Abstract

Motivation: Glycosylation is one of the most heterogenous and complex post-translational modifications, but.
Results: These are the resutls for this article.

Application of Network Smoothing to Glycan LC-MS Profiling

Joshua Klein

October 4, 2017

1 Introduction

Glycosylation is one of the most pervasive forms of post-translational modification (Varki (2017)).

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 glycosidase and glycosyltransferase reactions, the enumeration of which as shown in Akune *et al.* (2016) yields over a million of possible N-glycan topologies and epitopes. 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 observe N-glycan does not capture the full tree or graph structure of an N-glycan , so we reduce the topology to a count of each type of residue.

Starting with the core motif, we generate all combinations of monosaccharides ranging between the limits in Table 1. We created a copy of this database for native, reduced and permethylated, and deuteroreduced and permethylated. Let $n = 1240$ be the number of glycan compositions $\mathbf{g}$ in the database.

2.2 LC-MS Data Preprocessing

We used 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. All data were converted into 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. Scans were then subjected to iterative charge state deconvolution and deisotoping 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 cluster peaks whose neutral masses are within 15 parts-per-million error (PPM) of each other. When there are multiple candidate clusters for a single peak, we use the cluster with the lowest mass error. After all peaks are clustered, we sort each cluster by time, creating a list of aggregated chromatograms. To account for small mass differences, we find all chromatograms which are within 10 PPM of each other and which overlap in time and merge them.

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.* (2016)
20141031-07-Phil-82	QTOF	Native	Formate(3)	Khatri *et al.* (2016)
20141101-04-Phil-BS	QTOF	Native	Formate(3)	Khatri *et al.* (2016)
20141128-11-Phil-82[1]	QTOF	Deutero-reduced and Permethylated	Ammonium (3)	Khatri *et al.* (2016)
AGP-DR-Perm-glycans-1[1]	FTMS	Deutero-reduced and Permethylated	Ammonium (3)	Khatri *et al.* (2016)
AGP-permethylated-2ul-inj-55-SLens[1]	FTMS	Reduced and Permethylated	Ammonium (3)	Khatri *et al.* (2016)
Perm-BS-070111-04-Human-Serum[1]	FTMS	Reduced and Permethylated	Ammonium (3)	Yu *et al.* (2013)

[1] Included MS^n Scans

2.4 Glycan Composition Matching

For each chromatogram, we queried the target 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 target 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 chromatogramss 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. For cases where MS^n scans were present, these scans were mapped onto their precursor MS^1 features, and evaluated for glycan-like product ions, retaining only those which satisfied the signature ion criterion described in section S**??**. We term the remaining assigned and unassigned chromatograms *candidate features*.

2.5 Feature Evaluation

For each candidate feature, we compute 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 S**??**.

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 compostions 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)). This idea has been explored more generically under the name "Manifold Regularization" (Belkin *et al.* (2006)) and specifically "Laplacian Regularization" when the Laplacian matrix of a graph is used to influence the parameter scaling. We apply this idea to weighted networks of related glycans with arbitrarily defined and overlapping sub-populations.

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

2.6.1 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 the following neighborhoods for *N*-glycans:

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.

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

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. 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.

2.6.4 Parameter Estimation

We model the relationship between $\mathbf{s}$, $\phi_\mathbf{o}$, and τ as a set of gaussian distribution.

$$(\mathbf{s}|\phi_\mathbf{o},\tau) \sim \mathcal{N}(\phi_\mathbf{o},\Sigma) \tag{7}$$

$$\Sigma = \rho\mathbf{I} \tag{8}$$

$$\left(\begin{bmatrix}\phi_\mathbf{o}\\ \phi_\mathbf{m}\end{bmatrix}\middle|\tau\right) \sim \mathcal{N}(\mathbf{A}\tau, \lambda^{-1}\mathbf{L}^-) \tag{9}$$

$$(\phi_\mathbf{o}|\tau) \sim \mathcal{N}\left(\mathbf{A_o}\tau, \Sigma_{\phi_o}\right) \tag{10}$$

$$\Sigma_{\phi_o} = \lambda^{-1}\left(\mathbf{L_{oo}}-\mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)^{-1} \tag{11}$$

$$\tau \sim \mathcal{N}\left(0,\sigma^2\mathbf{I}\right) \tag{12}$$

Fully expanded, this becomes

$$\begin{bmatrix}\mathbf{s}\\ \phi_\mathbf{o}\\ \tau\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\Sigma+\Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \sigma^2\mathbf{A_o}\\ \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \sigma^2\mathbf{A_o}\\ \sigma^2\mathbf{A_o}^t & \sigma^2\mathbf{A_o}^t & \sigma^2\mathbf{I}\end{bmatrix}\right) \tag{13}$$

