Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

Systems Biology

Application of Network Smoothing to Glycan LC-MS Profiling

Joshua Klein[1], Luis Carvalho[1,2], and Joseph Zaia[1,3,*]

[1]Program for Bioinformatics, Boston University
[2]Department of Math and Statistics, Boston University
[3]Department of Biochemistry, Boston University

[1]Program for Bioinformatics, Boston University, Boston, Post Code, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Glycosylation is one of the most heterogeneous and complex protein post-translational modifications. Liquid chromatography coupled mass spectrometry (LC-MS) is a common high throughput method for analyzing complex biological samples. Accurate study of glycans require high resolution mass spectrometry. Mass spectrometry data contains intricate sub-structures that encode mass and abundance, requiring several transformations before it can be used to identify biological molecules, requiring automated tools to analyze samples in a high throughput setting. Existing tools for interpreting the resulting data do not take into account related glycans when evaluating individual observations, limiting their sensitivity.
Results: We developed an algorithm for assigning glycan compositions from LC-MS data by exploring biosynthetic network relationships among glycans. Our algorithm optimizes a set of likelihood scoring functions based on glycan chemical properties but uses network Laplacian regularization and optionally prior information about expected glycan families to smooth the likelihood and thus achieve a consistent and more representative solution. Our method was able to identify as many, or more glycan compositions compared to previous approaches, and demonstrated greater sensitivity with regularization. Our network definition was tailored to *N*-glycans but the method may be applied to glycomics data from other glycan families like *O*-glycans or heparan sulfate where the relationships between compositions can be expressed as a graph.
Availability:
Built Executable: http://www.bumc.bu.edu/msr/glycresoft/
Source Code: https://github.com/BostonUniversityCBMS/glycresoft
Contact:
jzaia@bu.edu

1 Introduction

Glycosylation modulates the structures and functions of proteins and lipids in a broad class of biological processes (Varki (2017)). Accurate mass measurement defines monosaccharide composition given assumptions regarding glycan class and biosynthesis (Zaia (2008)). For unseparated mixtures, mass spectrometry analysis determines the mass-to-charge ratio values for only the most abundant glycans; dynamic range for detection of glycans is poor because of ion suppression (Peltoniemi *et al.* (2013)). By contrast, online separations coupled with mass spectrometry improve dynamic range and reproducibility of glycan analysis, at the cost of increased analysis time and workflow 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. These programs address instrument-specific signal processing requirements. For example SysBioWare (Frank and Schloissnig (2010)) performs sophisticated baseline removal prior to fitting peaks, while GlyQ-IQ (Kronewitter *et al.* (2014)) was written for cleaner Fourier Transform MS (FTMS) that does not require such a baseline removal step. Tools that build on the THRASH implementation from Decon2LS (Jaitly *et al.* (2009); Yu *et al.* (2013); Maxwell *et al.* (2012)) are 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 and/or organisms, limiting their vocabulary of possible monosaccharides to just those commonly found in that subgroup (Yu *et al.* (2013); Kronewitter *et al.* (2014); Peltoniemi *et al.* (2013); Goldberg *et al.* (2009)). Often, these tools are tailored for analysis of a particular derivatization state, adduction conditions, or neutral loss pattern (Yu *et al.* (2013); Peltoniemi *et al.* (2013); Maxwell *et al.* (2012)). Work has been done to construct a standardized namespace and representation for glycans, glySpace, including both structures and compositions (Tiemeyer *et al.* (2017); Campbell *et al.* (2014)). This data is publicly accessible, including a programmatic query interface using SPARQL over HTTPS (Aoki-Kinoshita *et al.* (2015)). Tools that can communicate with these services have the potential to lead researchers to find deeper connections from cross-referenced information, and other researchers can more readily find and use their work.

