

Ion Mobility-Mass Spectrometry Analysis of Serum N-linked Glycans from Esophageal Adenocarcinoma Phenotypes

M. M. Gaye,[†] S. J. Valentine,[†] Y. Hu,[‡] N. Mirjankar,[§] Z. T. Hammoud,^{||} Y. Mechref,[‡] B. K. Lavine,[§] and D. E. Clemmer^{*,†}

[†]Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States

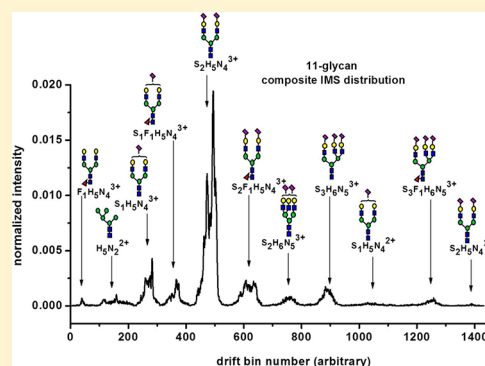
[‡]Department of Chemistry, Texas Tech University, Lubbock, Texas 79409, United States

[§]Department of Chemistry, Oklahoma State University, Stillwater, Oklahoma 74078, United States

^{||}Department of Surgery, Henry Ford Hospital, Detroit, Michigan 48202, United States

ABSTRACT: Three disease phenotypes, Barrett's esophagus (BE), high-grade dysplasia (HGD), esophageal adenocarcinoma (EAC), and a set of normal control (NC) serum samples are examined using a combination of ion mobility spectrometry (IMS), mass spectrometry (MS), and principal component analysis (PCA) techniques. Samples from a total of 136 individuals were examined, including 7 characterized as BE, 12 as HGD, 56 as EAC, and 61 as NC. In typical data sets, it was possible to assign ~20 to 30 glycan ions based on MS measurements. Ion mobility distributions for these ions show multiple features. In some cases, such as the $[S_1H_5N_4 + 3Na]^{3+}$ and $[S_1F_1H_5N_4 + 3Na]^{3+}$ glycan ions, the ratio of intensities of high-mobility features to low-mobility features vary significantly for different groups. The degree to which such variations in mobility profiles can be used to distinguish phenotypes is evaluated for 11 N-linked glycan ions. An outlier analysis on each sample class followed by an unsupervised PCA using a genetic algorithm for pattern recognition reveals that EAC samples are separated from NC samples based on 46 features originating from the 11-glycan composite IMS distribution.

KEYWORDS: ion mobility, electrospray ionization mass spectrometry, glycans, cancer, genetic algorithm, principal component analysis



INTRODUCTION

Recently, Isailovic et al. showed that patients with liver cancer, liver cirrhosis, and control groups could be delineated by analyzing serum N-linked glycans by a combination of ion mobility spectrometry (IMS), mass spectrometry (MS), and principal component analysis (PCA) techniques.^{1,2} In these studies, initial differentiation between patient groups was made on the basis of the ion mobility distribution for a single glycan ion¹ and was improved when distributions of 10 glycan ions were combined.² It seems likely that such an approach could be extended and used to distinguish other disease phenotypes. In this paper, we investigate disease state delineation between phenotypes associated with esophageal adenocarcinoma (EAC). The previously described methodology for data analysis was improved by performing an outlier analysis and by using a genetic algorithm for feature selection and pattern recognition.

EAC often starts as an asymptomatic state known as Barrett's esophagus (BE).^{3,4} As the disease progresses, the esophagus epithelium changes from a stratified structure to a simple columnar epithelium type.^{3,4} Along the progression pathway, low-grade dysplasia, high-grade dysplasia (HGD) and early stage esophagus adenocarcinoma are three different phenotypes (of many) associated with EAC.³ An individual diagnosed with BE has 100-fold greater risk for developing EAC; thus, these phenotypes appear to be highly connected.³ Ultimately, as in

other forms of cancer, the early diagnosis (preferably at an asymptomatic stage) is vital to ensure patient longevity.^{1,3–6} As an example, consider epithelial ovarian cancer (EOC) where ~80% of cases are detected in later developmental stages because the first two stages are asymptomatic.^{5,6} When EOC is diagnosed earlier, the five year survival rate rises from 45 to 94%.^{5,6} Currently, BE is diagnosed by means of an endoscopy, an invasive and expensive procedure.³ Significant efforts are underway to determine the feasibility of using biological fluids such as human blood to detect molecular signatures associated with disease states. One advantage of the latter approach is that blood samples are readily accessible and there is a potential for early stage (asymptomatic) detection of cancer.

