Relative Retention Time Estimation Improves *N*-Glycopeptide Identifications By LC-MS/MS

Joshua A. Klein[1] and Joseph Zaia[1,2]

[1]Bioinformatics Program, Boston University, Boston, MA 02215, USA
[2]Department of Biochemistry, Boston University School of Medicine, Boston, MA 02118, USA

Additional materials

List of Figures

List of Tables

S1 Supplement

S1.1 More Adduction Evidence

Adducts are visible readily in the MS1 spectra of the affected samples, as shown in Figures S-1 and S-2a. Although we did not search for them, we did spot-annotate them in MS/MS spectra in PXD009654 in Figure S-2.

S1.2 Monosaccharide Delta Histograms

To check our parameter estimates, we looked at the per-monosaccharide difference between glycopeptides from the same sample and same peptide backbone which differed by individual monosaccharides by one. We saw that the averaged differences for `NeuAc` and `Fuc` are close to the ammonium adduct corrected parameter estimates in Figure S-3, and those averages even hold when estimated without correcting for ammonium adducts in Figure S-4.

S1.3 Between Run Variation

Because the case study in PXD009654 used multiple LC-MS runs to estimate its parameters, it inherited another source of error. The sample-specific intercept terms can account for part of that error, but not all of it, as shown in Figure S-5.