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. Several methods have been developed using different facets of the observed data. Yu *et al.* (2013) used the isotopic pattern goodness-of-fit while Peltoniemi *et al.* (2013) used intensity features of associated MS^n scans to evaluate partial structure and composition match quality. Kronewitter *et al.* (2014) combined several features of the MS^1 evidence, including elution profile peak shape goodness-of-fit, isotopic fit, mass accuracy, scan count, and in-source fragmentation correlation. Some of these methods are well-defined and invariant from instrument to instrument in this era of high resolution mass spectrometry, but others are tightly coupled to the experimental equipment. Missing from this list are methods to target a glycan's intrinsic properties, such as charge state distribution or facility in acquiring adducts, which can increase the number of spurious assignments if not considered. We propose a new scoring function which is able to combine those properties which are independent of experimental setup with these glycan-aware features.

As observed by Goldberg *et al.* (2009), there is also value in including related glycan composition identifications in how much confidence one assigns to a given glycan composition assignment. They used a method to exploit the known biosynthetic rules of *N*-glycans to connect peaks in a MALDI spectrum assigned to a particular *N*-glycan by intact mass alone. Their method using the maximum weighted subgraph of the biosynthetic network had demonstrably better performance than chance with their expert system annotation method. Kronewitter *et al.* (2014) considered a similar idea with more emphasis on handling in-source fragmentation 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. We also apply our method using three different glycan composition search spaces to show how the underlying database can influence results. We present our method on typical *N*-glycans in humans, though our method can be applied to any variety of glycan composition whose monosaccharides can be described using IUPAC trivial names or or whose components can be described in terms of chemical formulae.

Table 1. Human N-glycan Composition Bounds Stanley et al. (2009)

Monosaccharide	Lower Limit	Upper Limit	Constraints
HexNAc	2	9	
Hex	3	10	
Fuc	0	4	$\mathbf{HexNAc} > \mathbf{Fuc}$
NeuAc	0	5	$(\mathbf{HexNAc} - 1) > \mathbf{NeuAc}$

2 Methods

2.1 Glycan Hypothesis Generation

In eukaryotes, a 14 monosaccharide *N*-glycan of composition **HexNAc2 Hex12** is transferred to a newly synthesized protein in the endoplasmic reticulum by the oligosaccharyl transferase protein complex. This glycan is trimmed to **HexNAc2 Hex9** during protein folding and quality control. As the glycoprotein transits the Golgi apparatus, *N*-glycans are trimmed to **HexNAc2 Hex5** before being elaborated into hybrid and complex *N*-glycan classes (Stanley *et al.* (2009)). Glycan structures are refined by a series of reactions that yield over a million possible *N*-glycan topologies, as shown in Akune *et al.* (2016). These topologies define the glycan's geometry and protein binding properties. Neither MS^1 nor collisional tandem MS of glycans can capture the full tree or graph structure of an *N*-glycan, so we reduced the topology to a count of each type of residue, a composition.

Starting with the core motif **HexNAc2 Hex3**, we generated all combinations of monosaccharides ranging between the limits in Table 1 to build a human *N*-glycan composition database, which produced 1240 distinct compositions. These rules are able to efficiently generate all glycan compositions from canonical branching patterns and lactosamine extensions, as well as rarer constructs such as LacdiNAc Goldberg *et al.* (2009) at the cost of including some wholly improbable compositions. To perform a side-by-side comparison we also extracted the glycan list from Yu *et al.* (2013) derived from the biosynthetic rules in Krambeck and Betenbaugh (2005) with 319 compositions, and another database using all curated *N*-glycans from glySpace via GlyTouCan (Tiemeyer *et al.* (2017)) containing only **[Hex, HexNAc, Fuc, Neu5Ac, sulfate]**, with 275 distinct compositions. As previous analysis of Influenza A virus samples detected sulfated *N*-glycans (Khatri *et al.* (2016)), we also created a combinatorial database with up to one sulfate included, for a total of 2480 compositions. As our algorithm treats **HexNAc** and **HexNAc(S)** as distinct entities, for all monosaccharides with post-attachment substituents such as **sulfate** and **phosphate**, we detached the substituent from the core monosaccharide. Our implementation is able to interpret IUPAC trivial names and compositions thereof with standard substituent and unambiguous backbone modifications, permitting a wide range of possible glycan compositions.

