Published in final edited form as:
Conf Proc IEEE Eng Med Biol Soc. 2008 ; 2008: 3791–3794. doi:10.1109/IEMBS.2008.4650034.

Integrated Peptide and Glycan Biomarker Discovery Using MALDI-TOF Mass Spectrometry

Rency S. Varghese,
Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC.

Lenka Goldman,
Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC.

Yanming An,
Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC.

Christopher A. Loffredo,
Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC.

Mohamed Abdel-Hamid,
Viral Hepatitis Research Laboratory, NHTMRI, Cairo, Egypt.

Zuzana Kyselova,
National Center for Glycomics & Glycoproteomics, Dept. of Chemistry, Bloomington, IN.

Yehia Mechref,
National Center for Glycomics & Glycoproteomics, Dept. of Chemistry, Bloomington, IN.

Milos Novotny,
National Center for Glycomics & Glycoproteomics, Dept. of Chemistry, Bloomington, IN.

Steve K. Drake,
Clinical Chemistry Service, Department of Laboratory Medicine, NIH, Bethesda, MD.

Radoslav Goldman, and
Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC.

Habtom W. Ressom[*]
Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC.

Abstract

Quantitative comparison of peptides and glycans in serum is conducted using matrix-assisted laser desorption/ionization–time of flight mass spectrometry (MALDI-TOF MS) to identify biomarkers. A peak selection algorithm is developed to identify a panel of integrated peptide and glycan peaks to distinguish hepatocellular carcinoma (HCC) cases from high-risk population of patients with chronic liver disease (CLD). Candidate peptide and glycan markers selected frequently in multiple runs of the algorithm are presented. The performance of these markers is evaluated in terms of their ability to distinguish HCC cases from patients with CLD in a blinded validation set.

I. INTRODUCTION

MASS spectrometry (MS) provides a rich source of information for molecular characterization of the disease process. The method has a great potential to identify a panel of biomarkers

[*] Corresponding author: hwr@georgetown.edu.

relevant for early diagnosis of complex diseases such as cancer. Several laboratories have demonstrated the feasibility of selecting peptide biomarkers in MALDI-TOF spectra for disease classification [1–3]. The characterization of glycans in serum of patients with liver disease is also a promising strategy for biomarker discovery [4]. An alternative strategy is to integrate both peptides and glycans for improved diagnostic capability.

Figure 1 illustrates our methodology for integrated peptide and glycan biomarker identification. The methodology is applied to discover a panel of candidate peptide and glycan biomarkers for the detection of HCC in a high-risk Egyptian population with chronic liver disease (CLD), consisting of fibrosis and cirrhosis patients [5,6]. Low molecular weight (LMW) enriched peptides and permethylated glycans are analyzed using a matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry (MALDI-TOF MS). Following spectral preprocessing, peak detection, and peak screening, the peptide and glycan peaks are combined. A peak selection algorithm called ACO-SVM is then applied to search for a panel of peaks that distinguishes HCC from CLD. ACO-SVM combines ant colony optimization (ACO) and support vector machine (SVM) methods to systematically search for the most useful panel of peaks from a large number of candidate peaks without performing an exhaustive search [3].

II. METHODS

A. Serum Samples

The participants in this study included 73 HCC cases, 78 healthy controls, and 52 CLD controls recruited in Cairo, Egypt [6]. Seventeen HCC cases were classified as early (Stage I and II) and 33 HCC cases as advanced (Stage III and IV) according to the staging system described in ref. [7]. For the remaining 23 HCC cases the available information was not sufficient to assign the stage. The CLD group included biopsy-confirmed 21 fibrosis and 25 cirrhosis patients. Six individuals in the CLD group did not have sufficient clinical information.

B. Sample Preparation and Spectral Generation

We utilized an enrichment procedure to analyze native peptides in the LMW fraction of serum using MALDI-TOF MS as described previously [2,8]. We acquired 203 mass spectra using the Flex Control and Flex Analysis software (Bruker Daltonics, Billerica, MA). Each spectrum consisted of ~136,000 *m/z* values with the corresponding intensities in the mass range of 900 to 10,000 Da.