We can form the conditional distribution $\tau|\mathbf{s}$ which has a mean

$$\mu_{\tau|\mathbf{s}} = 0 + (\sigma^2\mathbf{A_o}^t)\left(\Sigma+\Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o^t}\right)^{-1}\mathbf{s} \tag{14}$$

$$= \mathbf{A_o}^t\left(\tilde{\rho}\mathbf{I}+\frac{1}{\tilde{\lambda}}\mathbf{L_{oo}^-}+\mathbf{A_o}\mathbf{A_o^t}\right)^{-1}\mathbf{s} \tag{15}$$

We assume that $\sigma^2 \gg 1$, and treat λ and ρ as relative to σ^2, as $\tilde{\rho}$ and $\tilde{\lambda}$. This model gives us an estimate for τ given a value for ρ and λ. As ρ has no direct role in the central tendency of ϕ or $\mathbf{s}$, we choose to fix the value of $\tilde{\rho}=0.1$, which leaves only $\tilde{\lambda}$. We estimate the optimal $\tilde{\lambda}$ by grid search, minimizing the predicted residual error sum of squares (PRESS) statistic.

$$\arg\min_{\tilde{\lambda}} \frac{\mathbf{s}-\hat{\phi}_\mathbf{o}}{\left(1-\left(\mathbf{I}+\tilde{\lambda}\mathbf{L}\right)^{-1}\right)^2} \tag{16}$$

This formulation depends upon the value of $\mathbf{s}$ and is sensitive to low scoring matches, which can lead to incorrect estimates of τ and PRESS. We therefore perform a grid search over both $\tilde{\lambda}$ and a minimum threshold for $\mathbf{s}$, γ.

As we increase γ we remodel the graph $\mathcal{G}$, removing nodes whose score is below γ. For each pair of neighbors of removed node g_m, (g_u, g_v), if $L_1(g_u,g_v) > L_1(g_u,g_m)+L_1(g_m,g_v)$, we add an edge from g_u to g_v with weight $\frac{1}{L_1(g_u,g_m)+L_1(g_m,g_v)}$, up to a limit of $L_1(g_k,g_m)<5$. We give the result of this grid search the name $\mathbf{r}$. At each point, on the grid, we save the value of τ in $r_{\lambda_i,\gamma_j,\tau}$ and the PRESS in $r_{\lambda_i,\gamma_j,PRESS}$. To select the optimal parameters, we traverse the grid along γ, computing τ_γ:

$$\bar{\lambda}_j = \arg\min_{\lambda_i} r_{\lambda_i,\gamma_j,PRESS} \tag{17}$$

$$\tau_{\gamma_j} = |r_{\bar{\lambda}_j,\gamma_j,\tau}| * \left(\frac{\gamma_j}{b}+(1-\frac{1}{b})\right) \tag{18}$$

where b is a bias factor defining how much weight to give to higher values of γ which correspond to networks made up of higher confidence assignments. We chose $b = 4$. We define $\bar{\tau}_\gamma = \max \tau_\gamma$ and define the vector $\bar{\gamma} = \left[\gamma_j \leftarrow \tau_{\gamma_j} \geq \bar{\tau}_\gamma * 0.9\right]$. This favors values of γ where large values of τ are selected, meaning that the neighborhoods are well populated, while also giving an estimate for $\tilde{\lambda}$ that is non-zero. We term the values of γ in $\bar{\gamma}$ the *target thresholds* of $\mathbf{s}$.

