A Multi-PCA Approach to Glycan Biomarker Discovery using Mass Spectromtery Profile Data

Anoop M. Mayampurath, Chuan-Yih Yu

December 17, 2009

1 Abstract

Being one of the most common post-translational protein modifications occurring in humans, glycosylation plays a crucial role in the onset of various diseases such as cancer. Mass spectrometry is often used to acquire glycan profile data in order to provide a quantitative assessment of variations in glycan abundance between cancer and healthy patients, with the aim of identifying biomarkers of the disease. In this paper, we propose a simple computational method to accurately identify possible glycan biomarkers using profile data. Initially, we identify potential glycans from the profile data using spectral searching based on a dedicated list of glycan compositions. The method then identifies subsets of glycans that are significant in the fact that they exhibit similar patterns within a class (either disease or healthy) but exhibit large variance across classes. We illustrate the efficacy of the method using previously studied data for discovering biomarkers in hepatocellular carcinoma using N-glycan serum markers. We were able to identify the major biomarkers and were also able to identify several new glycans that could be potential glycan biomarkers for Hepatocelluar carcinoma.

2 Introduction

Mass spectrometry (MS) has been widely used to determine both sequence and structure of glycans in glycoproteins[6][9]. Matrix-Assisted Laser Desorption/Ionization (MALDI) - Time-of-flight (TOF) mass spectrometry platforms have been utilized to study profile of glycans. MALDI is used to ionize the glycan and the TOF reports accurate mass/charge (m/z) values for each glycan. The use of glycomic profile data from MALDI-TOF mass spectrometry platform for biomarker discovery has been previously explored in [5]. The method involved analysis of spectra using principal component analysis (PCA) to segregate prostate cancer from healthy. Region-of-convergence (ROC) plot for each manually annotated glycan was used to assign a confidence score indicative of utility of glycan as a biomarker for prostate cancer.

Figure 1 from [5] shows an example mass spectra for cancer and healthy data with annotated glycans.