The sample preparation for quantification of glycans involved enzymatic release of N-glycans from glycoproteins, extraction of N-glycans, and solid-phase permethylation as described previously [9]. The resulting permethylated glycans were spotted on a MALDI plate with DHB-matrix, MALDI plate was dried under vacuum, and mass spectra were acquired using a 4800 MALDI TOF/TOF analyzer [10]. The spectra were recorded in the positive-ion mode, since permethylation eliminates the negative charge normally associated with sialylated glycans. 203 raw spectra were exported as text files. Each spectrum consisted of ~121,000 *m/z* values with intensities in the mass range of 1,500–5,500 Da.

C. Spectral Preprocessing

Spectral preprocessing began by splitting the peptide and glycan spectra into a labeled set (training and validation sets) and a blinded set. The training and validation sets consisted of a subset of HCC cases, a subset of CLD controls, and all population controls (people without manifest liver disease). The blinded set comprised of masked HCC cases and CLD controls; this set was used to evaluate the generalization capability of the peaks selected on the basis of the training set. The spectral processing begins by binning the raw spectra. Fig. 2 shows examples of binned peptide and glycan spectra. Baseline correction, outlier screening, normalization, peak detection, peak calibration, peak screening, and peak selection were performed on the binned training and validation sets by subjecting the entire process to cross-validation [3,8,11].

For peak screening, logistic regression models were used to examine association of the peaks to known covariates including age, gender, smoking status, residency, HCV and HBV viral infections. This analysis was performed on the samples from healthy controls to unambiguously isolate peaks associated to the covariates [3]. Peaks that exhibit statistically significant association to the covariates were removed from subsequent analysis.

D. Integrated Marker Selection

Following peak screening, the peptide and glycan spectra were combined. The rationale for combining the peptide and glycan spectra was to allow the peptides and glycans to compete against each other during peak selection, so that a panel of integrated peptide and glycan biomarkers can be identified to obtain improved diagnostic capability.

The ACO-SVM algorithm searched for the optimal peak set consisting of a pre-specified number of peaks. The peak set was selected on the basis of its ability to distinguish HCC cases from patients with CLD in the validation set. Note that the spectra in the validation set were not involved in the peak detection, peak calibration, and peak screening steps. They were screened for outliers, binned, baseline corrected, normalized, and scaled on the basis of the parameters used to preprocess the spectra in the training set. These parameters included outlier screening factors (i.e., median record count and TIC) and a scaling factor that standardizes the peaks in the training set to have a maximum of 100. The peaks in the validation set were quantified at the selected markers. The quantified peaks were then presented to the SVM classifier previously trained using the peaks from the training set. The performance of the SVM classifier in predicting the disease state of the subjects in the validation set was used by ACO-SVM to guide its search for the optimal peak set.

E. Marker Evaluation

The spectral preprocessing, peak detection, peak calibration, peak screening, and peak selection steps were repeated multiple times by randomly splitting the spectra into training and validation sets with resubstitution. Note that the number of peaks detected and the size of the *m/z* windows could vary due to the change in the population set at each iteration. Due to this, the same peak can be represented by *m/z* windows with different widths. These are considered as overlapping *m/z* windows as long as they represent the same peak. We therefore merged these overlapping *m/z* windows and represented them by a new *m/z* window whose boundaries were determined as the minimum and maximum *m/z* window edges of the overlapping windows. The new *m/z* windows were summarized to determine the most frequently selected markers.

The spectra in the blinded set were outlier screened, binned, baseline corrected, normalized, and scaled on the basis of parameters used to preprocess the spectra in the training set. We built an SVM using the selected markers quantified in the training set and evaluated the capability of the SVM classifier in terms of sensitivity and specificity in distinguishing HCC from CLD in the blinded set.

For further evaluation of the selected peaks and to examine the consistency of our methodology, we repeated the entire peak selection process by randomly splitting the spectra into training, validation and blinded sets with resubstitution as illustrated in Fig. 1. The frequency of occurrence of the resulting summarized peaks and the corresponding sensitivity and specificity in multiple runs were evaluated.

F. Peptide and Glycan Identification

To ensure accurate biological interpretation of candidate biomarkers and to enable independent laboratory validation, peptides and glycans represented by the selected peaks need to be identified. MALDI-TOF/TOF and LC-MS/MS instruments were utilized for peptide and glycan identification. Sequencing of the peptides is under way. Glycan structures for 50% of the peaks detected in this study are available.