2.2 LC-MS Data Preprocessing

We analyzed samples from several sources, including both Quadrupole Time-of-Flight (QTOF) and Orbitrap instruments as shown in Table S1. For details on sample preparation and data acquisition, please see the source citations in the referenced table. We converted all datasets to mzML format (Martens *et al.* (2011)) using 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 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 $\delta_{mass} = 15$ 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 were within $\delta_{mass} = 10$ PPM of each other and which overlap in time and merge them. These mass tolerances were selected empirically, and can be adjusted as needed by the user.

2.4 Glycan Composition Matching

For each chromatogram, we searched each glycan database for compositions whose masses were within $\delta_{mass} = 10$ PPM for QTOF data, 5 PPM for FTMS data. We merged all chromatograms matching the same composition. Then, for each mass shift combination, we searched each glycan database for compositions whose neutral mass were within δ_{mass} of the observed neutral mass - mass shift 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 chromatograms with fewer than $k = 5$ data points. The same chromatogram may be given multiple assignments and designated multiple mass shifts, and chromatograms without glycan assignments may use chromatograms with glycan assignments as mass shifted components. This ambiguity information was propagated through each merge and split step. We termed these remaining assigned and unassigned chromatograms *candidate features*.

2.5 Feature Evaluation

We computed several metrics to estimate how distinguishable each candidate feature was from random noise. The metrics are mentioned in List 1, but for more information see Section S2.

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 missing peaks and interference.
5. Adduction states with respect to glycan composition and mass.

These metrics are bounded in $(-\infty, 1)$. Any observation for which any metric was observed below a feature specific threshold 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 $(-\infty, \infty)$, whose distribution we assume is asymptotically normal. A value of $s < 8$ 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 depend upon glycan abundance, sample quality and mass spectrometer resolution.

2.6 Glycan Composition Network Smoothing

Ideally, each glycan present in a sample under analysis would produce sufficient experimental evidence that they can be identified. In practice, glycan compositions with lower abundances may not present strong evidence, leading to those glycan compositions being discarded. Others have demonstrated that 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)). To improve performance, 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 its evaluation of glycan composition feature matching.

Previous approaches to using information regarding identification of one glycan composition to increase the confidence in another have been proposed by Goldberg *et al.* (2009) and Kronewitter *et al.* (2014) using different techniques. Goldberg et al. used random walks along the biosynthetic network between identified glycan compositions to increase the confidence of those connected compositions. This method works well but requires that the parameters of the random walk be properly tuned for the biosynthetic network being used. Laplacian regularized least squares is more robust to small changes to the network and is able to use the entire network. Kronwitter et al. included a term in their criterion for detection requiring the presence of another glycan composition with one more or one less monosaccharide to permit identification. This puts substantial weight on a boolean term, giving it the ability to overrule other experimental evidence. Similar methods could be devised using methods like ant colony optimization to traverse the biosynthetic graph, or a a database-specific belief network, but these methods would require considerable manual tuning for each new database to be tested.

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 $\mathcal{Z}^{+c}$ space where c is the number of components in any glycan composition, 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, representing the biosynthetic distance between the two compositions, an analog for the number of enzymatic steps needed to go from one glycan to the other. 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 how larger collections of glycan compositions are related. To this end, we extend the definition of neighborhoods for *N*-glycans using intervals over monosaccharide counts shown in Table 2. These neighborhoods are arranged to span particular epitopes or biosynthetically related subtypes of *N*-glycans, such as sialylation state or branching pattern. Neighborhoods overlap sets of glycan compositions which are also biosynthetically related. Each neighborhood spans the eponymous class of glycan compositions, as well as the preceding class and proceeding class. For example, For example, the Tri-Antennary neighborhood spans Bi-Antennary Tetra-Antennary compositions. This helps to channel the estimation of τ among related groups. The Hybrid, Bi-Antennary and Asialo-Bi-Antennary neighborhoods introduce complications because they are biosynthetically close to each other. For the

