A computational framework for discovery of glycoproteomic biomarkers

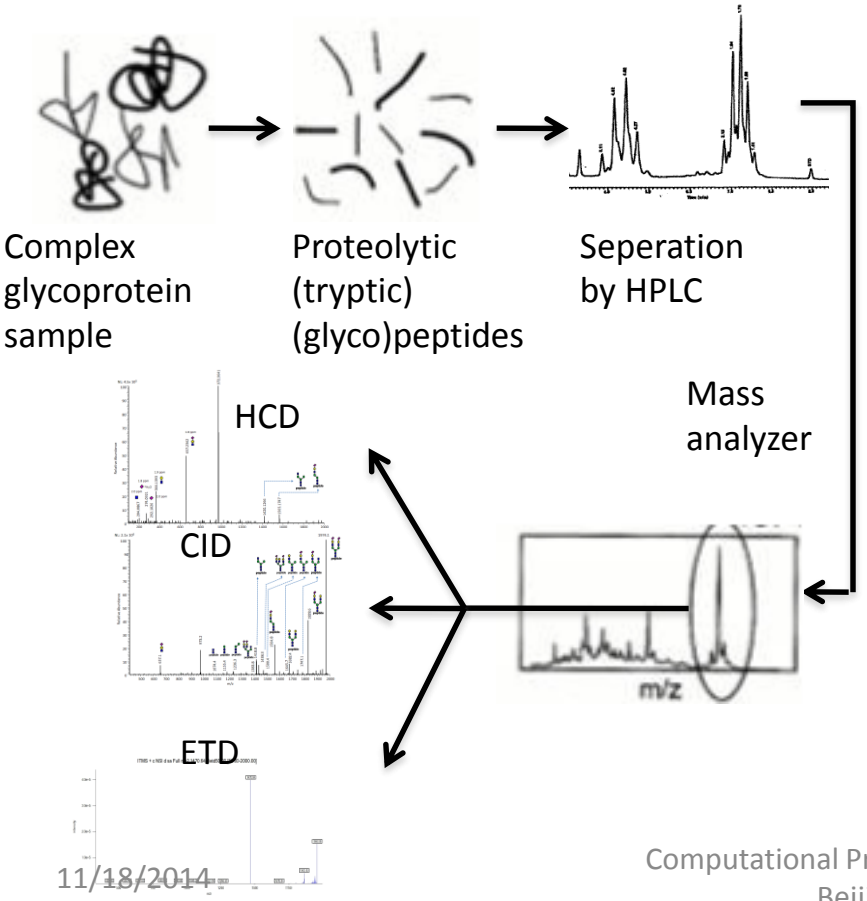
Haixu Tang, Anoop Mayampurath, Chuan-Yih Yu
Indiana University, Bloomington

Yehia Mechref, Erwang Song
Texas Tech University

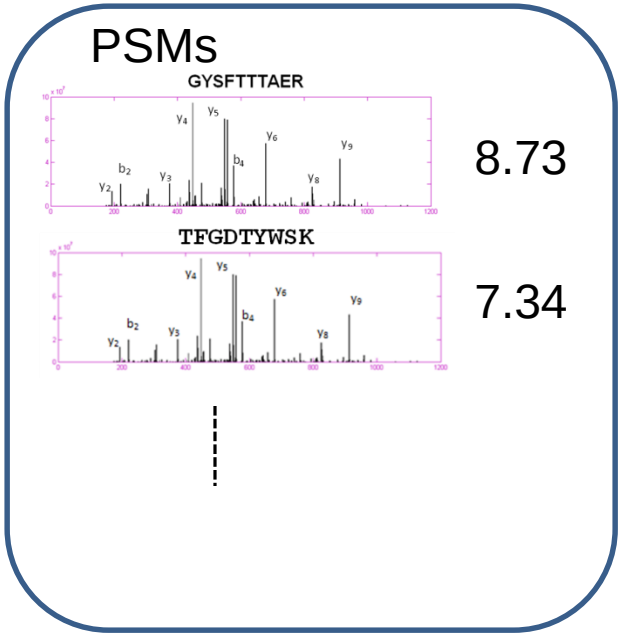
- Goal: to build computational tools that can identify (and quantify) intact glycopeptides in complex samples by using routine proteomic instruments and protocols
  - Mining glycoforms in massive proteomic datasets available for different biological samples (cell lines, tissues, bloods, animal models, etc)
  - *Accessible* by well established proteomics facilities for some applications (e.g., biomarker discovery)
- Expectation: instead of comprehensively characterizing all/most glycoforms, we target at the abundant glycopeptides (*major* glycoforms)

From proteomics to glycoproteomics

- High throughput data acquisition from complex samples (e.g., human blood samples) using conventional proteomics protocols



- Peptide **search engines** (>20): Assessment of peptide-spectrum matchings (PSMs)
  - Sequest, X!tandem – cross-correlation
  - Mascot, OMSSA – probabilistic scoring
  - InSpect – spectra alignment scoring
- Each experimental spectrum is compared against all putative peptides in a database with the matched precursor mass; only the 1st ranked PSM is considered to be correct.



Challenge: Assessment of PSMs

- To determine the correct PSMs among all PSMs
  - Each experimental spectrum is compared against all peptides in the database with the matched precursor mass; only the 1st ranked PSM is considered to be correct.
    - there is ≤1 correct PSM for each spectrum.

Industrial standard: to report identified peptides by controlled false discovery rate (FDR) – the target-decoy search strategy

•Search both the target and a decoy database (e.g. the reverse protein database)

•Use the peptide-spectrum matches (PSMs) in decoy database to estimate the false PSMs in the target database

- FDR = (# decoy PSMs) / (# target PSMs)

•Can be used for any search engine or scoring model

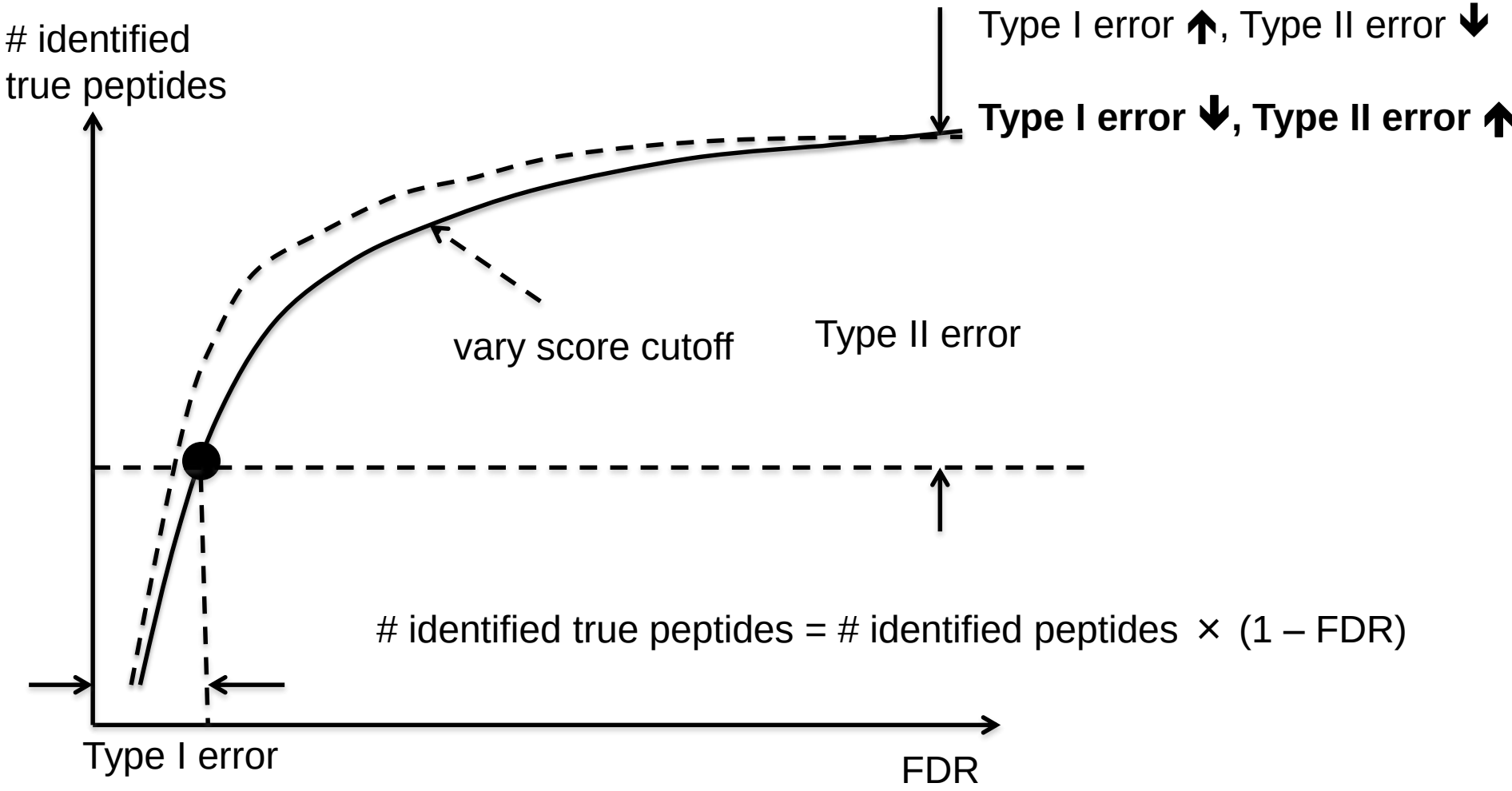
Toward the glycoproteomics for the identification of intact glycopeptides

