GPQuest Guidelines

Shadi Eshghi

March 2016

Software Overview

GPQuest is a software tool developed in MATLAB for identification of intact N- and O-linked glycopeptides from HCD fragmented proteomics and glycoproteomics mass spectrometry data in a shotgun experiment. The glycoprotein samples are digested using proteolytic enzymes and analyzed using LC-MS/MS. This tool takes the spectra files and the peptide and glycan databases of interest and provides the list of intact N- and O-linked glycopeptides in the glycoprotein mixture (Figure 1). The software has been particularly designed and extensively tested on data generated on a Thermo Scientific Q-Exactive instrument.

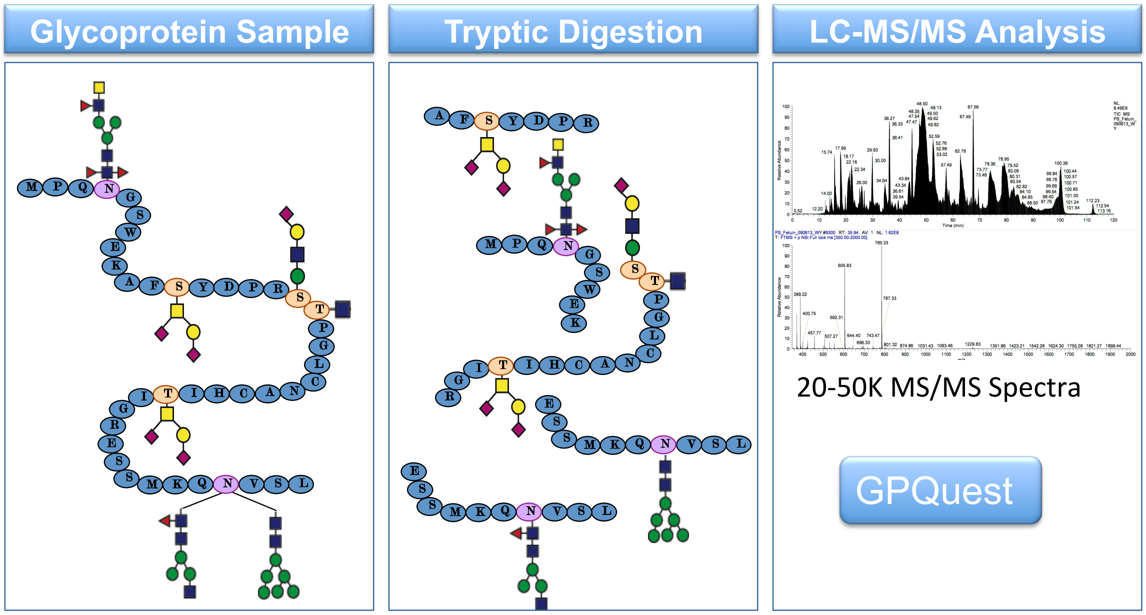
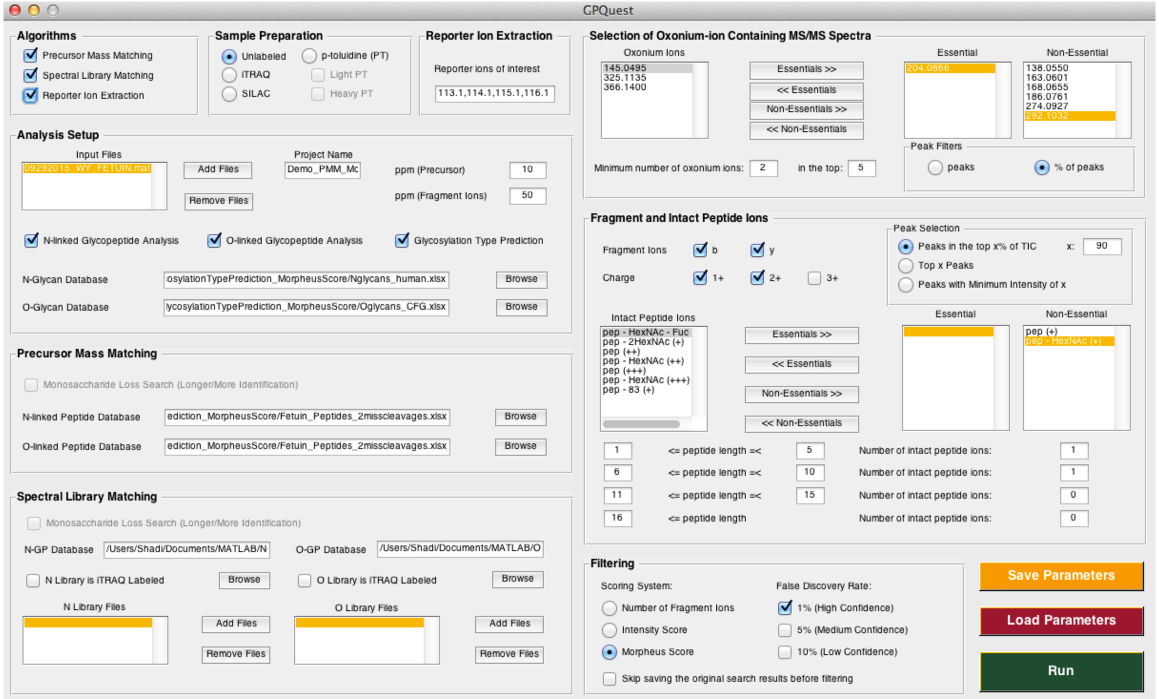


Figure 1. Experimental workflow for identification of intact glycopeptides using GPQuest.

Graphical User Interface

The graphical user interface (GUI) of GPQuest is shown in Figure 2. The GUI consists of nine panels of 1) Algorithms, 2) Sample preparation, 3) Reporter Ion Extraction, 4) Analysis setup, 5) Precursor mass matching, 6) Spectral library matching, 7) Selection of oxonium ion-containing MS/MS spectra, 8) Fragment and intact peptide ions and 9) Filtering. The function of each of these panels is explained in the forthcoming chapters.



Algorithms

☑ Precursor Mass Matching
☑ Spectral Library Matching
☑ Reporter Ion Extraction

Sample Preparation

● Unlabeled ○ p-toluidine (PT)
○ iTRAQ □ Light PT
○ SILAC □ Heavy PT

Reporter Ion Extraction

Reporter ions of interest: 113.1,114.1,115.1,116.1

Analysis Setup

Input Files: 09292015_WY_FETUIN.mat
Project Name: Demo_PMM_Mc
ppm (Precursor): 10
ppm (Fragment Ions): 50

☑ N-linked Glycopeptide Analysis ☑ O-linked Glycopeptide Analysis ☑ Glycosylation Type Prediction

N-Glycan Database: osylationTypePrediction_MorpheusScore/Nglycans_human.xlsx
O-Glycan Database: lycosylationTypePrediction_MorpheusScore/Oglycans_CFG.xlsx

Precursor Mass Matching

□ Monosaccharide Loss Search (Longer/More Identification)

N-linked Peptide Database: ediction_MorpheusScore/Fetuin_Peptides_2misscleavages.xlsx
O-linked Peptide Database: ediction_MorpheusScore/Fetuin_Peptides_2misscleavages.xlsx

Spectral Library Matching

□ Monosaccharide Loss Search (Longer/More Identification)

N-GP Database: /Users/Shadi/Documents/MATLAB/N
O-GP Database: /Users/Shadi/Documents/MATLAB/O
□ N Library is iTRAQ Labeled
□ O Library is iTRAQ Labeled
N Library Files
O Library Files

Selection of Oxonium-ion Containing MS/MS Spectra

Oxonium Ions	Essential	Non-Essential
145.0495 325.1135 366.1400	204.0866	138.0550 163.0601 168.0655 186.0761 274.0927 292.1032

Minimum number of oxonium ions: 2 in the top: 5

Peak Filters: ○ peaks ● % of peaks

Fragment and Intact Peptide Ions

Fragment Ions ☑ b ☑ y
Charge ☑ 1+ ☑ 2+ □ 3+

Peak Selection
● Peaks in the top x% of TIC x: 90
○ Top x Peaks
○ Peaks with Minimum Intensity of x

Intact Peptide Ions	Essential	Non-Essential
pep - HexNAc - Fuc pep - 2HexNAc (+) pep (++) pep - HexNAc (++) pep (+++) pep - HexNAc (+++) pep - 83 (+)		pep (+) pep - HexNAc (+)

1 <= peptide length =< 5 Number of intact peptide ions: 1
6 <= peptide length =< 10 Number of intact peptide ions: 1
11 <= peptide length =< 15 Number of intact peptide ions: 0
16 <= peptide length Number of intact peptide ions: 0

Filtering

Scoring System:
○ Number of Fragment Ions
○ Intensity Score
● Morpheus Score

False Discovery Rate:
☑ 1% (High Confidence)
□ 5% (Medium Confidence)
□ 10% (Low Confidence)

□ Skip saving the original search results before filtering

Figure 2. Graphical user interface of GPQuest.

