Network method for building a sample-specific glycan database for N-linked glycosylation from MS/MS data

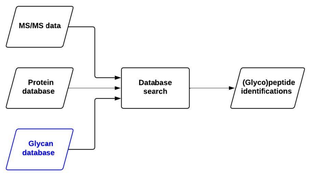
Aliana Tang[1]; Marshall Bern[2]
[1]Piedmont High School, Piedmont, CA; [2]Protein Metrics LLC, Boston, MA

Introduction

Glycans are critical for many biological processes, such as protein folding and intercellular communication

About half of all mammalian proteins are glycosylated

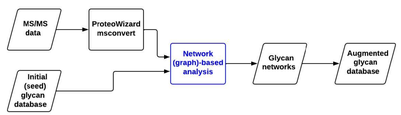
Mass spectrometry is widely used for studying proteins and glycoproteins



- Database search software is the predominant method for identifying proteins from tandem mass spectrometry data. Requirements are:
  - Complete protein database — often satisfied due to the ease of genome sequencing
  - For glycoproteins, additionally need complete glycan database — often not satisfied; usually only have imperfect knowledge
- **Goal is to develop software that builds improved N-glycan databases** by building a sample-specific glycan database based on the mass spectrometry data itself rather than relying solely on preexisting glycan databases

Methods

Wrote **network-based analysis software** to augment an initial ("seed") glycan database with additional glycans to construct a more complete glycan database



Tested the following algorithms:

- Algorithm 1 (no network): For each MS/MS spectrum, if there are peaks characteristic of N-glycosylation, infer glycan and add to the list of glycans
- Algorithm 2 (single network): From the glycans inferred in algorithm 1, construct a network, or graph, where each node is a glycan and each edge connects two nodes that differ in mass by a monosaccharide (HexNAc, Hex, Fuc, etc.). Construct a list of glycans from only those nodes that are in a cluster of size ≥ 3.
- Algorithm 3 (multiple networks): Separate the glycans inferred in algorithm 1 into bins, where everything in a bin has the same bare peptide mass. In each of the bins, follow the procedure of algorithm 2.

Conceptual examples of the algorithms

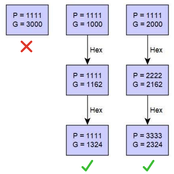
Algorithm 1 (no network)

Use approximate integer masses for convenience
P = mass of bare peptide ("bare" = without glycan)
G = mass of glycan
Hex residue mass ≅ 162

In algorithms 2 and 3, only clusters of size ≥ 3 are accepted

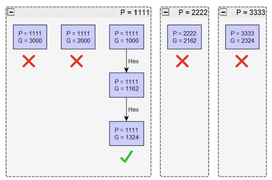
P	G
1111	1000
1111	1162
1111	1324
1111	2000
1111	3000
2222	2162
3333	2324

Algorithm 2 (single network)



P	G
1111	1000
1111	1162
1111	1324
1111	2000
2222	2162
3333	2324

Algorithm 3 (multiple networks)



P	G
1111	1000
1111	1162
1111	1324

Negative control

For a negative control, the software was tested on data that should have no N-glycosylation. Thus, the output should have no glycans. These samples were treated with PNGase F to detach all N-glycans (caveat: it is possible that there is some residual N-glycosylation left behind if the enzymatic reaction did not proceed to completion).

Number of glycans in output

Dataset ID	Sample	# spectra	Algorithm 1 (no network)	Algorithm 2 (single network)	Algorithm 3 (multiple networks)
MassIVE MSV000093894	1	14120	6 (0.04%)	0 (0.0%)	0 (0.0%)
	2	4286	5 (0.1%)	0 (0.0%)	0 (0.0%)
	3	37091	1 (0.003%)	0 (0.0%)	0 (0.0%)
	4	22549	5 (0.02%)	0 (0.0%)	0 (0.0%)
PRIDE PXD046405	1	16835	146 (0.9%)	0 (0.0%)	0 (0.0%)
	2	17075	164 (1%)	0 (0.0%)	0 (0.0%)
	3	17176	157 (0.9%)	3 (0.02%)	0 (0.0%)

Algorithm 3 had the fewest false positives → used for all subsequent work

Reanalysis of soybean root nodule data

- Dataset ID in MassIVE is MSV0000088754. Publication [ref 3] also has results from detached glycan MALDI experiments
- Samples are from soybean (Glycine max) root nodules infected with
  - Wild-type (WT) Bradyrhizobium bacteria that fixes nitrogen
  - Mutant (M) Bradyrhizobium bacteria that cannot fix nitrogen
- Our reanalysis started with the "seed" database "N-glycan 52 plants" from Byonic (original publication also used this database as a base)
- **Our reanalysis using Algorithm 3 discovered 9 additional glycans**

Samples with mutant bacteria	Samples with wild-type bacteria
HexNAc(2)Hex(4)Fuc(1)Pent(1)	HexNAc(2)Hex(4)Fuc(1)Pent(1)
HexNAc(2)Hex(5)Fuc(1)Pent(1)	HexNAc(2)Hex(2)Fuc(1)Pent(1)
HexNAc(3)Hex(5)Fuc(1)Pent(1)	HexNAc(2)Hex(12)
HexNAc(2)Hex(2)Fuc(1)Pent(1)	HexNAc(2)Hex(13)
HexNAc(2)Hex(1)Fuc(1)Pent(1)	HexNAc(2)Hex(14)
	HexNAc(2)Hex(15)

Red: *Glycans discovered by both our reanalysis and the detached glycan MALDI experiment*
Blue: *Glycans discovered by our reanalysis onlyHexNAc(*