- Main challenges
  - Data acquired from complex samples using routine proteomics protocols
    - A majority number of un-modified peptide ions and a considerable number of glycopeptide ions
    - Glycopeptide ions often show lower abundances than ions of non-modified peptide from the same protein due to microheterogenenity
  - target database is big
    - Various glycans may attach to the same glycosylation site: microheterogeneity
    - # putative glycopeptides = # peptides × # glycans

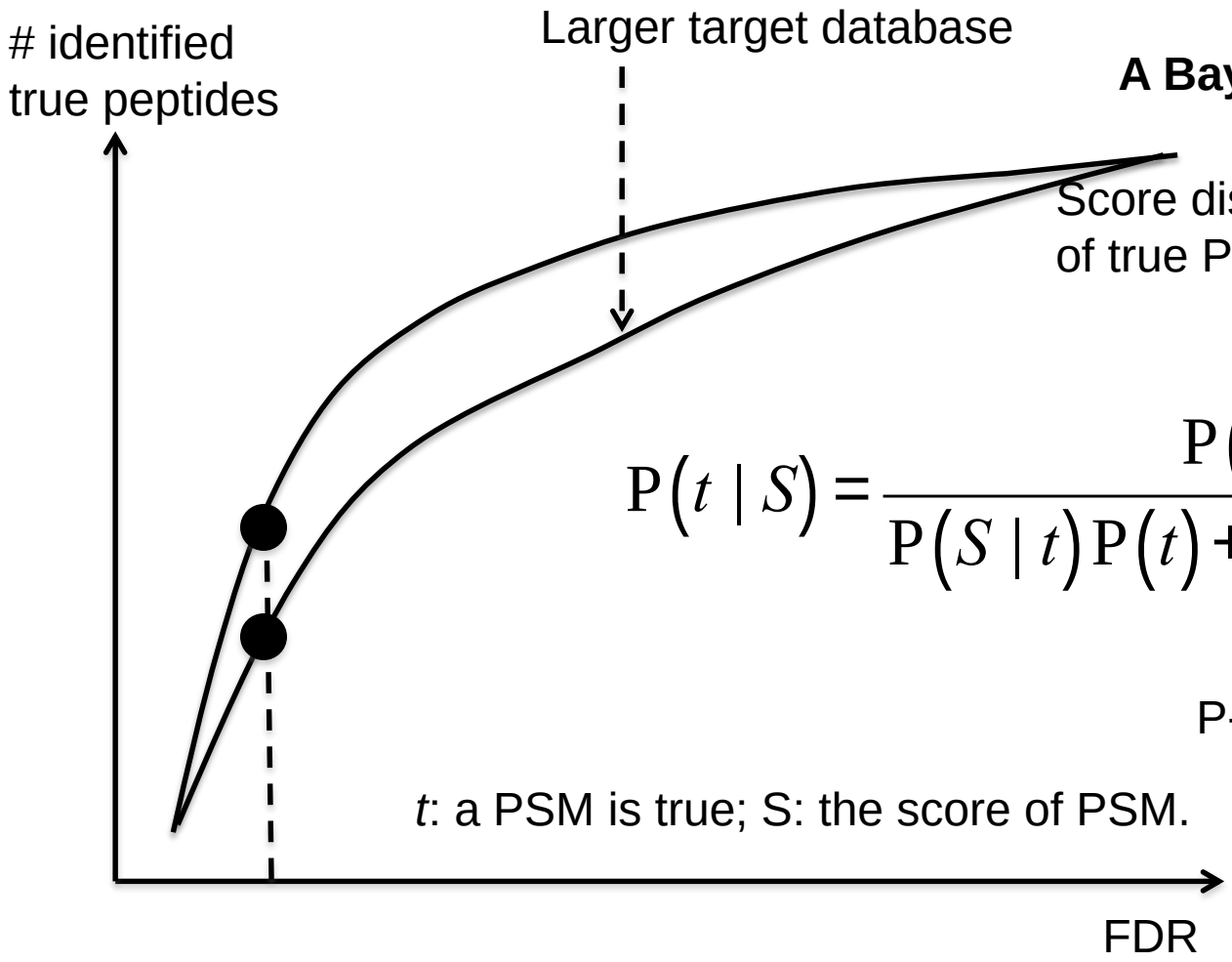
N-glycans: >150

  - More search time & false negatives (missing identifications)!

Higher false discovery rate with larger target database



Higher false discovery rate with larger target database



A Bayesian interpretation

Score distribution of true PSMs

Prior prob. for any PSM to be true

$$P(t \mid S) = \frac{P(S \mid t)P(t)}{P(S \mid t)P(t) + P(S \mid \text{not } t)[1 - P(t)]}$$

P-value (a null model)

t: a PSM is true; S: the score of PSM.

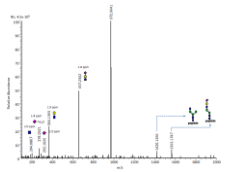
Higher false discovery rate with larger target database

$$\mathrm{P}(t \mid S) = \frac{\mathrm{P}(S \mid t)\mathrm{P}(t)}{\mathrm{P}(S \mid t)\mathrm{P}(t) + \mathrm{P}(S \mid \text{not } t)[1 - \mathrm{P}(t)]}$$

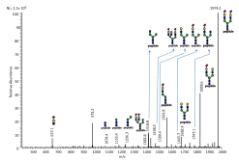
P(t) ~ # true PSMs / # potential PSMs
= (# peptides in the sample × # spectra per peptide) /
(# peptides in the database × # experimental spectra)

Therefore, the larger the database, and the more spectra, the higher FDR. → Larger effective search space (i.e., larger target database and more MS/MS spectra, e.g. in complex samples) introduces higher FDR.

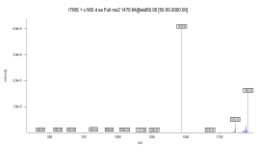
Goal: to reduce effective search space (fewer PSMs)!