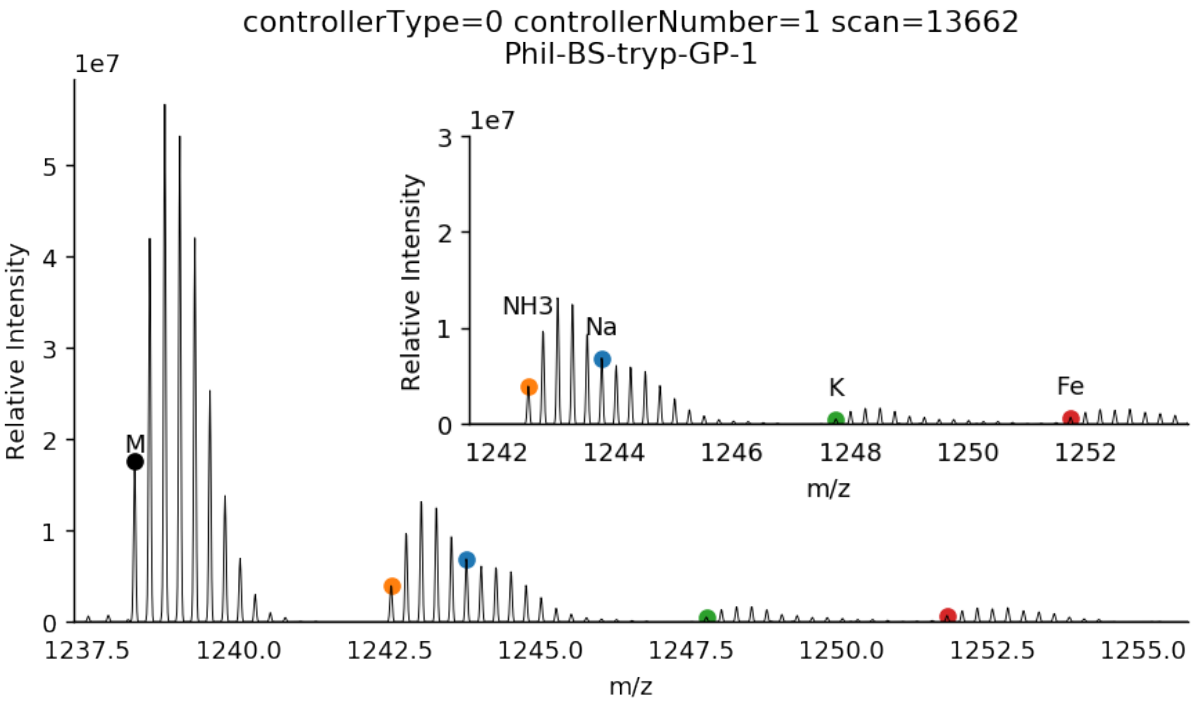


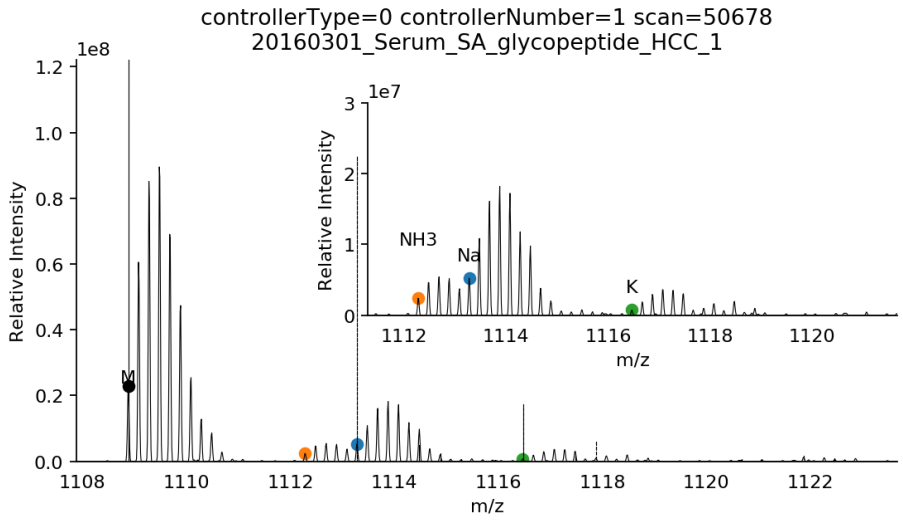
Figure S-1: The profile of multiple adduction states of SDAPIGTCSSECITPNGSIPNDKPFQNVNK{Hex:8; HexNAc:2}

S1.4 Chromatogram CSV Format

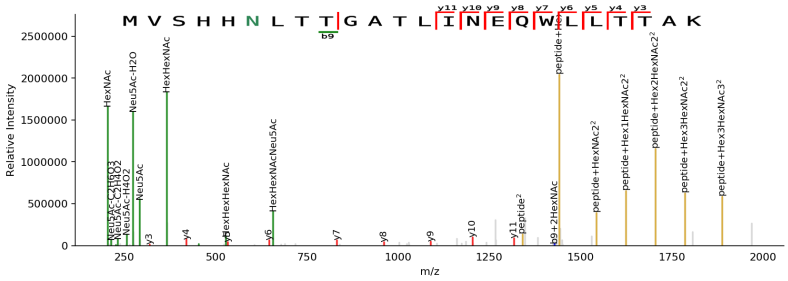
The CSV format that the retention time prediction software requires is described on Table S-1

Column Name	Description	Example
structure	The glycopeptide sequence encoding the amino acid sequence and the glycosylation	"N(N-Glycosylation)C(Carbamidomethyl)TL IDALLGDPHC(Carbamidomethyl)DG FQNEK{Hex:3; HexNAc:2}"
glycan_composition	The glycan composition of the glycopeptide to predict with	{Hex:3; HexNAc:2}
apex_time	The time at which the chromatographic peak was at its maximum height	42.67
total_signal	The total intensity or area under the chromatographic peak of this chromatogram	259660825.15
weighted_neutral_mass	The empirical neutral mass of the chromatogram	3308.39
analysis_name	The name of the analysis or replicate that this chromatogram came from	Phil-82-tryp-GP-1-1576192479

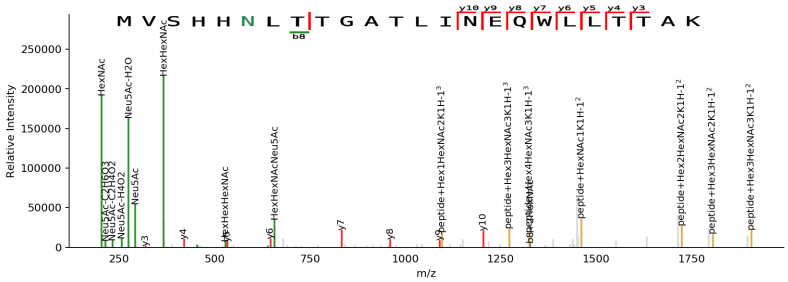
Table S-1: The columns required by the retention time modeling tool "glycresoft analyze retention-time fit-glycopeptide-retention-time"



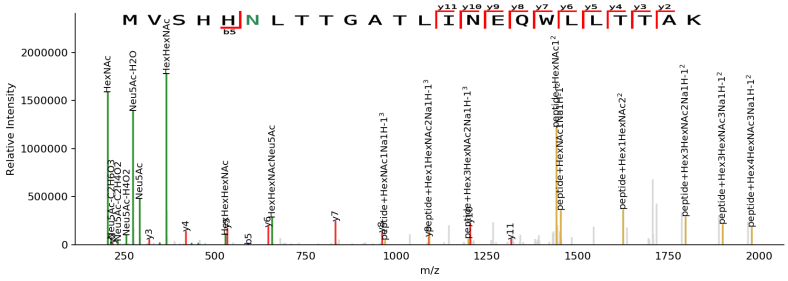
(a)



(b)



(c)



(d)

Figure S-2: Evidence for metallic cation adduction in PXD009654 as well, although it was not included in the search. S-2a shows MVSHHN(N-Glycosylation)LTTGATLINEQWLLTTAK{Hex:6; HexNAc:5; NeuAc:3} with multiple adduction states. S-2b shows an MS/MS spectrum in its native state. S-2c shows an MS/MS spectrum adducted with potassium from that structure. S-2d shows the same structure with a sodium adduct, with multiple abundant peptide+Y ions both with and without sodium adduction, suggesting that the adduct is either labile or partially associates with the glycan. The peptide sequence continued to fragment past the glutamic acid residue, suggesting that it is not the adduction site. Further, if the modification is localized there, no fragments spanning that residue match.