Algorithms

Three options of algorithms are available for each analysis: 1) Precursor mass matching, 2) Spectral library matching and 3) Reporter ion extraction. Any number of these three algorithms can be selected simultaneously for each analysis.

Precursor mass matching

Precursor mass matching (PMM) is a fundamental and efficient method for identification of glycopeptides from samples. PMM relies on databases of glycosite-containing peptides and glycans to build a database of theoretical glycopeptides in the sample by pairing up these peptides and glycans. For each MS/MS spectrum, PMM narrows down the list of potential glycopeptides by matching the corresponding precursor mass with that of glycopeptide database. Fragment ions and intact peptide ions are further used to identify the best glycopeptide match. Pros: PMM is a powerful, efficient and fast technique for identification of glycopeptides in biological samples. In particular, if simulation time is a constraint, PMM is faster compared to spectral library matching.

Cons: PMM heavily relies on the completeness of the glycan and peptide databases. So, potentially novel glycans or modified glycans that are not included in the glycan database will not be assigned or identified.

Spectral library matching

The application of spectral library matching (SLM) for identification of intact glycopeptides has been recently examined [1]. Spectral library matching takes advantage of the spectral similarity between the HCD fragmented glycopeptides and their deglycosylated counterparts. Using this algorithm, a spectral library of the glycosite-containing peptides in the sample is built by first analyzing the deglycosylated peptides extracted from the sample using LC-MS/MS. For identification of the glycopeptide corresponding to each MS/MS spectrum, the spectrum is first compared with the entries in the spectral library to find the closest peptide match. Knowing the peptide portion of the glycopeptide, the glycan portion can be identified by searching a glycan database for the mass difference of the precursor and the peptide.

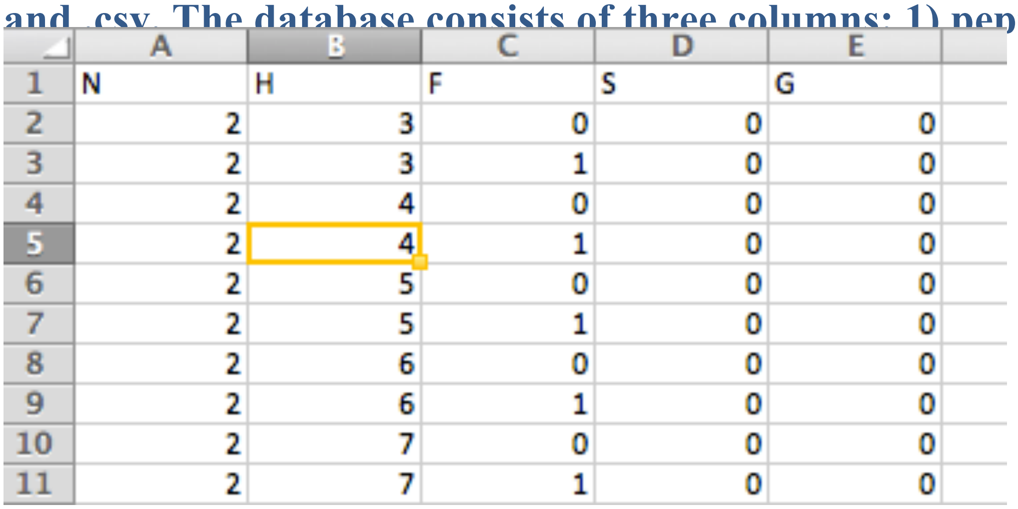
Pros: The main advantage of spectral library matching for identification of intact glycopeptides is that the identification of the peptide portion does not rely on the glycan database. Therefore, if the glycan portion is novel, or modified and therefore missing from the glycan database, the identification of the peptide portion is not affected. Additionally, this would create an opportunity for identification of novel glycans or novel glycan modifications.

Cons: Compared to PMM, SLM is a slow algorithm and therefore not optimal if the simulation time is a great constraint.

Reporter ion extraction

The objective of this function is to extract the intensities of reporter ions of interest. Checking this option would allow the user to enter the m/z of reporter ions in the reporter ion extraction panel. As a result, GPQuest will generate an excel file including the intensities of these reporter ions for each MS/MS spectrum that qualifies as an oxonium ion-containing spectrum. As an example, this function can be used for extracting the intensities of iTRAQ reporter ions for relative quantification. See

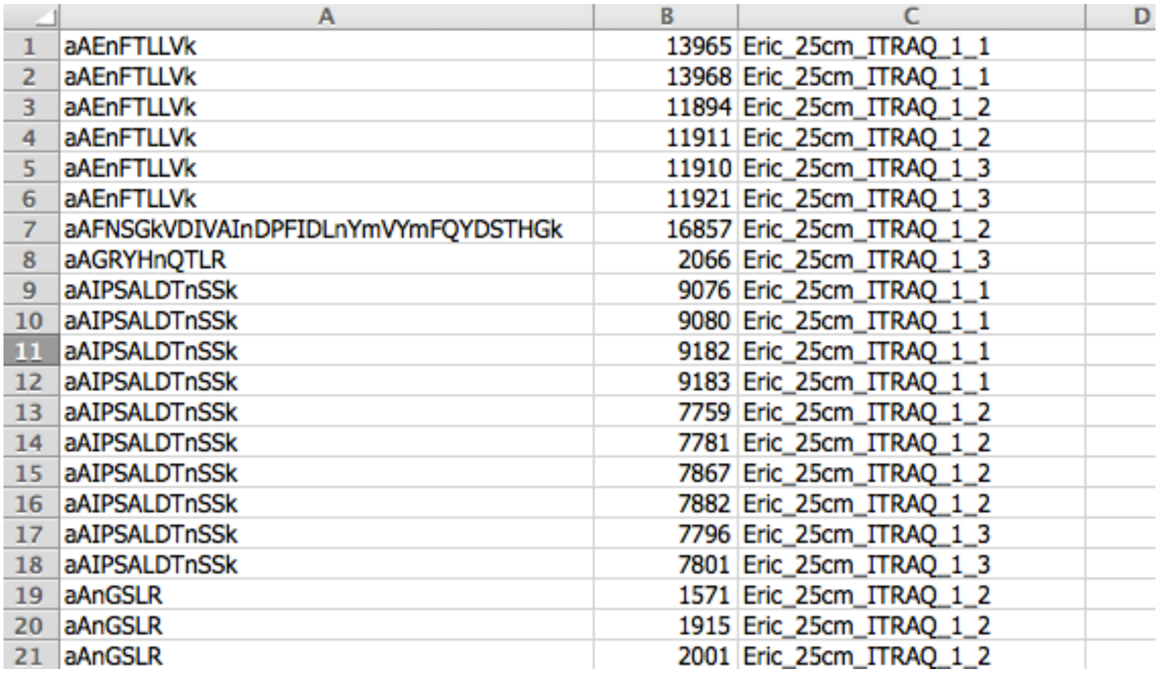
The implemented SLM algorithm in GPQuest first builds the spectral library based on the mzXML files corresponding to the glycosite-containing peptides and a peptide database that determines the target peptides in the glycoproteomics analysis. The path to the peptide database is specified for N- and/or O-linked peptides depending on the parameters selected in Analysis Setup. The accepted database formats are .xls, .xlsx and .csv. The database consists of three columns: 1) peptide,



	A	B	C	D	E
1	N	H	F	S	G
2	2	3	0	0	0
3	2	3	1	0	0
4	2	4	0	0	0
5	2	4	1	0	0
6	2	5	0	0	0
7	2	5	1	0	0
8	2	6	0	0	0
9	2	6	1	0	0
10	2	7	0	0	0
11	2	7	1	0	0

Figure 3. Glycan database format

Precursor Mass Matching, 2) name of the mzXML file containing the MS/MS spectrum of the deglycosylated peptide and 3) the scan number in the specified mzXML file that corresponds to the peptide. All the mzXML files listed in this database should be added to the Library files for N- and/or O-linked analysis (Figure 5). Independent of glycoproteomics input files, the user can specify whether the library files are iTRAQ-labeled. For more details on the accepted formats, refer to the provided Examples.