HCD



CID

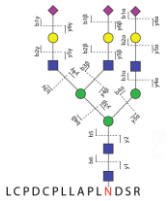


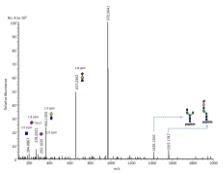
ETD

Glycoprotein database

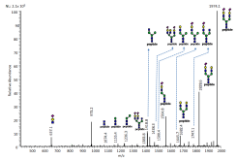
A

Putative glycan mass

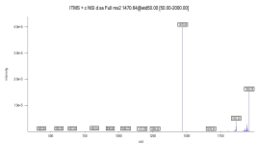




HCD



CID

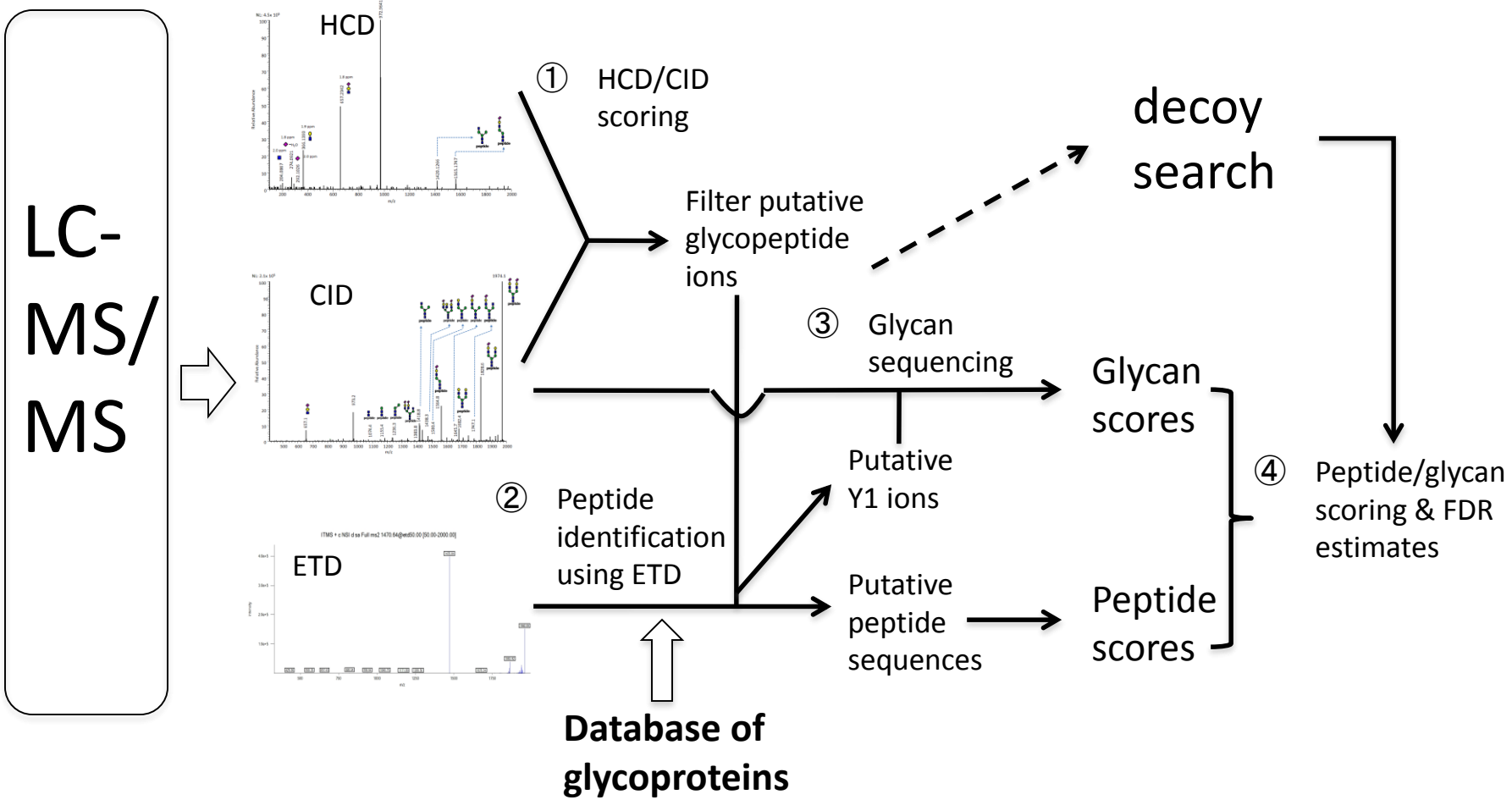


ETD

glycopeptide-spectrum matching (GPMs)

Strategies: 1) focus on only glycopeptides that are expected to be observed from the samples, e.g., from glycoproteins identified using un-modified peptide ions; using motifs of glycosylation sites;
2) Filter Ions that are likely derived from glycopeptides;
3) Orthogonal scoring of different fragmentation spectra, e.g. ETD for peptide fragmentation and CID for glycan fragmentation.

Overview of the computational framework

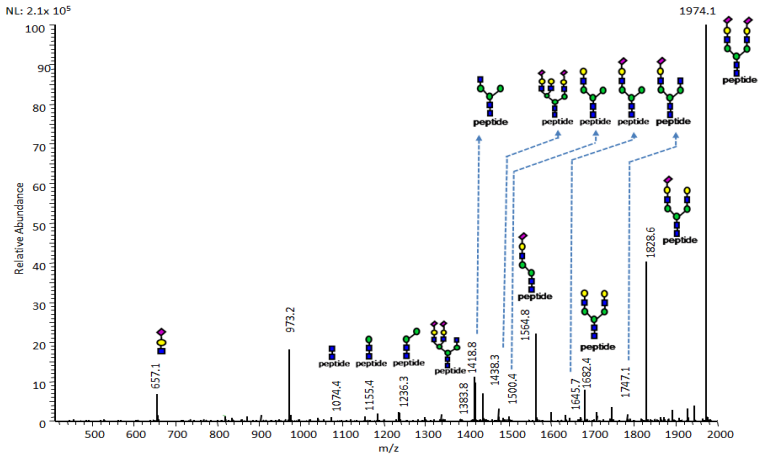


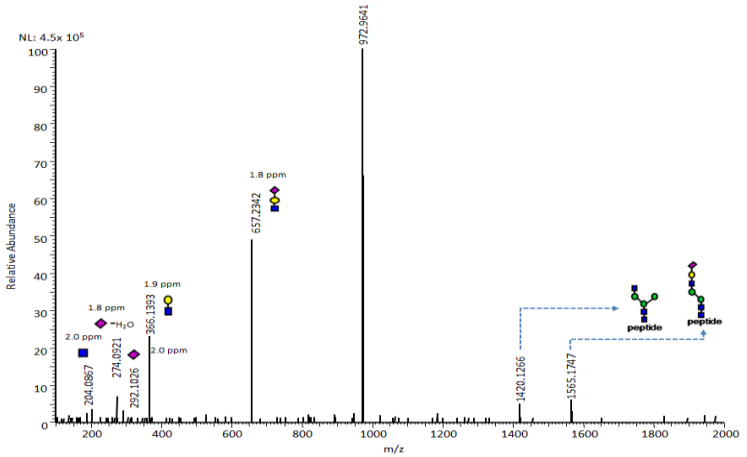
Testing datasets & target glycoprotein databases

Sample	Description	# datasets	# glycoproteins in database
Fetuin	Bovine Fetuin; 1 hr LC-MS/MS analysis	3	2
5 proteins mixture	Mixture of 5 model glycoproteins (bovine fetuin, human AGP, bovine pancreatic RNase B, porcine thyroglobulin (PTG), and human fibronectin ; 1 hr LC-MS/MS analysis	4	5
Serum from cancer patients	5 hrs LC-MS/MS analysis	6	105
Serum from control individuals	5 hrs LC-MS/MS analysis	6	105

For serum samples, a total (union) of 105 unique glycoproteins can be identified using Mascot against human IPI protein database in 12 cancer and control samples.

① HCD/CID scoring for filtering putative glycopeptide spectra





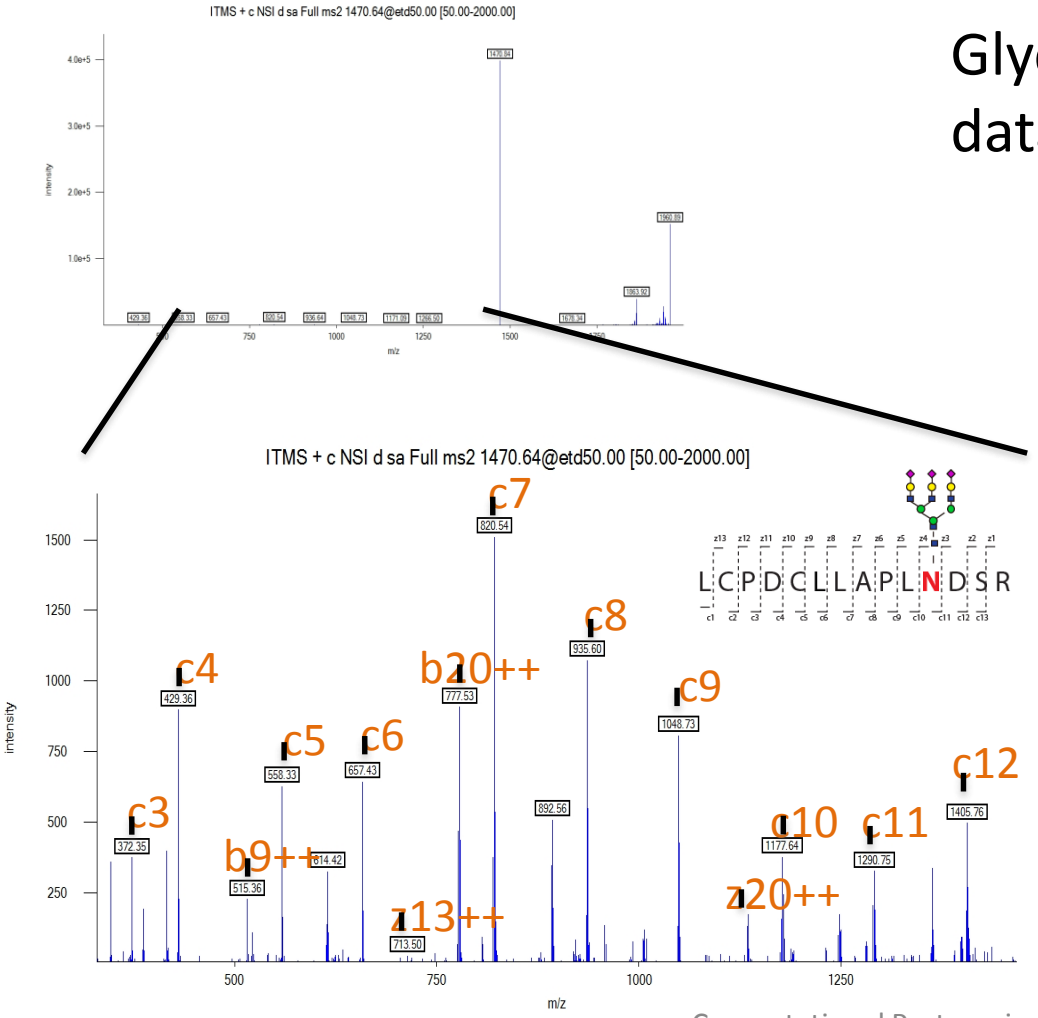
- count longest path from one peak to next whose inter-peak spacing corresponds to mono- or di-saccharide loss
- scores given by the length of longest pat