Name	HexNAC Min	HexNAC Max	Hex Min	Hex Max	NeuAc Min	NeuAc Max	Size
High Mannose	2	2	3	10	0	0	16
Hybrid	2	4	2	6	0	2	80
Bi-Antennary	3	5	3	6	1	3	104
Asialo-Bi-Antennary	3	5	3	6	0	1	96
Tri-Antennary	4	6	4	7	1	4	172
Asialo-Tri-Antennary	4	6	4	7	0	0	56
Tetra-Antennary	5	7	5	8	1	5	240
Asialo-Tetra-Antennary	5	7	5	8	0	0	60
Penta-Antennary	6	8	6	9	1	5	280
Asialo-Penta-Antennary	6	8	6	9	0	0	60
Hexa-Antennary	7	9	7	10	1	6	300
Asialo-Hexa-Antennary	7	9	7	10	0	0	60
Hepta-Antennary	8	10	8	11	1	7	150
Asialo-Hepta-Antennary	8	10	8	11	0	0	30

Table 2. N-Glycan Neighborhood Definitions. These define the ranges of monosaccharides which will be used to classify a glycan composition as being a member of each neighborhood, and the number of Combinatorial N-glycan compositions in each neighborhood.

simplicity, we chose to include all of Hybrid in Asialo-Bi-Antennary and permit up to one NeuAc in it.

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 $A_{i,k}$ is the degree to which g_i belongs kth neighborhood:

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

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

To accomplish our goal, we can use Laplacian regularized least squares to find a new score ϕ, based upon s and relationships among the observed glycans described by our biosynthetic graph $\mathcal{G}$. These relationships can be directed to move towards some central tendency τ using the Laplacian of $\mathcal{G}$ and some definitions of broad groups in $\mathcal{G}$.

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

$$\mathcal{S}(\mathbf{L}, \phi, \tau) = \begin{bmatrix}\phi_o - \tau_o, & \phi_m - \tau_m\end{bmatrix} \begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}} \\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix} \begin{bmatrix}\phi_o - \tau_o \\ \phi_m - \tau_m\end{bmatrix} \quad (2)$$

$$\ell = (\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda\mathcal{S}(\mathbf{L}, \phi, \tau) \quad (3)$$

Neighborhood τ	Phil-BS Combinatorial + Sulfate	Phil-BS glySpace	Phil-BS Krambeck	Serum Combinatorial	Serum glySpace	Serum Krambeck
high-mannose	18.008	15.061	17.089	20.328	19.392	19.720
hybrid	13.440	12.435	12.503	20.997	18.610	20.056
bi-antennary	0.000	0.000	0.000	15.901	16.826	17.593
asialo-bi-antennary	14.078	10.916	13.591	22.585	21.563	21.827
tri-antennary	0.000	0.000	0.000	26.420	19.605	23.644
asialo-tri-antennary	14.538	6.565	11.952	20.025	21.128	19.764
tetra-antennary	0.000	0.000	0.000	19.508	18.542	17.674
asialo-tetra-antennary	14.331	4.842	12.373	2.472	7.180	2.568
penta-antennary	0.000	0.000	0.000	11.878	15.035	11.682
asialo-penta-antennary	11.588	1.255	9.784	0.000	0.000	0.000
hexa-antennary	0.000	0.000	0.000	0.000	0.000	0.000
asialo-hexa-antennary	11.094	3.883	13.223	0.000	0.000	0.000
hepta-antennary	0.000	0.000	0.000	0.000	0.000	0.000
asialo-hepta-antennary	3.117	1.529	2.703	0.000	0.000	0.000
$\hat{\lambda}$	0.99	0.69	0.99	0.99	0.99	0.99
$\hat{\gamma}$	11.39	14.60	10.42	20.57	18.42	20.72

Table 3. Estimated values of smoothing parameters τ, λ, and γ for each dataset and database

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:

$$0 = \frac{\partial \ell}{\partial \phi_m}\left((\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda\mathcal{S}(\mathbf{L}, \phi, \tau)\right) \quad (4)$$

$$\hat{\phi}_m = -\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o - \tau_o) + \tau_m \quad (5)$$

and w.r.t. ϕ_o

$$0 = \frac{\partial \ell}{\partial \phi_o}\left((\mathbf{s} - \phi_\mathbf{o})^t(\mathbf{s} - \phi_\mathbf{o}) + \lambda\mathcal{S}(\mathbf{L}, \phi, \tau)\right) \quad (6)$$

$$\hat{\phi}_o = \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right]^{-1}(\mathbf{s} - \tau_o) + \tau_o \quad (7)$$

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.