III. RESULTS

The 203 serum samples that consisted of 73 HCC, 52 CLD, and 78 population control subjects were split into a training set (25 HCC, 25 CLD, and 78 population controls), a validation set (10 HCC, 10 CLD), and a blinded set (38 HCC and 17 CLD). Spectral preprocessing, peak detection, peak calibration and peak screening were performed as described in ref. [8]. The peptide spectra yielded ~240 *m/z* windows and the glycan spectra yielded ~90 *m/z* windows in mu006Ctiple runs with different training sets. These peptide and glycan *m/z* windows were then combined together before marker selection. The ACO-SVM was applied to search for a small peak set from a total of ~330 peaks (*m/z* windows). The capability of potential peak sets in predicting the spectra in the validation set was used by ACO-SVM as a criterion to search for the optimal peak set.

The above procedure was repeated 100 times by reshuffling the training and validation set and randomly selecting (with resubstitution) 25 HCC and 25 CLD spectra as a training set and the remaining 10 HCC and 10 CLD spectra as a validation set. The peaks selected in 100 runs were summarized by merging overlapping windows. Figure 3 shows the frequency plot of the summarized peptide and glycan peaks (represented by their *m/z* windows) over 100 runs. Only peaks that were selected in at least 10% of the runs are shown in the figure. The first seven peaks were selected in more than 45% of the runs. P1 and P2 are peptide peaks, while G1–G5 are glycan peaks. An SVM classifier built with these two peptides and five glycans yielded 87% sensitivity and 100% specificity in distinguishing HCC cases from patients with CLD in the blinded set.

To examine how the peak selection process is impacted by the subset of training and validation spectra used for peak selection, we repeated the entire peak selection process ten times by randomly splitting the spectra into training, validation and blinded sets with resubstitution. The average classification accuracy in distinguishing the samples in the blinded set over 10 runs was 98% with 2% standard deviation. Figure 4 presents a frequency plot of the peptides and glycans selected over the ten splits with each split run 100 times. The ten most frequent markers that consist of one peptide and nine glycans are shown in the figure. Fig. 5 shows the box plots for these markers after the HCC cases by stage (early stage HCC and late stage HCC) and the CLD controls by disease (fibrosis and cirrhosis). Sugar compositions for six of the nine glycans are presented in Fig. 1 along with mean glycan spectra for the HCC and CLD groups. Also, the mean peptide spectra are shown in the figure. The peptide sequence information for the selected peptide peak is unknown.

As a comparison, the peak selection process was performed by using the SVM-recursive feature elimination (SVM-RFE) method instead of ACO-SVM. The average accuracy of the peak sets obtained over 10 runs was 92% with 3% standard deviation, in distinguishing the cases from controls in the blinded set. Three glycan peaks overlap between the two peak selection methods (ACO-SVM and SVM-RFE) when the top 10 combined peaks are considered. We observe that the most frequently selected peak by ACO-SVM had a frequency of 95%, while the most frequently selected peak by SVM-RFE had only 65% frequency.

IV. CONCLUSION

This paper introduces an integrated approach for quantitative comparison of peptides and glycans in serum for biomarker discovery. Candidate peptide and glycan biomarkers of HCC were obtained by comparing MALDI-TOF spectra of LMW enriched peptides and permethylated glycans, respectively. The method has the potential to identify a panel of peptide and glycan biomarkers that may provide better diagnostic capability than those that consist of either peptide or glycan markers only.

The potential clinical utility of the selected candidate markers needs to be evaluated through independent laboratory experiments and samples from other populations. The computational methods presented in this paper provide researchers with a list of peaks - sorted by their capability to distinguish HCC patients from CLD cases - to allow a more targeted identification of peptide sequences and glycan structures as well as validation of the candidate biomarkers through independent laboratory experiments.

Acknowledgments

This work was supported in part by the National Cancer Institute (NCI) R21 CA130837 Grant, NCI R03 CA119313 Grant, NCI Early Detection Research Network Associate Membership Grant, and the Prevent Cancer Foundation Grant awarded to HWR; R01 CA115625 Grant awarded to RG.