Ions	m/z	Description
1	138.05	Oxonium Ion (HexNAc2H_2O)
2	163.06	Hex
3	204.09	HexNAc
4	274.09	NeuAc-H_2O
5	292.08	NeuAc
6	366.14	Hex+HexNAc
7	657.23	NeuAc+Hex+HexNAc

- Presence/absence of 7 characteristic peaks
- P-value computed from binomial distribution

Mayampurath, et. al., 2011, RCM, 25 (14), 2007-2019

② Peptide identification by matching ETD spectra with peptides in database



Glycoprotein database

A

Putative glycan mass

Matched precursor mass

Theoretical fragments: c/z, b/y, neutral losses

Scoring: # matched peaks

This has been commonly used in peptide search engines.

How accurate is the ETD scoring? A target-decoy search for estimating FDR

Target search

Glycoprotein database **A** Putative glycan mass

Glycan decoy

Glycoprotein database **A** False glycan mass (e.g mass-40)

Peptide decoy

Reverse glycoprotein database (preserving the motifs) **A** Putative glycan mass

Combined target-decoy search, FDR=(# top-decoy-hits) / (# top-target-hits);

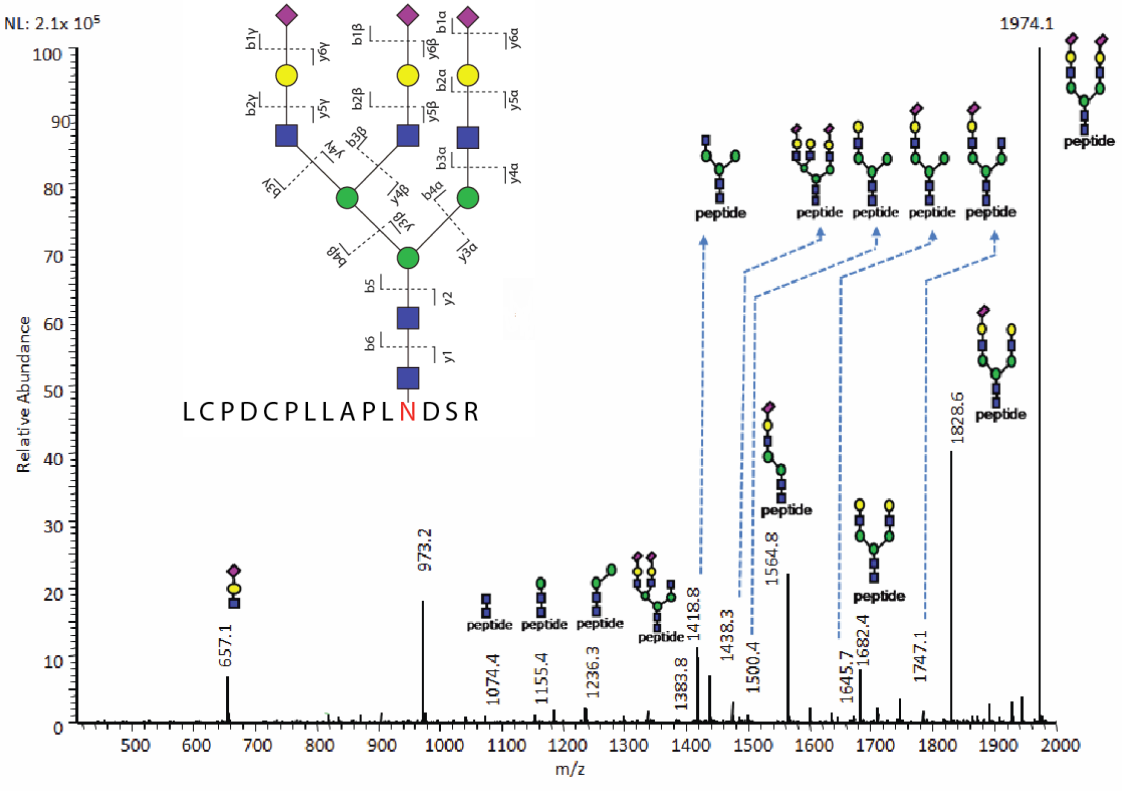
Note: in both cases, the number of glycopeptides in the decoy database is equal to the number of glycopeptides in the target database.

ETD scoring is not sufficient to identifying intact N-linked glycopeptides in complex samples

	Glycan decoy search (FDR<0.05)				Peptide decoy search (FDR<0.05)			
	Spectra	Glyco-peptides	Sites	Proteins	Spectra	Glyco-peptides	Sites	Proteins
Fetuin	25	15	5	2	16	7	3	2
Cancer serum	28	20	13	12	23	16	10	9
Healthy serum	25	16	10	9	9	4	2	2

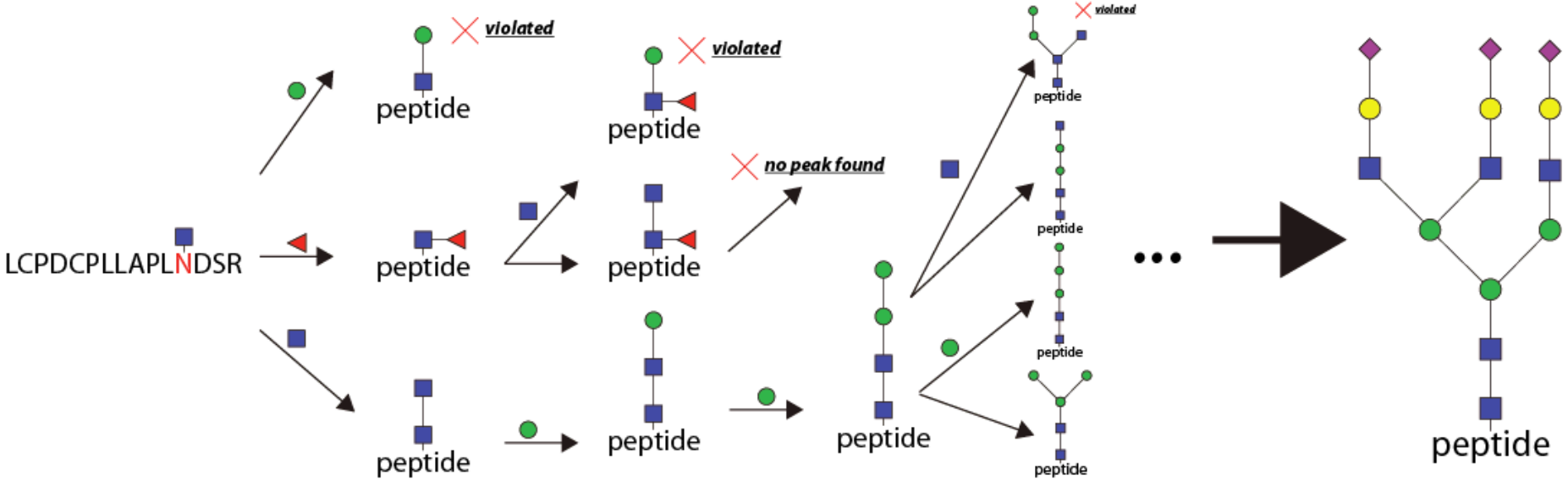
Note: the quality of ETD spectra is not sufficient to distinguish the true from false glycopeptide-spectrum matching (GPMs); therefore, we consider a unified scoring for both peptide identification (using ETD) and glycan sequencing (using CID) for the characterization of glycopeptides.

