Abstract

Motivation: Glycosylation is one of the most heterogeneous and complex post-translational modifications.
Results: These are the results for this article.

Application of Network Smoothing to Glycan LC-MS Profiling

Joshua Klein, Luis Carvalho, Joseph Zaia

October 13, 2017

1 Introduction

Glycosylation is one of the most pervasive and diverse forms of post-translational modification (Varki (2017)). Their study is of great importance for understanding broad classes of biological processes. Mass spectrometry (MS) is a powerful tool for glycan analysis (Zaia (2008)). While unseparated MS experiments using methods like MALDI provide strong signal, but cannot interpret complex mixtures Peltoniemi *et al.* (2013). An online separation method like liquid chromatography (LC) or capillary electrophoresis (CE) makes analyzing such complex samples possible, at the cost of increased analytical complexity.

There are many tools for interpreting glycan mass spectral datasets (Yu *et al.* (2013); Peltoniemi *et al.* (2013); Kronewitter *et al.* (2014); Goldberg *et al.* (2009); Maxwell *et al.* (2012); Ceroni *et al.* (2008); Frank and Schloissnig (2010)) for both unseparated and separated experimental protocols. The different types of instrumentation these programs were written to accommodate introduces different types of signal processing approaches, for example SysBioWare (Frank and Schloissnig (2010)) performs sophisticated baseline removal prior to fitting peaks, while tools like GlyQ-IQ (Kronewitter *et al.* (2014)) was written for much cleaner Fourier Transform MS (FTMS), and so does not accommodate these well. Tools that build on the THRASH implementation from Decon2LS (Jaitly *et al.* (2009); Yu *et al.* (2013); Maxwell *et al.* (2012)) are likewise unable to deal with variable baseline noise or extreme dynamic range.

Each tool also has its own format for defining glycan structures or compositions, some even bundling a large database with their software to remove the burden from the user to build a list of candidates themselves (Yu *et al.* (2013); Kronewitter *et al.* (2014); Goldberg *et al.* (2009)) while others define methods for building glycan databases as part of the program (Maxwell *et al.* (2012); Ceroni *et al.* (2008)). Many of these tools are designed for specific glycan subclass such as N-glycans or glycosaminoglycans, limiting their vocabulary of possible monosaccharides to just those found in that subclass (Yu *et al.* (2013); Kronewitter *et al.* (2014); Peltoniemi *et al.* (2013); Goldberg *et al.* (2009)), Often, these tools are tailored to analysis of a particular derivatization state, adduction conditions, or neutral loss pattern (Yu *et al.* (2013); Peltoniemi *et al.* (2013); Maxwell *et al.* (2012)).

These spectral processing and glycan library properties are reflected in the scoring function that each program uses to discriminate glycan signal from the background noise and contaminants. As observed by Goldberg *et al.* (2009), there is value in including related glycan composition identifications in how much confidence one assigns to a another glycan composition assignment. They use a method to exploit the known biosynthetic rules to connect peaks in a MALDI spectrum which could be assigned to a particular N-glycan by intact mass alone. Their method using the maximum weighted subgraph of the biosynthetic network in one of their three had demonstrably better performance than chance with their expert system annotation method. Kronewitter and colleagues considered a similar idea with more emphasis on handling in-source fragmentation (Kronewitter *et al.* (2014)) observed in LC-MS and LC-MS/MS experiments.

We extend this notion of a glycan family to cover more sectors of the biosynthetic landscape which we term "neighborhoods", and present an algorithm for learning the importance of each neighborhood from observed data, which can in turn be used to improve glycan composition assignment performance.

2 Methods

2.1 Glycan Hypothesis Generation

In eukaryotes, N-glycans start with a common, conserved core of **HexNAc2 Hex3**, building up to **HexNAc2 Hex9** (Stanley *et al.* (2009)). This structure is refined by sequentially removing monosaccharides and replacing them with more complex structures through a series of glycosylase and glycosyltransferase reactions, the enumeration of, which as shown in Akune *et al.* (2016), yields over a million of possible N-glycan topologies. These topologies define the geometry of the glycan, affecting the glycan's binding affinities and how the glycan may influence protein folding and accessibility, the glycan's functional aspects. The medium through which we observed glycans did not capture the full tree or graph structure of an N-glycan, so we reduced the topology to a count of each type of residue.

Starting with the core motif, we generated all combinations of monosaccharides ranging between the limits in Table 1 to build a glycan composition database, which produced 1240 distinct compositions. We created a copy of this database for native, reduced and permethylated, and deuteroreduced and permethylated for each experimental protocol we analyzed in

Table 1: Glycan Composition Rule Table

Monosaccharide	Lower Limit	Upper Limit	Constraints
HexNAc	2	9	
Hex	3	10	
Fuc	0	4	**HexNAc** > **Fuc**
NeuAc	0	5	(**HexNAc** − 1) > **NeuAc**

