Chemometrics and Intelligent Laboratory Systems 176 (2018) 74–81

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Wavelet based classification of MALDI-IMS-MS spectra of serum N-Linked glycans from normal controls and patients diagnosed with Barrett's esophagus, high grade dysplasia, and esophageal adenocarcinoma

B.K. Lavine [a,*], C.G. White [a], T. Ding [a], M.M. Gaye [b], D.E. Clemmer [b]

[a] Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, United States
[b] Department of Chemistry, Indiana University, Bloomington, IN 47405, United States

ARTICLE INFO

Keywords:
Variable selection
Discrete wavelet transform
Pattern recognition
Genetic algorithms
Profile analysis
N-linked serum glycans

ABSTRACT

Profiling of complex biological samples (e.g., serum) using mass spectrometry continues to be an active area of research with a large and growing literature. Pattern recognition techniques can be effective methods for the analysis of complex data sets generated in these types of studies. Currently, we are investigating the discrimination of disease phenotypes associated with esophageal adenocarcinoma by analysis of single N-linked glycans using matrix assisted laser desorption ionization-ion mobility spectrometry-mass spectrometry (MALDI-IMS-MS). The glycans were extracted from sera of healthy (normal) controls (NC) and patients diagnosed with Barrett's Esophagus (BE), high grade dysplasia (HGD), and esophageal adenocarcinoma (EAC). MALDI-IMS-MS spectral images were collected in duplicate for these 58 serum samples: BE (14 individuals), HGD (7 individuals), EAC (20 individuals) and NC (17 individuals). Ion mobility distributions of N-linked glycans that possessed sufficient signal to noise in all 116 spectra were extracted from the images by box selection across a specific drift bin and m/z range corresponding to a single linked N-glycan ion. A composite ion mobility distribution profile was obtained for each image by sequentially splicing together the mobility distributions of each N-linked glycan across an arbitrary drift bin axis. Wavelet preprocessing of the composite ion mobility distribution profiles was performed using the discrete wavelet transform, which was coupled to a genetic algorithm for variable selection to identify a subset of wavelet coefficients within the data set that optimized the separation of the four classes (BE, HGD, EAC, and NC) in a plot of the two largest principal components of the wavelet transformed data. A discriminant developed from the wavelet coefficients identified by the pattern recognition GA correctly classified all ion mobility distribution profiles in the training set (45 individuals and 87 distribution profiles) and 23 of 26 blinds (13 individuals and 26 distribution profiles) in the prediction set. The proposed MALDI-IMS-MS and pattern recognition methodology has the potential to exploit molecules in serum samples that can serve as the basis of a potential method for cancer prescreening.

1. Introduction

Alterations in the glycosylation of proteins have been implicated in various types of cancers [1,2]. Changes in glycosylation have been associated with the expression of glycosyltransferases and glycosidases within a malignant cell [3]. In an attempt to exploit aberrations in glycosylation for purposes of disease diagnosis, Isailovic and coworkers [4,5] have shown that patients with liver cancer, liver cirrhosis, and healthy individuals can be differentiated by analyzing glycomic profiles of sera obtained from ion mobility spectrometry (IMS) and mass spectrometry (MS). Although the initial discrimination between liver cancer and the two control groups was made on the basis of the ion distribution profile of a single glycan, the discrimination between these three groups was further enhanced when the distribution profiles of ten glycans were combined. Because IMS-MS allows for the separation of molecules according to their mass-to-charge ratio and their shape, isomeric and/or conformeric separations of individual glycans can be obtained. This is not the case with matrix assisted laser desorption ionization-time of flight-mass spectrometry (MALDI-TOF-MS) as it is restricted to molecular separations based solely on the mass-to-charge ratio.

It is plausible that an approach similar to the one used by Isailovic could successfully be applied to delineate other disease phenotypes. In this study, discrimination between phenotypes associated with esophageal adenocarcinoma was undertaken. Analysis of N-linked glycan serum

* Corresponding author.
E-mail address: bklab@chem.okstate.edu (B.K. Lavine).

https://doi.org/10.1016/j.chemolab.2018.03.008
Received 6 June 2017; Received in revised form 5 January 2018; Accepted 13 March 2018
Available online 15 March 2018

samples in duplicate obtained from healthy individuals and patients with Barrett's esophagus (BE), high grade dysplasia (HGD), and early stage esophagus adenocarcinoma (EAC) was performed using MALDI in combination with IMS-MS. BE and HGD are two phenotypes that appear connected to EAC as individuals diagnosed with BE or HGD are at greater risk to develop EAC. As in other forms of cancer, early diagnosis is vital to ensure successful patient outcomes.

In this paper, a new procedure for discrimination of disease phenotypes using serum N-linked glycan profiles is proposed based on: (1) the preprocessing of the MALDI-IMS-MS glycan profiles using the discrete wavelet transform, (2) identification of wavelet coefficients correlated to disease phenotype by a genetic algorithm (GA) for feature selection, and (3) the development of discriminants based on principal component analysis (PCA) plots of the wavelet coefficients identified by a pattern recognition GA [6–12] with particular attention paid to the evaluation of the results (specifically, the total classification success-rate, sensitivity, and specificity). This new procedure was assessed using a data set consisting of 116 MALDI-IMS-MS serum N-linked glycan profiles from 58 individuals, some healthy which served as normal controls (NC) and others diagnosed with EAC, HGD, or BE. The advantage of using PCA [13] to evaluate the information content of the wavelet coefficients identified by the pattern recognition GA in this application is that problems arising from overfitting are mitigated.