REFERENCES

1. Schwegler EE, Cazares L, Steel LF, Adam BL, Johnson DA, Semmes OJ, Block TM, Marrero JA, Drake RR. SELDI-TOF MS profiling of serum for detection of the progression of chronic hepatitis C to hepatocellular carcinoma. Hepatology 2005;vol. 41:634–642. [PubMed: 15726646]

2. Orvisky E, Drake SK, Martin BM, Abdel-Hamid M, Ressom HW, Varghese RS, An Y, Saha D, Hortin GL, Loffredo CA, Goldman R. Enrichment of low molecular weight fraction of serum for MS analysis of peptides associated with hepatocellular carcinoma. Proteomics 2006;vol. 6:2895–2902. [PubMed: 16586431]

3. Ressom HW, Varghese RS, Drake SK, Hortin GL, Abdel-Hamid M, Loffredo CA, Goldman R. Peak selection from MALDI-TOF mass spectra using ant colony optimization. Bioinformatics 2007;vol. 23:619–626. [PubMed: 17237065]

4. Callewaert N, Van Vlierberghe H, Van Hecke A, Laroy W, Delanghe J, Contreras R. Noninvasive diagnosis of liver cirrhosis using DNA sequencer-based total serum protein glycomics. Nat Med 2004;vol. 10:429–434. [PubMed: 15152612]

5. Nada O, Abdel-Hamid M, Ismail A, El Shabrawy L, Sidhom KF, El Badawy NM, Ghazal FA, El Daly M, El Kafrawy S, Esmat G, Loffredo CA. The role of the tumor necrosis factor (TNF)--Fas L and HCV in the development of hepatocellular carcinoma. J Clin Virol 2005;vol. 34:140–146. [PubMed: 16157266]

6. Ezzat S, Abdel-Hamid M, Eissa SA, Mokhtar N, Labib NA, El-Ghorory L, Mikhail NN, Abdel-Hamid A, Hifnawy T, Strickland GT, Loffredo CA. Associations of pesticides, HCV, HBV, and hepatocellular carcinoma in Egypt. Int J Hyg Environ Health 2005;vol. 208:329–339. [PubMed: 16217918]

7. AJCC Cancer Staging Manual, 6th Edition. American College of Surgeons. Philadelphia: Lippincott-Raven; 2002.

8. Ressom HW, Varghese RS, Goldman L, An Y, Loffredo CA, Abdel-Hamid M, Kyselova Z, Mechref Y, Novotny M, Drake SK, Goldman R. Analysis of MALDI-TOF mass spectrometry data for discovery of peptide and glycan biomarkers of hepatocellular carcinoma. J Proteome Res 2008;vol. 7:603–610. [PubMed: 18189345]

9. Kang P, Mechref Y, Klouckova I, Novotny MV. Solid-phase permethylation of glycans for mass spectrometric analysis. Rapid Commun Mass Spectrom 2005;vol. 19:3421–3428. [PubMed: 16252310]

10. Kyselova Z, Mechref Y, Al Bataineh MM, Dobrolecki LE, Hickey RJ, Vinson J, Sweeney CJ, Novotny MV. Alterations in the serum glycome due to metastatic prostate cancer. J Proteome Res 2007;vol. 6:1822–1832. [PubMed: 17432893]

11. Ressom HW, Varghese RS, Goldman L, Loffredo CA, Abdel-Hamid M, Kyselova Z, Mechref Y, Novotny M, Goldman R. Analysis of MALDI-TOF mass spectrometry data for detection of glycan biomarkers. Pac Symp Biocomput 2008;vol. 13:216–227. [PubMed: 18229688]