Table 2: Samples Used

Sample Name	Instrument	Derivatization	Adduction	Source
20150930-06-AGP	QTOF	Native	Formate (1)	Khatri *et al.* (2016a)
20141031-07-Phil-82	QTOF	Native	Formate(3)	Khatri *et al.* (2016a)
20141103-02-Phil-BS	QTOF	Native	Formate(3)	Khatri *et al.* (2016a)
20151002-02-IGG	QTOF	Native	Formate (2)	Khatri *et al.* (2016b)
20141128-11-Phil-82[1]	QTOF	Deutero-reduced and Permethylated	Ammonium (3)	Khatri *et al.* (2016a)
AGP-DR-Perm-glycans-1[1]	FTMS	Deutero-reduced and Permethylated	Ammonium (3)	Khatri *et al.* (2016a)
AGP-permethylated-2ul-inj-55-SLens[1]	FTMS	Reduced and Permethylated	Ammonium (3)	Khatri *et al.* (2016a)
Perm-BS-070111-04-Serum[1]	FTMS	Reduced and Permethylated	Ammonium (3)	Yu *et al.* (2013); Hu and Mechref (2012)

[1] Included MS^n Scans

this study. We chose to use a combinatorial database for simplicity. The later algorithms can be used with an arbitrary glycan composition list. This places the burden of finding or creating such a list on the user. The glycan database is stored in a SQLite3 (v3.15.2) database file (Hipp and Et Al. (2016)).

2.2 LC-MS Data Preprocessing

We analyzed samples from several sources, including both QTOF and FTMS instruments as shown in Table 2. For details on sample preparation and data acquisition, please see their source citation. We converted all datasets to mzML format (Martens *et al.* (2011)) prior to analysis with Proteowizard (Kessner *et al.* (2008)) without any data transforming filters. We applied a background reduction method based upon (Kaur and O'Connor (2006)), using a window length of 2 m/z. Next, we picked peaks using a simple Gaussian model and iteratively charge state deconvoluted and deisotoped using an averagine (Senko *et al.* (1995)) formula appropriate to the molecule under study. For native glycans, the formula was **H 1.690 C 1.0 O 0.738 N 0.071**, for permethylated glycans, the formula was **H 1.819 C 1.0 O 0.431 N 0.042**. We used an iterative approach which combines aspects of the dependence graph method (Liu *et al.* (2010)) and with subtraction. All samples were processed using a minimum isotopic fit score of 20 with an isotopic strictness penalty of 2.

2.3 Chromatogram Aggregation

We clustered peaks whose neutral masses were within 15 parts-per-million error (PPM) of each other. When there were multiple candidate clusters for a single peak, we used the cluster with the lowest mass error. Next, we sorted each cluster by time, creating a list of aggregated chromatograms. To account for small mass differences, we found all chromatograms which are within 10 PPM of each other and which overlap in time and merge them.

2.4 Glycan Composition Matching

For each chromatogram, we searched the glycan database for compositions whose masses were within $\delta_{mass} = 10$ PPM for QTOF data, 5 PPM for FTMS data. We merged all features matching the same composition. Then, for each adduct combination, we searched the glycan database for compositions whose neutral mass were within δ_{mass} of the observed neutral mass - adduct combination mass, followed by another round of merging chromatograms with the same assigned composition. We reduced the data by splitting each feature where the time between sequential observation was greater than $\delta_{rt} = 0.25$ minutes and removed features with fewer than $k = 5$ data points.We term the remaining assigned and unassigned chromatograms *candidate features*.

2.5 Feature Evaluation

For each candidate feature, we computed several metrics to estimate how distinguishable the observed signal was from random noise. The features are mentioned in List 1, but for more information see Section S1.

List 1: Chromatographic Feature Metrics

1. Goodness-of-fit of chromatographic peak shape to a model function (Yu and Peng (2010); Kronewitter *et al.* (2014)).
2. Goodness-of-fit of isotopic pattern to glycan composition weighted by peak abundance (Maxwell *et al.* (2012)).
3. Observed charge states with respect to glycan composition and mass.
4. Time gap between MS^1 observations detecting measuring missing peaks and interference.
5. Adduction states with respect to glycan composition and mass.

These metrics are bounded in $[0, 1)$. Any observation for which any metric was observed below 0.15 was discarded as having insufficient evidence for consideration. The *observed score* s for each candidate feature is the sum of the logit-transformation of these metrics. This produces a single value bounded in $[0, \infty)$, whose distribution we assume is asymptotically normal. $s < 8$ reflects a low confidence match, with confidence increasing as s does. As these metrics are tied to reliable detection of the the glycan by the mass spectrometer, they are dependent upon glycan abundance and sample quality and the resolution of the mass spectrometer used.

2.6 Glycan Composition Network Smoothing

Evidence for individual glycan compositions can often be enough to claim that composition had been detected. Lower abundance may score poorly in one or more features, leading to the glycan composition being discarded. Other methods have demonstrated it is advantageous to use relationships between glycans based on biosynthetic or structural rules to adjust the score of a single glycan assignment (Goldberg *et al.* (2009); Kronewitter *et al.* (2014)). We propose a method based on Laplacian Regularized Least Squares (Belkin *et al.* (2006)) to use evidence from glycan compositions related over a network to smooth it's classification of glycan composition feature matching.

2.6.1 Glycan Composition Graph

For each database of theoretical glycan compositions we create, we define each composition to be a coordinate vector in a $\mathcal{Z}^{+4}$ space, and represented by a node in an undirected glycan composition graph $\mathcal{G}$. Under this interpretation, we can compute the L_1-distance between two glycan compositions. For any two glycan compositions g_u, g_v, if $L_1(g_u, g_v) = 1$ we add an edge connecting g_u and g_v to $\mathcal{G}$ with weight $w = 1$.

