GlycoMaster DB: A Software Tool for Automated Glycopeptide Identification from MS/MS

Bioinformatics Solutions Inc.

University of Waterloo

Lin He[1], Baozhen Shan[2], Lei Xin[2], Gilles A. Lajoie[3], Bin Ma[1]
University of Waterloo, Waterloo, Ontario, Canada[1]
Bioinformatics Solutions Inc., Waterloo, Ontario, Canada[2]
University of Western Ontario, London, Ontario, Canada[3]

Overview

We propose a software tool, GlycoMaster DB, for the automated and high-throughput identification of glycopeptides, including the glycosylation site and the glycan structure, from large scale MS/MS data generated by HCD fragmentation. The software takes a MS/MS data file as input, searches in a protein sequence database and a glycan structure database simultaneously, and reports the optimal peptide and glycan structure pair that best matches each spectrum. Testing the software on a large MS/MS data set demonstrated promising performance of our software.

Introduction

MS/MS has become one of the most versatile and powerful tools in glycoproteome analysis because of its high sensitivity and selectivity. However, the automated interpretation of glycopeptide-generated MS/MS spectra still faces two major challenges: 1) the fragmentation pattern of a glycopeptide is more complex than a non-glycopeptide, 2) the combination of the protein sequence database and the glycan structure database greatly increases the search space. To meet these challenges, our new software package, GlycoMaster DB, is developed to confidently identify the optimal peptide and glycan structure pair from MS/MS data through simultaneously searching a protein database and the GlycomeDB database [1].

Methods

GlycoMaster DB takes a MS/MS data file as input. It also requires a protein sequence database and a glycan structure database. The protein sequence database can be any database in the simple FASTA file format. The glycan structure database needs to be in the condensed GlycoCT format, and a default N-linked glycan database extracted from GlycomeDB has been integrated into GlycoMaster DB. The MS/MS data is analyzed in four major steps:

1. Protein identification. MS/MS spectra are searched against the given protein database to identify a list of proteins, through the non-glycosylated peptides existing in the sample.
2. Glycopeptide spectra filtration. Spectra generated from non-glycosylated peptides are filtered out according to the existence of glycan signature ions (with m/z values 204 and 366) and monosaccharide tags.
3. Glycan structure assignment. The glycan structure database is searched for the best matching glycan for each spectrum. A scoring function is defined to evaluate the match between a glycan structure and a spectrum.
4. Glycopeptide identification. The peptide sequences are identified primarily from two sources of information: the precursor mass of the glycopeptide and the existence of a N-glycan motif.

The identification result of the input MS/MS data is presented as a table in a HTML page. Figure 1 shows a screenshot of such a result page. Each identified glycan-spectrum match (GSM) is displayed in a row. The images of the glycan structures are also displayed in the table. Additional links are provided to show the annotated spectrum, error map, and score distribution for each identified GSM, as shown in Figure 2, and redirect to the corresponding entry of the GlycomeDB database.

No.	Scan#	PrecMz	PrecZ	RT	Glycan#	Glycan Structure	GSM Score	Composition	Peptides	Err (ppm)
1	6089	1006.76733	3	-	10673		45.06	(HexNAc)4(Hex)5(Fuc)1(NeuAc)1	-	-
2	16595	1006.7675	3	-	10673		44.67	(HexNAc)4(Hex)5(Fuc)1(NeuAc)1	-	-
3	20978	1001.07623	3	-	13322		44.32	(HexNAc)4(Hex)5(NeuAc)2	GHVNITR	0.24
4	20608	1355.5647	2	-	12670		42.7	(HexNAc)4(Hex)5(NeuAc)1	GHVNITR	-1.17

Figure 1. A screenshot of the GlycoMaster DB result page. The reported glycan information, GSM score, and the best matched glycopeptide are displayed. The hyperlink of the score directs to a detailed GSM result page, partially shown as Figure 2.

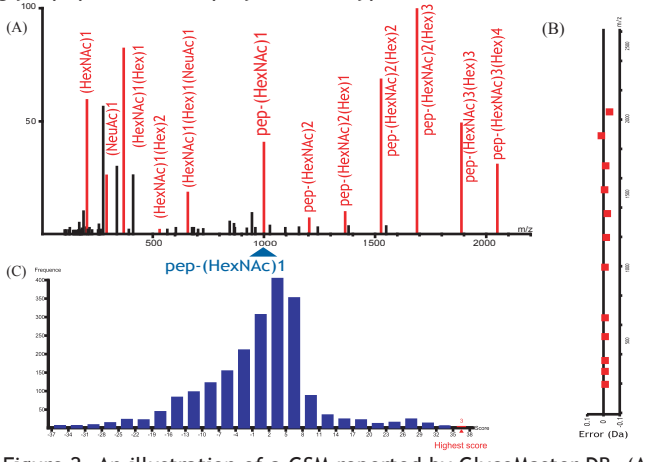

Figure 2. An illustration of a GSM reported by GlycoMaster DB. (A) Annotated spectrum using the reported glycan structure. Red peaks are matched by B- and Y-ions fragmented from the glycan. (B) Error map of the matches between theoretical ions and peaks. (C) A histogram to show the distribution of the matching scores between the spectrum and all possible glycan candidates from GlycomeDB database. The red bar locates the reported GSM score, and the number above denotes the number of glycan candidates with the same matching score.

Software Performance Evaluation

MS/MS data of five fractions from a human urinary proteome experiment [2] were used to evaluate the performance of our software. The sample was preprocessed using lectin affinity enrichment and then analyzed by an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher ScientificTM, Bremen, Germany). The data set contains 25,451 HCD MS/MS spectra.

Among the 25,451 MS/MS spectra analyzed by our software, 2,130 have glycan signature ions, and 675 of them were selected as glycopeptide-generated spectra according to our glycan tag de novo sequencing algorithm. These 675 spectra were searched against the GlycomeDB database, and 422 spectra were identified with high confidence (-10lgP ≥ 15). Glycopeptide sequences were also searched in the protein list, which contains 638 proteins reported by PEAKS 6 database search. 359 out of 422 spectra, with glycans reported, were identified with glycopeptide sequences. Figure 2 illustrates a glycan-spectrum match reported by GlycoMaster DB. Most of the high-intensity peaks are explained by the B- and Y-ions of the glycan, and most of the Y-ions of the glycan find corresponding peaks in the spectrum. This indicates a highly confident identification of the glycan structure.

Summary

GlycoMaster DB automatically and simultaneously determine the glycan structure and the glycopeptide sequence from an HCD spectrum of an intact glycopeptide. The software is currently available as an online web server at http://www-novo.cs.uwaterloo.ca:8080/GlycoMasterDB.

References

1. R. Ranzinger, S Herget, T. Wetter, C. Lieth. GlycomeDB - Integration of Open-Access Carbohydrate Structure Databases. BMC Bioinformatic, 9:384, 2008.
2. A. Marimuthu, R. O'Meally, R. Chaerkady, Y. Subbannayya, et. al. A Comprehensive Map of the Human Urinary Proteome. Journal of Proteome Research, 10:2734, 2011.