	A	B	C	D
1	aAEnFTLLVk	13965	Eric_25cm_ITRAQ_1_1	
2	aAEnFTLLVk	13968	Eric_25cm_ITRAQ_1_1	
3	aAEnFTLLVk	11894	Eric_25cm_ITRAQ_1_2	
4	aAEnFTLLVk	11911	Eric_25cm_ITRAQ_1_2	
5	aAEnFTLLVk	11910	Eric_25cm_ITRAQ_1_3	
6	aAEnFTLLVk	11921	Eric_25cm_ITRAQ_1_3	
7	aAFNSGkVDIVAInDPFIDLnYmVYmFQYDSTHGk	16857	Eric_25cm_ITRAQ_1_2	
8	aAGRYHnQTLR	2066	Eric_25cm_ITRAQ_1_3	
9	aAIPSALDTnSSk	9076	Eric_25cm_ITRAQ_1_1	
10	aAIPSALDTnSSk	9080	Eric_25cm_ITRAQ_1_1	
11	aAIPSALDTnSSk	9182	Eric_25cm_ITRAQ_1_1	
12	aAIPSALDTnSSk	9183	Eric_25cm_ITRAQ_1_1	
13	aAIPSALDTnSSk	7759	Eric_25cm_ITRAQ_1_2	
14	aAIPSALDTnSSk	7781	Eric_25cm_ITRAQ_1_2	
15	aAIPSALDTnSSk	7867	Eric_25cm_ITRAQ_1_2	
16	aAIPSALDTnSSk	7882	Eric_25cm_ITRAQ_1_2	
17	aAIPSALDTnSSk	7796	Eric_25cm_ITRAQ_1_3	
18	aAIPSALDTnSSk	7801	Eric_25cm_ITRAQ_1_3	
19	aAnGSLR	1571	Eric_25cm_ITRAQ_1_2	
20	aAnGSLR	1915	Eric_25cm_ITRAQ_1_2	
21	aAnGSLR	2001	Eric_25cm_ITRAQ_1_2	

Figure 5. Spectral library matching peptide database format

Reporter Ion Extraction and Selection of Oxonium Ion-Containing Spectra for more details.

Sample Preparation

GPQuest 2.0 supports the following sample modifications: 1) iTRAQ 2) P-toluidine (CH_3 C_6H_4 NH_2) and 3) heavy isotope of P-toluidine (CD_3 C_6D_4 ND_2).

iTRAQ is routinely used for isobaric quantification in proteomics experiments and can be similarly applied for quantification of glycopeptides. iTRAQ tags react with the N-terminal of the peptide and side chain of lysine, shifting the monoisotopic peak by 144.102063 Da for each modification.

P-toluidine is a compound that reacts with the carboxyl groups of sialic acid residues on glycans as well as aspartic and glutamic acid residues of peptides. Therefore, each modification by light p-toluidine and heavy p-toluidine shifts the mass by 89.0629346 and 96.1068718 Da, respectively. Labeling of sialic acid with p-toluidine has been shown to reduce post-source decay of sialylated glycans [2].

Analysis Setup

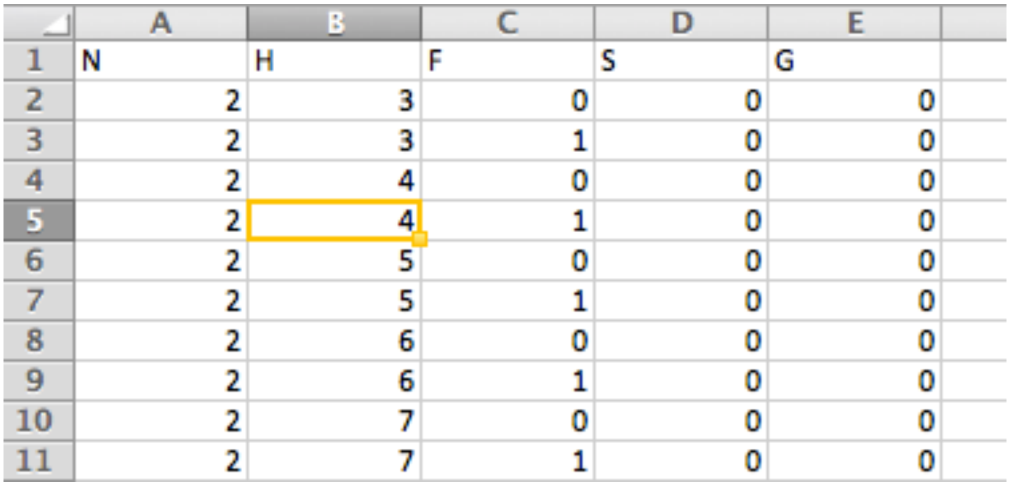
This panel is designed for setting the basic parameters used throughout the analysis. The input files for glycoproteomics analysis are accepted in mzXML format. Typically, the raw mass spectrometry files are converted to mzXML using the **msconvert tool in the Trans-Proteomic Pipeline (TPP)**, with the "centroid all scans" option selected. The converted mzXML files are added to the list of input files. Users can enter a project name, which will be used to create a folder and organize the output files. The tolerance used for matching the peaks at MS1 level and MS2 level can be determined in the ppm (precursor) and ppm (fragment ions) fields, respectively.

Note: To ensure compatibility of the input files with GPQuest, we strongly suggest using the msconvert tool for generating the mzXML files.

N-linked and O-linked glycosylation analysis

GPQuest can be used for analysis of both N- and O-linked glycopeptides using similar algorithms. Users can select the type of glycosylation analysis by checking the appropriate option. Both analyses can be performed concurrently and the user can specify distinct N-linked and O-linked glycan and peptide databases accordingly. The function of the glycosylation type prediction is to predict the type of glycosylation corresponding to each MS/MS spectrum based on its spectral features (manuscript in preparation). If this option is selected, the spectra that are predicted to belong to N-linked (O-linked) glycopeptides are only matched against the N-linked (O-linked) glycan and peptides databases. Otherwise, if both N-linked and O-linked databases are provided, both options are considered for each MS/MS spectrum regardless of the glycosylation type predicted by the algorithm.

Depending on which glycosylation type is selected, the path to N-linked and/or O-linked glycan databases should be specified in this panel. The accepted formats for the glycan databases are .xls, .xlsx and .csv and the database is a 5 column file with the following headers: N, H, F, S and G, where each column corresponds to the number of HexNAc, Hexose, Fucose, Neu5Ac and Neu5Gc residues, respectively (Figure 3). For more details on the accepted formats for glycan databases, refer to Examples.

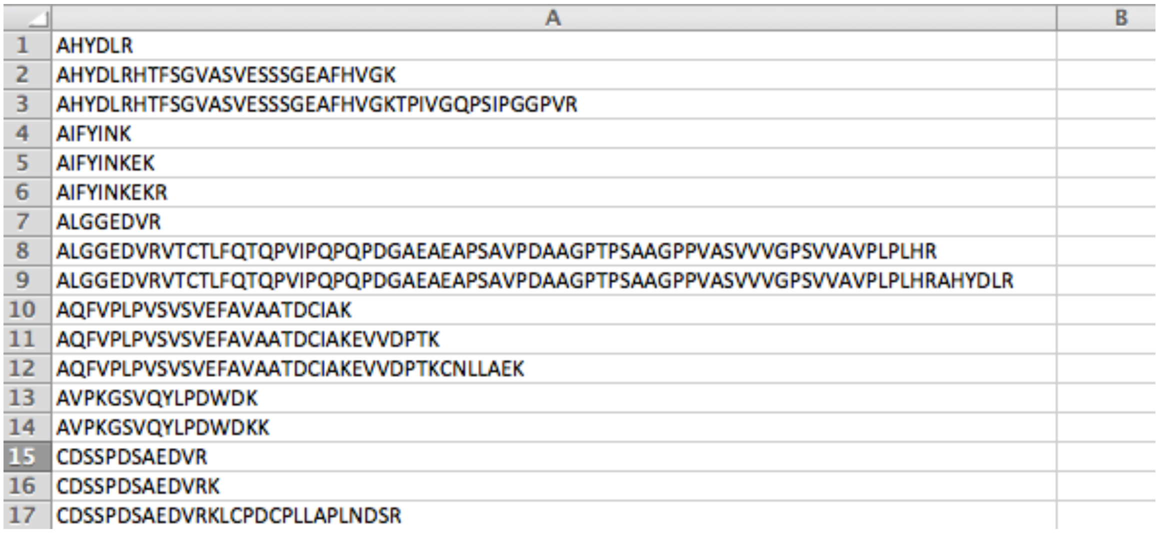


	A	B	C	D	E
1	N	H	F	S	G
2	2	3	0	0	0
3	2	3	1	0	0
4	2	4	0	0	0
5	2	4	1	0	0
6	2	5	0	0	0
7	2	5	1	0	0
8	2	6	0	0	0
9	2	6	1	0	0
10	2	7	0	0	0
11	2	7	1	0	0

Figure 3. Glycan database format

Precursor Mass Matching

The PPM algorithm uses the database of glycosite-containing peptides specified in this panel and the glycan database specified in the Analysis Setup panel to build the list of theoretical peptides in the sample. Depending on the parameters selected in the Analysis Setup panel, databases of N- and/or O-linked peptides should be specified here. These databases are single column .xls, .xlsx or .csv files containing the list of target peptides (Figure 4). Each peptide is a string of single letter amino acid codes listed in Table 1. Please refer to Examples for instances of the accepted format of the peptide database.