2.6.2 Neighborhood Definition

Our definition of distance connects glycan compositions which differ by a single monosaccharide, but we can assert larger collections of glycan compositions are related. We define neighborhoods for N-glycans using intervals over monosaccharide counts defined in Table 3.

Glycan compositions may belong to zero or more neighborhoods, as there are unusual glycan compositions which do not satisfy any neighborhood's rules, and several neighborhoods intentionally overlap to express broad relationships between groups. We define a matrix $\mathbf{A}$ as an $n \times k$ matrix where $A_{i,k}$ to be the degree to which g_i belongs kth neighborhood:

$$A_{i,k} = \frac{1}{|\text{neighborhood}_k|} \sum_{g^* \in \text{neighborhood}_k} L_1(g_i, g^*) \tag{1}$$

To reduce the impact of neighborhood size on the elements of $\mathbf{A}$, the columns of $\mathbf{A}$ are first normalized to sum to 1, and then the rows of $\mathbf{A}$ are normalized to sum to 1[1].

We assume that members of the same neighborhood will share a central tendency, τ.

[1] The stated reduction is not well tested, and the change may well be minimal because all that really happens is the weight of the column for each row is weighted by a shrinking function of column size. It may be better if we don't manipulate A at all.

Name	Bounds
High Mannose	$\mathbf{HexNAc} = 2 \wedge \mathbf{Hex} \in [3, 10] \wedge \mathbf{NeuAc} = 0$
Hybrid	$\mathbf{HexNAc} \in [2, 4] \wedge \mathbf{Hex} \in [2, 6] \wedge \mathbf{NeuAc} \in [0, 2]$
Bi-Antennary	$\mathbf{HexNAc} \in [3, 5] \wedge \mathbf{Hex} \in [3, 6] \wedge \mathbf{NeuAc} \in [1, 3]$
Asialo-Bi-Antennary	$\mathbf{HexNAc} \in [3, 5] \wedge \mathbf{Hex} \in [3, 6] \wedge \mathbf{NeuAc} \in [0, 1]$
Tri-Antennary	$\mathbf{HexNAc} \in [4, 6] \wedge \mathbf{Hex} \in [4, 7] \wedge \mathbf{NeuAc} \in [1, 4]$
Asialo-Tri-Antennary	$\mathbf{HexNAc} \in [4, 6] \wedge \mathbf{Hex} \in [4, 7] \wedge \mathbf{NeuAc} \in [0, 0]$
Tetra-Antennary	$\mathbf{HexNAc} \in [5, 7] \wedge \mathbf{Hex} \in [5, 8] \wedge \mathbf{NeuAc} \in [1, 5]$
Asialo-Tetra-Antennary	$\mathbf{HexNAc} \in [5, 7] \wedge \mathbf{Hex} \in [5, 8] \wedge \mathbf{NeuAc} \in [0, 0]$
Penta-Antennary	$\mathbf{HexNAc} \in [6, 8] \wedge \mathbf{Hex} \in [6, 9] \wedge \mathbf{NeuAc} \in [1, 5]$
Asialo-Penta-Antennary	$\mathbf{HexNAc} \in [6, 8] \wedge \mathbf{Hex} \in [6, 9] \wedge \mathbf{NeuAc} \in [0, 0]$
Hexa-Antennary	$\mathbf{HexNAc} \in [7, 9] \wedge \mathbf{Hex} \in [7, 10] \wedge \mathbf{NeuAc} \in [1, 6]$
Asialo-Hexa-Antennary	$\mathbf{HexNAc} \in [7, 9] \wedge \mathbf{Hex} \in [7, 10] \wedge \mathbf{NeuAc} \in [0, 0]$
Hepta-Antennary	$\mathbf{HexNAc} \in [8, 10] \wedge \mathbf{Hex} \in [8, 11] \wedge \mathbf{NeuAc} \in [1, 7]$
Asialo-Hepta-Antennary	$\mathbf{HexNAc} \in [8, 10] \wedge \mathbf{Hex} \in [8, 11] \wedge \mathbf{NeuAc} \in [0, 0]$

Table 3: N-Glycan Neighborhoods

2.6.3 Laplacian Regularization

We combine the observed score $\mathbf{s}$ and the structure of $\mathcal{G}$ to estimate a smoothed score ϕ that combines the evidence for each individual glycan composition as well as its relatives. As $\mathbf{s}$ is the size of the set of observed glycan composition p while ϕ is of size n, we partition ϕ into a block vector $\begin{bmatrix}\phi_o \\ \phi_m\end{bmatrix}$ with dimensions $\begin{bmatrix}p \\ n-p\end{bmatrix}$.

Let $\mathbf{L}$ be the weighted Laplacian matrix of $\mathcal{G}$, which is an $n \times n$ matrix. To ensure $\mathbf{L}$ is invertible, we add $\mathbf{I}_n$ to $\mathbf{L}$. We partition $\mathbf{L}$ into blocks $\begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix}$. We also partition $\mathbf{A}$ into $\begin{bmatrix}\mathbf{A}_o \\ \mathbf{A}_m\end{bmatrix}$ and $\tau_o = \mathbf{A}_o\tau$, $\tau_m = \mathbf{A}_m\tau$.

