

Generation of Asparagine-Linked Glycan Structure Databases and their Use

Hui Ying Gao

BioPharmaSoft, Boston, Massachusetts, USA

Asparagine linked glycans (N-glycans) are important in biological processes. Yet, their structural complexity and lack of databases hinder progress in glycomics and glycobiology. We present a way for *in silico* generation of very large N-glycan structure databases and their use in high throughput composition and primary structure determination of N-glycans attached to peptides, based on CID (collision induced dissociation) MS/MS (tandem mass spectrometric data). The database and the integrated search engine is called Glyquest and is available to the glycomics community. (*J Am Soc Mass Spectrom* 2009, 20, 1739–1742) © 2009 American Society for Mass Spectrometry

It is estimated that about 50% of all mammalian proteins are glycosylated [1], 90% of which contains N-glycans (asparagine-linked glycans) [1]. Glycans participate in important biological processes [2], and they are potential cancer biomarkers [3] and drug targets [4]. It has been difficult to characterize N-glycan structures at high throughput [5].

The success of genomics and proteomics may partly be due to the availability of databases. Lack of databases may be the main factor hindering our characterization of glycans. N-glycans are produced by joined functions of many different enzymes, and there are no templates to predict the monosaccharide sequences in the outer branches. Here we show a way for *in silico* generation of large N-glycan structure databases and their use in high throughput composition and primary structure determination of N-glycans attached to peptides, based on collision induced dissociation tandem mass spectrometric data.

Experimental

To allow readers to verify our findings, we chose to use raw mass spectrometric data available in the Open Proteomics Database (web site <http://bioinformatics.icmb.utexas.edu/OPD>) by E. M. Marcotte's laboratory at the University of Texas, also see reference [6]. The raw data file is found under the Miscellaneous section, acc#: opd00099_PROTS, 6prot_75pmol_each.

The sample was a mixture of tryptic digests of six proteins at 75 pmol each. The mass spectrometer used was LCQ (Thermo Fisher Scientific, Waltham, MA, USA) with positive ESI (electrospray ionization). The normalized collision energy used was 32 and the activation time was 30 ms. The peptides and glycopeptides were eluted from a C18 column using a gradient of

0%–65% acetonitrile in 140 min. We point out that the mass spectrometer and the analytical column used here are available in most laboratories doing proteomic analysis. Thus, N-glycan structure characterization could be within reach of most of these laboratories.

Theory

To generate large N-glycan structure databases *in silico*, we divide an N-glycan structure into the core part and the outer branch structure part and consider them separately. All mammalian N-glycans contain a pentasaccharide core [7] (three mannoses and two GlcNAc (N-Acetylglucosamine)). Alternatively, the pentasaccharide core may contain a bisecting GlcNAc, or a fucose attached to the innermost GlcNAc, or both, for a total four different core structures [7].

The notorious complexity of N-glycan structures originates from the high variability of the outer branch structures. However, we notice that there are actually a very limited number of unique, individual outer branch structure types. From this limited number of branch structure types, we could generate a very large number of combinations, and each combination may be considered as the representation of the outer branch structures of a unique N-glycan structure. As an example, assume that a mammalian glycoprotein contains an N-glycan structure that has two outer branches with known structures, one branch having the monosaccharide sequence, sialic acid-Gal-GlcNAc, and the other, fucose-Gal-GlcNAc. By sialic acid, we mean N-acetylneurameric acid. Let S = sialic acid, G = Gal, T = GlcNAc, and F = fucose. The two outer branch structures can be represented by SGT and FGT.

If a biological system can synthesize one N-glycan with SGT and FGT as the outer branch structures, it is logical to assume that it may also be able to synthesize other N-glycans whose outer branch structures are various combinations of SGT and FGT. If we also

Address reprint requests to Ms. Hui Gao, BioPharmaSoft, Boston, MA 02478, USA. E-mail: h.gao@biopharmasoft.com

assume the number of outer branches to be four and ignore differences in linkage, we can easily generate the following five unique combinations of SGT and FGT by simple mathematics: (SGT, SGT, SGT, SGT), (SGT, SGT, SGT, FGT), (SGT, SGT, FGT, FGT), (SGT, FGT, FGT, FGT), and (FGT, FGT, FGT, FGT). Each combination represents the outer branch structures of one unique N-glycan structure. Since there are possibly four different types of core structures, we may have a total of $5 \times 4 = 20$ unique N-glycan structures (based on two known outer branch structures). All these N-glycans may actually exist in the biological system.