	A	B
1	AHYDLR	
2	AHYDLRHTFSGVASVESSSGEAFHVGK	
3	AHYDLRHTFSGVASVESSSGEAFHVGKTPIVGQPSIPGGPVR	
4	AIFYINK	
5	AIFYINKEK	
6	AIFYINKEKR	
7	ALGGEDVR	
8	ALGGEDVRVTCTLFQTQPVIPQPQPDGAEAEAPSAVPDAAGPTPSAAGPPVASVVVGPSVVAVPLPLHR	
9	ALGGEDVRVTCTLFQTQPVIPQPQPDGAEAEAPSAVPDAAGPTPSAAGPPVASVVVGPSVVAVPLPLHRAHYDLR	
10	AQFVPLPVSVSVEFAVAATDCIAK	
11	AQFVPLPVSVSVEFAVAATDCIAKEVVDPTK	
12	AQFVPLPVSVSVEFAVAATDCIAKEVVDPTKCNLLAEK	
13	AVPKGSVQYLPDWDK	
14	AVPKGSVQYLPDWDKK	
15	CDSSPDSAEDVR	
16	CDSSPDSAEDVRK	
17	CDSSPDSAEDVRKLCPDCPLLAPLNDSR	

Figure 4. Precursor mass matching peptide database format

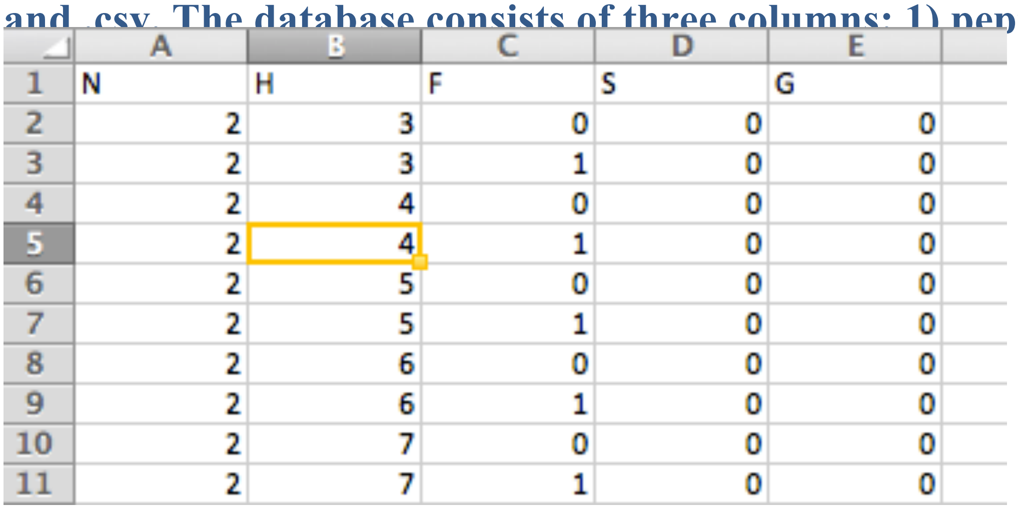
Note: Presence of amino acids except the ones listed in Table 1 could result in disruption of the analysis and is a common source of error.

Table 1. List of accepted amino acids in the peptide databases.

Amino Acid	Letter Code	Monoisotopic Mass
Alanine	A	71.037114
Cysteine	C	103.009184
Aspartic acid	D	115.026943
Glutamic acid	E	129.042593
Phenylalanine	F	147.068414
Glycine	G	57.021464
Histidine	H	137.058912
Isoleucine	I	113.084064
Lysine	K	128.094963
Leucine	L	113.084064
Methionine	M	131.040485
Asparagine	N	114.042927
Proline	P	97.052764
Glutamine	Q	128.058578
Arginine	R	156.101111
Serine	S	87.032028
Threonine	T	101.047678
Valine	V	99.068414
Tryptophan	W	186.079313
Tyrosine	Y	163.063329

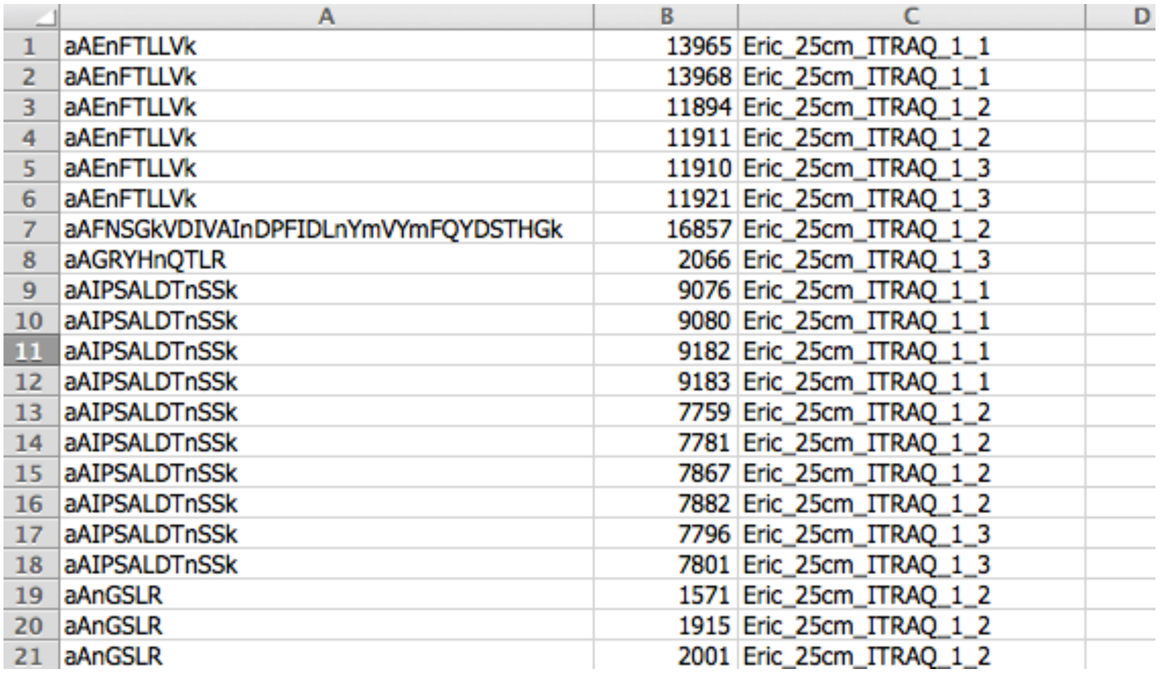
Spectral Library Matching

The implemented SLM algorithm in GPQuest first builds the spectral library based on the mzXML files corresponding to the glycosite-containing peptides and a peptide database that determines the target peptides in the glycoproteomics analysis. The path to the peptide database is specified for N- and/or O-linked peptides depending on the parameters selected in Analysis Setup. The accepted database formats are .xls, .xlsx and .csv. The database consists of three columns: 1) peptide, Precursor Mass Matching, 2) name of the mzXML file containing the MS/MS spectrum of the deglycosylated peptide and 3) the scan number in the specified mzXML file that corresponds to the peptide. All the mzXML files listed in this database should be added to the Library files for N- and/or O-linked analysis (Figure 5). Independent of glycoproteomics input files, the user can specify whether the library files are iTRAQ-labeled. For more details on the accepted formats, refer to the provided Examples.



	A	B	C	D	E
1	N	H	F	S	G
2	2	3	0	0	0
3	2	3	1	0	0
4	2	4	0	0	0
5	2	4	1	0	0
6	2	5	0	0	0
7	2	5	1	0	0
8	2	6	0	0	0
9	2	6	1	0	0
10	2	7	0	0	0
11	2	7	1	0	0

Figure 3. Glycan database format



	A	B	C	D
1	aAEnFTLLVk	13965	Eric_25cm_ITRAQ_1_1	
2	aAEnFTLLVk	13968	Eric_25cm_ITRAQ_1_1	
3	aAEnFTLLVk	11894	Eric_25cm_ITRAQ_1_2	
4	aAEnFTLLVk	11911	Eric_25cm_ITRAQ_1_2	
5	aAEnFTLLVk	11910	Eric_25cm_ITRAQ_1_3	
6	aAEnFTLLVk	11921	Eric_25cm_ITRAQ_1_3	
7	aAFNSGkVDIVAInDPFIDLnYmVYmFQYDSTHGk	16857	Eric_25cm_ITRAQ_1_2	
8	aAGRYHnQTLR	2066	Eric_25cm_ITRAQ_1_3	
9	aAIPSALDTnSSk	9076	Eric_25cm_ITRAQ_1_1	
10	aAIPSALDTnSSk	9080	Eric_25cm_ITRAQ_1_1	
11	aAIPSALDTnSSk	9182	Eric_25cm_ITRAQ_1_1	
12	aAIPSALDTnSSk	9183	Eric_25cm_ITRAQ_1_1	
13	aAIPSALDTnSSk	7759	Eric_25cm_ITRAQ_1_2	
14	aAIPSALDTnSSk	7781	Eric_25cm_ITRAQ_1_2	
15	aAIPSALDTnSSk	7867	Eric_25cm_ITRAQ_1_2	
16	aAIPSALDTnSSk	7882	Eric_25cm_ITRAQ_1_2	
17	aAIPSALDTnSSk	7796	Eric_25cm_ITRAQ_1_3	
18	aAIPSALDTnSSk	7801	Eric_25cm_ITRAQ_1_3	
19	aAnGSLR	1571	Eric_25cm_ITRAQ_1_2	
20	aAnGSLR	1915	Eric_25cm_ITRAQ_1_2	
21	aAnGSLR	2001	Eric_25cm_ITRAQ_1_2	

Figure 5. Spectral library matching peptide database format

Reporter Ion Extraction

The objective of this panel is to provide the m/z values of the reporter ions of interest and extracting their intensities. This panel is activated upon checking the Reporter ion extraction function in the Algorithms panel. The m/z values provided here are searched in the MS/MS spectra of oxonium ion-containing spectra with a tolerance provided in the ppm (Fragment Ions) in the Analysis Setup panel.