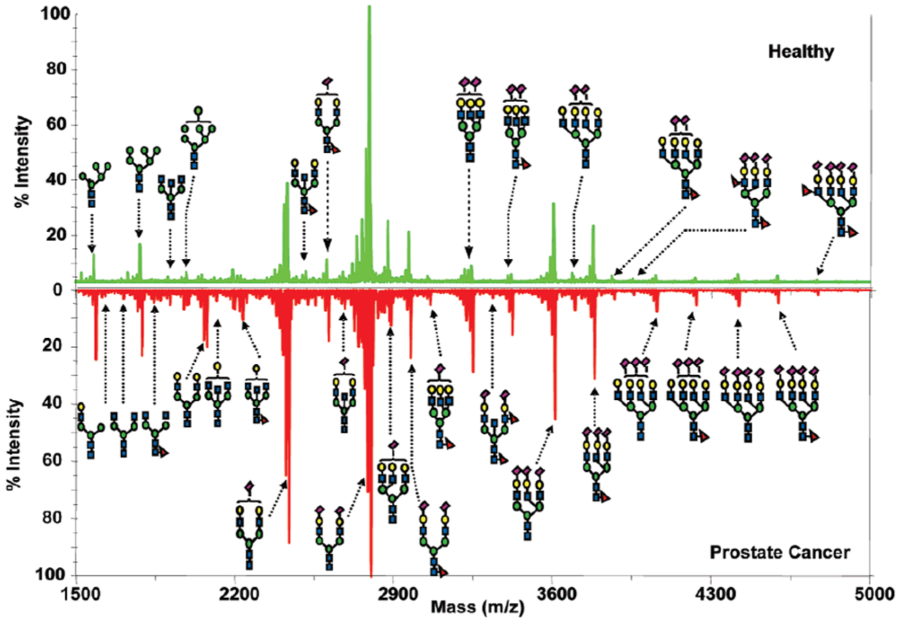


Figure 1: Example of utility of glycan profile data for biomarker discovery

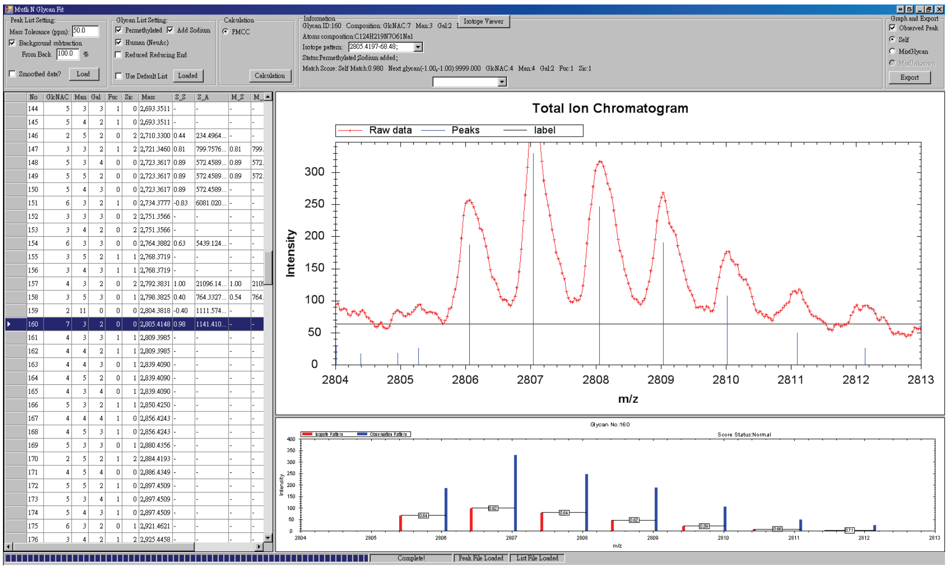
Recently [8] and [7] developed a computational method for identifying potential biomarkers for hepatocellular carcinoma (HCC) and chronic liver disease (CLD). Hepatocellular carcinoma (HCC) is a form of cancer associated with the liver, typically associated with alcoholism [2]. HCC is difficult to diagnose due to its heterogenity along with low sensitivity nature of current biomarkers used, which includes alphafetoprotein [3]. The glycoproteins associated with HCC often display aberrant variations in glycan profiles between cancer and healthy patients. In order to capitalize on this, [8] and [7] build Support Vector Models (SVMs) to isolate importance spectra and to identify glycans that show considerable change among HCC, CLD and healthy. Using these methods, the authors were able to identify seven potential glycan biomarkers. However, the use of SVMs along with a complicated glycan peak-picking algorithm makes the use of this methodology tedious.

In this project, a simple computational method is proposed that involves automatic annotation of glycan spectra following which a multi-PCA to identify 'significant' glycan biomarkers. Significance implies that these glycans show a cohesive profile pattern among either disease or healthy, but show a large variance between healthy and disease. One hundred and fifty one mass spectra that included HCC (73) and normal (78) spectra from [8] was used as test datasets. The original seven identified glycans were accounted for and several new potential glycan biomarkers are also reported. The details of the method are outlined below in the methods section, and results acquired are described in the results section.

3 Methods

3.1 DataProcessing

An in-house developed software tool, MultiNGlycan, was used to annotate all mass spectra. A screen shot of the tool is given in Figure 2. MultiNGlycan uses different glycan compositions to construct theoretical glycan isotope distributions, following which, the tool tries to find the best correlation score between the generated theoretical distributions and the observed distributions in a spectrum to annotate all observed glycans. Detailed working of the tool and the steps it uses are outlined below.



Mutli N Glycan Fit

Peak List Setting:
Mass Tolerance (ppm): 50.0
☑ Background subtraction
From Back 100.0 %
☐ Smoothed data? Load

Glycan List Setting:
☑ Permethylated ☑ Add Sodium
☑ Human (NeuAc)
☐ Reduced Reducing End
☐ Use Default List Loaded

Calculation
● PMCC
Calculation

Information
Glycan ID:160 Composition: GlcNAC:7 Man:3 Gal:2
Atoms composition:C124H219N7O61Na1
Isotope pattern: 2805.4197-68.48;
Status:Permethylated;Sodium added;
Match Score: Self Match:0.980 Next glycan(-1.00,-1.00):9999.000 GlcNAC:4 Man:4 Gal:2 Fuc:1 Sic:1