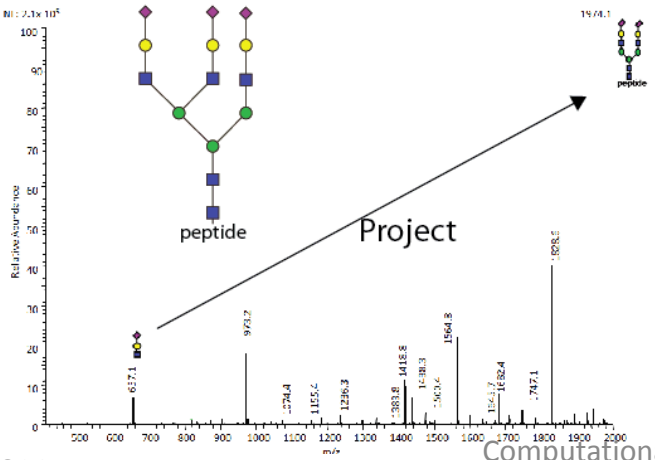
③ Glycan sequencing by using CID spectra of glycopeptides



- CID spectra of glycopeptide contain intensive peaks resulted from glycan fragmentation;
- Glycan sequencing from a CID spectrum of a glycopeptide is equivalent to that from a glycan spectrum, **if Y1 ion in the glycopeptide spectrum is given**;
- Y1 ion can be predicted from the peptide identified by using ETD scoring .

③ Glycan sequencing by using CID spectra of glycopeptides





Add pseudo Y-ions by subtracting corresponding b-ion mass from precursor mass.

Score: # matched peaks in the CID spectrum; **additive** to ETD score of the same ion

Purpose: 1) to assess how likely the predicted Y1 ion derive a N-glycan (and thus correct); 2) derive the most likely N-glycan structure.

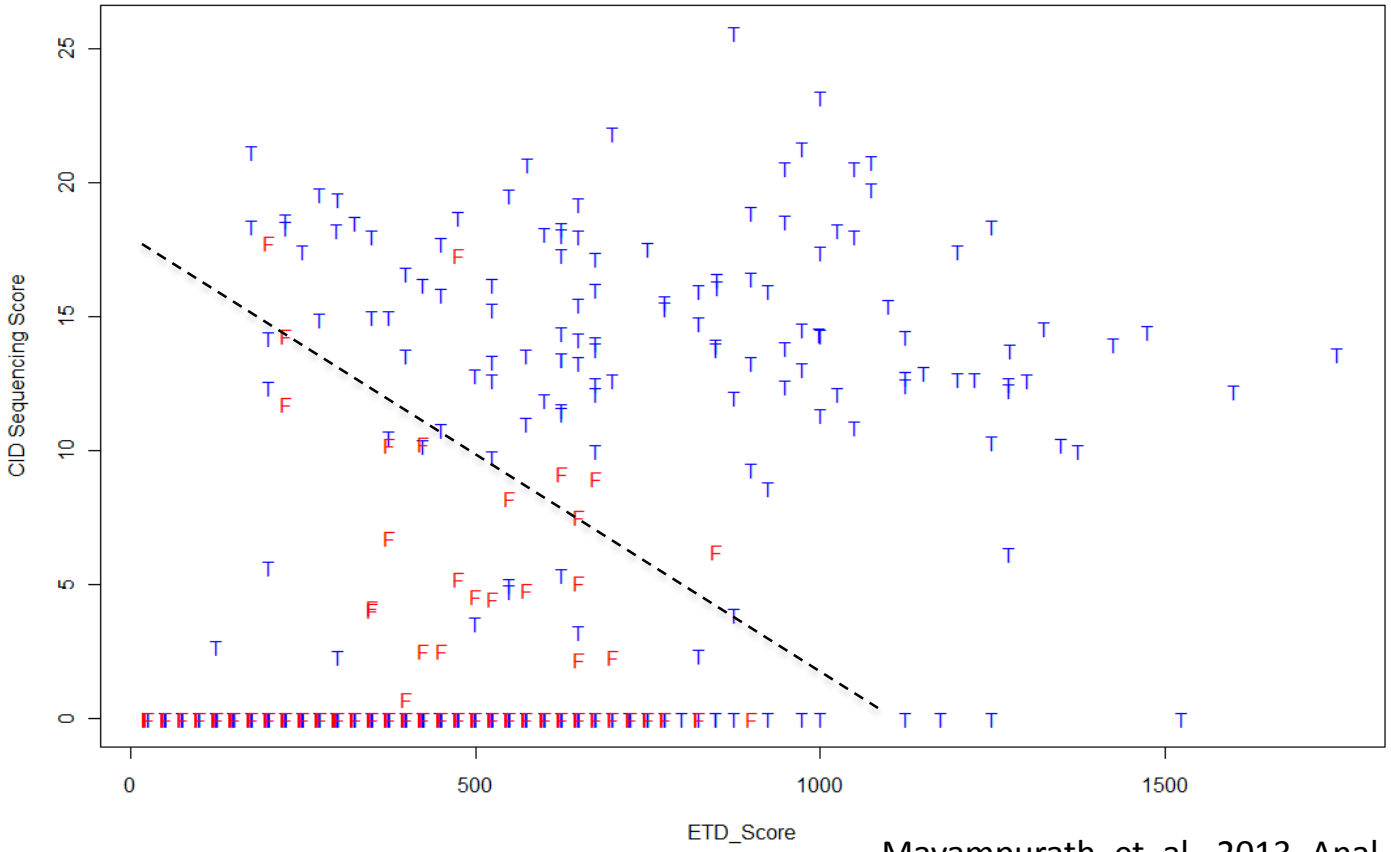
Tang, et. al., 2005, Yu, et. al., 2014.

④ FDR estimation based on a unified scoring in glycan/peptide sequencing

Target search			Peptide decoy		
Glycoprotein database	A	Putative glycan mass	Reverse glycoprotein database	A	Putative glycan mass

A *glycopeptide-spectrum matching* (GPMs) is defined as the triplet (E, C, P), where E and C represent the ETD and CID spectrum of a precursor ion, respectively, and P is a peptide. A GPM is scored as the total score: S(E,C,P)=S(E,P)+S(C,P). Because here we used the peak counts in both ETD/CID scoring, they are directly additive. In general, one can train a linear or non-linear function of these two scores. In a combined target-decoy search, the GPM score can be computed for sorting each target or decoy glycopeptide, and FDR can be estimated by FDR=(# top-decoy-hits) / (# top-target-hits). Note: in most cases, the top-hit of an ETD spectrum to decoy peptide will NOT be the reverse peptide of the true glycopeptide, indicating a false Y1-ion will be assigned in this case, which will often lead to a low S(C,P) in glycan sequencing. Therefore, the total score S(E,C,P) is lower, providing higher distinguishing power between true and false GPMs.