To estimate $\tilde{\lambda}$ and τ from these results, we select the columns of the grid $\mathbf{r}$ at each $\gamma_j \in \bar{\gamma}$.

$$\bar{\tau}_\gamma = \max \tau_\gamma \tag{19}$$

$$\bar{\gamma} = \left\{\gamma_j \leftarrow \tau_{\gamma_j} \geq \bar{\tau}_\gamma * 0.9\right\} \tag{20}$$

$$\bar{\lambda} = \left\{\bar{\lambda}_j \leftarrow \gamma_j \in \bar{\gamma}\right\} \tag{21}$$

$$\mathbf{s}_{\gamma_j} = \left\{s_i \leftarrow s_i > \gamma_j\right\} \tag{22}$$

$$\bar{\tau}_\mathbf{j} = \mu_{\tau|\mathbf{s}_{\gamma_j}, \bar{\lambda}_j} \tag{23}$$

$$\hat{\lambda} = \frac{1}{|\bar{\lambda}|} \sum_j \bar{\lambda}_j \tag{24}$$

$$\hat{\tau} = \frac{1}{|\bar{\tau}|} \sum_j \bar{\tau}_\mathbf{j} \tag{25}$$

$$\hat{\gamma} = \frac{1}{|\bar{\gamma}|} \sum_j \bar{\gamma}_j \tag{26}$$

where $\mathbf{s}_{\gamma_j}$ is the set of observed scores which are greater than γ_j, but where the estimation of is carried out with the complete Laplacian $\mathbf{L}$, not the reduced network used to compute $\mathbf{r}$. This set of averaged estimates of $\hat{\lambda}$ and $\hat{\tau}$ are then used to estimate $\hat{\phi}_o$ by (6).

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 their mzXML alternative did not produce matches consistent with their previously published results on *Perm-BS-070111-04-Human-Serum*, and when ran on *AGP-permethylated-2ul-inj-55-SLens* it ran out of memory. GlyQ-IQ was made available for testing by its authors, but required data be in Thermo-Fisher's binary format, it 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 *20141101-04-Phil-BS* and *Perm-BS-070111-04-Human-Serum* . Please refer to section S**??** 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 *20141101-04-Phil-BS*

The fitted parameters for the network constructed for *20141101-04-Phil-BS* are shown in Table 4. 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.* (2016)).

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 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.

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

The fitted parameters for the network constructed for *Perm-BS-070111-04-Human-Serum* are shown in Table 5. The assigned chromatograms are shown in Figure 3.

Table 4: Fitted λ, γ, and τ for *20141101-04-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

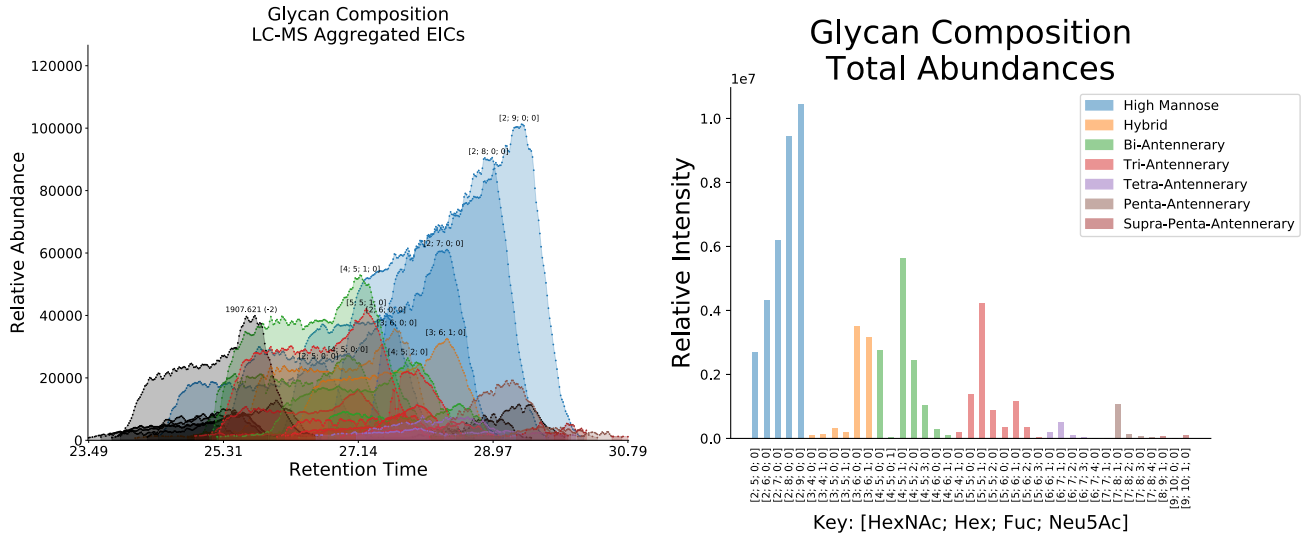
Fitted $\lambda = 0.99$ and $\gamma = 11.12$.