MALDI-IMS-MS profiles of N-linked serum glycans from healthy individuals or patients with documented phenotypes of esophageal cancer (BE, HGD, and EAC) may have as many as 30,000 observations per image. To concentrate this information and to reduce the noise, an alternative and more compact representation of the data which describes the important mass spectral features is needed. The goal of this study is to demonstrate that data compression of mass spectral profiles of N-linked serum glycans can be achieved using the discrete wavelet transform to preprocess the data and a genetic algorithm for feature selection that identifies the smallest subset of serum samples in a plot of the two or three largest principal components of the wavelet preprocessed data.

The discrete wavelet transform [14–16] partitions a signal into its constituent frequencies. MALDI-IMS-MS data can be thought of as conveying a signal consisting of components at different frequencies that are not observable in the time domain. Wavelets transform the mass spectral data into components called scales, with each scale corresponding to a different frequency contribution. Application of the discrete wavelet transform (DWT) converts a signal from its original domain to the wavelet domain. The original data is represented as wavelet coefficients. Since wavelets can often extract information from the data using a small number of coefficients, it has been suggested that wavelet coefficients are superior to the original data for modeling. Since each wavelet coefficient is computed using a weighted sum of several points from the original data, smoothing occurs which causes a reduction in noise after wavelet preprocessing has occurred as unwanted variation has been removed. Compression of the data is achieved by removing wavelet coefficients that do not contain information about the class membership of the serum samples.

Although several features extraction/selection methods for mass spectra of serum samples have been previously published [17–24], most of them have been applied in the original frequency domain and ignore the fact that MALDI-IMS-MS spectra contain peaks with different scales. In this study, the discrete wavelet transform, which is able to address the multiscale nature of MALDI-IMS-MS spectral data, is proposed as a method to improve both feature extraction and classification of MALDI-IMS-MS spectra. Classification models developed from wavelet coefficients selected by a GA that utilizes a pattern recognition approach based on identifying the smallest subset of wavelet coefficients within the data set that optimize the separation of the classes in a plot of the two or three largest principal components of the data yielded lower misclassification rates than models developed from the original data.

2. Experimental

Serum samples were collected at the Henry Ford Health Clinic (Detroit, MI) from fifty-eight volunteers. Seventeen were from disease free individuals (i.e., NC), fourteen were from BE patients, seven were from HGD patients, and twenty were from EAC patients. Sera from healthy individuals and from patients diagnosed with BE, HGD, and EAC were randomized and treated with a mixture of Endo M and PNGase F [25] to release the N-glycans from the tryptic digested samples. The enzymatic reaction was allowed to run overnight at 37 °C in a phosphate buffer (pH 7.5).

After digestion of the serum samples, the N-glycans were preconcentrated using C_{18} Sep-Pak cartridges that were preconditioned with ethanol and deionized water [26,27], followed by passing the released N-glycans through a home-packed activated micro-spin column (preconditioned by acetonitrile and equilibrated with 0.1% trifluoroacetic acid solution) to further purify the glycans. Each purified sample was dried under vacuum and permethylated using a previously published procedure [28].

The samples were analyzed by a Waters Synapt G2-S traveling wave ion mobility mass spectrometer (Waters Corporation, Manchester, UK) coupled to a MALDI source operated in the positive ion mode. Each sample was dissolved in 2 μl of a 1:1 (v: v) water/methanol solution mixed with 2 μl of the MALDI matrix (2, 5 dihydrobenzoic acid at 10 mg ml^{-1} in 1:1 water: methanol with 2 mM sodium acetate). 2 μl of each sample/matrix mixture was spotted in duplicate (one following the other) on two 96-well MALDI plates (referred to as Plate 1 and Plate 2). The laser used to ionize each sample was a frequency-tripled Nd YAG laser (355 nm) firing at a rate of 1000 Hz in a reverse-spiral pattern as it is well known that carbohydrates are preferentially localized at the edges of a MALDI spot. For every sample, mass spectra were acquired from 1000 to 5000 m/z for 3 min. An external calibration was performed between runs using a MassPREP™ calibration mix containing polyethylene glycol (Waters Corporation, Milford, MA). All mass spectral data used in this study were collected within a 24-h time period.

Data for the N-glycans obtained in a diagonal selection from the two-dimensional mass spectra were extracted by Driftscope software (Waters Corporation, Manchester, UK). For each sample, a box selection was performed across a specific drift bin and m/z range corresponding to a single N-linked glycan ion. Although twelve N-linked glycan ions were detected, only nine N-linked glycan ions had sufficient S/N across all samples. Mobility distributions for each of the nine glycan ions were obtained by exporting the data in the box selection to MassLynx software (Waters Corporation, Manchester, UK) with the retained drift time function enabled which allowed generation of the arrival time distributions. For each sample run, a composite IMS distribution was obtained by sequentially splicing together the mobility distribution from these nine N-linked glycan ions across an arbitrary drift bin axis [29]. These nine glycan composite ion mobility distributions were obtained for the entire set of 116 MALDI-IMS-MS spectra.

3. Pattern recognition methodology

The data set was divided into a training set of forty-five serum

Table 1
Training set.

Disease Phenotype	Number of Serum Samples (Donors)	Number of Spectra
Normal Controls	14	28
Barrett's Esophagus	10	20
High Grade Dysplasia	5	10
Esophageal Adenocarcinoma	16	32
Total	45	90