3 Results

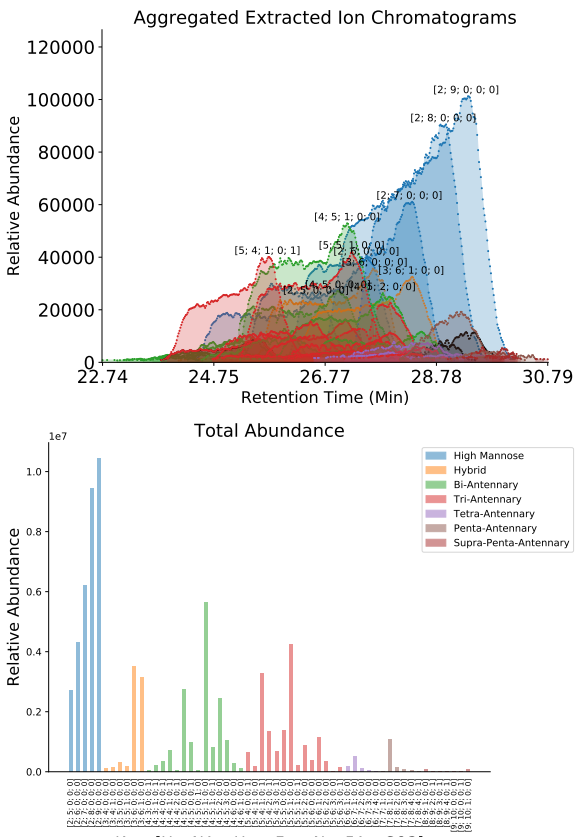
We demonstrated the performance of our algorithm using released influenza hemagglutinin data set *20141103-02-Phil-BS* and a serum glycan data set *Perm-BS-070111-04-Serum*. Please refer to section S6 for all other datasets. For each comparison, the unregularized case is not smoothed, effectively $\lambda = 0$, 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 3. 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 for each database are shown in Figure 2 and Table 4. We used the Receiver Operator Characteristic (ROC) Area Under the Curver

Fig. 1: Chromatogram Assignments and Quantification for *20141103-02-Phil-BS* Using the *Combinatorial + Sulfate* database. . The Retention Time (Min) axis shows the retention time in minutes from the experiment, and the Relative Abundance axis shows the intensity of the signal from the particular aggregated ion species for each analyte. The identified glycan compositions are labeled with a tuple describing the number of each component of the form [**HexNAc**, **Hex**, **Fuc**, **NeuAc**, **SO3**]



(AUC) to measure performance, using manually validated compositions as ground truth. We observed the greatest number of assignments using the *Combinatorial + Sulfate* database, and the greatest ROC AUC in the partially regularized condition.

Name	ROC AUC	True Matches[1]
Combinatorial Unregularized	0.882	56
Combinatorial Partial	0.995	57
Combinatorial Grid	0.991	57
GlySpace Unregularized	0.811	40
GlySpace Partial	0.808	38
GlySpace Grid	0.802	31
Krambeck Unregularized	0.742	28
Krambeck Partial	0.742	29
Krambeck Grid	0.742	29
Khatri *et al.* (2016)	-	46

[1] Selected at $\phi_o > 5.0$

Table 4. Performance Comparison for 20141103-02-Phil-BS , using Receiver Operator Characteristic (ROC) Area Under the Curver (AUC). The Combinatorial Partial Regularization approach performed best.

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 3. The assigned chromatograms are shown in Figure 4.

The comparison of assignment performance with differing degrees of smoothing is shown in Figure 3. We observe the greatest number of total true identifications using the partially regularized Combinatorial database. However, the Combinatorial database also has many more false positives, with a ROC AUC of 0.816. These false positives do not appear in the biosynthetically constraind Krambeck database which maximizes its ROC AUC in the partially regularized condition at 0.883. After removing all ambiguous matches, the Krambeck database also has nearly the same number of true matches as the Combinatorial database.