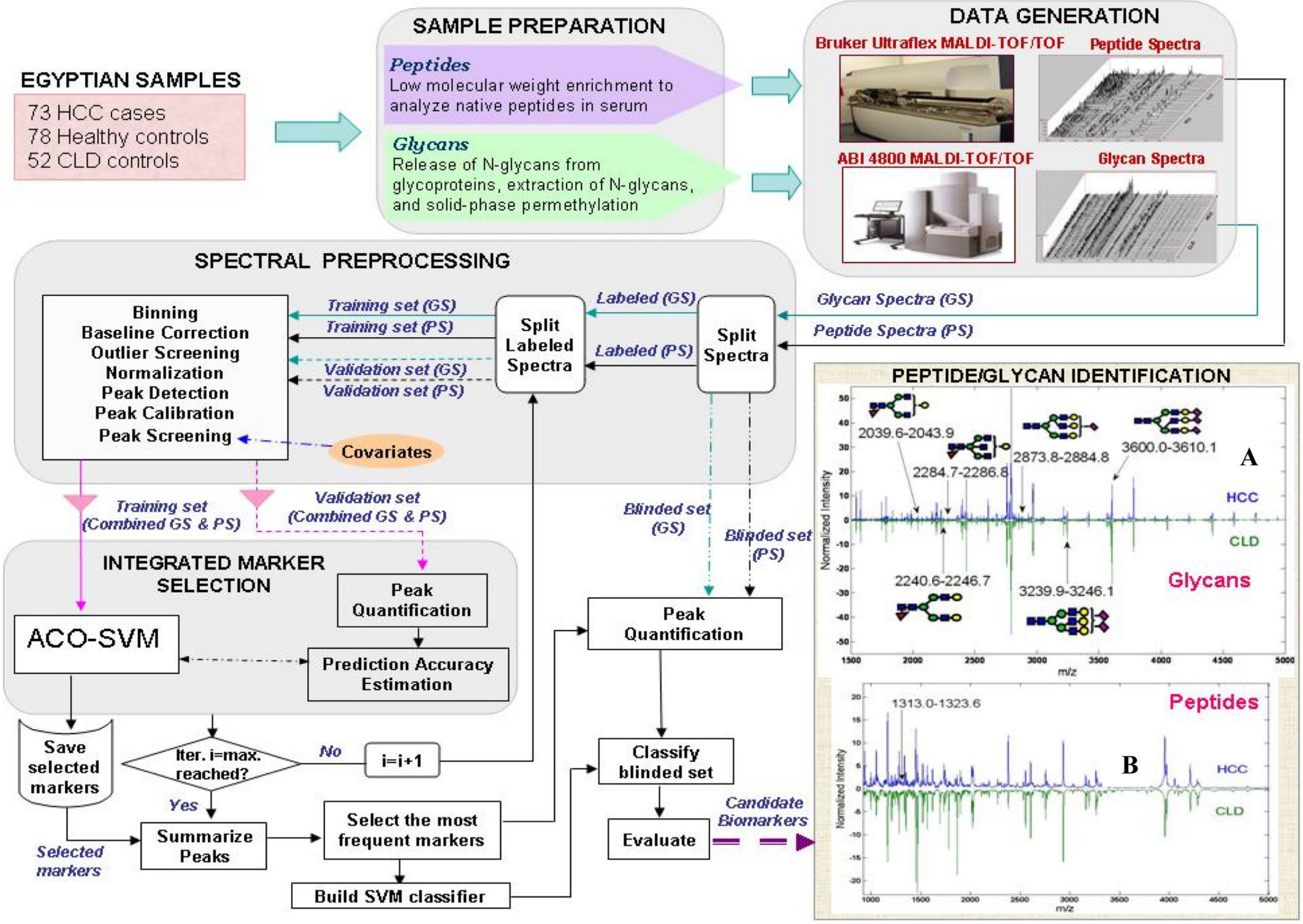

Fig. 1.

Methodology for identification of a panel of integrated peptide and glycan biomarkers. **A**: Glycan structures of six candidate biomarkers and average glycan spectra for HCC and CLD patients. **B**: Average peptide spectra and a candidate peptide biomarker (see Results section).