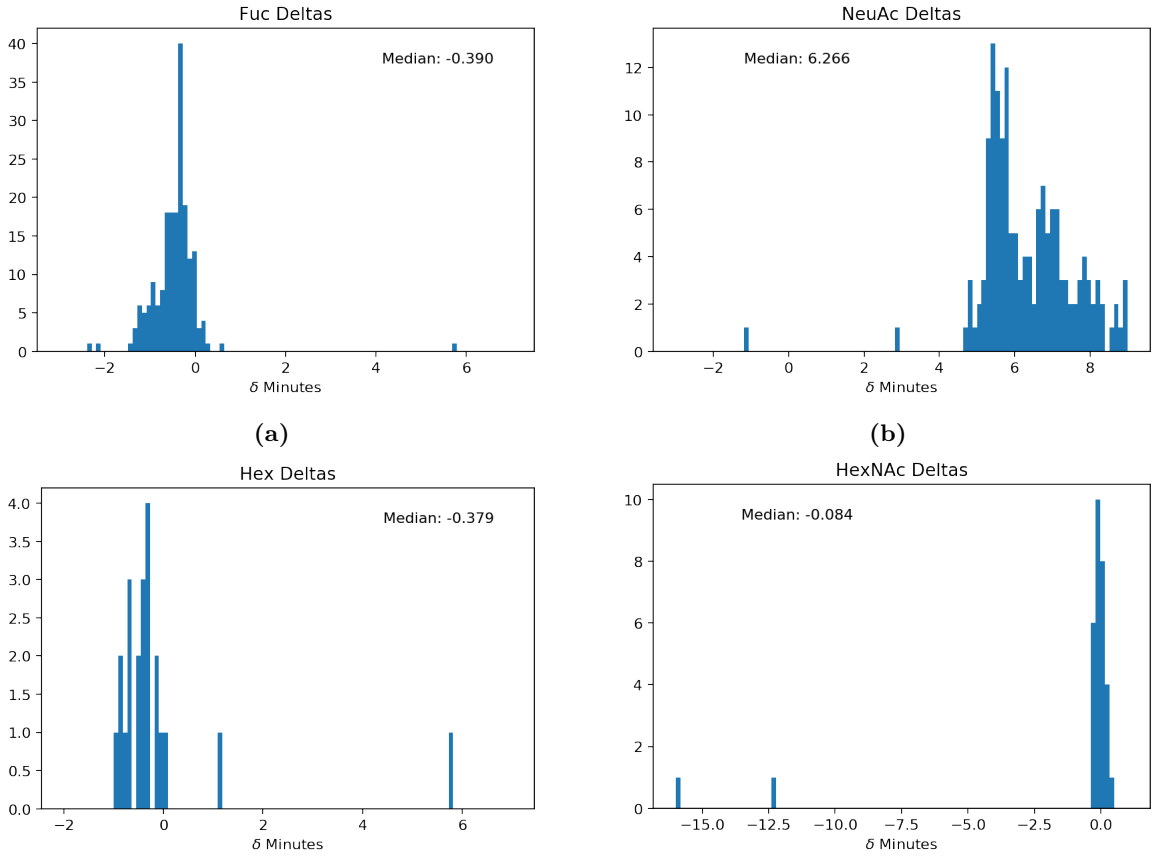


Figure S-3: The distribution of single monosaccharide difference retention time shifts within LC-MS runs from the early portion of the ammonium adduct corrected results for PXD009654. We observe multimodal distributions for `Fuc` and `NeuAc` which may reflect some underlying process or interaction not captured by our model.