Selection of Oxonium Ion-Containing Spectra

Oxonium ions are low mass ions that belong to monosaccharides or disaccharides cleaved off glycans during HCD fragmentation and are signature of MS/MS of glycopeptides. This panel provides the means to set the criteria for distinguishing the MS/MS spectra of glycopeptides from those of non-glycosylated peptides. A list of the most common oxonium ions are provided: 138 (internal fragment of HexNAc), 145 (Hex− H2O), 163 (Hex), 168 (HexNAc−2H2O), 186 (HexNAc− H2O), 204 (HexNAc), 325 (Hex2), 366 (HexHexNAc), 274 (Neu5Ac−H2O), and 292 (Neu5Ac).

The presence of ions added to the essential list are considered mandatory by GPQuest for further analysis by PMM, SLM or Reporter Ion Extraction. The ions added to the non-essential list are deemed to be optional. The users can also determine the minimum number of oxonium ions required in spectra for glycoproteomics analysis.

Here is an example of how to use this panel to select the oxonium ion spectra: Assume that 138, 204 and 163 are added to the essential and 168, 186 and 274 are added to non-essential oxonium ion lists. Minimum number of required oxonium ions in the top 10% of the peaks is set to 4. According to these criteria, GPQuest first sorts the peaks in each MS/MS spectrum based on their intensities and searches the top 10% for ions at 138, 204, 163, 168 and 186. Only the spectra that contain all three ions at 138, 204 and 163 and at least one of the two ions at 168 and 186 will be analyzed by the matching algorithm for identification of glycopeptides.

Fragment and Intact Peptide Ions

PMM and SLM normally result in multiple glycopeptide spectral matches (GPSM) for each oxonium ion-containing spectrum. To find the most likely GPSM for each spectrum, additional criteria are needed. At this step, matching the fingerprint of peptides in each GPSM with mass spectral peaks in the MS/MS spectrum could eliminate the unlikely GPSMs.

Peptide fragment ions

Fragment b and y ions are the more dominant forms of peptide fragment ions generated using HCD fragmentation. Fragment b and y ions are generated when the charge is retained on the N-terminal and C-terminal of the peptide, respectively. This panel allows the user to indicate which fragment ions should be used for refining the search and scoring the GPSMs. Singly, doubly and triply charge b and y ions can be selected for constructing the theoretical fingerprint of each peptide. GPQuest matches the list of ions in the peptide fingerprint with the deisotoped peaks in the MS/MS spectrum.

Intact peptide ions

Intact peptide ions are fragment ions generated by partial cleavage of glycans from glycopeptides. Some intact peptide ions such as pep(+) and pep-HexNAc(+) are among dominant ions in the mass spectra of HCD fragmented glycopeptides. Number of matching intact peptide ions is another indication of how well a GPSM matches its corresponding spectrum. A list of nine forms of intact peptide ions are provided in this panel: pep(+), pep-83(+), pep-HexNAc(+), pep-HexNAc-Fuc(+), pep-2HexNAc(+), pep(++), pep-HexNAc(++), pep(+++) and pep-HexNAc(+++). The intact peptide ions can be flagged as essential or non-essential. The set of essential and non-essential intact ions will be used to score each GPSM as described in Scoring of glycopeptide-spectral matches. Additionally, GPQuest retains a GPSM only if the corresponding spectrum contains all the essential intact ions.

The number of observed intact peptide ions for each peptide depends on the length of that peptide. More specifically, shorter peptides are more likely to generate intense intact peptide ion peaks. In this panel, GPQuest gives the user the opportunity to set a minimum on the required number of matching intact peptide ions for each GPSMs and to program this number as a function of the length of the peptide. All GPSMs that do not meet these requirements will be removed from the output.

Peak selection

To minimize the potential of matching the b, y and intact peptide ions with the background noise, low intensity peaks could be eliminated from the MS/MS spectrum using the options provided in the peak selection subpanel. Three options are provided:

1) Peaks in the top x% of TIC: the mass spectral peaks in the MS/MS spectrum are sorted based on the intensity and only the peaks contributing to the top x% of the total ion current (TIC) are searched for b, y and intact peptide ions, where x could theoretically range from 1 to 100.
2) Top x peaks: the mass spectral peaks in the MS/MS spectrum are sorted based on the intensity and only the top x peaks are searched for b, y and intact peptide ions, where x could theoretically be any number greater than 1.

3) Peaks with minimum intensity of x: all the mass spectral peaks with intensities less than x are removed from the MS/MS spectrum. The user can specify the x in the intensity unit of the spectra in mass spectrometry input files.

Filtering

Precursor mass matching and spectral library matching each generate a list of glycopeptide-spectral matches (GPSM). At this point and before filtering, each MS/MS spectrum could correspond to multiple GPSMs. The objective of the filtering panel is to 1) assign unique GPSMs to each oxonium ion-containing MS/MS spectrum and 2) remove the false matches to achieve the target false discovery rate.

Scoring of glycopeptide-spectral matches

In order to narrow down the list of GPSMs generated by matching algorithms to one GPSM per MS/MS spectrum, a score is assigned. Three scoring formula are provided in this panel and only one of the three scores can be selected for each analysis:

1) Number of fragment ions: The number of matching peptide fragment b and y ions

2) Intensity score: The percentage of the intensity of non-oxonium ion peaks covered by the matching peptide b and y and intact peptide ions plus the number of matching peptide b and y and intact peptide ions

3) Morpheus score: The number of matching peptide b and y and intact peptide ions plus the fraction of the intensity of non-oxonium ion peaks covered by the matching b and y intact peptide ions

Number of fragment ion is the most basic formulation for scoring of GPSMs and in essence yields similar results to the Morpheus score. Intensity score is a helpful measure for assessing the match to sparse MS/MS spectra that have resulted due to low abundance of the precursor ion or poor fragmentation. In general, Morpheus score, which is adapted for HCD fragmented glycopeptides from the score introduced by Wenger et al [3], results in more identification compared to the other scoring schemes.

False discovery rate estimation

The false discover rate (FDR) provides a measure of the number of false assignments of glycopeptides to MS/MS spectra. To estimate the FDR, first a database of decoy peptides is generated and included in the matching search. After the matching algorithm generates the list of GPSMs, the FDR is given by twice the number of matches to the decoy peptides divided by the total number of matches. In GPQuest, the decoy database originates from the database of glycosite-containing peptides. The peptides in this database are first stringed together, the amino acids randomly shuffled and the string cut into decoy peptides with the same length as the target peptides in the database. This strategy results in a decoy database that has exactly the same number of peptides as the target peptide database. In addition, the m/z distribution of the decoy peptides resembles that of the target peptides. In proteomics assignment algorithms, it is customary to use the reverse of the target database as decoy. However, since in glycoproteomics the intact peptide ions play a role in scoring of the glycopeptides, using the reversed database would in a biased assessment. This decoy strategy is discussed in more detail by Eshghi et al [1].

On the GUI, the user is given three choices for cut off FDR: 1) 1%: high confidence, 2) 5%: medium confidence and 3) 10%: low confidence. Multiple FDR selections are allowed and each will result in a separate output excel file.

The original search results contain all the possible GPSMs for each MS/MS spectrum, which results in a large output file. The user is given the choice to skip writing this information into the output file to save time by checking the “Skip saving the original search results before filtering” option.

Monoisotopic mass identification

The m/z reported in the mzXML file as the precursor mass is not necessarily the monoisotopic peak. Since correct assignment of the MS/MS spectra relies on precise identification of the precursor monoisotopic mass, GPQuest uses an algorithm to identify the correct monoisotopic peak. This is accomplished by averaging several MS1 spectra in the vicinity of the target spectrum to gain a more accurate estimate of the isotopic mass spectral cluster for the precursor. The first peak present in the revised cluster, with the lowest mass, is selected as the monoisotopic peak. For more details on this algorithm please refer to the work by Eshghi et al [1].

