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

June 13, 2017

1 Introduction

Glycosylation is one of the most pervasive forms of post-translational modification.

2 Methods

2.1 Glycan Hypothesis Generation

N-glycans start with a common core, but growing outwards, they can become quite complex Stanley *et al.* (2009). Starting with the core motif of **HexNAc2 Hex3**, all combinations of monosaccharides ranging between:

Monosaccharide	Lower Limit	Upper Limit
HexNAc	2	9
Hex	3	10
Fuc	0	4
NeuAc	0	5

subject to to the limitation $\mathbf{HexNAc} > \mathbf{Fuc}$ and $(\mathbf{HexNAc} - 1) > \mathbf{NeuAc}$. We created a copy of this database for native, reduced and permethylated, and deuteroreduced and permethylated.

2.2 LC-MS Preprocessing

Raw LC-MS data must be preprocessed in a number of ways before it can readily be compared to theoretical glycan compositions. 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.

2.3 LC-MS Feature Aggregation

We aggregated deconvoluted peaks over time to construct LC-MS features. 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. After all peaks were clustered, we sorted each cluster by time, creating a list of LC-MS features.

2.4 Glycan Composition Matching

For each LC-MS feature, we queried the target glycan database for compositions whose masses were within $\delta_{mass} = 10$ PPM mass error for QTOF data, 5 PPM mass error for Orbitrap 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 LC-MS features 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 termed the remaining assigned and unassigned LC-MS features *candidate features*.

2.5 Feature Evaluation

For each candidate feature, we computed several statistics to estimate how distinguishable the observed signal was from random noise. We use the following quantities from each LC-MS feature:

neutral mass_i	The neutral mass of the ith chromatogram
intensity_i	The total intensity array assigned to the ith chromatogram
node $\text{intensity}_{i,j}$	The sum of all peak intensities for peaks observed in the jth scan for the ith chromatogram
$\text{intensity}_{i,\text{charge}=j}$	The total intensity assigned to the ith chromatogram with charge state j
$\text{time}_{i,j}$	The time of the jth scan of the ith chromatogram
charges_i	The set of charge states observed for the ith chromatogram
$\text{peak}_{i,j}$	The jth deconvoluted MS peak assigned to the ith chromatogram
peak $\text{intensity}_{i,j}$	The intensity assigned to $\text{peak}_{i,j}$
$\text{envelope}_{i,j}$	The normalized experimental isotopic envelope composing $\text{peak}_{i,j}$, a sequence of size K, whose members sum to 1
adducts_i	The set of adduction states observed for the ith chromatogram
$\text{intensity}_{i,\text{adduct}=j}$	The total intensity assigned to the ith chromatogram with adduct j

2.5.1 Chromatographic Peak Shape

An LC-MS elution profile should be composed of one or more peak-like components, each following a bi-Gaussian peak shape model Yu and Peng (2010) or in less ideal chromatographic circumstances, a skewed Gaussian peak shape model. We fit these models using non-linear least squares (NLS). As measures of goodness of fit are not generally available for NLS, we used the following criterion:

$$\hat{y}_i = NLS(\text{node intensity}_i, \text{time}_i) \tag{1}$$

$$e_{i,NLS} = \text{node intensity}_i - \hat{y}_i \tag{2}$$

$$\bar{y}_i = \frac{1}{n}\sum_{j}^{n}(\text{node intensity}_{i,j}) \tag{3}$$

$$e_{i,null} = \text{node intensity}_i - \bar{y}_i \tag{4}$$

$$\text{line score}_i = 1 - \frac{\sum e_{i,NLS}^2}{\sum e_{i,null}^2} \tag{5}$$

where line score describes how much the peak shape fit improves on a flat line fit null model.

We applied two competitive peak fitting strategies to address distorted, overlapping, or multimodal elution profiles. The first worked iteratively by finding a best-matching peak shape using non-linear least squares, subtracting the fitted signal and checked if there was another peak with at least half as tall as the removed peak, if so repeating the process until no peak can be found, saving each peak model so constructed. The second approach started by locating local minima between putative peaks, and partitioning the LC-MS feature into sub-groups which would be fitted independently. This method generates a candidate list of minima, and selects the case which has the greatest difference between the minimum and its pair of maxima to split the feature at. The strategy which produced the maximum *line score* was chosen.

2.5.2 Composition Dependent Charge State Distribution

As the number of monosaccharides composing a glycan increases, the number of possible sites for charge localization increases. Under normal conditions, we would expect to observe the same molecule in multiple charge states Maxwell *et al.* (2012). Which charge states are expected would depend upon the size of the molecule and it's constituent units' electronegativity. In it's native state, **NeuAc**'s acidic group causes glycans with one or more **NeuAc** to have a propensity for higher negative charge statesVarki and Schauer (2009). To capture this relationship, we modeled the probability of observing a glycan composition for sialylated and unsialylated compositions separately.

$$m_i = (\lfloor(\text{neutral mass}_i/w)/10\rfloor + 1) * 10 \tag{6}$$

$$\text{charged intensity}_{i,j} = \frac{\text{intensity}_{i,\text{charge}=j}}{\text{intensity}_i} \tag{7}$$

$$P(c,m) = |m| \sum_{m_i \in m} \text{charged intensity}_{i,j} \tag{8}$$

$$\text{charge score}_i = \sum_{c_{i,j} \in \text{charges}_i} P(c_{i,j}, m_i) \tag{9}$$

$$\tag{10}$$

where w is the width of the mass bin divided by 10 and $P(c,m)$ is defined as part of the model estimation procedure.

2.5.3 Isotopic Pattern Consistency

Our ahead-of-time deconvolution procedure uses an averagine isotopic model and does not capture the consistency of the isotopic pattern that was fit with the isotopic pattern of the glycan composition that matched that peak. The criterion

$$\text{isotope score}_i = 1 - \frac{2}{\text{intensity}_i} \sum_{j}^{J} \text{peak intensity}_{i,j} \sum_{k}^{K} |\text{envelope}_{i,j,k}(\ln(\text{envelope}_{i,j,k}) - \ln(\text{tid}_{i,k}))| \tag{11}$$

where *tid* is the theoretical isotopic pattern derived from either the ith glycan composition or an averagine interpolated for neutral mass$_i$. This computes an peak intensity weighted mean G-test comparing the goodness of fit between the experimental envelope and the theoretical isotopic pattern.

2.5.4 Observation Spacing Score

The less time between observations of a glycan composition the less likely the LC-MS feature is to contain peaks missing or caused by isotopic pattern interference or missing information.

$$\text{spacing score}_i = 1 - \frac{2}{\text{intensity}_i} \sum_{j=1}^{J} \text{node intensity}_{i,j}(\text{time}_{i,j} - \text{observed time}_{i,j-1}) \tag{12}$$

References

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.

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.

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. and Schauer, R. (2009). *Sialic Acids*. Cold Spring Harbor Laboratory Press.

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.