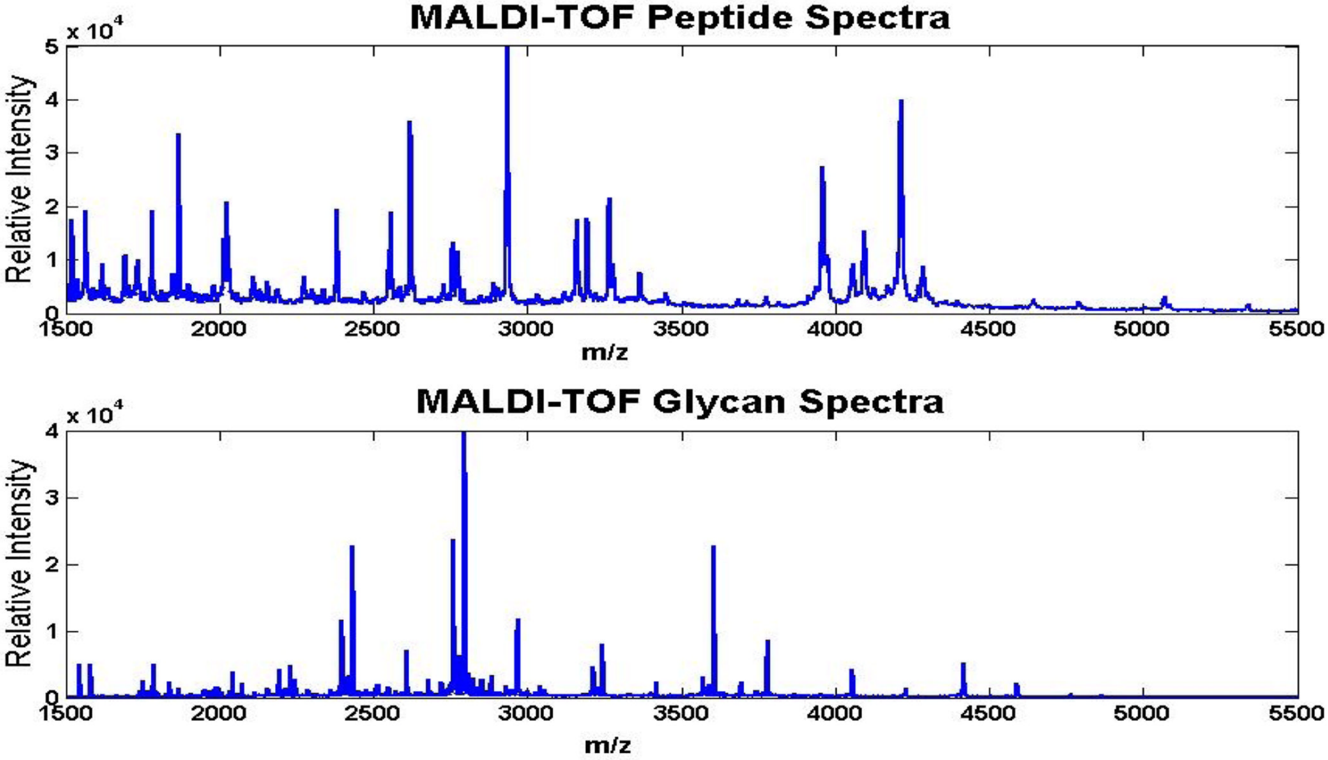

Fig. 2.

Binned MALDI-TOF peptide spectrum (top) and glycan spectrum (bottom) for the same sample