Table 5: Fitted λ, γ, and τ for *Perm-BS-070111-04-Human-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 *20141101-04-Phil-BS*



The comparison of assignment performance with differing degrees of smoothing is shown in Figure 4. The ROC AUC for the unregularized condition is 0.719, for the partially regularized condition is 0.835, and for the fully regularized condition is 0.791. The Precision-Recall AUC for the unregularized condition is 0.834, for the partially regularized condition is 0.896, and for the fully regularized condition is 0.852. 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^nthis cannot be resolved.

4 Discussion

We demonstrate that this regularization method can improve the sensitivity and specificity of glycan composition assignment for LC-MS based experiments. The method we proposed uses 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.

Figure 2: Performance Comparison with and without Network Smoothing for *20141101-04-Phil-BS*

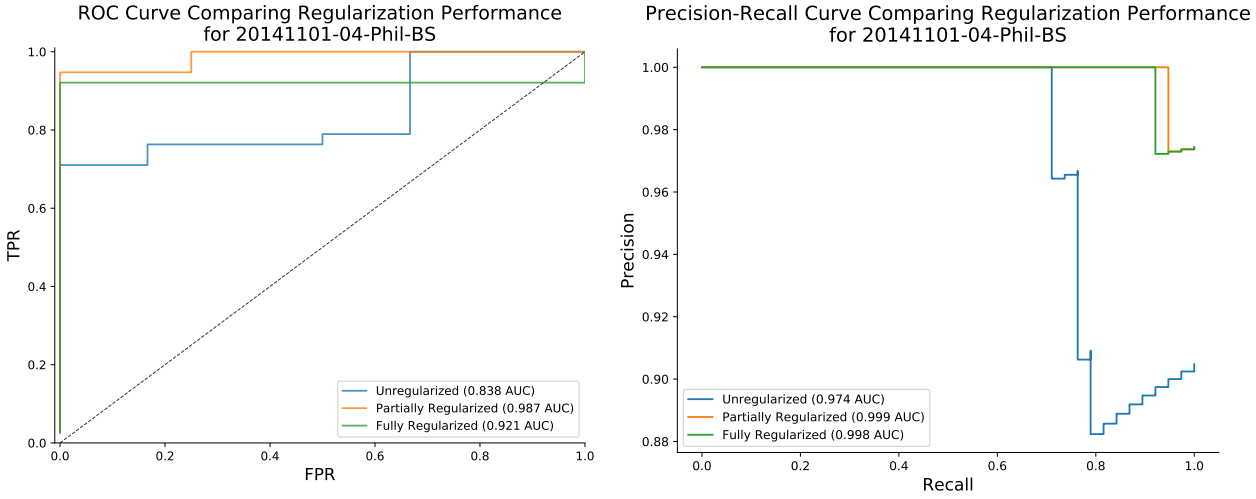
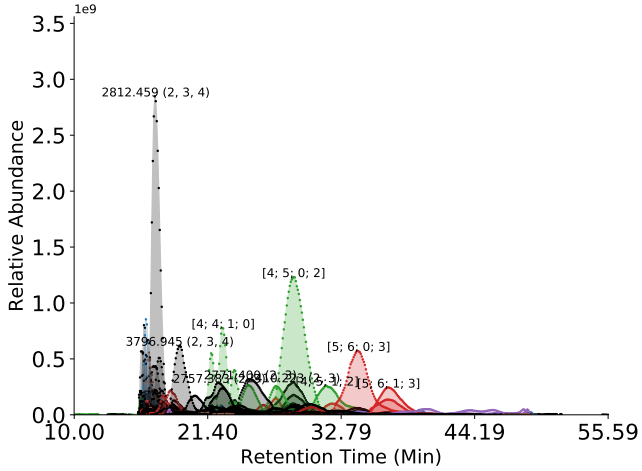
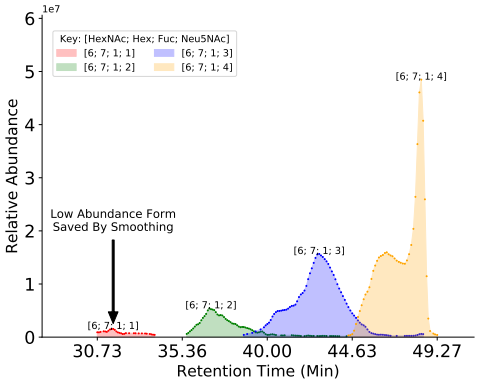


