Mass Spectrometry Data Needed to Distinguish Between 2'-3'/2'-6' Sialic Acid Linkage for Machine Learning

Nathan Murray[1], Xinyi Ni[2], Stephanie Archer-Hartmann[1], Lauren E. Pepi[1], Richard Helm[3], Pengyu Hong[2] Parastoo Azadi[1]

[1]Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia 30602

[2]Department of Computer Science, Brandeis University, Waltham, Massachusetts 02453

[3]Department of Biochemistry, Virginia Tech, Blacksburg, Virginia 24061

Introduction

Sequencing and linkage determination of glycopolymers is extremely challenging, labor intensive, and requires specialized personnel. The aim of this project is to make mass spectrometry data analysis more amenable to larger number laboratories by using machine learning to assist with the data interpretation. Currently, requisite quantities of sample as well as the amount of time trained personnel must spend collecting and analyzing data makes mass spectrometry prohibitive for glycosidic linkage characterization in many contexts. Sialic acids (SA) are a broad class 9-carbon carbohydrates with more than 20 sialyltransferases contributing to their prevalence on various glycomaterials. Here we propose a framework to deliver appropriate MS^n protocols for the characterization of sialic acid linkages in addition to the development of machine learning techniques for the automatic inference of glycosidic linkages when provided with pre-collected MS^n data.

Research Question

Can a method involving machine learning techniques make sialic acid linkage characterization by mass spectrometry more accessible and/or can the data analysis be automated?

Methods

- 2-3' and 2-6' sialyllactose standards and purified bovine fetuin were purchased from Sigma, and glycan standards were purchased from NIST
- Sialyllactose (SL) and NIST standards were suspended in water, and various known ratios of 2-3' to 2-6' sialyllactose were created before all samples were lyophilized dry
- Purified fetuin was resuspended in NH_4HCO_3 and various samples were subjected 3 exoglycosidase digests: **1.** No enzyme **2.** α2-3 Neuraminidase **3.** α2-3,6,8,9 Neuraminidase
- Feutin samples were then reduced with 25mM DTT and desalted with 10kDa MWCO filters before PNGase F incubation at 37°C for 24h
- PNGase F released N-glycans were purified using C18 SPE cartridges and lyophilized
- The dried Fetuin N-glycans, sialyllactose standards, and NIST standards were permethylated with DMSO/NaOH base and iodomethane.
- Following the reaction, permethylated glycans were isolated and desalted with a water/DCM phase extraction. The DCM layer containing permethylated glycans was dried under N_2 and resuspended in 50% $MeOH_{(aq)}$ with 1mM Li_2CO_3 for analysis
- Samples were run by direct infusion in positive mode on a Thermo Oribtrap Eclipse Tribrid. 60 unique runs were performed for each sample with all possible combinations of isolation window (1, 2, 3) and CID percentage (30, 35, 40, 45, 50) at MS^n scan levels of MS^2 through MS^5
- Collected data files were converted from .raw format to .mzXML format using Proteowizard msConvert software and were used for model development

Results

Figure 1: MSn fragmentation pathway for fetuin released glycans with example MS5 spectrum.

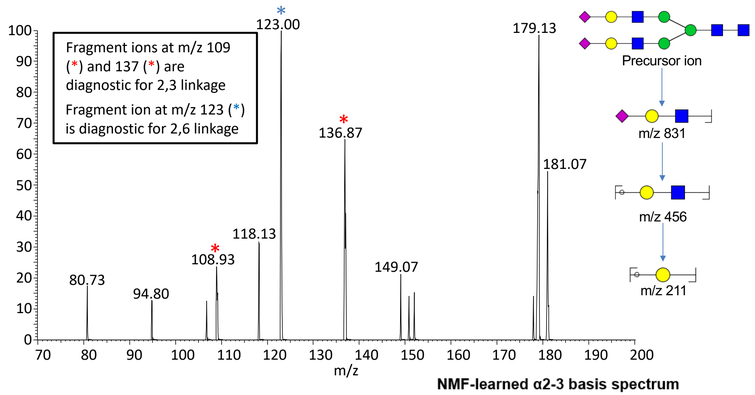
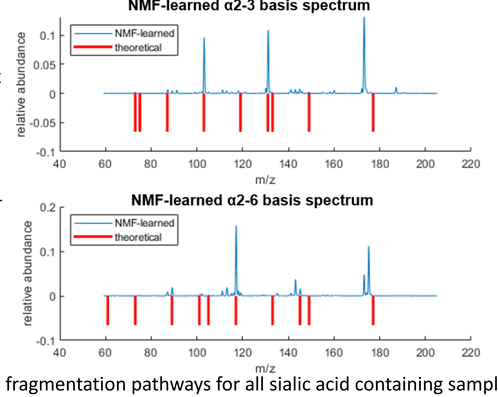