Interpreting the output files

For each analysis three output files are generated:

1) Original search file, which includes all GPSMs before filtering the results to achieve the desired FDR. This file is named in the following way: InputFileName_Algorithm_Raw_Output_Date_Time.csv
2) Filtered list of GPSMs, which includes the GPSMs achieved after filtering the original search file based on the desired FDR. This file is named in the following way: InputFileName_Algorithm_FDR_FDRvalue_Data_Time.csv
3) Parameters, which saves all the parameters that were used for the analysis. This file is named in the following way: parameters_Data_Time.txt

In addition to the output files, the parameters that are used for the analysis can be saved as a template in a .mat file using the Save Parameters button and retrieved using the Load Parameters button.

The output files are in .csv format and depending on the algorithm used include all or some of the following columns:

MS1 Scan Number: Precursor scan number

MS2 Scan Number: MS/MS spectrum scan number

Retention Time: Retention time reported in the mzXML file

Peptide Sequence: Amino acid sequence of the peptide in the GPSM

Peptide Modification: Any modifications on the peptide in the GPSM

Glycosylation Type (Glycan Datbase/Predicted [Probability]: Type of the glycan in the GPSM based on the database of origin (N for N-linked, O for O-linked, B if the glycan belongs to both of the databases and ? if the glycan could not be found in any of the databases, which only could happen in case of the SLM algorithm)/ Glycosylation type predicted by the algorithm based on the spectral features of the MS/MS spectrum (N for N-linked and O for O-linked). The predicted glycosylation type is followed by the numerical value of the probability of a correct prediction. In the output of SLM algorithm, the spectral library that the peptide in the GPSM originates from is also shown in this column.

Candidate Glycan Exact Mass: Exact mass of the glycan in the GPSM (Da)

Candidate Glycan Composition: Composition of the glycan in the GPSM

Candidate Glycan Modification: Any modifications on the glycan in the GPSM

Mass Difference Corresponding to Glycan: Precursor Mass – Peptide Mass (SLM only)

Precursor mz: Precursor ion m/z reported in the mzXML file

Precursor Mass: Precursor monoisotopic mass (after monoisotopic mass correction)

Precursor Charge: Precursor charge reported in the mzXML file. If the charge is not reported in the mzXML file, GPQuest attempts to calculate the charge estate by analyzing the MS1 signal.

Mass Shift: Difference between the precursor mass reported in the mzXML file and the monoisotopic mass calculated after peak correction (Da)

M Oxidation: Indicates whether the peptide has any modifications

Library Y+: Singly charged peptide y ions in the spectral library (Only in SLM)

Library Y++: Doubly charged peptide y ions in the spectral library (Only in SLM)

Library Y+++: Triply charged peptide y ions in the spectral library (Only in SLM)

Library B+: Singly charged peptide b ions in the spectral library (Only in SLM)

Library B++: Doubly charged peptide b ions in the spectral library (Only in SLM)

Library B+++: Triply charged peptide b ions in the spectral library (Only in SLM)

Matched Experimental Y+: Library y+ ions identified in the spectrum (Only in SLM)

Matched Experimental Y++: Library y++ ions identified in the spectrum (Only in SLM)

Matched Experimental Y+++: Library y+++ ions identified in the spectrum (Only in SLM)

Matched Experimental B+: Library b+ ions identified in the spectrum (Only in SLM)

Matched Experimental B++: Library b++ ions identified in the spectrum (Only in SLM)

Matched Experimental B+++: Library b+++ ions identified in the spectrum (Only in SLM)

Matched Experimental % Y+: Percentage of library y+ ions identified in the spectrum (Only in SLM)

Matched Experimental % Y++: Percentage of library y++ ions identified in the spectrum (Only in SLM)

Matched Experimental % Y+++: Percentage of library y+++ ions identified in the spectrum (Only in SLM)

Matched Experimental % B+: Percentage of library b+ ions identified in the spectrum (Only in SLM)

Matched Experimental % B++: Percentage of library b++ ions identified in the spectrum (Only in SLM)

Matched Experimental % B+++: Percentage of library b+++ ions identified in the spectrum (Only in SLM)

Matched Theoretical Y+: Theoretical y+ ions identified in the spectrum

Matched Theoretical Y++: Theoretical y++ ions identified in the spectrum

Matched Theoretical Y+++: Theoretical y+++ ions identified in the spectrum

Matched Theoretical B+: Theoretical b+ ions identified in the spectrum

Matched Theoretical B++: Theoretical b++ ions identified in the spectrum

Matched Theoretical B+++: Theoretical b+++ ions identified in the spectrum

Matched Theoretical % Y+: Percentage of theoretical y+ ions identified in the spectrum

Matched Theoretical % Y++: Percentage of theoretical y++ ions identified in the spectrum

Matched Theoretical % Y+++: Percentage of theoretical y+++ ions identified in the spectrum

Matched Theoretical % B+: Percentage of theoretical b+ ions identified in the spectrum

Matched Theoretical % B++: Percentage of theoretical b++ ions identified in the spectrum

Matched Theoretical % B+++: Percentage of theoretical b+++ ions identified in the spectrum

Peptide +: Intensity of intact peptide ions pep(+)

Peptide-HexNAc/2 +: Intensity of intact peptide ions pep-83(+)

Peptide-HexNAc +: Intensity of intact peptide ions pep-HexNAc(+)

Peptide-HexNAc-Fuc +: Intensity of intact peptide ions pep-HexNAc-Fuc(+)

Peptide-2HexNAc +: Intensity of intact peptide ions pep-2HexNAc(+)

Peptide ++: Intensity of intact peptide ions pep(++)

Peptide-HexNAc ++: Intensity of intact peptide ions pep-HexNAc(++)

Peptide +++: Intensity of intact peptide ions pep(+++)

Peptide-HexNAc +++: Intensity of intact peptide ions pep-HexNAc(+++)

Decoy: Set to 1 for peptides that belong to the decoy database

PSM Score Number: GPSM score based on the Number of Fragment Ions formula

PSM Score Intensity: GPSM score based on the Intensity formula

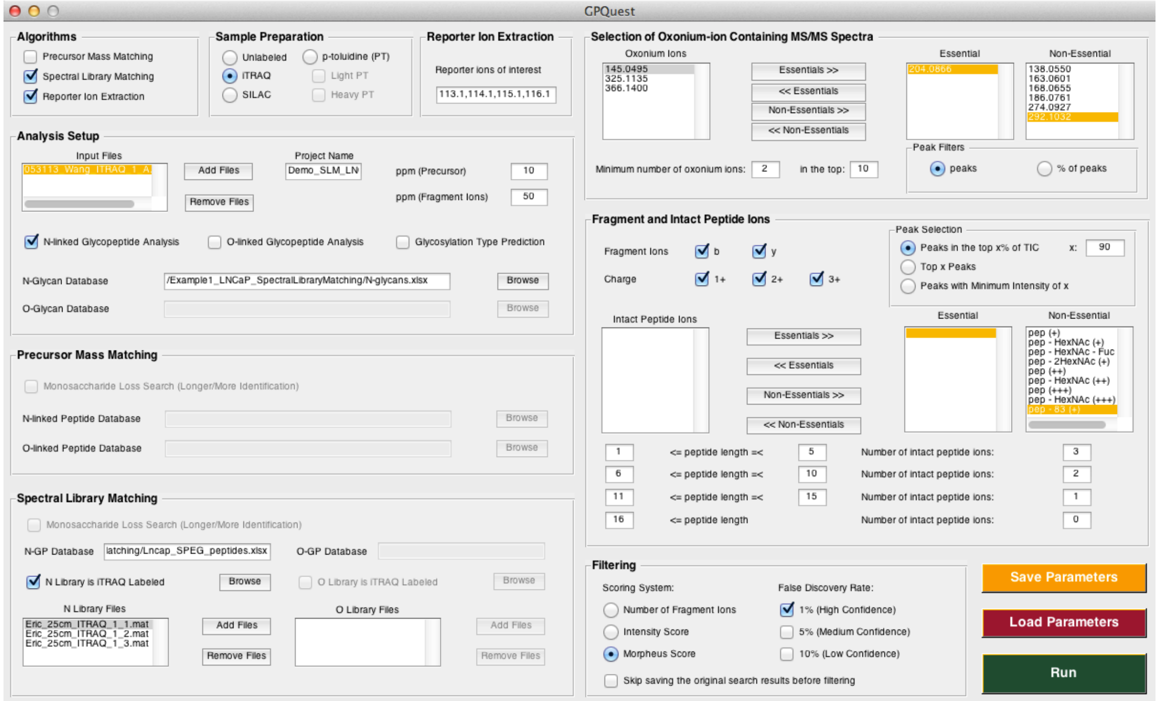
PSM Score Morpheus: GPSM score based on the Morpheus formula

Matching Error in ppm: ppm error between the glycopeptide in the GPSM and the precursor mass

Examples

Example 1: N-glycoproteomics analysis of LNCaP cell lysates using spectral library matching

LNCaP is an androgen dependent prostate cancer cell line. The tryptic peptides extracted from the cells are fractionated first using bRPLC and one fraction is analyzed in GPQuest using the spectral library matching for identification of intact N-glycopeptides. For details on sample preparation, please refer to the publication by Eshghi et al [1]. Figure 6 shows the parameters used in this example.