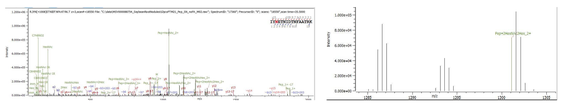
Reanalysis of grass carp data

- Dataset ID in PRIDE is PXD010308. Publication [ref 4] also has results from detached glycan MALDI experiments
- Sample is IgM from grass carp (Ctenopharyngodonidella idella)
- Our reanalysis started with the "seed" database "N-glycan 182 human no multiple fucose" from Byonic (original publication also used this database)
- **Our reanalysis using Algorithm 3 discovered 18 additional glycans**

HexNAc(5)Hex(9)	HexNAc(6)Hex(10)Fuc(1)
HexNAc(5)Hex(7)NeuAc(1)	HexNAc(6)Hex(10)Fuc(2)
HexNAc(5)Hex(8)NeuAc(1)	HexNAc(6)Hex(9)NeuAc(1)
HexNAc(6)Hex(8)	HexNAc(6)Hex(10)NeuAc(1)
HexNAc(6)Hex(10)	HexNAc(6)Hex(11)NeuAc(1)
HexNAc(6)Hex(11)	HexNAc(6)Hex(9)Fuc(1)NeuAc(1)
HexNAc(6)Hex(12)	HexNAc(6)Hex(10)Fuc(1)NeuAc(1)
HexNAc(6)Hex(8)Fuc(1)	HexNAc(7)Hex(11)
HexNAc(6)Hex(9)Fuc(1)	HexNAc(7)Hex(13)

Validation by database search

- Glycans discovered by our reanalyses can be validated (as best as possible) by running database search (Byonic) using the augmented glycan database
- The more tall peaks in the spectrum that can be labeled, the more confident the glycopeptide identification
- Example here shows HexNAc(2)Hex(2)Fuc(1)Pent(1) from soybean (M)
- Zoomed view shows pentose peak



Reanalysis of filamentous fungus data

- Dataset ID in PRIDE is PXD041208
- Sample is a monoclonal IgG1 antibody produced in a genetically modified filamentous fungus expression system Thermothelomyces heterothallica (C1)
- Data analysis in original publication [ref 5] has a series of oligomannose N-glycans HexNAc(2)Hex(n), 1 ≤ n ≤ 11, as well as HexNAc(3)Hex(n), 2 ≤ n ≤ 6
- **Our reanalysis using Algorithm 3 discovered 6 additional glycans**

HexNAc(3)Hex(7)	HexNAc(3)Hex(10)
HexNAc(3)Hex(8)	HexNAc(3)Hex(11)
HexNAc(3)Hex(9)	HexNAc(3)Hex(12)

- We also inspected the MS/MS of these additional glycans more closely
  - No peak at m/z 773 → suggests the 3rd HexNAc is not a bisecting GlcNAc
  - Hex peaks (at m/z 163, 145, 127) are significantly smaller in HexNAc(3)Hex(…) compared to HexNAc(2)Hex(…) → suggests the 3rd HexNAc is terminal and not in a LacNAc unit
- **Hypothesis:** The HexNAc(3)Hex(…) may be better described as oligomannose N-glycans with a terminal GlcNAc rather than as hybrid N-glycans

- Another interesting result from our reanalysis is that there is an additional large glycan "cluster" where each element in the cluster is 28 Da heavier (almost exactly) than a known glycan
- Hypothesis: Artifact resulting from formylation [ref 6]

Conclusions

- Database search is arguably the most effective software, from a practical perspective, for identifying proteins or glycoproteins from MS/MS data
- However, database search is blind to proteins or glycans that are not in the database. The goal of this project is to eliminate or mitigate the blind spot by making the glycan database more complete
- **Using a network was critical to the effectiveness of our software**
- The network model mimics the in vivo process of glycan synthesis
- **Our network algorithm discovered additional glycans in previously published data**
  - Validated subsequently by database search
  - Consistent with experimental results (detached glycan MALDI)

References

1. *D. Goldberg, M. Bern, S. J. North, S. M. Haslam, and A. Dell, "Glycan family analysis for deducing N-glycan topology from single MS," Bioinformatics, vol. 25, no. 3, pp. 365–371, Feb. 2009, doi: 10.1093/bioinformatics/btn636.2.*
2. *A. Guthals, J. D. Watrous, P. C. Dorrestein, and N. Bandeira, "The spectral networks paradigm in high throughput mass spectrometry," Mol. BioSyst., vol. 8, no. 10, p. 2535, 2012, doi: 10.1039/c2mb25085c.3.*
3. *D. Veličković et al., "Spatial Mapping of Plant N-Glycosylation Cellular Heterogeneity Inside Soybean Root Nodules Provided Insights Into Legume-Rhizobia Symbiosis," Front. Plant Sci., vol. 13, p. 869281, May 2022, doi: 10.3389/fpls.2022.869281.4.*
4. *Y.-L. Su et al., "Site-Specific N-Glycan Characterization of Grass Carp Serum IgM," Front. Immunol., vol. 9, p. 2645, Nov. 2018, doi: 10.3389/fimmu.2018.02645.5.*
5. *F. K. Kaiser et al., "Filamentous fungus-produced human monoclonal antibody provides protection against SARS-CoV-2 in hamster and non-human primate models," Nat Commun, vol. 15, no. 1, p. 2319, Mar. 2024, doi: 10.1038/s41467-024-46443-0.6.*
6. *Y. Zhi et al., "Formylation: an undesirable modification on glycopeptides and glycans during storage in formic acid solution," Anal Bioanal Chem, vol. 414, no. 11, pp. 3311–3317, May 2022, doi: 10.1007/s00216-022-03989-6.*

The authors declare no competing financial interest.