Re-assessing GSMs in LC-MS/MS data from human serum samples



T: GSMs from target database;
F: GSMs from decoy database;

Mayampurath, et. al., 2013,,Anal. Chem., 25 (14), 2007-2019

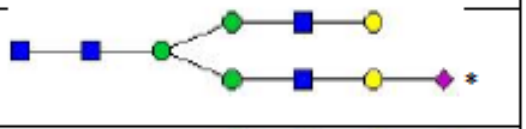
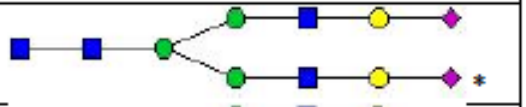
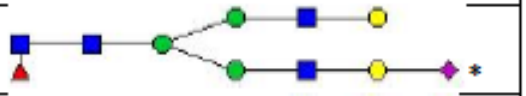
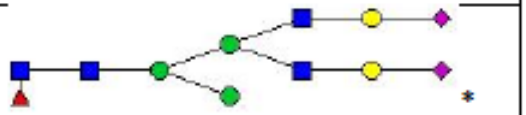
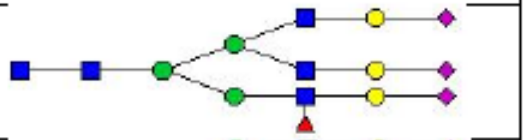
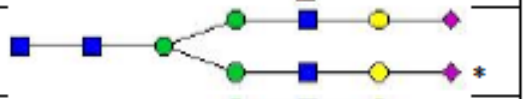
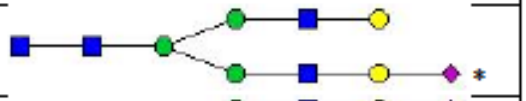
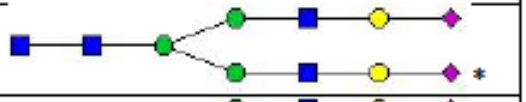
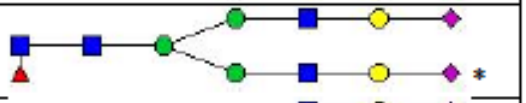
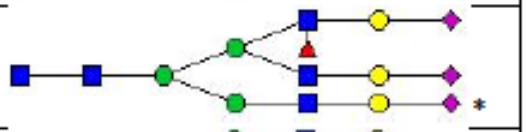
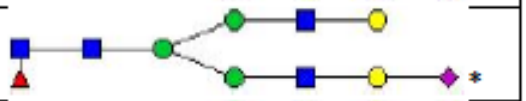
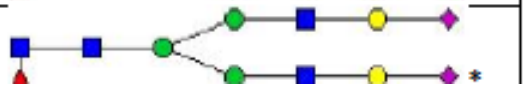
Identification of glycopeptides using CID/ETD combined scoring

GlycoMap Analysis	# protein IDs	# sites detected	# intact glycopeptides	# glycopeptides with glycan sequences completely sequenced	Glycan class distribution for completely sequenced glycans		
					# complex	# high mannose	# hybrid
Fetuin	2	5	22	8	8	0	0
5 protein mixture	4	5	11	6	6	0	0
Cancer serum	32	50	101	93	83	4	6
Control serum	29	44	92	89	82	1	5
Serum (total)	33	53	103	94	84	4	6

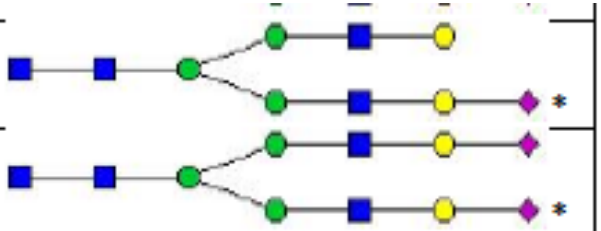
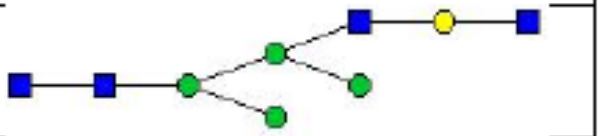
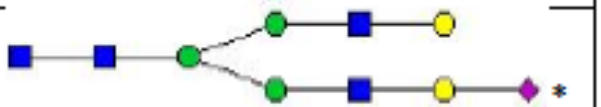
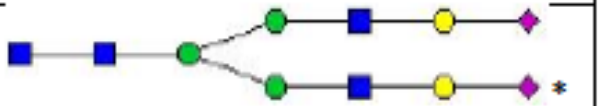
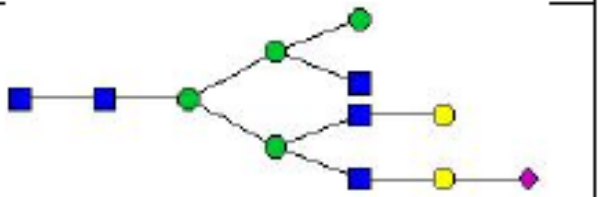
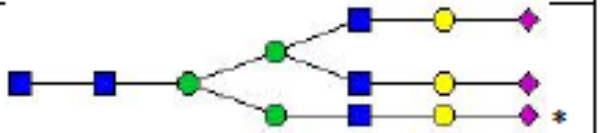
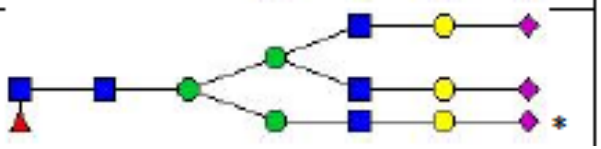
*FDR < 0.05

Mayampurath, et. al., 2013,,Anal. Chem., 25 (14), 2007-2019

ceruloplasmin precursor

Protein	Site	Glycan composition	Top ranking glycan sequence derived
sp|P00450|CERU_HUMAN	N-138	$HexNAc_4Hex_5NeuAc_1$	
		$HexNAc_4Hex_5NeuAc_2$	
		$HexNAc_4Hex_5DeHex_1NeuAc_1$	
		$HexNAc_4Hex_5DeHex_1NeuAc_2$	
		$HexNAc_5Hex_6DeHex_1NeuAc_3$	
	N-358	$HexNAc_4Hex_5NeuAc_2$	
	N-397	$HexNAc_4Hex_5NeuAc_1$	
		$HexNAc_4Hex_5NeuAc_2$	
		$HexNAc_4Hex_5DeHex_1NeuAc_2$	
		$HexNAc_5Hex_6DeHex_1NeuAc_3$	
	N-762	$HexNAc_4Hex_5DeHex_1NeuAc_1$	
		$HexNAc_4Hex_5DeHex_1NeuAc_2$	

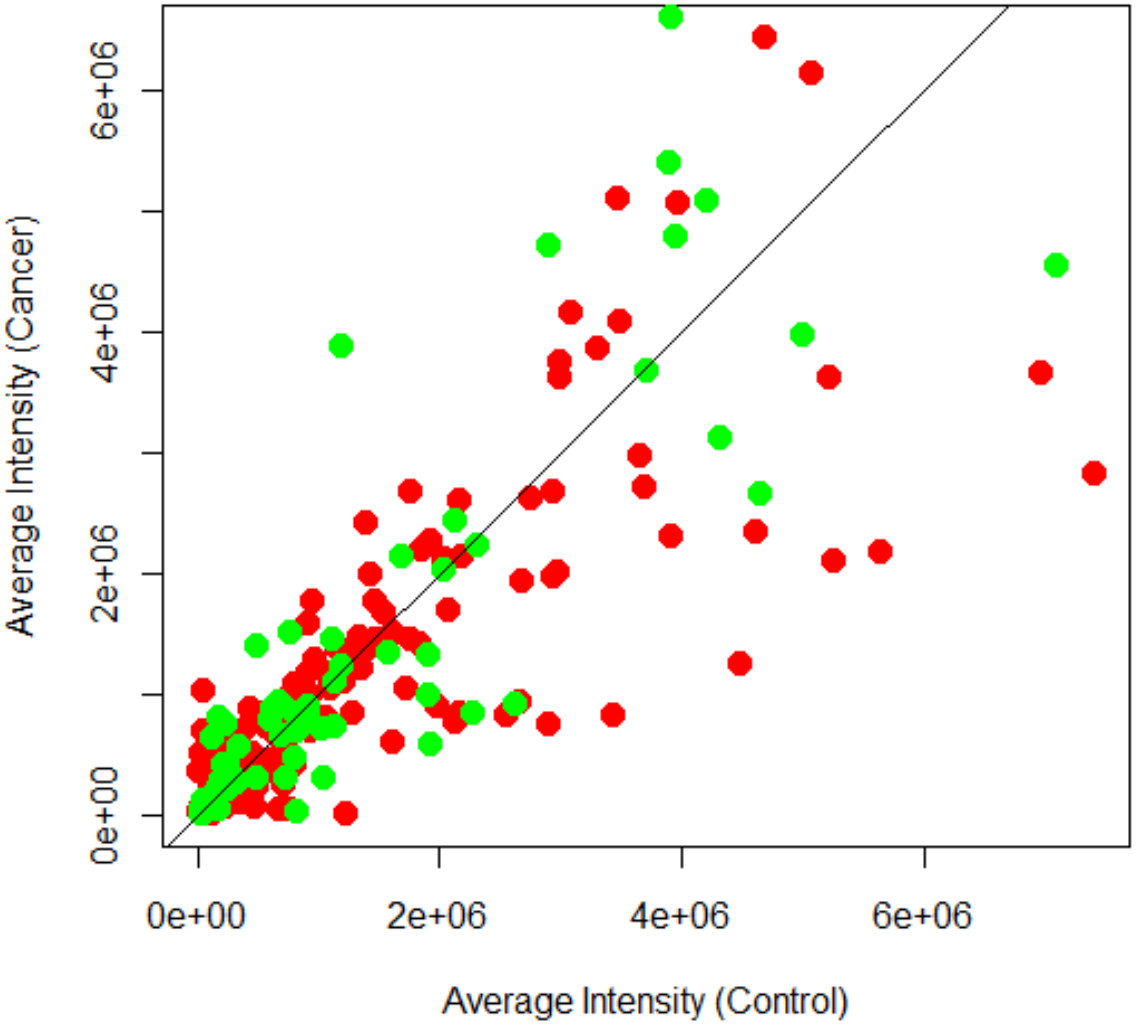
Haptoglobin

sp|P00738|HPT_HUMAN	N-184	$HexNAc_4Hex_5NeuAc_1$	
		$HexNAc_4Hex_5NeuAc_2$	
	N-241	$HexNAc_4Hex_5$	
		$HexNAc_4Hex_5NeuAc_1$	
		$HexNAc_4Hex_5NeuAc_2$	
		$HexNAc_5Hex_6NeuAc_1$	
		$HexNAc_5Hex_6NeuAc_3$	
		$HexNAc_5Hex_6DeHex_1NeuAc_3$	