We find the ϕ that minimizes the expression

$$\ell = (\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda \begin{bmatrix}\phi_o - \tau_o, & \phi_m - \tau_m\end{bmatrix} \begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix} \begin{bmatrix}\phi_o - \tau_o \\ \phi_m - \tau_m\end{bmatrix} \tag{2}$$

where λ controls how much weight is placed on the network structure and τ.

To obtain the optimal ϕ, we take the partial derivative of ℓ w.r.t ϕ_m

$$0 = \frac{\partial \ell}{\partial \phi_m}\left((\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda \begin{bmatrix}\phi_o - \tau_o, & \phi_m - \tau_m\end{bmatrix} \begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix} \begin{bmatrix}\phi_o - \tau_o \\ \phi_m - \tau_m\end{bmatrix}\right) \tag{3}$$

$$\hat{\phi}_m = -\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o - \tau_o) + \tau_m \tag{4}$$

and w.r.t. ϕ_o

$$0 = \frac{\partial \ell}{\partial \phi_o}\left((\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda \begin{bmatrix}\phi_o - \tau_o, & \phi_m - \tau_m\end{bmatrix} \begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix} \begin{bmatrix}\phi_o - \tau_o \\ \phi_m - \tau_m\end{bmatrix}\right) \tag{5}$$

$$\hat{\phi}_o = \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right]^{-1}(\mathbf{s} - \tau_o) + \tau_o \tag{6}$$

To use this method, we must provide values for λ and τ. While these values could be chosen based on the expectations of the user for a given experiment, we provide an algorithm for selecting their values in Section S 4. These methods use the topology of the glycan composition graph and the distribution of observed scores, and cannot fully capture boundary cases or related but disconnected parts of the graph.

2.7 Performance Comparison

We compare the performance of the described algorithm with and without network smoothing. State of the art glycan LC-MS profiling software has been designed around Thermo-Fisher Scientific instrumentation, with support for their binary format (Kronewitter *et al.* (2014), Yu *et al.* (2013)) but not open community formats. MultiGlycan-ESI, though publicly available, was unable to be applied to the majority of our datasets because they were not acquired on that vendor's instruments, and neither their RAW nor their mzXML input method produced matches consistent with their previously published results on *Perm-BS-070111-04-Human-Serum*. When we ran MultiGlycan-ESI on *AGP-permethylated-2ul-inj-55-SLens*, the program ran out of memory with 16 GB of RAM available. GlyQ-IQ was made available for testing by its authors, but required data be in Thermo-Fisher's binary format, assumed that glycans were in native form, and did not make its test data publicly available.

3 Results

The performance of our algorithm is demonstrated on *20141103-02-Phil-BS* and *Perm-BS-070111-04-Serum*. Please refer to section S5 for all other datasets. For each comparison, the unregularized case is not smoothed, effectively $\lambda = 0$, the partially regularized case uses the grid search fitted values of τ but uses a fixed $\lambda = 0.2$, and the fully regularized case uses the grid search fitted values of both τ and λ.

3.1 Chromatogram Assignment Performance for *20141103-02-Phil-BS*

The fitted parameters for the network constructed for *20141103-02-Phil-BS* are shown in Table 5. The assigned chromatograms are shown in Figure 1. We observe up to seven branch structures in this sample, consistent with these N-glycans being derived from an avian context (Stanley *et al.* (2009); Khatri *et al.* (2016a)).

Table 4: Fitted λ, γ, and τ for *20141103-02-Phil-BS*

Neighborhood Name	τ_i
high-mannose	17.615084
hybrid	13.599120
bi-antennary	0.0
asialo-bi-antennary	13.919251
tri-antennary	0.0
asialo-tri-antennary	12.906467
tetra-antennary	0.0
asialo-tetra-antennary	14.723146
penta-antennary	0.0
asialo-penta-antennary	11.226188
hexa-antennary	0.0
asialo-hexa-antennary	10.696785
hepta-antennary	0.0
asialo-hepta-antennary	3.071313

Fitted $\lambda = 0.99$ and $\gamma = 11.12$.

Table 5: Fitted λ, γ, and τ for *20141031-07-Phil-82*

Neighborhood Name	τ_i
high-mannose	17.076439
hybrid	13.706498
bi-antennary	0.0
asialo-bi-antennary	16.624051
tri-antennary	0.0
asialo-tri-antennary	15.671840
tetra-antennary	0.0
asialo-tetra-antennary	7.949766
penta-antennary	0.0
asialo-penta-antennary	0.0
hexa-antennary	0.0
asialo-hexa-antennary	0.0
hepta-antennary	0.0
asialo-hepta-antennary	0.0