For the last 40 years, it has been known that aberrant glycosylation of proteins can be associated with cancer;^{1,4,5,7–15} thus, with the development of new analytical techniques, numerous analyses of glycoproteins and glycans from sera have aimed to characterize these molecules. A number of specific structural changes occur in metastatic cells, including: variation in the abundances of specific glycans; changes in sialylation or fucosylation of glycans;^{14,16} and, the formation of incomplete or

Received: August 8, 2012

truncated structures, originating from variations in the expression of glycosyltransferases.^{14,16}

In general, structural characterization of glycans is challenging because these molecules often exist in very low abundances, and unlike other linear biopolymers, glycans can form a vast number of different structures (including isomers) that arise due to alternative branching and linkage possibilities.^{14–17} MS provides a powerful means of determining structures for trace levels of materials. However, often MS alone cannot distinguish isomeric forms (often because fragmentation patterns of multiple isomers are similar or because the pattern leads to ambiguities in assignments).¹⁸ Therefore, it is of interest to utilize a hybrid IMS-MS analysis. An ion's mobility is governed by its overall size and charge. Thus, in favorable cases, IMS allows the resolution of isomers prior to MS analysis.^{18–20} This work presents preliminary efforts for characterizing serum N-linked glycans associated with EAC using IMS coupled with time-of-flight (TOF) MS and applying an algorithm for pattern recognition on the generated data set.

■ EXPERIMENTAL SECTION

Materials

Peptide-N-glycosidase F (PNGase F, ≥95% purity), ammonium bicarbonate (≥99.0% purity), sodium hydroxide (NaOH, 97% purity) beads and iodomethane (99% purity) were purchased from Sigma (St. Louis, MO). Chloroform (99.8% purity), trifluoroacetic acid (TFA, 99+% purity) and solid-phase extraction cartridges (Discovery, DSC-pH) were obtained from Aldrich (Milwaukee, WI). Dimethyl sulfoxide (DMSO, 99.9% purity) and sodium chloride (99.0% purity) were purchased from J. T. Baker (Phillipsburg, NJ) and Merck (Darmstadt, Germany) respectively. Sodium phosphate monobasic and dibasic monohydrate were obtained from EM Science (Gibbstown, NJ). Microspin columns of active charcoal and empty microspin columns were purchased from Harvard Apparatus (Holliston, MA).

Samples (Populations and Collection)

Serum samples from patients with documented phenotypes (7 with BE, 12 HGD, and 56 EAC) and disease-free volunteers [also referred to as normal control (NC, 61 individuals)] were obtained from the Henry Ford Health clinic (Detroit, MI). These samples were acquired under an IRB approved protocol. EAC samples were obtained from both male and female subjects, their age spanning 46 to 91 years. The 56 individuals diagnosed with EAC were not at the same disease stage. That is, patients diagnosed with a small tumor size and no regional lymph node involvement (T1N0) as well as patients diagnosed with a more advanced stage [bigger tumor size, involvement of regional lymph nodes, presence of distant metastasis (T3N1M1)] and various stages in between are comprised in the EAC group. Venous blood samples were taken in the morning's fasting state. The samples were collected with minimal stasis in evacuated tubes. Within two hours, the tubes were centrifuged at –20 °C for 12 min at 1200× g. Finally, the serum samples were stored frozen in plastic vials at –80 °C until analysis.

Release of N-Glycans from Human Serum

The method used for release of N-linked glycans was adapted from previously described procedures.^{21,22} Briefly, 90.0 μL of a solution composed of sodium dodecyl sulfate, 2-mercaptoethanol and phosphate buffer (pH 7.5) was added to 10.0 μL of serum. After 45 min of incubation at 60 °C, 5.0 μL of 10-fold

diluted NP-40 was added and the mixture was left for 5 min in the dark. Finally 1.2 μL of 10-fold diluted PNGase F (final activity of 5 mU) was added to the mixture and incubated overnight at 37 °C.

Purification of N-Glycans

A solid-phase extraction cartridge (Discovery, DSC-pH) was conditioned with ethanol and water as described elsewhere.^{21,23} After digestion of the serum sample by PNGase F, the sample volume was adjusted to 1.0 mL with water (HPLC grade) and applied to the column after centrifugation. After five circulations the column was washed with 1.0 mL water; the resulting pass-through and the fifth circulation were combined and then dried under vacuum at 50 °C. Next, N-glycans were further purified using a microspin column of activated charcoal. The microspin column was washed, conditioned and glycans were eluted with acetonitrile/water/TFA solutions [85:15:0.1%, 5:95:0.1% and 40:60:0.1% (v:v:v), respectively]. Finally, glycans were dried under vacuum prior to solid-phase permethylation.