Besides the above two outer branch structures, there are many other known branch structures [7] in mammals. We can generate N-glycan structure databases based on all these known outer branch structures. Again, assuming the number of outer branches to be four, we could get 280, 2860, 35420, 1171300, 17685100, and 274740200 unique N-glycan structures from combinations of 5, 10, 20, 50, 100, and 200 different outer branch structure types and the four different core structures. So, our databases can be very large and comprehensive. As an approximate comparison, the CCSd (Complex Carbohydrate Structure Database, the largest glycan structure database), also known as CarbBank, contains about 23,118 distinct glycan structures. It took years to build.

Of course, not all N-glycans have exactly four outer branches. By including an outer branch structure with no monosaccharide residue, we could get N-glycan structures with zero to four outer branches. By including an outer branch structure containing mannoses only, we could get N-glycan structures with high mannoses and hybrid structures, besides the complex types. The molecular weights of the N-glycans are recorded in the databases. A computer program has been developed to do all the above as well as the searches of the databases (see below) automatically.

When an N-glycan structure with a new outer branch structure is identified, we can easily generate many other N-glycan structures incorporating the new outer branch structure. Similarly, if any new core structure is identified, it can be combined with all the outer branch structures to form new N-glycan structures.

We now turn to the use of the databases generated. While the databases may find other uses in glycomics and glycobiology, their immediate use is for the composition and primary structure determinations of N-glycan structures attached to peptides. Here is one possible approach. After a glycoprotein is purified and digested with a protease (typically, trypsin), the digest is injected on to a C18 column for liquid chromatography-mass spectrometry (LC-MS) analysis and collision induced dissociation (CID) MS/MS, as is done in a typical proteomics analysis. Because glycosidic bonds in glycopeptides are weaker than those between amino acids, the glycosidic bonds are preferably fragmented during CID [8, 9]. To find all the N-glycan structures attached to one specific peptide, we first calculate the

molecular weight of the protonated peptide, MW_{pep}. Then, for a given MS/MS spectrum, we calculate the molecular weight of any N-glycan (MW_{gly}) attached to the peptide based on the molecular weight of the protonated precursor (MW_{pre}) of the MS/MS spectrum, and the molecular weight of the protonated peptide, MW_{gly} = MW_{pre} – MW_{pep}. Then, we search our N-glycan structure database for a list of candidate N-glycan structures whose molecular weights are within a predetermined window of the calculated MW_{gly}. Each candidate N-glycan structure is fragmented to generate a theoretical spectrum and compared with the experimental MS/MS spectrum for similarity. As an example, to generate peaks corresponding to the loss of one silac acid, S, from one of the four outer branches (SGT, SGT, SGT, SGT), we simply remove one S. Thus, the outer branches become (–GT, SGT, SGT, SGT). We calculate the molecular weight of the fragment and *m/z* values at different charges. By sequentially removing other monosaccharide residues, we get other theoretical fragment peaks. Similarly, we can get the *m/z* values of B ions and the *m/z* values due to the fragmentation of the core structure.

Importantly, for each theoretical peak, we record the structure of the N-glycan structure fragment, e.g., (–GT, SGT, SGT, SGT). Later on, we display these structures graphically when we plot the match between the experimental tandem mass spectrum and the theoretical mass spectrum of the N-glycan structure identified. This graphical display is valuable for the validation and visual inspection of any possible matches.

Computational algorithms for matching theoretical mass spectra with experimental mass spectra are well known in the art of mass spectrometry based proteomics, such as those used in SEQUEST, OMSSA, and X!Tandem. Any of these algorithms will work here. We chose to use spectra dot product [10] to score the similarity between the experimental and theoretical spectra.