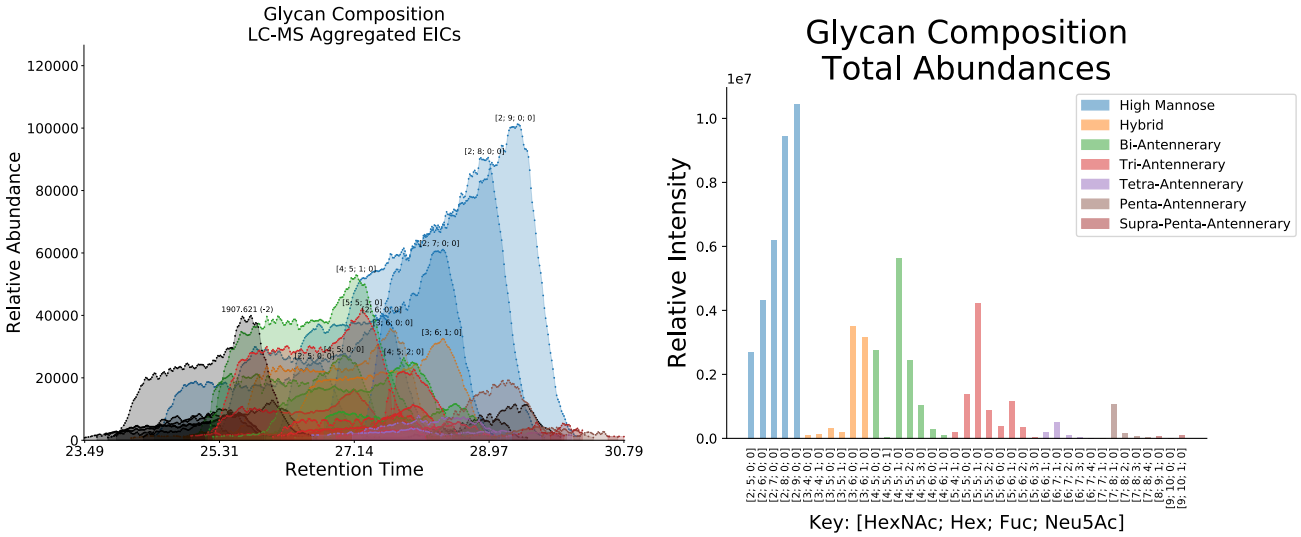
Fitted $\lambda = 0.99$ and $\gamma = 16.26$.

Table 6: Fitted λ, γ, and τ for *Perm-BS-070111-04-Serum*

Neighborhood Name	τ_i
high-mannose	19.839321
hybrid	20.154402
bi-antennary	19.288264
asialo-bi-antennary	21.339077
tri-antennary	22.502435
asialo-tri-antennary	19.923334
tetra-antennary	15.617820
asialo-tetra-antennary	3.509291
penta-antennary	8.764010
asialo-penta-antennary	4.244343
hexa-antennary	0.000000
asialo-hexa-antennary	0.000000
hepta-antennary	0.000000
asialo-hepta-antennary	0.000000

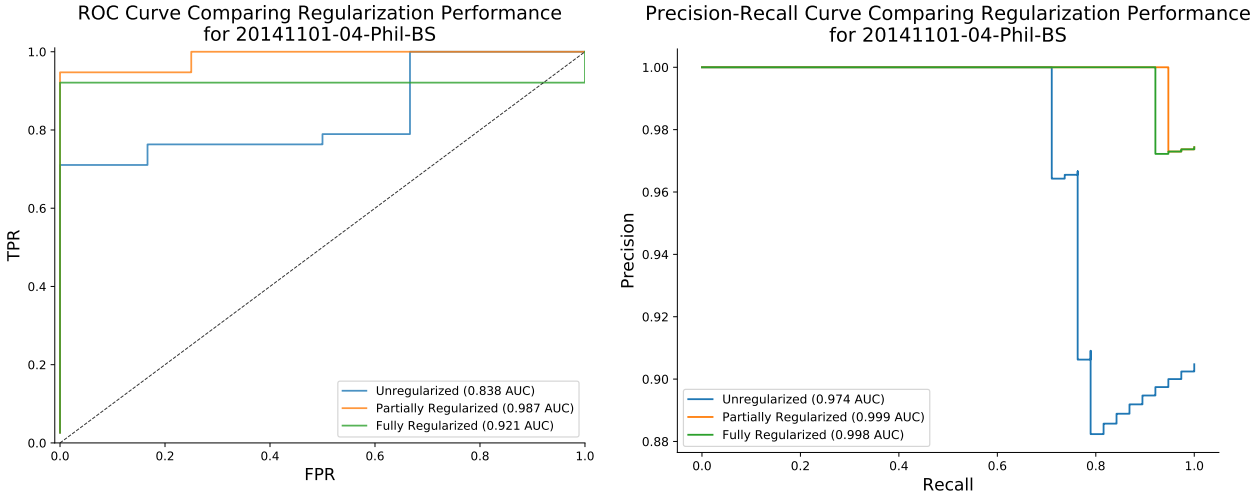
Fitted $\lambda = 0.99$ and $\gamma = 18.962$.

Figure 1: Chromatogram Assignments and Quantification for *20141103-02-Phil-BS*



The comparison of assignment performance with differing degrees of smoothing is shown in Figure 2. The ROC AUC for the unregularized condition is 0.838, for the partially regularized condition is 0.987, and for the fully regularized condition

Figure 2: Performance Comparison with and without Network Smoothing for *20141103-02-Phil-BS*



is 0.921. This demonstrates a higher true positive rate at the same false positive rate for both regularization conditions compared to the unregularized condition. In this condition, the Precision-Recall curve does not show a substantial difference in performance between conditions.