4 Discussion

We demonstrated that the 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),

Name	ROCAUC	True Matches[1]	Non-Ambiguous Matches
Combinatorial Unregularized	0.679	86	61
Combinatorial Partial	0.816	87	62
Combinatorial Grid	0.804	86	61
GlySpace Unregularized	0.788	59	51
GlySpace Partial	0.803	60	52
GlySpace Grid	0.809	60	52
Krambeck Unregularized	0.866	70	60
Krambeck Partial	0.883	70	60
Krambeck Grid	0.882	69	59
Yu *et al.* (2013)	-	72[2]	59

[1] Selected at $\phi_o > 5.0$

[2] Only includes cases with sufficient MS1 scans available for comparison

Table 5. Performance Comparison for Perm-BS-070111-04-Serum , using Receiver Operator Characteristic (ROC) Area Under the Curver (AUC) and number of non-ambiguous matches. While the Krambeck database had a better ROC AUC, the Combinatorial database had more true matches.

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.* (2016)). In our findings shown in Figure 1, we recapitulate these results while reducing the number of false assignments, Table 4. 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 chicken eggs, we 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^n data. The τ fit for Phil-BS (shown) and Phil-82 (supplement) 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 observed the best performance with the *Combinatorial + Sulfate* database, which produced more than half-again as many true matches than the other two databases. It produced several false matches as well, but the smoothing process removed these while boosting the score of other low abundance matches which were consistent with higher scoring matches.

The Krambeck database performed identically in all smoothing conditions as it was only able to match the common species, not including cases that were multiply fucosylated or sulfated. It had no false matches ranked alongside its true matches so smoothing could not change its performance. The glySpace-derived database produced more true matches, but also lacked some of these more fucosylated and complex compositions. Some of the compositions included by the glySpace-derived database were lower scoring, but the chosen value of γ for that database was greater than 18, causing the fitted values of τ to omit the larger, less abundant complex-type *N*-glycans. This caused smoothing to lower the scores of these real matches rather than raise them, as with the *Combinatorial + Sulfate* database.

As we show in Figure 3, regularization improves the predictive performance of the identification algorithm on *Perm-BS-070111-04-Serum* for all databases. We reproduce the majority of the glycan assignments from Yu *et al.* (2013), but the ambiguity caused by ammonium adduction as shown in Figure 4 makes a direct comparison of composition assignment lists difficult. Our algorithm requires a minimum amount of MS1 information in order to compute a score, which some of the assignments in the original published results do not possess, and are omitted from the count in Table 5. After accounting for ambiguity, we were able to assign all of the compositions previously reported using the Krambeck database, which was used by Yu *et al.* (2013), and with the combinatorial database. The glySpace-derived database did not contain all of these compositions, but performed competitively with the combinatorial database's ROC AUC. The combinatorial database matched a small number of glycan compositions which were not in Krambeck but which were consistent with other glycan compositions observed nearby in retention time. The combinatorial database also benefited most substantially from smoothing, discarding many false positives while retaining many more true positives at the same false positive rate compared to the other databases. 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 biosynthetically valid elute later on in the experiment. Therefore a for a retention-time aware approach to evaluating glycan composition assignments, as described in Hu *et al.* (2016) could also be useful, but this is likely dependent upon the experimental workup and separation technique used.

While the biosynthetically constrained Krambeck database performed better on *Perm-BS-070111-04-Serum*, it did not contain all of the reasonably assignable glycan compositions, and it performed poorly on *20141103-02-Phil-BS* with a false negative rate of 50% compared to the combinatorial database. This is because the necessary enzymatic pathways were either not considered in the original authors' model because either the enzyme was excluded for simplicity (Krambeck *et al.* (2009)) or because the particular enzymes used were not within the scope of the model used (Spiro and Spiro (2000); Ichimiya *et al.* (2014)). This highlights the importance of selecting a good reference database, though a post-processing step such as the we described here can help mitigate using too large a database, but not a too small one.