Figure 2 : An example of the basis spectra learned by Non-negative Matrix Factorization (NMF)



- The top basis spectrum contains the learned α2-3' peaks (m/z 103 & 131) while the bottom contains learned α2-6' peaks (m/z 89 & 117)
- These learned basis spectra align with their respective theoretical SA linkage types

Table 1: MS^n Experimental Parameters and fragmentation pathways for all sialic acid containing samples

Sample	Fragmentation Pathway (m/z)	Scan Levels Collected	Isolation Windows (m/z)	CID Collision Energies (%)
Sialyllactose Standards				
100% α2'-3' SL	822 > 447 > 211	MS^4	1,2,3	30, 35, 40, 45, 50
90% α2'-3' SL, 10% α2'-6' SL	822 > 447 > 211	MS^4		
50% α2'-3' SL, 50% α2'-6' SL	822 > 447 > 211	MS^4		
10% α2'-3' SL, 90% α2'-6' SL	822 > 447 > 211	MS^4		
100% α2'-6' SL	822 > 447 > 211	MS^4		
NIST Standards				
G2S1(6)	1211 > 831 > 456 > 211	MS^{2-5}	1,2,3	30, 35, 40, 45, 50
G2S2(3)	930 > 831 > 456 > 211	MS^{2-5}		
G2S2(6)	930 > 831 > 456 > 211	MS^{2-5}		
G2FS1(6)	867 > 831 > 456 > 211	MS^{2-5}		
G2FS2(3)	988 > 831 > 456 > 211	MS^{2-5}		
G2FS2(6)	988 > 831 > 456 > 211	MS^{2-5}		
Glycoprotein Glycans				
Fetuin	930 > 831 > 456 > 211	MS^{2-5}	1,2,3	30, 35, 40, 45, 50
α2-3' Neur. Fetuin	959 > 831 > 456 > 211	MS^{2-5}		

Conclusion

- SA linkage characterization with mass spec via MS^5 fragmentation has been established here, however it requires sensitive instrumentation (i.e., ITMS or FTMS), large quantities of sample, and time.
- To conserve resources and eliminate prohibitive aspects of existing methods, we have proposed a machine learning model which can predict the linkage of sialic acid from $MS^{2/3/4}$
- The ability to distinguish SA linkage at these MS^n levels could allow researchers without ITMS or FTMS instrumentation to conduct linkage experiments, providing this resource to a broader scientific audience
- Additionally, a machine learning model could automate the data analysis at all MS^n levels, thereby reducing the time required for trained personnel.
- We plan to expand on this work with additional mass spectrometry data from biologically relevant samples, such as cell lines and tissues in order to train the machine learning model with a greater diversity of samples We hope to expand this model to other linkages that are not conserved within N-glycans, as well as other glycomaterials

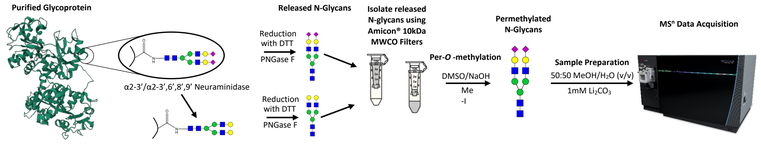


Figure 3 : Graphical Absctract

Acknowledgement

This work was supported by GlycoMIP, a National Science Foundation Materials Innovation Platform funded through Cooperative Agreement DMR-1933525.

References

[1] Ma, X. *et al*. Slc35a1 deficiency causes thrombocytopenia due to impaired megakaryocytopoiesis and excessive platelet clearance in the liver. *Haematologica 2021*, ***106*** (3), 759-769. (2021)

[2] Anthony, R. M. *et al*. Recapitulation of IVIG anti-inflammatory activity with a recombinant IGG FC. *Science*, ***320*** (5874), 373–376 (2008).

[3] Adusumilli, R.; Mallick, P. Data Conversion with ProteoWizard msConvert. In *Methods in Molecular Biology*, Springer New York, 339-368. (2017)

[4] Shajahan, A. *et al*. High-Throughput Automated Micro-permethylation for Glycan Structure Analysis. *Anal Chem*, ***91*** (2), 1237-1240 (2019)