Table 7: Comparison of Performance on *20141103-02-Phil-BS* with Previous Analysis

Method	With Asialo-sulfated		Without Asialo-sulfated	
	Reported	Validated	Reported	Validated
Current Method [1]	62	62	42	42
Khatri *et al.* (2016a)	145	46	136	40

[1] Using partial regularization with $\lambda = 0.2$ and selected at $\phi_o > 5.0$

When compared to previously published results,

3.2 Chromatogram Assignment Performance for *Perm-BS-070111-04-Serum*

The fitted parameters for the network constructed for *Perm-BS-070111-04-Serum* are shown in Table 6. The assigned chromatograms are shown in Figure 3.

The comparison of assignment performance with differing degrees of smoothing is shown in Figure 4. The ROC AUC for the unregularized condition is 0.687, for the partially regularized condition is 0.831, and for the fully regularized condition is 0.797. The Precision-Recall AUC for the unregularized condition is 0.863, for the partially regularized condition is 0.913, and for the fully regularized condition is 0.838. This demonstrates that the partially regularized condition has superior performance to the unregularized and fully regularized conditions. This is complicated by the redundancies caused by ammonium adduction, but without high quality MS^n this cannot be resolved.

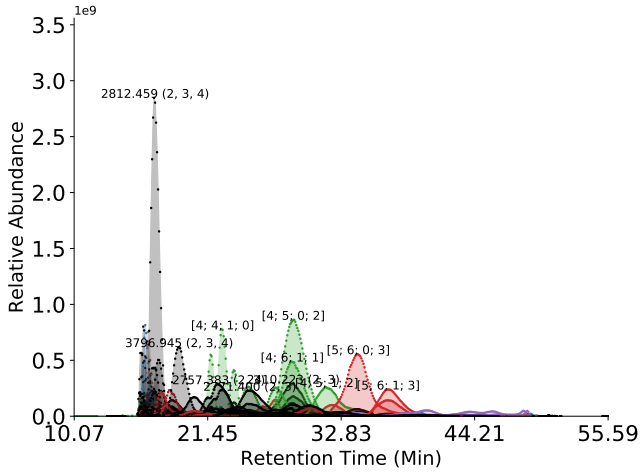
4 Discussion

We demonstrate that this regularization method improved the sensitivity and specificity of glycan composition assignment for LC-MS based experiments. The method used similar assumptions about the importance of common substructural elements of N-glycans to Goldberg *et al.* (2009), but we extend this concept with the addition of a procedure for learning the relationship strengths and use broader groups of structures.

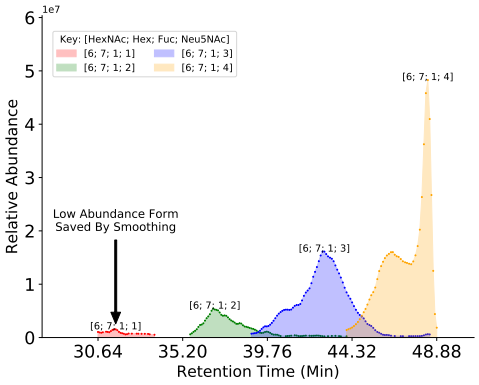
The experimental results from the original analysis of *20141103-02-Phil-BS* and *20141031-07-Phil-82* 82 demonstrated that while both strains expressed predominantly high-mannose glycosylation, *20141103-02-Phil-BS* expressed more larger complex-type structures (Khatri *et al.* (2016a)). In our findings shown in Figure 1, we recapitulate these results while reducing the number of false positive assignments, after controlling for sulfated species as shown in Table 7[2]. There are substantial

[2] Should the description of this procedure be shunted to the supplement or should I describe the creation of a special database for Phil-BS permitting up to one sulfation not bound to a single residue and show the updated annotated chromatograms? The plotted performance does change, as do the estimates of τ change just due to the addition of more glycans to contribute to each neighborhood, but the chosen λ does not.
I am leaning towards rewriting the earlier sections to specifically mention this.

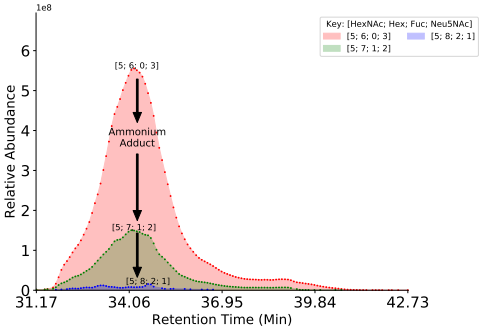
Figure 3: Chromatogram Assignments for *Perm-BS-070111-04-Serum*



(a) Features Assigned After Grid Regularization of *Perm-BS-070111-04-Serum* . Layout not finalized



(b) Low scoring features which may be discarded based on individual evidence alone may be more reasonable to accept given evidence from related composition, such as our network smoothing method.



(c) This sample contains heavy ammonium adduction which introduces ambiguity in intact mass based assignments.

