

Multivariate Analysis of Glycomic Mass Spectrometry Data

Parul Purohit¹, David Rocke², Hyunjoo An³, Suzanne Miyamoto⁴, Carlito Lebrilla⁵

Keywords: glycomics, mass spectrometry, multivariate analysis

1 Introduction.

Glycans (oligosaccharides) are fundamental to many biologic processes. They are attached to proteins as a post-translational modification and are important for protein function. Glycosylation of proteins can change in response to a range of diseases, especially cancer. In particular, in the case of ovarian cancer, a leading cause of death, it is imperative to find a more robust diagnostic test than the unreliable CA125 test currently used. A study of the glycosylation changes due to the cancer thus provides a novel approach to finding alternate biomarkers for cancer.

Mass spectrometry is an effective way to detect and measure glycans (400-1500AMU). The data however tends to be complicated with thousands of mass measurements per patient and a relatively small number of patients and data analyses remains a great challenge. Use of multivariate analysis has proved invaluable for patient classification and biomarker identification. Principal Component Analysis (PCA) has traditionally been used for such data. However for complicated data, a combination of supervised and unsupervised learning provides a more effective approach.

By using a combination of pre-processing, dimension reduction and discriminant analysis techniques, we attempt to classify a set of 43 patients that were tested for ovarian cancer. Twenty of these were considered 'normal' and the other 23 as having cancer as indicated by their high CA125 levels. By using a leave one out cross-validated approach, we obtain mis-classifications which we show are consistent with the mass spectrometry data.

2 Methods.

Mass spectrometry was conducted using patient serum using matrix-assisted laser desorption ionization (MALDI) Fourier Transformed spectrometry (FTMS). Oligosaccharides were chemically cleaved from glycoproteins. The released glycans were then subjected to solid phase extraction and three separate fractions eluted (10%, 20% and 40% ACN in H₂O). Each fraction was analyzed by MALDI-FTMS in the positive mode.

¹ Institute for Data Analysis and Integrated Computing, University of California, Davis, CA, USA. E-mail: pvpuurohit@ucdavis.edu

² Division of Biostatistics, Department of Applied Science and Institute for Image Analysis and Integrated Computing, University of California, Davis, CA, USA. E-mail: dmrocke@ucdavis.edu

³ Department of Chemistry, University of California, Davis, CA, USA. E-mail: jooan@ucdavis.edu

⁴ Department of Hematology and Oncology, University of California, Davis, CA, USA. E-mail: smiyamoto@ucdavis.edu

⁵ Department of Chemistry, University of California, Davis, CA, USA. E-mail: cblebrilla@ucdavis.edu

Initially, a spectrum of each fraction had a total of 507,765 data points; we reduced the dimension by a binning technique of 500 points per window which preserved the salient features of each spectrum while reducing each spectrum to 1015 predictors. Other pre-processing steps involved background correction using a scoring approach and normalization of each spectrum to an integrated intensity of 1.0. Finally, all 3 fractions per patient were additively combined to obtain one spectrum per patient giving us a data matrix of 43 x 1015. T-tests were used to choose predictors, with the criterion set at a p-value of 0.005. Using these, we performed further dimension reduction by using partial least squares regression (PLS) and PCA. Using 9 factors to describe the data set (estimated using a scree plot), we performed logistic regression to predict the classification of each patient. All calculations were performed using a cross-validated approach predicting the classification of each patient using the rest of the data set as a 'training' set.

3 Results.

Dimension reduction by PCA yielded 3 mis-classifications whereas PLS yielded 6. However 2 of the patients were consistently mis-classified by both methods as not having cancer even though they had high CA125 levels. We plotted the raw normalized binned spectra of these 2 patients against all the other high CA125 patients in the glycan region of mass 500-600AMU as shown in Figure 1. As a comparison, a plot of all the 'normal' patients is shown as well. The figures indicate that the glycan peak at ~ 507AMU is absent for the 'normal' patients whereas the peak at 567 AMU is absent for the high CA125 patients. However, in the case of the mis-classified patients, the spectra are closer to the normal spectra exposing the unreliable nature of the CA125 test. With such statistical analysis, we hope to achieve a more robust classification than the CA125 test. The analysis can be further explored to find and validate new serum biomarkers of ovarian cancer and has the potential to improve early detection of the disease.

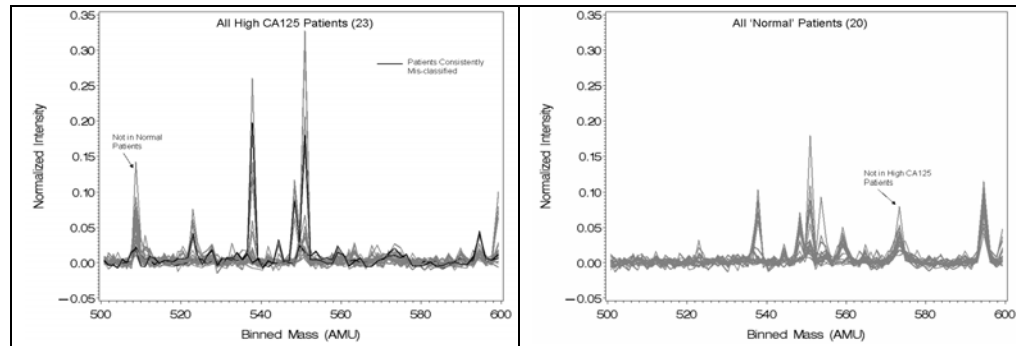


Figure 1. Spectra for all the high CA125 patients and for all the patients classified as 'normal'.

4 References.

- [3] Campa, M. J., Fitzgerald, M. C., and Patz, F., editors. 2003. Mining MALDI Data. *Proteomics 3 (9)*. Wiley-VCH.
- [1] Stimpfl, M., Schmid, Schiebl, I, Tong, D., Leodolter, S., Obermair, A., and Zeillinger, R. 1999. Expression of mucins and cytokeratins in ovarian cancer cell lines. *Cancer Lett. 145*, pp. 133-141
- [2] Xi, Y., and Rocke, D. M. Baseline correction for NMR spectroscopic metabolomics data analysis. To be published.