Graph and Export
☑ Observed Peak
● Self
○ MixGlycan
○ MixUnknown
Export

	No	GlcNAC	Man	Gal	Fuc	Sic	Mass	S_S	S_A	M_S	M_
	144	5	3	3	1	0	2,693.3511	-	-	-	-
	145	5	4	2	1	0	2,693.3511	-	-	-	-
	146	2	5	2	0	2	2,710.3300	0.44	234.4964...	-	-
	147	3	3	2	1	2	2,721.3460	0.81	799.7576...	0.81	799.
	148	5	3	4	0	0	2,723.3617	0.89	572.4589...	0.89	572.
	149	5	5	2	0	0	2,723.3617	0.89	572.4589...	0.89	572.
	150	5	4	3	0	0	2,723.3617	0.89	572.4589...	-	-
	151	6	3	2	1	0	2,734.3777	-0.83	6081.020...	-	-
	152	3	3	3	0	2	2,751.3566	-	-	-	-
	153	3	4	2	0	2	2,751.3566	-	-	-	-
	154	6	3	3	0	0	2,764.3882	0.63	5439.124...	-	-
	155	3	5	2	1	1	2,768.3719	-	-	-	-
	156	3	4	3	1	1	2,768.3719	-	-	-	-
	157	4	3	2	0	2	2,792.3831	1.00	21096.14...	1.00	210
	158	3	5	3	0	1	2,798.3825	0.40	764.3327...	0.54	764.
	159	2	11	0	0	0	2,804.3818	-0.40	1111.574...	-	-
▶	160	7	3	2	0	0	2,805.4148	0.98	1141.410...	-	-
	161	4	3	3	1	1	2,809.3985	-	-	-	-
	162	4	4	2	1	1	2,809.3985	-	-	-	-
	163	4	4	3	0	1	2,839.4090	-	-	-	-
	164	4	5	2	0	1	2,839.4090	-	-	-	-
	165	4	3	4	0	1	2,839.4090	-	-	-	-
	166	5	3	2	1	1	2,850.4250	-	-	-	-
	167	4	4	4	1	0	2,856.4243	-	-	-	-
	168	4	5	3	1	0	2,856.4243	-	-	-	-
	169	5	3	3	0	1	2,880.4356	-	-	-	-
	170	2	5	2	1	2	2,884.4193	-	-	-	-
	171	4	5	4	0	0	2,886.4349	-	-	-	-
	172	5	5	2	1	0	2,897.4509	-	-	-	-
	173	5	3	4	1	0	2,897.4509	-	-	-	-
	174	5	4	3	1	0	2,897.4509	-	-	-	-
	175	6	3	2	0	1	2,921.4621	-	-	-	-
	176	3	4	2	1	2	2,925.4458	-	-	-	-

Complete! Peak File Loaded List File Loaded

Figure 2: Screenshot of MultiNGlycan Software Tool

The first step is signal processing. MultiNGlycan lets the user specify the mode of background subtraction. The user can set the percentages of spectrum intensities to be considered as background, which is then subtracted from the original spectrum. For example, if the user chooses a percentage level of 70, the program will start form the highest mass and sum up 70% of intensities and divide by total number of peaks. Once the background has been removed, peak picking is performed in order to detect glycan composition peaks. Peak picking is done by first calculating the slope of two adjacent points. If the slopes have a dramatic change from positive to negative or vice versa, then the point is chosen to a peak or a valley depending on the change.

It will not only consider two adjacent points but also a wide range of slope change. This will prevent identification of noise as a true peak.

The second step is to load a glycan composition list, which contains putative glycan compositions. Here a list generated by a Ph.D. Student, Daniel Schrider was used. For each chemical composition, we generate theoretical isotope pattern. MultiNGlycan also has the capability to find glycans that have overlapping isotope patterns with each other.