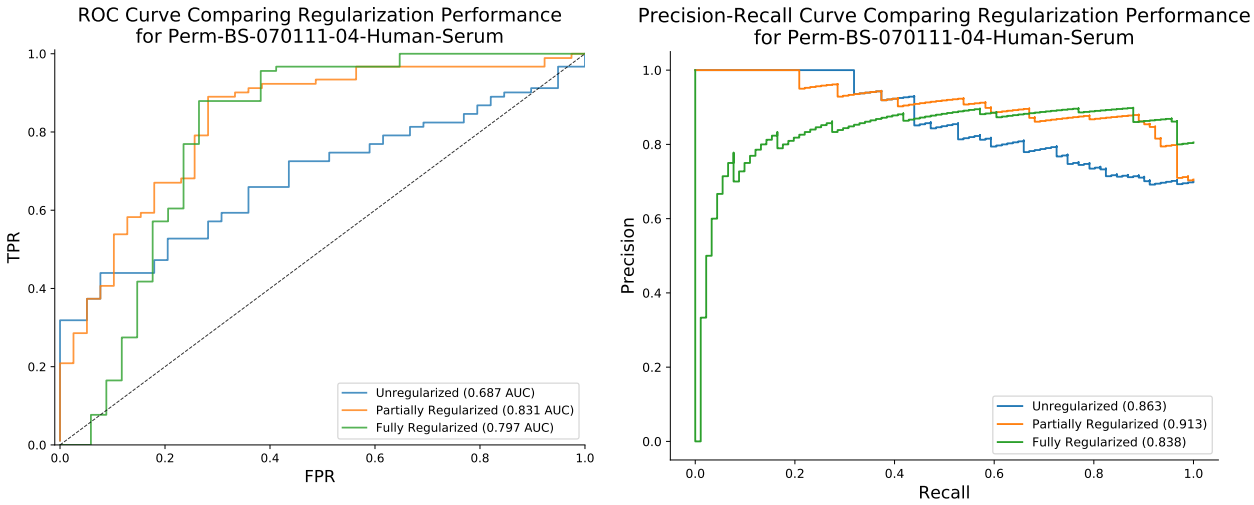
differences in both the mass spectral processing and scoring schemes which contribute to these results, but the regularization procedure is responsible for recovering many low abundance features from this comparison. As these samples are derived from chicken eggs, we may have observed larger branching patterns than are observed in normal mammalian tissue (Stanley *et al.* (2009)). There is evidence for this in the *20141103-02-Phil-BS* with **HexNAc9 Hex10**-based compositions suggesting a seven branch pattern, though this cannot be determined without high quality MS^ndata. The τ fit for both strains have smaller values in the neighborhoods of their largest glycan compositions as these features tended to be low in abundance and not high scoring in their own right, but were partially supported by the overlap with the next largest neighborhood, as expected.

As we show in Figure 4, regularization improves the predictive performance of the algorithm. We reproduce the majority of the glycan assignments from Yu *et al.* (2013), but the ambiguity caused by ammonium adduction as shown in Figure 3c makes a direct comparison of composition assignment lists difficult. Out of the eight missed compositions, four were missing because of insufficient data points to fit a peak shape, which requires at least five points. The other four were not detected either due to mass error, Yu *et al.* (2013) used 10 ppm while we used 5 ppm for FT-MS data, or due to low level signal processing decisions. Since we were unable to reproduce the published results from Yu *et al.* (2013) using their software and accompanying database, it was not reasonable to adapt our composition database to work with their software and run a side-by-side test to demonstrate how many additional glycan compositions one algorithm identifies compared to another without bias.

Of the compositions assigned by our algorithm that were not mentioned in Yu *et al.* (2013) but were annotated in the original publication of this dataset in Hu and Mechref (2012) include **HexNAc3 Hex4**, **HexNAc3 Hex4 NeuAc1**, and **HexNAc5 Hex3**. Because our database was constructed based on combinatorial rules that did not take into account all biosynthetic constraints, we include infeasible compositions in our search space, such as **HexNAc2 Hex10 Fuc1** and **HexNAc5 Hex3 Fuc1 NeuAc2**. Future work could be done to restrict the database to only biosynthetically feasible glycan compositions. This would also have benefits for the construction of the composition network where only those compositions which have an enzymatic reaction to from one to the other would have an edge connecting them, such that **HexNAc5 Hex6 NeuAc2** would not have an edge to **HexNAc5 Hex7 NeuAc2** as in our current model.

These invalid glycan compositions can match LC-MS features at any point in the elution profile, though in this dataset the majority of these matches appear to be in the time range between 10 and 22 minutes, and similar glycan compositions that are

Figure 4: Performance Comparison with and without Network Smoothing for *Perm-BS-070111-04-Serum*



biosynthetically valid elute later on in the experiment. This indicates a need for a retention-time aware approach to evaluating glycan composition assignments, as described in Hu *et al.* (2016), but this is likely dependent upon the experimental workup and separation technique used. The definition of our composition graph and its neighborhoods also mitigates this to some extent, though it depends upon the feature scoring metrics to determine whether a feature is eligible for smoothing. These metrics were based upon a limited sampling of data and could be improved by acquiring more training sets to build more granular models in the case of the charge state and adduct scores.

5 Conclusions

In this study, we demonstrated the advantages of our application of Laplacian Regularization to smooth LC-MS assignments of glycan compositions across multiple experimental protocols (Hu and Mechref (2012); Khatri *et al.* (2016a)). Our algorithm's performance is competitive with existing tools for analyzing the same type of data, while being more flexible.