Solid-phase Permethylation

When N-glycans are permethylated, the negative charge on the sialic acid residues is neutralized. As a consequence, acidic and neutral structures can be detected at the same time as protonated or sodiated adducts in positive-ion mode electrospray ionization with an enhanced sensitivity.^{15,24} Permethylation was conducted as previously described.^{22,24} An empty microspin column was packed with NaOH beads suspended in acetonitrile and conditioned with DMSO prior to the application of sample. A reaction mixture composed of 60.0 μL DMSO, 2.4 μL water, 44.0 μL iodomethane and the dry extract of N-glycans was applied to the column. After 15 min a second aliquot of 44.0 μL iodomethane was added. Finally, the N-glycans were purified by liquid–liquid extraction with chloroform:aqueous sodium chloride solution (1 M) and then chloroform:water. The sample was dried under vacuum and was conserved at –20 °C until analyzed. The dry extract of N-linked glycans was dissolved in 60.0 μL methanol/water (50:50) with 0.2% formic acid and 2 mM sodium acetate prior to injection.










IMS-MS Analysis

Experimental and theoretical aspects of IMS-MS measurements,^{25–28} instrumental modes of operation,^{28–30} and utility for biomolecular analyses^{29,31,32} are discussed elsewhere. The home-built IMS-TOFMS instrument used to analyze N-linked glycans, released from 136 serum samples has been previously described.²⁸ Briefly, the instrument can be divided into three main parts: an electrospray ionization source; a 2 m long drift tube; and, a TOF mass analyzer. An automated injection system (NanoMate TriVersa, Advion, Ithaca, NY) is used to introduce electrosprayed glycan ions into the source. Ions are accumulated in an ion funnel,^{33,34} periodically pulsed into the drift tube and separated according to their mobilities in ~3 Torr of 300 K helium buffer gas. Finally, they are extracted into a TOF source for mass analysis. Glycans are identified according to their mass-to-charge ratios (m/z). A total acquisition time of five minutes for each sample has been used for the analyses presented here.

Pattern Recognition Analysis

As discussed in more detail below, an 11-glycan composite IMS distribution is created for each of the 136 samples and is used as an input for a statistical evaluation of the data set. Table 1 summarizes 11 glycan ions that we have chosen for further statistical analysis, based on insight that comes from prior work.^{1,2} For example, 9 of the glycan ions (doubly charged H_5N_2 ,

Table 1. Glycan Ions Used to Perform the Principal Component Analysis

glycan composition ^a	<i>m/z</i> measured	charge	<i>m/z</i> calculated	Tentative structure ^{47,49}
F ₁ H ₅ N ₄	763.6	3+	763.4	
H ₅ N ₂	801.5	2+	801.4	
S ₁ H ₅ N ₄	826.4	3+	825.7	
	1227.3	2+	1227.1	
S ₁ F ₁ H ₅ N ₄	884.1	3+	883.8	
S ₂ H ₅ N ₄	946.7	3+	946.1	
	1408.2	2+	1407.7	
S ₂ F ₁ H ₅ N ₄	1004.8	3+	1004.2	
S ₂ H ₆ N ₅	1096.3	3+	1095.9	
S ₃ H ₆ N ₅	1216.9	3+	1216.3	
S ₃ F ₁ H ₆ N ₅	1274.6	3+	1274.3	

^aS represents sialic acid (diamonds); F represents fucose (triangles); H represents hexose (mannose open circles, galactose solid circles); N represents N-acetyl glucosamine (solid squares).

S₁H₅N₄, S₂H₅N₄, and triply charged S₁H₅N₄, S₁F₁H₅N₄, S₂H₅N₄, S₂H₆N₅, S₃H₆N₅, and S₃F₁H₆N₅ that we have selected here were observed to be important in distinguishing between patients having liver cancer or cirrhosis by Isailovic et al.¹ Most of these glycan structures showed some degree of clustering of the disease phenotypes (with the exception of the doubly- and triply charged glycan ions S₂H₅N₄). Recently, the fucosylated form (triply charged F₁H₅N₄) of a promising ion among the former set of glycans was added² to the pool of surveyed ions because variations in fucosylation are frequently involved in the development and progression of cancer.^{14,16} As a result, the data set clustering was improved.² Finally, we add here the triply charged glycan ion S₂F₁H₅N₄ (fucosylated equivalent of S₂H₅N₄), as the majority of promising markers previously reported for EAC were fucosylated.⁴ Overall, the number and the nature of glycans used for further analysis are somewhat arbitrary; that is, fewer, more, or different combinations could be used and improve the data set clustering with respect to disease groups. Combinations, other than the 11 that we have chosen, are not discussed further in the present work.