The final step is identifying the glycan compositions in spectra confidently. Pattern matching is done between observed glycan distribution and the theoretical glycan distribution using a fit correlation score. In MultiNGlycan, three different models are used to calculate the correlation coefficient score. The first is the 'self-glycan' model, in which the glycan is assumed to be non-overlapping. A correlation coefficient score is calculated by matching theoretical and observed patterns through correlation methods. The second is the 'mix-glycan model' in which the glycan is assumed to be overlapping with another glycan. In this case, a composite theoretical pattern is generated from the two individual glycan patterns and matched to the observed. The third model, the 'unknown-mix-glycan' model, the glycan is assumed to be overlapping with an unknown compound such as a contamination. The mass value of the unknown compound is estimated from the spectra and a composite theoretical pattern is again generated and matched to the observed. Linear regression methods are then used to identify the best correlation coefficient value. In this project, the single-glycan model was used along with its corresponding correlation score as a metric of confidence.

In order to filter out inaccurate annotations, filtering steps were undertaken. Glycan annotations that were present in 30% of total spectra and with a correlation fit score > 0.5 were retained. Also, duplicated compositions that matched on mass (e.g. compositions that differ only in mannose composition in that one had glucose and the other had galactose) were removed. After filtering, a total of 307 potential glycans were identified. The glycan shaving algorithm, described next, was used to identify the ten most significant glycans that showed maximum variation between healthy and disease.

3.2 Glycan shaving

Analysis of 300 glycans across 151 spectra can be quite tedious, and thus necessitates identification of significant manageable number of glycans. The reduction of the number of glycans, based on techniques used in [4], is done by correlating the data with its first principal component. The glycans with the lowest correlation score are indicative of glycans that show minimal difference across the two classes. These can be removed. The method is iterated until the desired numbers of glycans are reached. Step-by-step details of the algorithm are given is section 3.2.1. A pictorial representation of the algorithm is shown in Figure 3. The algorithm was iterated until 10 glycans were acquired. These glycans are supposed to be coherent in intensity changes while having high variance between cancer and no-cancer (healthy). Another utility of the shaving technique is to reduce the cardinality of the number of spectra by transposing X to shave off spectra instead of glycans. This was done to acquire 10 significant spectra.

3.2.1 Multi-PCA Algorithm

Consider that after annotation and filtering, a Nxp matrix X is obtained where N represents the number of glycans and p represents the number of spectra

- PCA is performed on X resulting in generation on N principal components 1. ψ_1 is the first principal component and accounts for the most variance in the data among all other principal components.

$$X \longmapsto \psi_1, \psi_2, \psi_3 \ldots . \psi_N \tag{1}$$

- ψ_1 is correlated with X using inner product $< X, psi_1 >$.
- Sort glycans in X by inner product
- Shave of 10% of glycans with the lowest innert product score
- Repeat

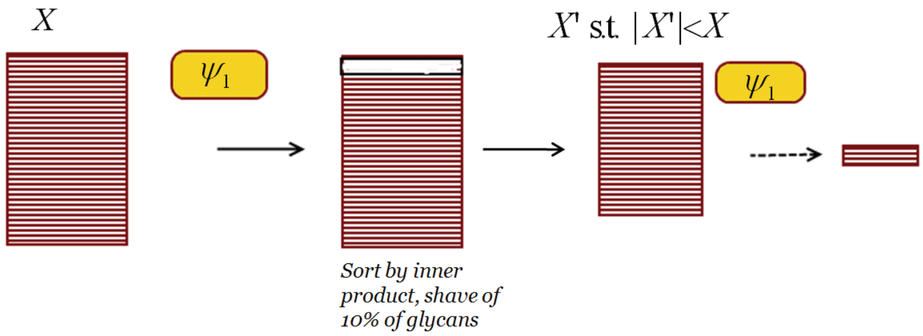


Figure 3: Multi-PCA algorithm

4 Results

Figure 4 depicts the comparison of results from both the SVM method from [8] and the simpler multi-PCA approach described in this paper. As can be seen, the multi-PCA approach identified the most significant biomarker