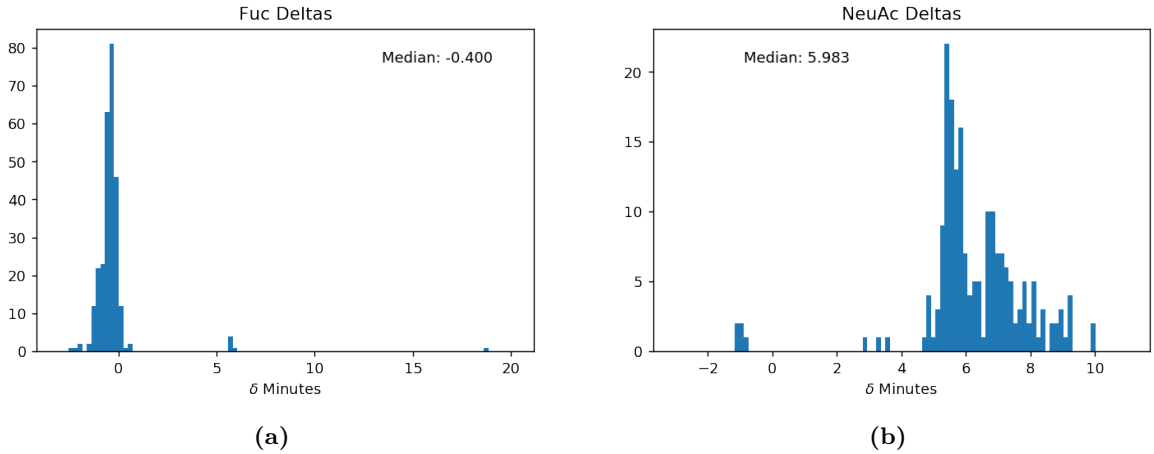
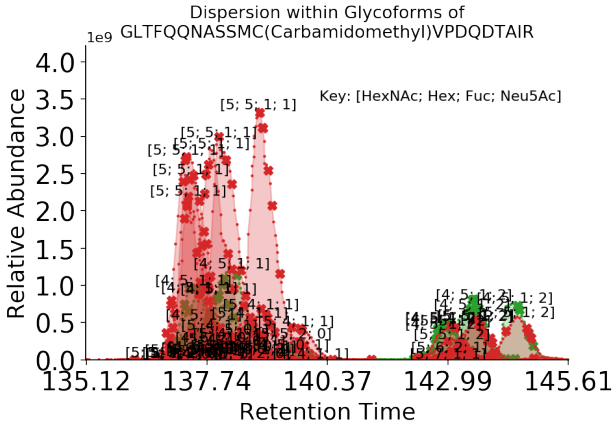
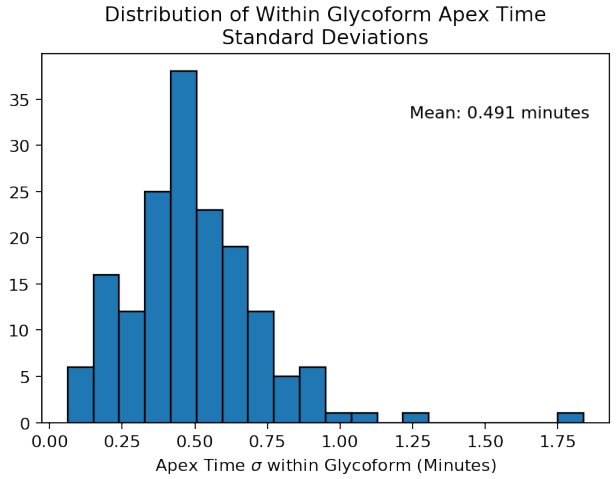


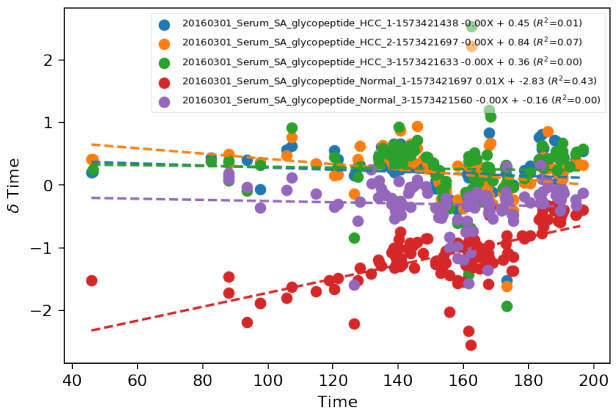
Figure S-4: The distribution of single monosaccharide difference retention time shifts within LC-MS runs from the early portion of the ammonium adduct uncorrected results for PXD009654.



(a)



(b)



(c)

Figure S-5: All figures shown are based upon PXD009654's samples. The dispersion of retention times within the same glycoform of the same peptide across runs varied. S-5a shows the distribution of chromatograms for a single peptide sequence with a total of five glycoforms across runs. S-5b shows the distribution of within glycoform apex time standard deviations. It is likely that an alignment of LC-MS runs would produce narrower error distributions, but adds several layers of complexity that may mask the true model parameters. S-5c shows the deviation between all other runs to "20160301_Serum_SA_glycopeptide_Normal_2"