Results and Discussions

Figure 1 shows the output of our software for an N-glycan found on the peptide NEEYNK of human α -1-acid glycoprotein. In the figure, the peak on the experimental MS/MS spectrum is plotted red if it is matched to a theoretical fragment peak of the N-glycan identified, and plotted green otherwise. As can be seen, almost all the major peaks are matched to theoretical N-glycan fragment peaks. Drawings in the figure are done automatically by the software. The scan number is 1559 with a charge of 4+. The same N-glycan is also identified by other spectra in the same raw data file (scan number 1543 and 1552 with a precursor charge of 3+, and scan number 1554 with a precursor charge of 4+).

The approach is also applicable to some chemically derivatized N-glycans (such as those using 2-amino benzamide). However, our approach is not applicable to peptides containing other unknown modifications. Some isomers of N-glycan structures may have very si-

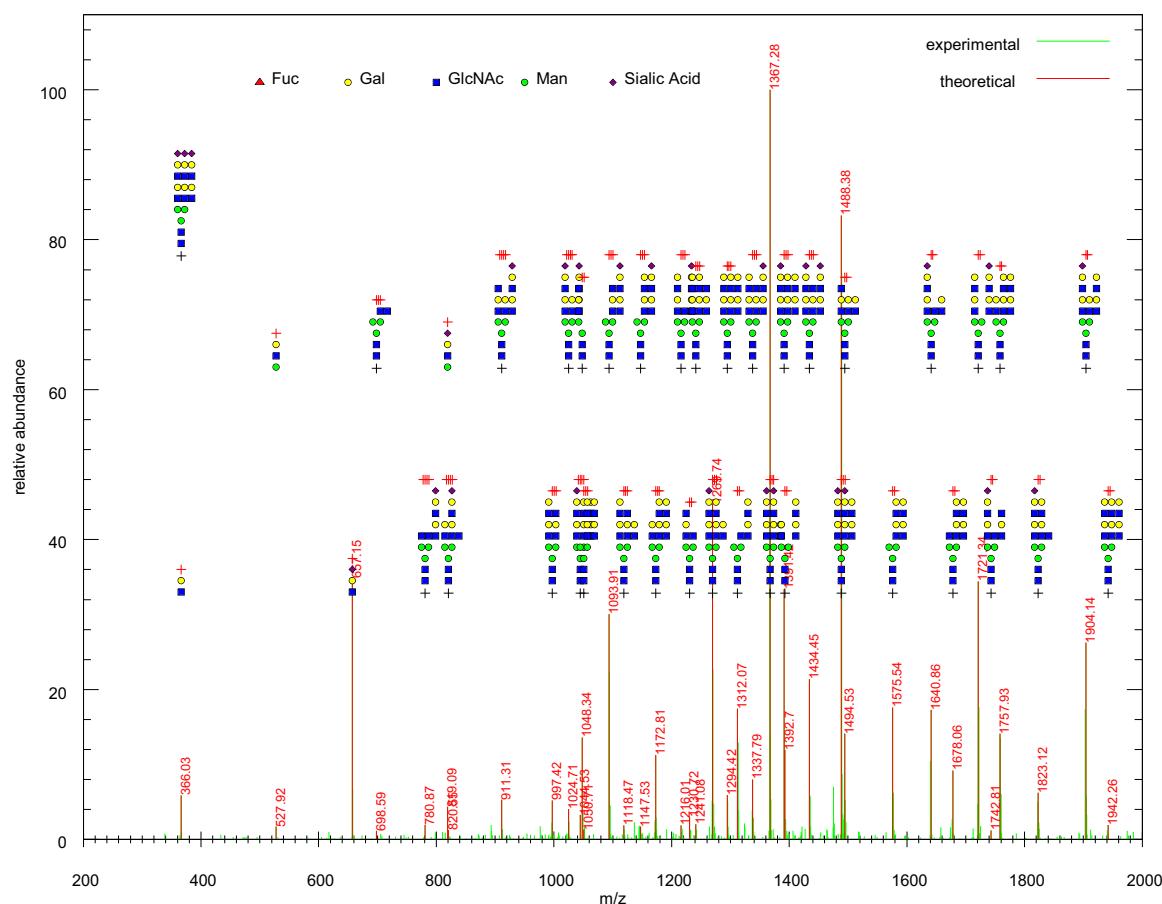


Figure 1. Structure of an N-glycan found on a peptide whose MH^+ is 796.8 Da. Most likely, the peptide is NEEYNK from human α -1-acid glycoprotein, which was one of the six proteins in the sample. The intact glycan structure is shown in the upper left corner of the plot. The glycan fragment structures are aligned with the experimental peaks (plotted red) they match. Experimental peaks not matched are plotted green. The number of red pluses (+) equals to the charge of the fragment. A black plus (+) indicates that the peptide is attached to the fragment. The scan number of the spectrum is 1559 (with a precursor charge of 4).

milar MS/MS spectra, therefore they have very close similarity scores and could not be differentiated by our software. This is a limitation of mass spectrometry technology itself. Another limitation of the approach is that it is best used for simple mixtures of a few glycoproteins. For more complex mixtures, the identities of the peptide part become less certain. Future development may allow the identification of these peptides attached to the glycans.

It is possible to include the most commonly seen linkage type into the databases. In the future, this linkage information could potentially be used for the confirmation of linkages based on relative intensities of fragment peaks. But it would not be practical or of any value to include all possible linkage types in the database.

Conclusions