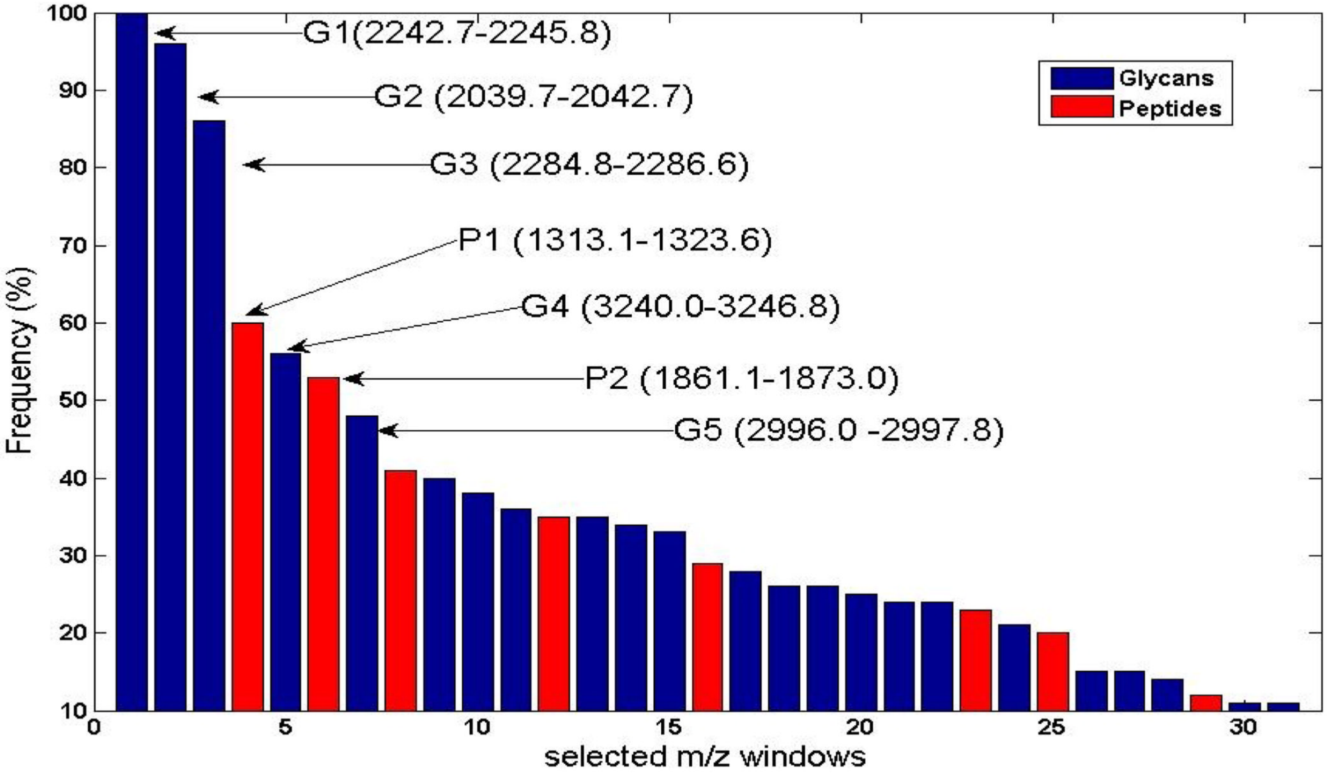

Fig. 3.

Frequency of occurrence of peptide and glycan peaks selected by ACO-SVM over 100 runs. The top seven *m/z* windows are shown

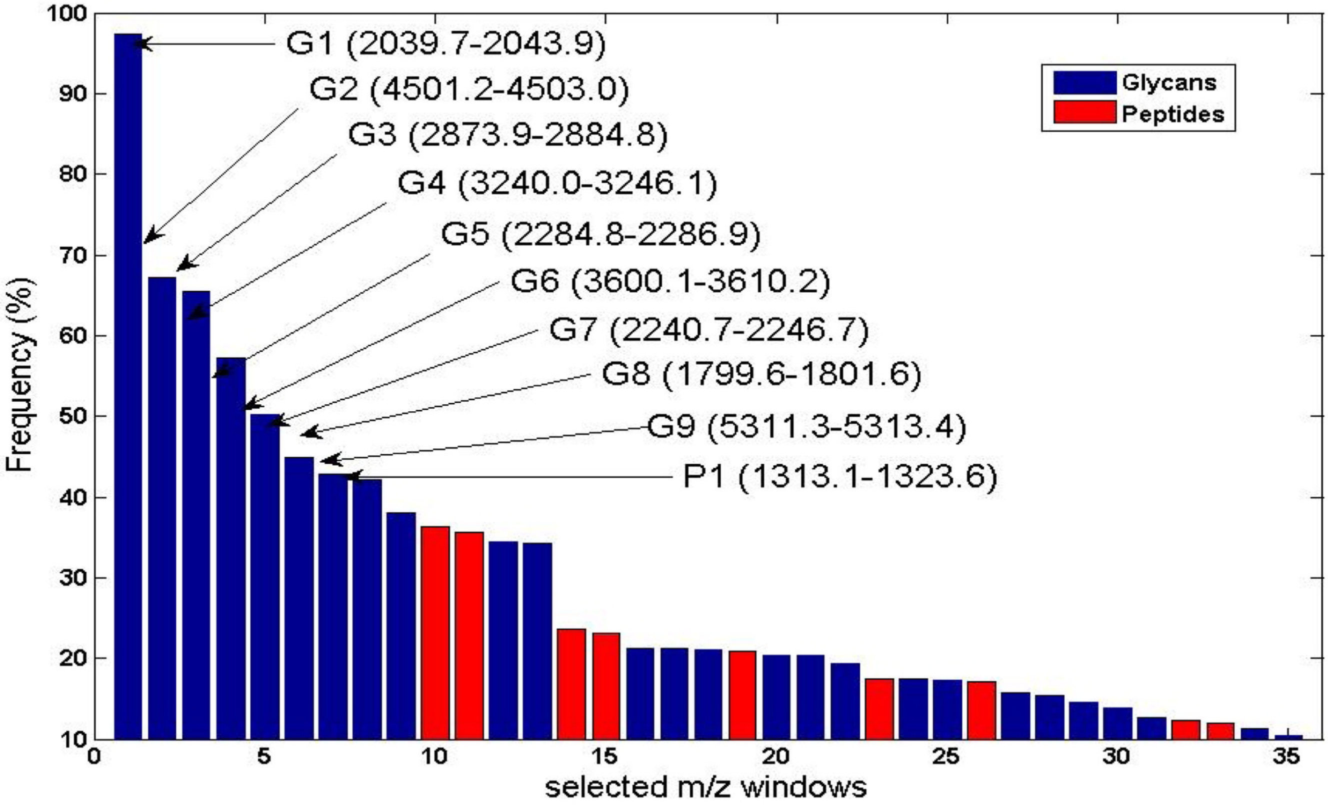

Fig. 4.