Implementation

- GlycoFragwork implemented in C#
  - http://darwin.informatics.indiana.edu/col/GlycoFragwork/
  - Source code: http://sourceforge.net/projects/glycofragwork/
  - Input: mzXML or .raw file;
  - output: identified glycopeptides, cartoon of glycans, CID & ETD scores, FDR
  - Glycan sequencing algorithm is implemented as an independent .dll

Glycopeptide quantification for biomarker discovery

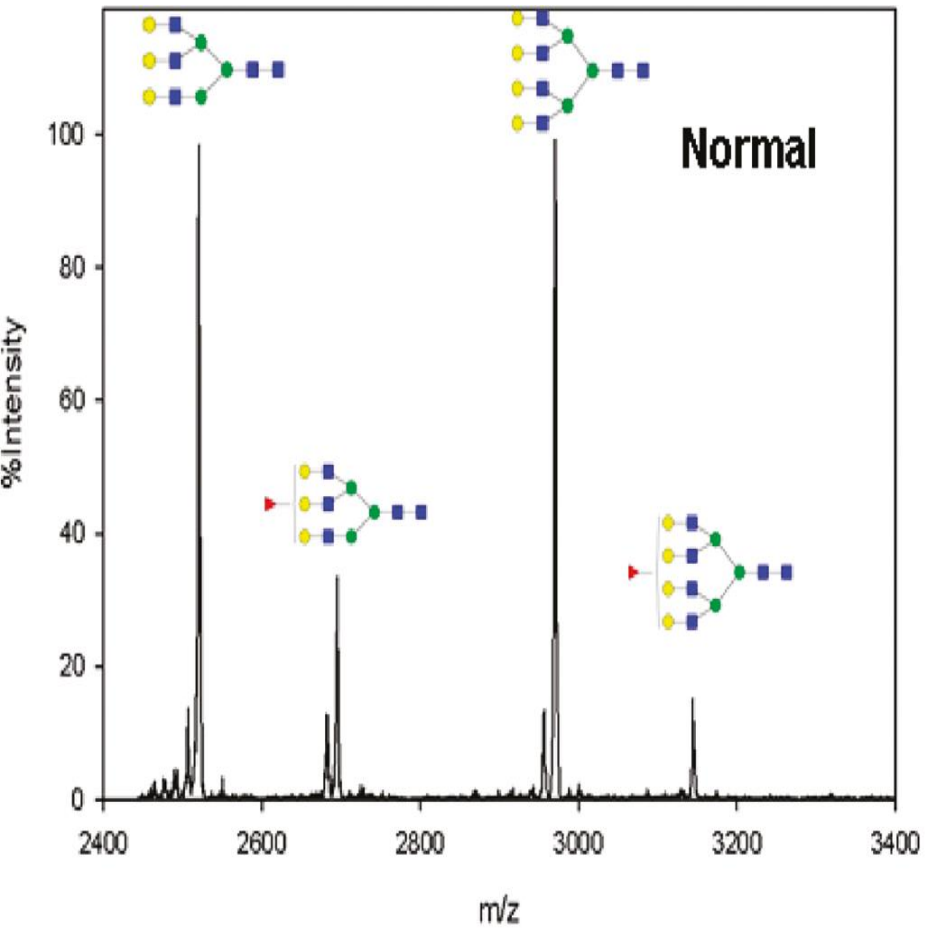


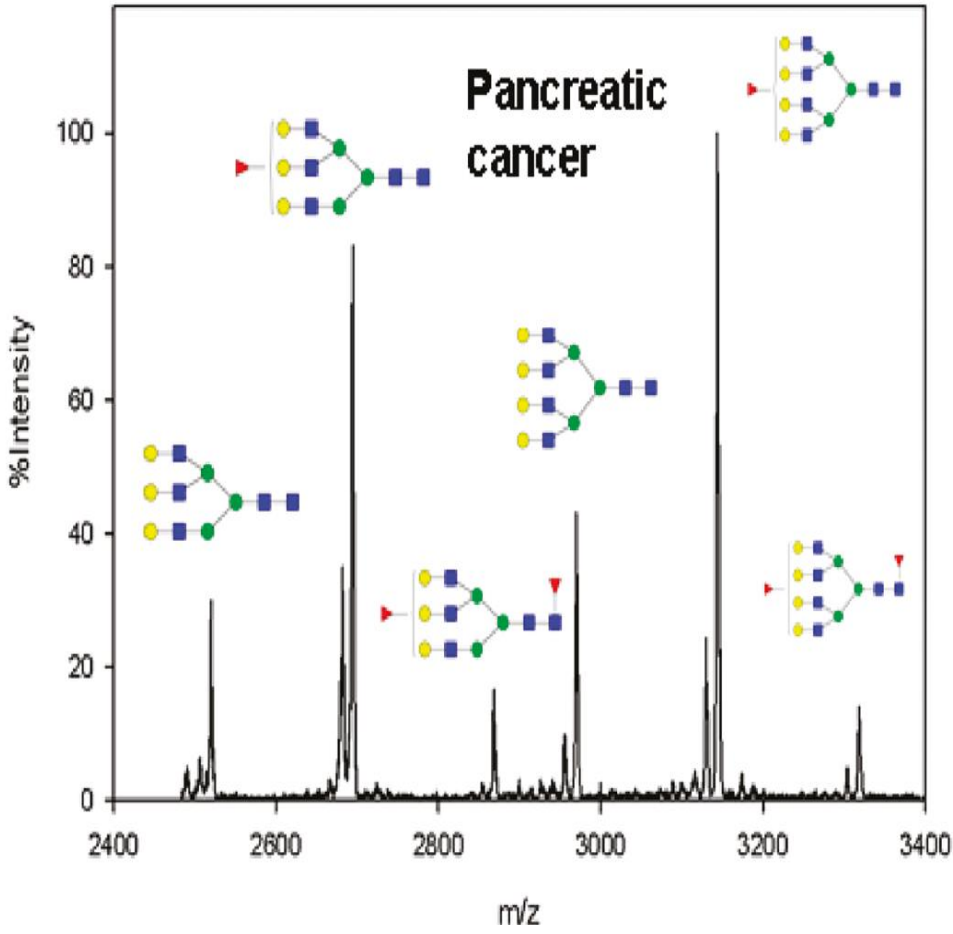
- Asialylated glycopeptides
- Sialylated glycopeptides

Integrated peak area of identified intact glycopeptides are indicative for their abundances

Site-specific protein glycosylations (in cancer)

Haptoglobin : N-184, N-207, N-211, N-241





Lin et al. J. Proteome Res. 2011 ; 10:2602-2611

A linear ANOVA model for discovery of disease associated site-specific glycosylations

$$y_{i,j(i),k(j(i)),c} = p_i + r_{i,c} + r_c + f_{j(i)} + g_{k(j(i))} + b_q + e_{i,j(i),k(j(i)),c}$$

p_i - Abundance of ith protein

$$g_{k(j(i))} = \frac{\text{abun. of k}^{th}\text{ glycan, c=1}}{\text{abun of j}^{th}\text{peptide, c=1}}$$

r_c - Class effect

$$f_{j(i)} = \frac{\text{abun. of j}^{th}\text{peptide, c=1}}{\text{abun of i}^{th}\text{protein, c=1}}$$

$$r_{i,c} = \frac{P_{i,c\in\{1,2\}}}{P_{i,c=1}}$$

b_q - Experimental effect

$e_{i,j(i),k(j(i)),c,q}$ - Error

$$\sum_c r_{i,c} = 0 \qquad \sum_{j(i)} f_{j(i)} = 0 \qquad \sum_{k(j(i))} g_{k(j(i))} = 0$$

Mayampurath, et. al., J. Proteome Res, 2014.

