examining the spectrum and always leave some tolerance for novel structure discovering.

References

1. Martin-Rendon E, Blake DJ. Protein glycosylation in disease: new insights into the congenital muscular dystrophies. *Trends Pharmacol Sci.* 2003; 24(4):178–183. [PubMed: 12707004]
2. Scanlin TF, Glick MC. Terminal glycosylation and disease: influence on cancer and cystic fibrosis. *Glycoconj J.* 2000; 17(7–9):617–626. [PubMed: 11421353]
3. Turner GA. N-glycosylation of serum proteins in disease and its investigation using lectins. *Clin Chim Acta.* 1992; 208(3):149–171. [PubMed: 1499135]
4. Raman R, et al. Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology.* 2006; 16(5):82R–90R.
5. PMSV: PARC Mass Spectrum Viewer. Available from: http://bio.parc.com/mass_spec.
6. Doubet S, et al. The Complex Carbohydrate Structure Database. *Trends Biochem Sci.* 1989; 14(12):475–477. [PubMed: 2623761]
7. Ranzinger R, et al. GlycomeDB - integration of open-access carbohydrate structure databases. *BMC Bioinformatics.* 2008; 9:384. [PubMed: 18803830]
8. Lutteke T, et al. GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology.* 2006; 16(5):71R–81R.
9. Krambeck FJ, Betenbaugh MJ. A mathematical model of N-linked glycosylation. *Biotechnol Bioeng.* 2005; 92(6):711–728. [PubMed: 16247773]
10. Hastie T, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000; 1(2) RESEARCH0003.
11. Kyselova Z, et al. Alterations in the serum glycome due to metastatic prostate cancer. *J Proteome Res.* 2007; 6(5):1822–1832. [PubMed: 17432893]
12. Tang Z, et al. Identification of N-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data. *J Proteome Res.* 2010; 9(1):104–112. [PubMed: 19764807]
13. Lapadula AJ, et al. Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSⁿ data. *Anal Chem.* 2005; 77(19):6271–6279. [PubMed: 16194088]

Figure 1. PARC Mass Spectrum Viewer screenshot. The glycan structures annotated in the input mass spectra is shown on the top.

Figure 2.

GlycoWorkbench screenshot. There are four major panels in GlycoWorkbench, workspace, canvas, spectrum view, and result list. The data can be loaded via right click on the workspace tree node. Canvas panel provides a GUI interface for user to draw glycan structures. Users also can view their raw spectrum in the spectrum view panel. Result list panel contains peaks, fragments, annotation and profile list. All the results will be shown in this panel.

Figure 3.
Multi N-Glycan screenshot

Figure 4.
Number of glycans between software tools

Table 1

Three levels of representation for glycans. The basic composition representation of a glycan gives us only the mass of the glycan. Sequence representations (or the cartoons) inform us monosaccharide and topology information indicative of branching. Addition of type of linkage gives us the comprehensive view of the glycan. The glycans shown are N-linked glycans from the CFG database [4]. Note the core of these glycans with two GlcNAc and three Mannose residues.

Level	Representation Example	Note
Composition	9 Hex 2 HexNAc (or) Hex ₉ HexNAc ₂ (or) 2 GlcNAc 9 Man	Indicates number and the type of monosaccharide Abbreviations : Hex : Hexose HexNAc : N-acetylhexosamine Man: Mannose GlcNAc : N-acetylglucosamine
Sequence		Called 'cartoon graphs'. Indicates monosaccharide and topology. Blue squares indicate GlcNAc and green circles indicate Mannose. More symbol nomenclatures in (Varki 2008)
Sequence with Linkage		Cartoon graphs also indicating linkage information. More information on different types of linkage available in (Varki 2008)

Table 2

Comparison table for different invoking method.

	Running time (second)	Memory usage (MB)
command line	50	1,800
jar directly	215	300

Table 3

Total number of identified glycans in two different data sets (HC-146, NC-33)

	Cartoonist	GlycoWorkbench	Multi N-Glycan
	HC-146		
Total identified glycan	48	691	120
	NC-33		
Total identified glycan	50	653	114