(1580 m/z). Please note that multi-PCA results report mass and so will be different from the m/z value in the SVM method by about 1. The multi-PCA approach also identified 2851 with a m/z error or 0.5. This appears to be down-regulated in cancer. This might due to two reasons - one, the results for the SVM-method show intensities from selected spectra, whereas the multi-PCA shows glycan summed intensity for all spectra. Second, it might be that there is no variation between HCC and NC (no-cancer) at all to begin with. Out of the remaining 5 peaks, m/zs 1996, 4311 and 4501 were not present in the original composition file and were thus available for annotation. 2040 m/z was filtered out due to bad correlation score. All these four glycans missing from the multi-PCA result appear to not show any aberrant intensity changes between cancer and non-cancer. The case of the remaining glycan, 2187, is curious. The SVM-method identified the glycan to be down regulated. The multi-PCA approach gave a bad correlation score for this glycan, so in essence it got filtered out. However, on detailed analysis, it was found to be overlapping with another glycan present at 2192 m/z, which the multi-PCA reported to be significant and up-regulated for HCC. This indicates that our assumption of single-glycan patterns while annotation is insufficient due to the presence of overlapping glycans. This can be amended by choosing the 'glycan-mix' correlation score and is left as one of the future directions of this research.

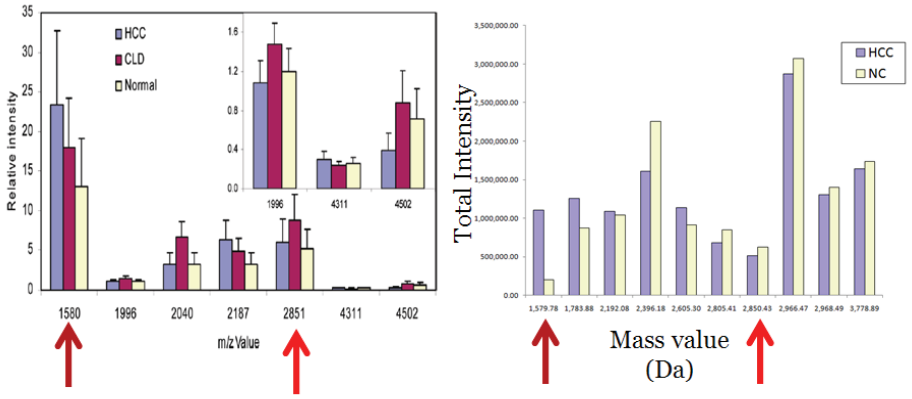
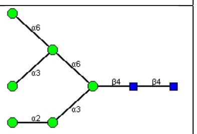
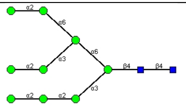
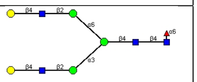
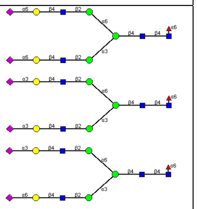


Figure 4: Comparison of results from SVM-method (left) and multi-PCA method (right). The SVM-method results plots seven glycan peaks as m/z and their corresponding intensity in selected spectra. The multi-PCA method results plots ten glycan peaks as mass (Da) and their summed intensity across all spectra. Comparison reveals two common identifications, which are indicated.

The multi-PCA approach was also able to identify additional glycan peaks that showed remarkable variation between HCC and NC. Details of some of these glycans including structural nature is given in Table 1. The structures were acquired by searching the Functional Glycomics database [1] using the matched glycan composition and looking for structures seen in human serum which is the sample origin for this dataset. Of particular interest are glycans at 2605 Da which indicate fucosylation and the glycan at 2966 Da whose all three possible structure show sialic-acid termination.

The shaving method was also applied on spectra to acquire 10 significant spectra. 7 HCC and 3 NC spectra were obtained and their sum is plotted in Figure 5 against each other.

Table 1: Glycan List

Mass	Glucose	Mannose	GalNAc	Fucose	Neu5Ac	Structure
1783.88	0	6	2	0	0	
2396.18	0	9	2	0	0	
2605.30	2	3	4	1	0	
2966.47	2	3	4	1	2	

5 Conclusion & Future Direction

This project outlines a simple computational method to identify potential glycan biomarkers using mass spectra acquired from a MALDI-TOF platform. The method was applied to data from previous research to validate glycan biomarkers for hepatocellular carcinoma. Some reported biomarkers were validated, and the missing ones were all accounted for. Additionally, several new glycan composition and structures were identified that could be further potential biomarkers.