All of the methods demonstrated in this paper are available as part of the open source, cross-platform glycomics and glycoproteomics software `GlycReSoft`, freely available at http://www.bumc.bu.edu/msr/glycresoft/.

References

Akune, Y., Lin, C.-H., Abrahams, J. L., Zhang, J., Packer, N. H., Aoki-Kinoshita, K. F., and Campbell, M. P. (2016). Comprehensive analysis of the N-glycan biosynthetic pathway using bioinformatics to generate UniCorn: A theoretical N-glycan structure database. *Carbohydrate Research*, **431**, 56–63.

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, **7**(2006), 2399–2434.

Ceroni, A., Maass, K., Geyer, H., Geyer, R., Dell, A., and Haslam, S. M. (2008). GlycoWorkbench: A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans. *Journal of Proteome Research*, **7**(4), 1650–1659.

Frank, M. and Schloissnig, S. (2010). Bioinformatics and molecular modeling in glycobiology. *Cellular and Molecular Life Sciences*, **67**(16), 2749–2772.

Goldberg, D., Bern, M., North, S. J., Haslam, S. M., and Dell, A. (2009). Glycan family analysis for deducing N-glycan topology from single MS. *Bioinformatics*, **25**(3), 365–371.

Hipp, R. and Et Al. (2016). SQLite.

Hu, Y. and Mechref, Y. (2012). Comparing MALDI-MS, RP-LC-MALDI-MS and RP-LC-ESI-MS glycomic profiles of permethylated N-glycans derived from model glycoproteins and human blood serum. *Electrophoresis*, **33**(12), 1768–1777.

Hu, Y., Shihab, T., Zhou, S., Wooding, K., and Mechref, Y. (2016). LC-MS/MS of permethylated N-glycans derived from model and human blood serum glycoproteins. *ELECTROPHORESIS*, **37**(11), 1498–1505.

Jaitly, N., Mayampurath, A., Littlefield, K., Adkins, J. N., Anderson, G. A., and Smith, R. D. (2009). Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC bioinformatics*, **10**(1), 87.

Kaur, P. and O'Connor, P. B. (2006). Algorithms for automatic interpretation of high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, **17**(3), 459–468.

Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008). ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics*, **24**(21), 2534–2536.

Khatri, K., Klein, J. A., White, M. R., Grant, O. C., Lemarie, N., Woods, R. J., Hartshorn, K. L., Zaia, J., Leymarie, N., Woods, R. J., Hartshorn, K. L., and Zaia, J. (2016a). Integrated omics and computational glycobiology reveal structural basis for Influenza A virus glycan microheterogeneity and host interactions. *Molecular & cellular proteomics : MCP*, **13975**(615).

Khatri, K., Klein, J. A., and Zaia, J. (2016b). Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation. *Analytical and Bioanalytical Chemistry*.

Kronewitter, S. R., Slysz, G. W., Marginean, I., Hagler, C. D., LaMarche, B. L., Zhao, R., Harris, M. Y., Monroe, M. E., Polyukh, C. A., Crowell, K. L., Fillmore, T. L., Carlson, T. S., Camp, D. G., Moore, R. J., Payne, S. H., Anderson, G. a., and Smith, R. D. (2014). GlyQ-IQ: Glycomics quintavariate-informed quantification with high-performance computing and glycogrid 4D visualization. *Analytical Chemistry*, **86**(13), 6268–6276.

Liu, X., Inbar, Y., Dorrestein, P. C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J. P., Bafna, V., and Pevzner, P. A. (2010). Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Molecular & cellular proteomics : MCP*, **9**(12), 2772–2782.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-a., and Deutsch, E. W. (2011). mzML–a community standard for mass spectrometry data. *Molecular & cellular proteomics : MCP*, **10**(1), R110.000133.

Maxwell, E., Tan, Y., Tan, Y., Hu, H., Benson, G., Aizikov, K., Conley, S., Staples, G. O., Slysz, G. W., Smith, R. D., and Zaia, J. (2012). GlycReSoft: a software package for automated recognition of glycans from LC/MS data. *PloS one*, **7**(9), e45474.

Peltoniemi, H., Natunen, S., Ritamo, I., Valmu, L., and Räbinä, J. (2013). Novel data analysis tool for semiquantitative LC-MS-MS2 profiling of N-glycans. *Glycoconjugate journal*, **30**(2), 159–70.

Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, **6**(4), 229–233.

Stanley, P., Schachter, H., and Taniguchi, N. (2009). *N-Glycans*. Cold Spring Harbor Laboratory Press.

Varki, A. (2017). Biological roles of glycans. *Glycobiology*, **27**(1), 3–49.

Yu, C.-Y. C.-Y., Mayampurath, A., Hu, Y., Zhou, S., Mechref, Y., and Tang, H. (2013). Automated annotation and quantification of glycans using liquid chromatography-mass spectrometry. *Bioinformatics*, **29**(13), 1706–1707.

Yu, T. and Peng, H. (2010). Quantification and deconvolution of asymmetric LC-MS peaks using the bi-Gaussian mixture model and statistical model selection. *BMC bioinformatics*, **11**(1), 559.

Zaia, J. (2008). Mass spectrometry and the emerging field of glycomics. *Chemistry & biology*, **15**(9), 881–92.