GPQuest

Algorithms

☐ Precursor Mass Matching
☑ Spectral Library Matching
☑ Reporter Ion Extraction

Sample Preparation

○ Unlabeled ○ p-toluidine (PT)
● iTRAQ ☐ Light PT
○ SILAC ☐ Heavy PT

Reporter Ion Extraction

Reporter ions of interest: 113.1,114.1,115.1,116.1

Analysis Setup

Input Files: 053113_Wang_ITRAQ_1_A
Project Name: Demo_SLM_LN
ppm (Precursor): 10
ppm (Fragment Ions): 50

☑ N-linked Glycopeptide Analysis ☐ O-linked Glycopeptide Analysis ☐ Glycosylation Type Prediction

N-Glycan Database: /Example1_LNCaP_SpectralLibraryMatching/N-glycans.xlsx
O-Glycan Database:

Precursor Mass Matching

☐ Monosaccharide Loss Search (Longer/More Identification)

N-linked Peptide Database:
O-linked Peptide Database:

Spectral Library Matching

☐ Monosaccharide Loss Search (Longer/More Identification)

N-GP Database: latching/Lncap_SPEG_peptides.xlsx
O-GP Database:

☑ N Library is iTRAQ Labeled ☐ O Library is iTRAQ Labeled

N Library Files:
Eric_25cm_ITRAQ_1_1.mat
Eric_25cm_ITRAQ_1_2.mat
Eric_25cm_ITRAQ_1_3.mat

O Library Files:

Selection of Oxonium-ion Containing MS/MS Spectra

Oxonium Ions:
145.0495
325.1135
366.1400

Essential:
204.0866

Non-Essential:
138.0550
163.0601
168.0655
186.0761
274.0927
292.1032

Minimum number of oxonium ions: 2 in the top: 10

Peak Filters: ● peaks ○ % of peaks

Fragment and Intact Peptide Ions

Fragment Ions ☑ b ☑ y
Charge ☑ 1+ ☑ 2+ ☑ 3+

Peak Selection:
● Peaks in the top x% of TIC x: 90
○ Top x Peaks
○ Peaks with Minimum Intensity of x

Intact Peptide Ions:

Essential:

Non-Essential:
pep (+)
pep - HexNAc (+)
pep - HexNAc - Fuc
pep - 2HexNAc (+)
pep (++)
pep - HexNAc (++)
pep (+++)
pep - HexNAc (+++)
pep - 83 (+)

1 <= peptide length =< 5 Number of intact peptide ions: 3
6 <= peptide length =< 10 Number of intact peptide ions: 2
11 <= peptide length =< 15 Number of intact peptide ions: 1
16 <= peptide length Number of intact peptide ions: 0

Filtering

Scoring System:
○ Number of Fragment Ions
○ Intensity Score
● Morpheus Score

False Discovery Rate:
☑ 1% (High Confidence)
☐ 5% (Medium Confidence)
☐ 10% (Low Confidence)

☐ Skip saving the original search results before filtering

Figure 6. Spectral library matching for N-glycoproteomics analysis of LNCaP sample. GPSMs are scored using the Morpheus formula.

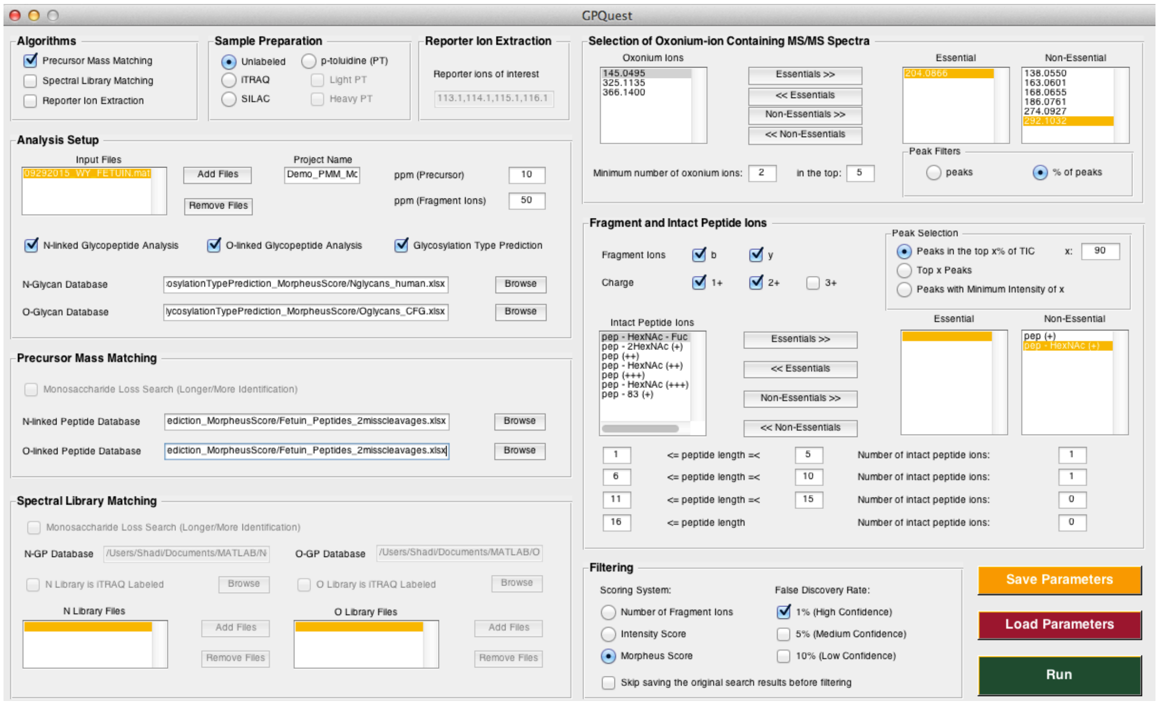
Example 1 includes the following files:

1. Input: 053113_Wang_ITRAQ_1_A.mat
2. N-glycan database: N-glycans.xls
3. Peptide database with reference to the library files: Lncap_SPEG_peptides.xls
4. Parameters template file: Example1_parameters.mat
5. The generated output files are stored in the Demo_SLM_LNCap folder. In order to recreate this analysis, you could load the template file (Example1_parameters.mat) using the Load Parameters option and change all the directory paths accordingly.

Example 2: Analysis of Fetuin using precursor mass matching and Morpheus scoring with glycosylation type prediction

Fetuin is a glycoprotein containing both N- and O-linked glycosylation and is used here as an example of conducting simultaneous N- and O-linked glycoproteomics analysis.

For this example, precursor mass matching is applied for matching the spectra to intact glycopeptides. Figure 7 depicts the parameters used in this example.



Algorithms
☑ Precursor Mass Matching
□ Spectral Library Matching
□ Reporter Ion Extraction

Sample Preparation
● Unlabeled ○ p-toluidine (PT)
○ iTRAQ □ Light PT
○ SILAC □ Heavy PT

Reporter Ion Extraction
Reporter ions of interest: 113.1,114.1,115.1,116.1

Analysis Setup
Input Files: 09292015_WY_FETUIN.mat
Project Name: Demo_PMM_Mc
ppm (Precursor): 10
ppm (Fragment Ions): 50
☑ N-linked Glycopeptide Analysis ☑ O-linked Glycopeptide Analysis ☑ Glycosylation Type Prediction
N-Glycan Database: osylationTypePrediction_MorpheusScore/Nglycans_human.xlsx
O-Glycan Database: lycosylationTypePrediction_MorpheusScore/Oglycans_CFG.xlsx

Precursor Mass Matching
□ Monosaccharide Loss Search (Longer/More Identification)
N-linked Peptide Database: ediction_MorpheusScore/Fetuin_Peptides_2misscleavages.xlsx
O-linked Peptide Database: ediction_MorpheusScore/Fetuin_Peptides_2misscleavages.xlsx

Spectral Library Matching
□ Monosaccharide Loss Search (Longer/More Identification)
N-GP Database: /Users/Shadi/Documents/MATLAB/N
O-GP Database: /Users/Shadi/Documents/MATLAB/O
□ N Library is iTRAQ Labeled
□ O Library is iTRAQ Labeled
N Library Files
O Library Files

Selection of Oxonium-ion Containing MS/MS Spectra
Oxonium Ions: 145.0495, 325.1135, 366.1400
Essential: 204.0866
Non-Essential: 138.0550, 163.0601, 168.0655, 186.0761, 274.0927, 292.1032
Minimum number of oxonium ions: 2 in the top: 5
Peak Filters: ○ peaks ● % of peaks

Fragment and Intact Peptide Ions
Fragment Ions ☑ b ☑ y
Charge ☑ 1+ ☑ 2+ □ 3+
Peak Selection:
● Peaks in the top x% of TIC x: 90
○ Top x Peaks
○ Peaks with Minimum Intensity of x
Intact Peptide Ions: pep - HexNAc - Fuc, pep - 2HexNAc (+), pep (++), pep - HexNAc (++), pep (+++), pep - HexNAc (+++), pep - 83 (+)
Essential
Non-Essential: pep (+), pep - HexNAc (+)