An important component of glycan analysis that is missing here is effect of linkage between saccharide molecules. These can be studied by applying the same method to fragmentation data that represent glycan fragments. Also, as indicated before, the effect of overlapping glycans can be studied in

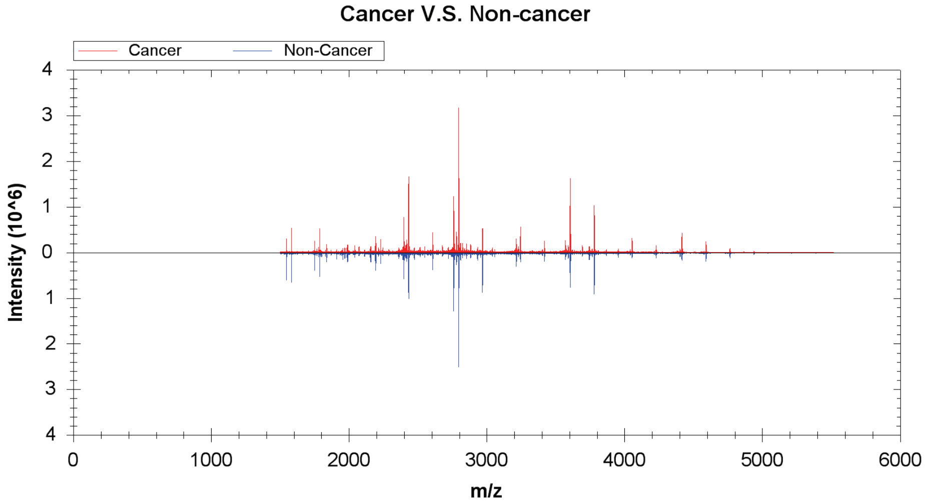


Figure 5: Summed intensities of 10 significant spectra across two classes

detail. Another possible area of future research is through orthogonalizing the dataset in order to identify other patterns.

6 Acknowledgements

We thank Prof. Haixu Tang from the School of Informatics & Computing for advising our project and Prof. Yehia Mechref from the Dept of Chemistry for providing the glycan mass spectra.

References

[1] Functional glycomics database. `http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp`.

[2] H. B. El-Serag, J. A. Marrero, L. Rudolph, and K. R. Reddy. Diagnosis and treatment of hepatocellular carcinoma. *Gastroenterology*, 134:1752–63, 2008.

[3] J. Filmus and M. Capurro. Glypican-3 and alphafetoprotein as diagnostic tests for hepatocellular carcinoma. *Mol Diagn*, 8:207–12, 2004.

[4] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown. 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*, 1, 2000.

[5] Z Kyselova, Y. Mechref, M. M. Al Bataineh, L. E. Dobrolecki, R. J. Hickey, J. Vinson, C. J. Sweeney, and M. V. Novotny. Alterations in the serum glycome due to metastatic prostate cancer. *Journal of Proteome Research*, 6:1822–1832, 2007.

[6] Y. Mechref, N.V. Novotny, and C. Krishnan. Stuctural characterization of oligosaccharides using maldi-tof/tof tandemn mass spectrometry. *Analytical Chemistry*, 75(18):4895–4903, 2003.

[7] H. W. Ressom, R. S. Varghese, L. Goldman, Y. An, C. A. Loffredo, M. Abdel-Hamid, Z. Kyselova, Y. Mechref, M. Novotny, S. K. Drake, and R. Goldman. Analysis of maldi-tof mass spectrometry data for discovery of peptide and glycan biomarkers of hepatocellular carcinoma. *J Proteome Re*, 7:603–10, 2008.

[8] Z. Tang, R. S. Varghese, S. Bekesova, C. A. Loffredo, M. A. Hamid, Z. Kyselova, Y. Mechref, M. V. Novotny, R. Goldman, and H. W. Ressom. Identification of n-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data. *J Proteome Res*, 2009.

[9] W. Wuhree, A.M. Deelder, and C.H. Hokke. Protein glycosylation analysis be liquid chromatography and tandem mass spectrometry. *Journal of Chromatography*, 825:124–133, 2005.