In this work, we used the same network neighborhood imposed over different underlying sets of composition nodes, and the connectivity of those networks did not take into account the constraints of the biosynthetic process. It may be possible to obtain better performance by defining network connectivity according to concrete enzymatic relationships. This may also alter how the neighborhoods are defined and how $\mathbf{A}$ is parameterized, and in turn how τ is learned. Similarly, this procedure depends upon the scoring functions used, so selecting another set of functions for the data to fit may lead to different parameter values.

Lastly, while these case studies have demonstrated the algorithm's ability to learn network parameters from the data, an expert can define τ

and $\mathbf{A}$ themselves or obtain a model fitted on related data and apply it directly without a fitting step. An expert could use this model specification to impose prior beliefs on the evaluation process, and adjust λ to control the importance of the these beliefs. Similarly, one could also use the derivation of $\hat{\phi}_m$ to estimate the score for an unobserved glycan composition, given $\mathbf{A}$ and τ.

We used our glycoinformatics toolkit to produce a richer abstraction of glycans and monosaccharides, including producing standard-compliant textual representations of these structures and compositions. We produced a text file containing all of the glycan compositions found in the Krambeck and Combinatorial database but not the glySpace-derived database in the above samples (see supplemental information 8), and have submit it to GlyTouCan (Tiemeyer *et al.* (2017)) for registration so that future researchers can use these structures.

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.* (2016)). Our algorithm's performance is competitive with existing tools for analyzing the same type of data, with the added benefit of more flexible evaluation process and broader range of understood monosaccharides. Our tools integrate with glySpace and allows users to leverage existing glycomics repositories to build databases where applicable.

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/.

6 Acknowledgements

This work was supported by National Institute of Health 1U01CA221234

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.

Aoki-Kinoshita, K., Agravat, S., Aoki, N. P., Arpinar, S., Cummings, R. D., Fujita, A., Fujita, N., Hart, G. M., Haslam, S. M., Kawasaki, T., Matsubara, M., Moreman, K. W., Okuda, S., Pierce, M., Ranzinger, R., Shikanai, T., Shinmachi, D., Solovieva, E., Suzuki, Y., Tsuchiya, S., Yamada, I., York, W. S., Zaia, J., and Narimatsu, H. (2015). GlyTouCan 1.0 âŁ“ The international glycan structure repository. *Nucleic Acids Research*, page gkv1041.

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.

Campbell, M. P., Peterson, R., Mariethoz, J., Gasteiger, E., Akune, Y., Aoki-Kinoshita, K. F., Lisacek, F., and Packer, N. H. (2014). UniCarbKB: Building a knowledge platform for glycoproteomics. *Nucleic Acids Research*, **42**(D1), D215–21.

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.

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.

Ichimiya, T., Nishihara, S., Takase-Yoden, S., Kida, H., and Aoki-Kinoshita, K. (2014). Frequent glycan structure mining of influenza virus data revealed a sulfated glycan motif that increased viral infection. *Bioinformatics*, **30**(5), 706–711.

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

Krambeck, F. J. and Betenbaugh, M. J. (2005). A mathematical model of N-linked glycosylation. *Biotechnology and Bioengineering*, **92**(6), 711–728.

Krambeck, F. J., Bennun, S. V., Narang, S., Choi, S., Yarema, K. J., and Betenbaugh, M. J. (2009). A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology*, **19**(11), 1163–1175.

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.

Spiro, M. J. and Spiro, R. G. (2000). Sulfation of the N-linked oligosaccharides of influenza virus hemagglutinin: temporal relationships and localization of sulfotransferases. *Glycobiology*, **10**(11), 1235–42.

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

Tiemeyer, M., Aoki, K., Paulson, J., Cummings, R. D., York, W. S., Karlsson, N. G., Lisacek, F., Packer, N. H., Campbell, M. P., Aoki, N. P., Fujita, A., Matsubara, M., Shinmachi, D., Tsuchiya, S., Yamada, I., Pierce, M., Ranzinger, R., Narimatsu, H., and Aoki-Kinoshita, K. F. (2017). GlyTouCan: an accessible glycan structure repository. *Glycobiology*, **27**(10), 915–919.