Our approach represents a step toward automatic interpretation of tandem mass spectrometric data for N-glycan structure determination. The software searches thousands of tandem mass spectra and presents a

selected few for manual inspection by the user. Graphical display of fragmented N-glycan structures matched to experimental peaks greatly facilitates this manual inspection. The database and the integrated search engine (called Glyquest) is available at BioPharmaSoft.com.

Yet, high throughput characterization of glycoproteins based on mass spectrometric data is still at its early developing stage. Large, well characterized datasets are needed before confidence levels of the results can be established.

Acknowledgments

The author acknowledges that the raw data used here was made available by the open proteomics database at <http://bioinformatics.icmb.utexas.edu/OPD> (E. M. Marcotte's laboratory at University of Texas).

References

- Apweiler, R.; Hermjakob, H.; Sharon, N. On the Frequency of Protein Glycosylation, as Deduced from Analysis of the SWISS-PROT Database. *Biochim. Biophys. Acta.* 1999, 1473, 4–8.

2. Ohtsubo, K.; Marth, J. D. Glycosylation in Cellular Mechanisms of Health and Disease. *Cell* **2006**, *126*, 855–867.
3. Fuster, M. M.; Esko, J. D. The Sweet and Sour of Cancer: Glycans as Novel Therapeutic Targets. *Nat. Rev. Cancer* **2005**, *5*, 526–542.
4. Dube, D. H.; Bertozzi, C. R. Glycans in Cancer and Inflammation—Potential for Therapeutics and Diagnostics. *Nat. Rev. Drug Discov.* **2005**, *4*, 477–488.
5. Packer, N. H.; von der Lieth, C. W.; Aoki-Kinoshita, K. F.; Lebrilla, C. B.; Paulson, J. C.; Raman, R.; Rudd, P.; Sasisekharan, R.; Taniguchi, R.; York, W. S. Frontiers in Glycomics: Bioinformatics and Biomarkers in Disease. *Proteomics* **2008**, *8*, 8–20.
6. Prince, J. T.; Carlson, M. W.; Wang, R.; Lu, P.; Marcotte, E. M. The Need for a Public Proteomics Repository. *Nat. Biotechnol.* **2004**, *22*, 471–472.
7. Kornfeld, R.; Kornfeld, S. Assembly of Asparagine-Linked Oligosaccharides. *Annu. Rev. Biochem.* **1985**, *54*, 631–664.
8. Ren, J. M.; Rejtar, T.; Li, L.; Karger, B. L. N-Glycan Structure Annotation of Glycopeptides Using a Linearized Glycan Structure Database (GlyDB). *J. Proteome Res.* **2007**, *6*, 3162–3173.
9. Demelbauer, U. M.; Zehl, M.; Pleimat, A.; Allmaier, G.; Rizzi, A. Determination of Glycopeptide Structures by Multistage Mass Spectrometry with Low-Energy Collision-Induced Dissociation. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 1575–1582.
10. Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859–866.