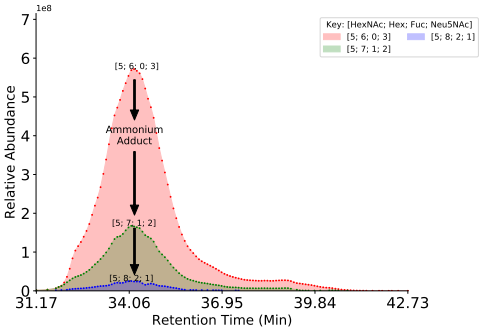
Figure 3: Chromatogram Assignments for *Perm-BS-070111-04-Human-Serum*



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



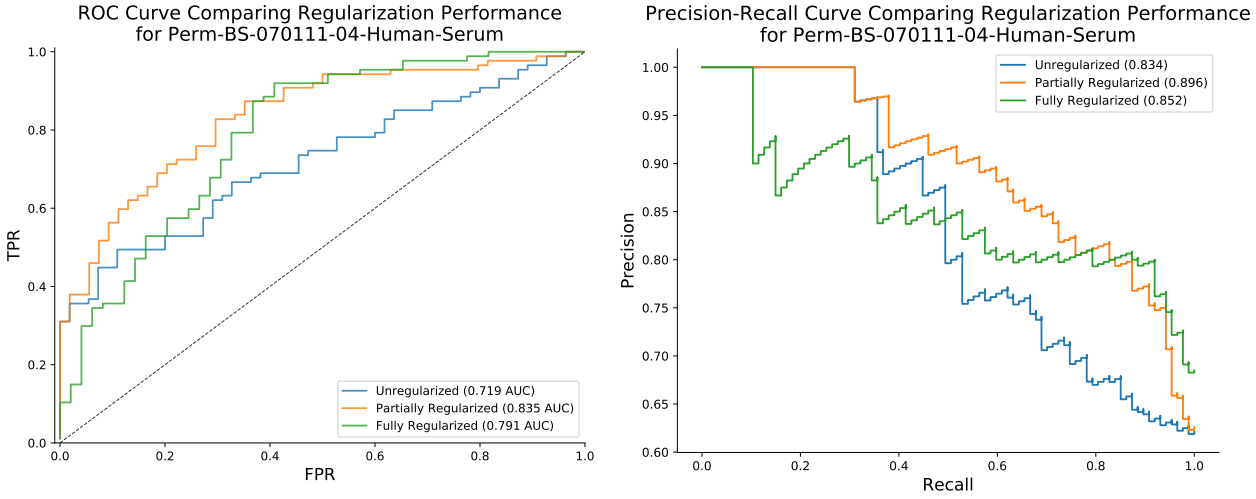
(b) Low scoring features which may be discarded based on individual evidence alone may be more reasonable to accept given evidence from related composition, such as our network smoothing method.



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

The experimental results from the original analysis of *20141101-04-Phil-BS* and *20141031-07-Phil-82* 82 demonstrated that while both strains expressed predominantly high-mannose glycosylation, *20141101-04-Phil-BS* expressed more larger complex-type structures (Khatri *et al.* (2016)). In our findings, we recapitulate these results while reducing the number of false positive assignments. There are substantial 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 an avian tissue, we may be able to observe larger branching patterns than are observed in normal mammalian tissue (Stanley *et al.* (2009)). There is evidence for this in the *20141101-04-Phil-BS* with HexNAc9Hex10-based compositions suggesting a seven branch pattern, though this cannot be determined without high

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



quality MS^n. 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.

We reproduce the majority of the glycan assignments from Yu *et al.* (2013). however 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, 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 this 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 NeuAc**, 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 Fuc** 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 match in the time range between 10 and 22 minutes, and similar glycan compositions that are 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 eligble 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.

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/.

5 Conclusions

In this study, we demonstrated the advantages of our application of Laplacian Regularization to smooth LC-MS assignments of glycan compositions

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.

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.

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.

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. (2016). Integrated omics and computational glycobiology reveal structural basis for Influenza A virus glycan microheterogeneity and host interactions. *Molecular & cellular proteomics : MCP*, **13975**(615).

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.

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.