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.

Fig. 2: Performance Comparison with and without Network Smoothing for *20141103-02-Phil-BS* . The Receiver Operator Characteristic Curve (ROC) comparing True Positive Rate (TPR) to False Positive Rate (FPR) shows how each database performed under different regularization conditions, summarized with the Area Under the Curve (AUC) in the legend. The *Combinatorial + Sulfate* database showed the best performance, and improved with regularization.

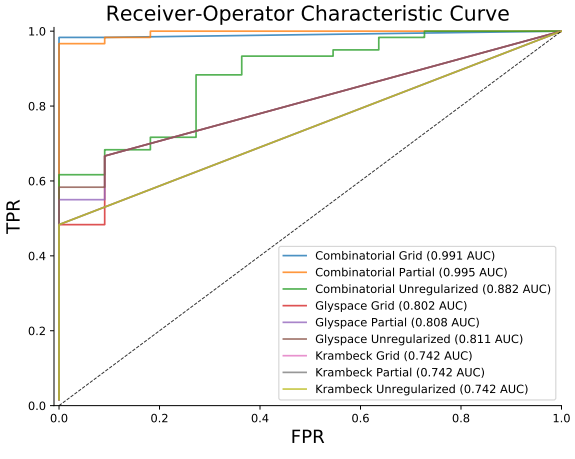
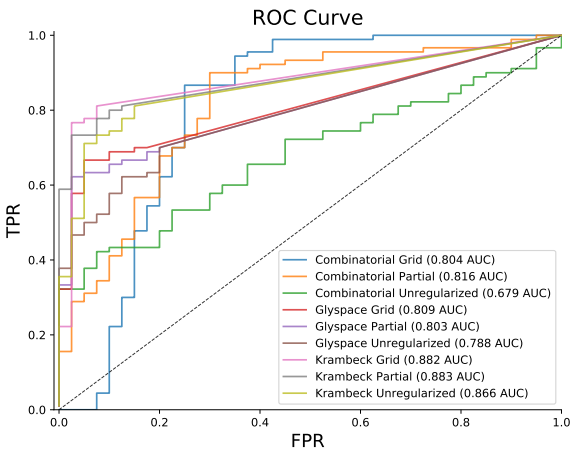


Fig. 3: Performance Comparison with and without Network Smoothing for *Perm-BS-070111-04-Serum* . The Receiver Operator Characteristic Curve (ROC) comparing True Positive Rate (TPR) to False Positive Rate (FPR) shows how each database performed under different regularization conditions, summarized with the Area Under the Curve (AUC) in the legend



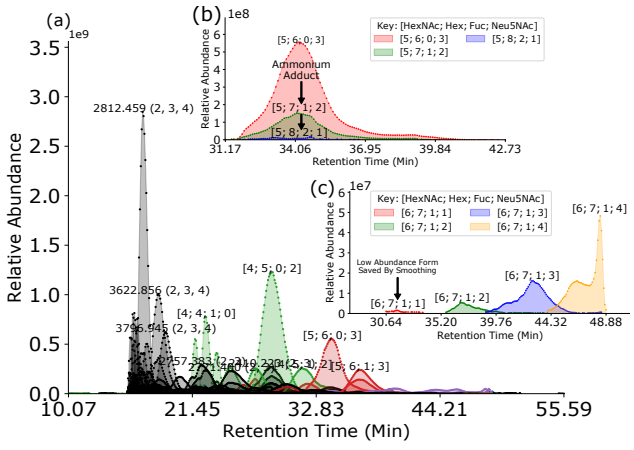


Fig. 4: Chromatogram Assignments for *Perm-BS-070111-04-Serum*. In all panels, the Retention Time (Min) axis shows the retention time in minutes from the experiment, and the Relative Abundance axis shows the intensity of the signal from the particular aggregated ion species for each analyte. The identified glycan compositions are labeled with a tuple describing the number of each component of the form [**HexNAc**, **Hex**, **Fuc**, **NeuAc**] (a) Features Assigned After Grid Regularization of *Perm-BS-070111-04-Serum* (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