Log-Likelihood ratio test

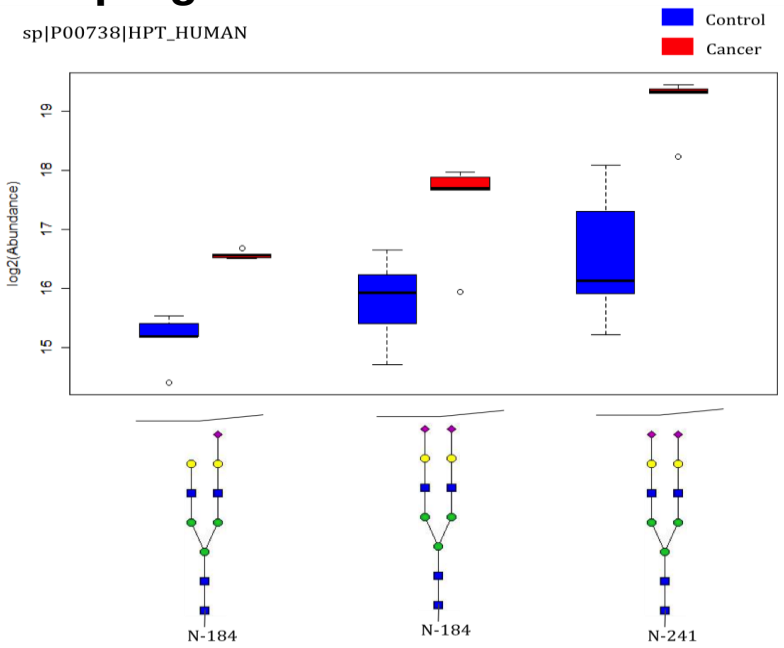
$$H_0 : r_{i,c=1} = 0, H_a : r_{i,c=1} \neq 0$$

protein	p values
sp|P04004|VTNC_HUMAN	8.80×10^{-17}
sp|P02790|HEMO_HUMAN	1.29×10^{-11}
sp|P01024|CO3_HUMAN	1.21×10^{-6}
sp|P00738|HPT_HUMAN	0.000372122
sp|P00450|CERU_HUMAN	0.606276481
sp|P02749|APOH_HUMAN	0.999999995

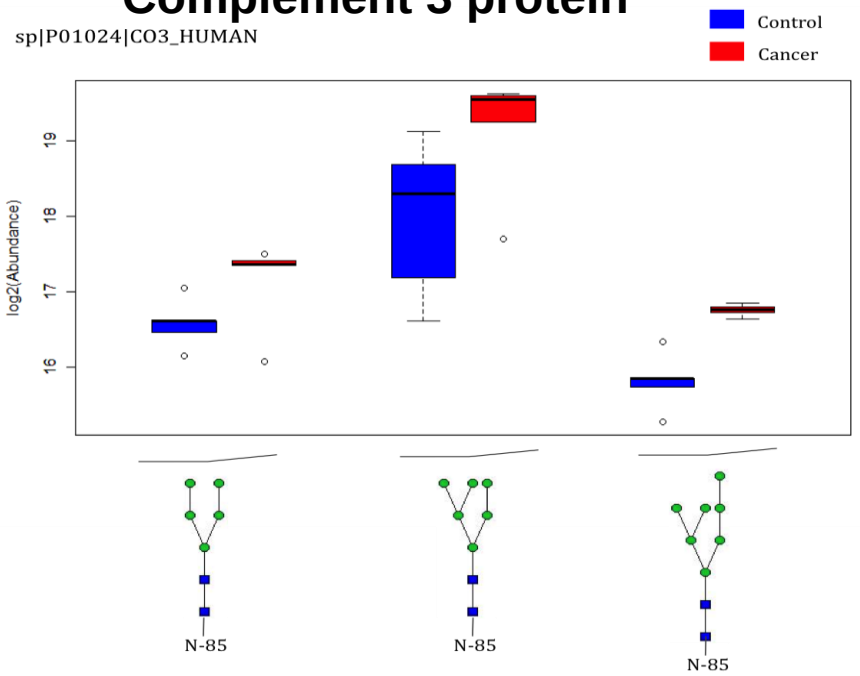
Note: None of these four glycoproteins shown as significant (p-value<0.01) when t-test was applied to the quantities of individual glycopeptides from these proteins.

Mayampurath, et. al., J. Proteome Res, 2014.

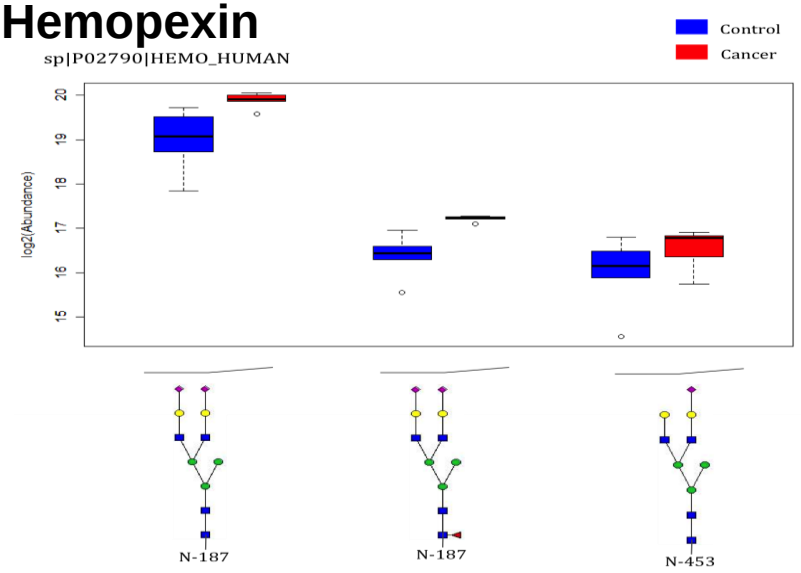
Haptoglobin



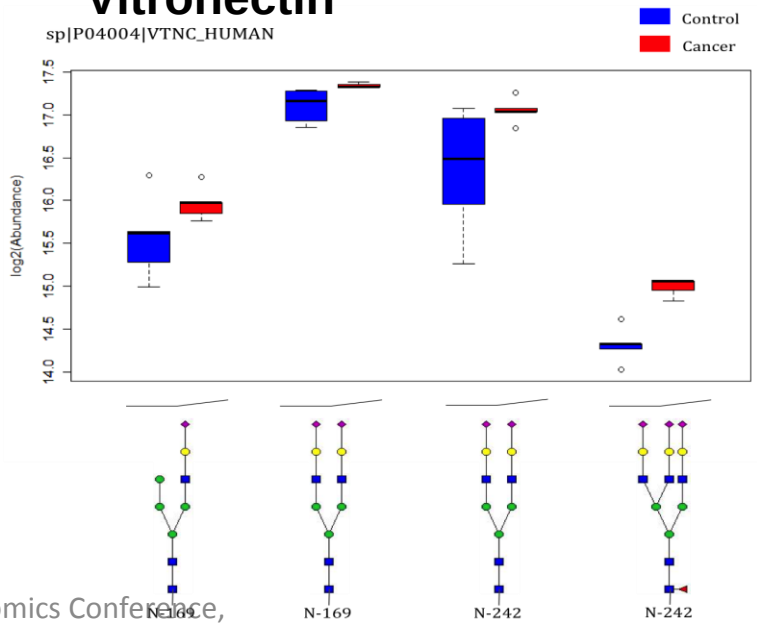
Complement 3 protein



Hemopexin



Vitronectin



Conclusions

- We developed a computational framework for discovery of glycoproteomic biomarkers using LC-MS/MS data
- The framework can be applied to clinical proteomics without specific sample preparation and analytical protocols
  - Routine proteomic approaches can be used for data collection
- We expect glycopeptides can be identified and quantified from existing proteomic data by using the framework

Acknowledgements

IU

- **Anoop Mayampurath**
- Chuan-Yih Yu
- Abhinav Mathur
- Jagadeshwar Balan
- Yin Wu

Texas Tech

- **Prof. Yehia Mechref**
- Ehwang Song
- Yunli Hu

PNNL

- Brian LaMarche

Funding: NSF: DBI-0642897

Persistent Fellowship (AM)