Frequency of occurrence of peptide and glycan peaks selected by ACO-SVM over 10 splits, with each split run 100 times (10*100 runs). The top ten *m/z* windows are shown.

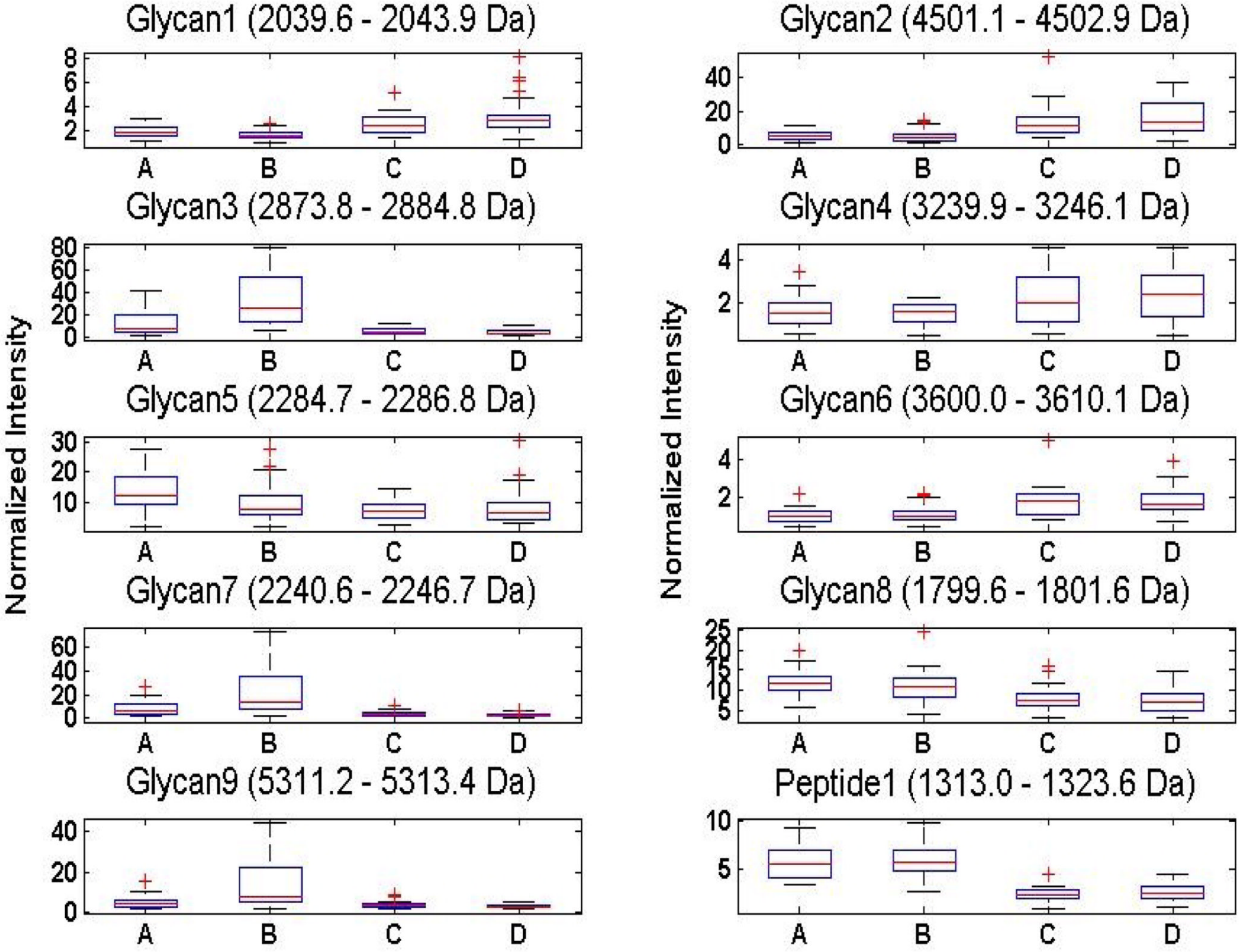

Fig. 5.

Boxplots for the 9 glycan and 1 peptide markers selected. **A**. Fibrosis (n=20); **B**. Cirrhosis (n=25); **C**. Early HCC (stage I/II) (n=17); **D**. Late HCC (stage III/IV) (n=33)