1	<= peptide length =<	5	Number of intact peptide ions:	1
6	<= peptide length =<	10	Number of intact peptide ions:	1
11	<= peptide length =<	15	Number of intact peptide ions:	0
16	<= peptide length		Number of intact peptide ions:	0

Filtering
Scoring System:
○ Number of Fragment Ions
○ Intensity Score
● Morpheus Score
False Discovery Rate:
☑ 1% (High Confidence)
□ 5% (Medium Confidence)
□ 10% (Low Confidence)
□ Skip saving the original search results before filtering

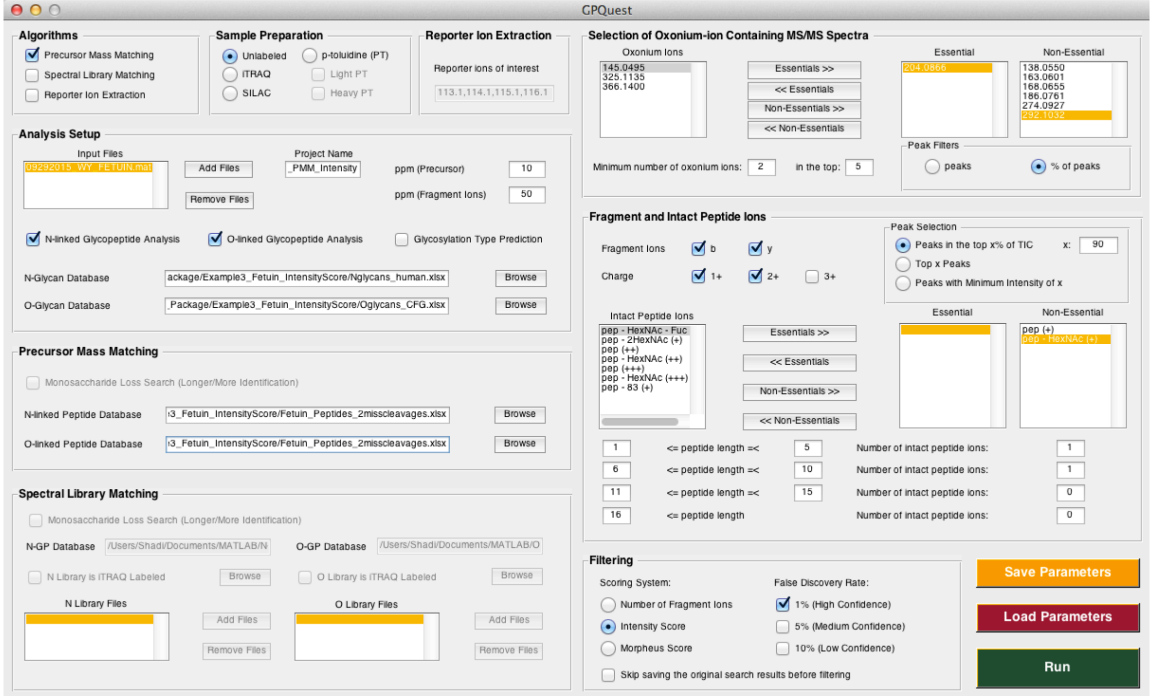
Figure 7. Precursor mass matching for glycoproteomics analysis of Fetuin sample. GPSMs are scored using the Morpheus formula and the glycosylation type prediction feature is activated.

Example 2 includes the following files:

6. Input: 09292015_WY_FETUIN.mat
7. N-glycan database: Nglycans_human.xlsx
8. O-glycan database: Oglycans_CFG.xlsx
9. Peptide database (used as both N- and O-linked peptide database): Fetuin_Peptides_2misscleavages.xlsx
10. Parameters template file: Example2_parameters.mat
11. The generated output files are stored in the Demo_PMM_Morpheus folder. In order to recreate this analysis, you could load the template file (Example2_parameters.mat) using the Load Parameters option and change all the directory paths accordingly.

Example 3: Analysis of Fetuin using precursor mass matching and Intensity scoring without glycosylation type prediction

This example is similar to example 2 with the exception that instead of the Morpheus formula, GPSMs are scored based on the Intensity formula. Also, glycosylation type prediction is deactivated, meaning that each MS/MS spectrum, regardless of its predicted glycosylation type, is matched with both N- and O-linked glycopeptide databases. Figure 8 depicts the parameters used in this example.



Algorithms
☑ Precursor Mass Matching
☐ Spectral Library Matching
☐ Reporter Ion Extraction

Sample Preparation
● Unlabeled ○ p-toluidine (PT)
○ iTRAQ ☐ Light PT
○ SILAC ☐ Heavy PT

Reporter Ion Extraction
Reporter ions of interest: 113.1,114.1,115.1,116.1

Analysis Setup
Input Files: 09292015_WY_FETUIN.mat
Project Name: _PMM_Intensity
ppm (Precursor): 10
ppm (Fragment Ions): 50
☑ N-linked Glycopeptide Analysis ☑ O-linked Glycopeptide Analysis ☐ Glycosylation Type Prediction
N-Glycan Database: ackage/Example3_Fetuin_IntensityScore/Nglycans_human.xlsx
O-Glycan Database: _Package/Example3_Fetuin_IntensityScore/Oglycans_CFG.xlsx

Precursor Mass Matching
☐ Monosaccharide Loss Search (Longer/More Identification)
N-linked Peptide Database: :3_Fetuin_IntensityScore/Fetuin_Peptides_2misscleavages.xlsx
O-linked Peptide Database: :3_Fetuin_IntensityScore/Fetuin_Peptides_2misscleavages.xlsx

Spectral Library Matching
☐ Monosaccharide Loss Search (Longer/More Identification)
N-GP Database: /Users/Shadi/Documents/MATLAB/N
O-GP Database: /Users/Shadi/Documents/MATLAB/O
☐ N Library is iTRAQ Labeled ☐ O Library is iTRAQ Labeled
N Library Files
O Library Files

Selection of Oxonium-ion Containing MS/MS Spectra
Oxonium Ions: 145.0495, 325.1135, 366.1400
Essential: 204.0866
Non-Essential: 138.0550, 163.0601, 168.0655, 186.0761, 274.0927, 292.1032
Minimum number of oxonium ions: 2 in the top: 5
Peak Filters: ○ peaks ● % of peaks

Fragment and Intact Peptide Ions
Fragment Ions ☑ b ☑ y
Charge ☑ 1+ ☑ 2+ ☐ 3+
Peak Selection
● Peaks in the top x% of TIC x: 90
○ Top x Peaks
○ Peaks with Minimum Intensity of x
Intact Peptide Ions: pep - HexNAc - Fuc, pep - 2HexNAc (+), pep (++), pep - HexNAc (++), pep (+++), pep - HexNAc (+++), pep - 83 (+)
Essential
Non-Essential: pep (+), pep - HexNAc (+)
1 <= peptide length =< 5 Number of intact peptide ions: 1
6 <= peptide length =< 10 Number of intact peptide ions: 1
11 <= peptide length =< 15 Number of intact peptide ions: 0
16 <= peptide length Number of intact peptide ions: 0

Filtering
Scoring System:
○ Number of Fragment Ions
● Intensity Score
○ Morpheus Score
False Discovery Rate:
☑ 1% (High Confidence)
☐ 5% (Medium Confidence)
☐ 10% (Low Confidence)
☐ Skip saving the original search results before filtering

Figure 8. Precursor mass matching for glycoproteomics analysis of Fetuin sample. GPSMs are scored using the Intensity formula and the glycosylation type prediction feature is deactivated.

Example 3 includes the following files:

1. Input: 09292015_WY_FETUIN.mat
2. N-glycan database: Nglycans_human.xlsx
3. O-glycan database: Oglycans_CFG.xlsx
4. Peptide database (used as both N- and O-linked peptide database): Fetuin_Peptides_2misscleavages.xlsx
5. Parameters template file: Example3_parameters.mat
6. The generated output files are stored in the Demo_PMM_Intensity folder. In order to recreate this analysis, you could load the template file (Example3_parameters.mat) using the Load Parameters option and change all the directory paths accordingly.

References and Additional Reading

1. Toghi Eshghi, S., Shah, P., Yang, W., Li, X. & Zhang, H. GPQuest: A Spectral Library Matching Algorithm for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides. *Anal. Chem.* **87,** 5181–5188 (2015).
2. Shah, P. *et al.* Mass spectrometric analysis of sialylated glycans with use of solid-phase labeling of sialic acids. *Anal. Chem.* **85,** 3606–13 (2013).
3. Wenger, C. D. & Coon, J. J. A Proteomics Search Algorithm Specifically Designed for High- Resolution Tandem Mass Spectra. *J. Proteome Res.* **12,** 1377–1386 (2013).