The composite IMS distribution of 11 glycan ions is created by first extracting ion mobility distributions for each one of the 11 glycans (from the IMS-MS data sets, intensities centered about the nominal *m/z* value are summed across a narrow *m/z* range for each drift time bin) and by then combining the distributions in equal time bins to create a single ion mobility axis that contains every peak that is observed for each of the 11 selected distributions. A composite IMS distribution is created for each sample. This is further explained in the Results and Discussion section and an example composite IMS distribution is shown in

this section. Within each composite IMS distribution the intensities are scaled uniformly for the data analysis. All combined mobility distributions used in this study contained 1431 points spanning 11 bins with the *m/z* values of these bins corresponding to the nominal values associated with the different glycan ions (763.4, 801.4, 825.7, 883.8, 946.1, 1004.2, 1095.9, 1216.3, 1227.1, 1274.3, and 1407.7). For pattern recognition analysis, each 11-glycan composite IMS distribution was represented as a data vector $X = (x_{763.4,1}, \dots, x_{763.4,83}, \dots, x_{801.4,84}, \dots, x_{801.4,217}, \dots, x_{1407.7,1431})$ where $x_{763.4,83}$ is the intensity of the 83rd point in the composite IMS distribution for the glycan with a nominal *m/z* value of 763.4.

In this study, PCA³⁵ was used to analyze the data. PCA attempts to reduce the dimensionality of the data by finding a set of orthogonal axes that represent the directions of maximum variance in the data. Each axis is called a principal component (PC). PCA is performed by decomposing the data matrix X ($n \times p$) into a score matrix T ($n \times f$), a loading matrix P ($f \times p$), and a residual matrix E ($n \times p$) where n is the number of samples, p is the number of variables, and f is the number of principal components extracted from the data with f less than p . The matrix equation for PCA is

$$X = 1x_{\text{mean}}^T + TP + E \quad (1)$$

where 1 is a column vector ($n \times 1$) of ones, and x_{mean}^T is the transpose of a column vector ($p \times 1$) corresponding to the sample mean. The score matrix defines the coordinates of the data points in the PC space of the data, and the loading matrix defines the relationship between the original measurement variables and the principal components. Both the score and loading matrices describe the signal in the data whereas the residual matrix describes the noise. PCA as well as other soft modeling techniques³⁶ possess two key attributes which ensure a successful analysis of this data: dimensionality reduction resulting in simplification of the multivariate data, and separation of signal from noise in the data.

Because outliers can adversely affect the performance of statistical and pattern recognition methods,^{37,38} the first step in this study was to perform outlier analysis on each class in the data set (56 EAC samples, 61 NC, 7 BE, and 12 HGD). A battery of tests was used including the generalized distance test,³⁹ PC plots, and the sample leverage.⁴⁰

The next step was feature selection. A genetic algorithm (GA) for feature selection and pattern recognition^{41–46} was used to identify spectral features that can differentiate EAC samples from NC. The pattern recognition GA identified features that optimized the separation of the combined mobility distributions by sample type in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by these features is about sample type. A principal component plot that shows separation of the spectra by sample type can only be generated using features whose variance or information is primarily about differences between the classes in the data set. This fitness criterion reduces the size of the search space as it limits the search to these types of features. In addition, the pattern recognition GA focuses on those classes and (or) samples that are difficult to classify by boosting the relative importance of those classes and (or) samples that are consistently misclassified as it trains. Over time, the algorithm learns its optimal parameters in a manner similar to that of a neural network. The pattern recognition GA integrates aspects of artificial intelligence and

evolutionary computations to yield a “smart” one-pass procedure for feature selection and pattern classification.

RESULTS AND DISCUSSION

Example Nested IMS-MS Distributions of N-Linked Glycans from Serum

Figure 1 shows a typical nested ion mobility-mass spectrum for one NC sample. Glycan ions are observed across a drift time

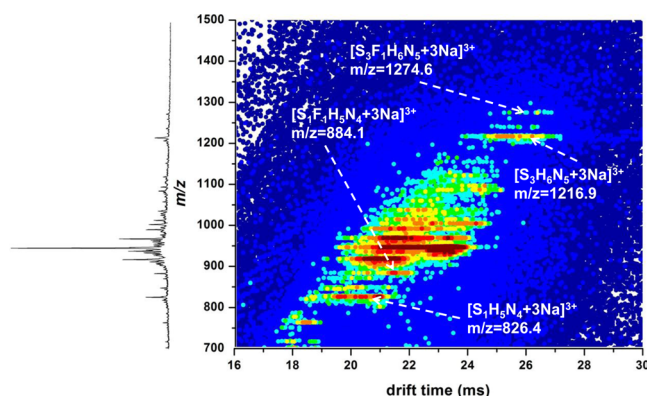


Figure 1. Two-dimensional dot plot of N-linked glycans from patient serum (NC) showing ion intensity as a function of drift time and m/z values is shown. Ion intensities are represented by a color code in which blue represents the lowest intensity and red indicates the highest intensity. Features corresponding to the $[S_1H_5N_4 + 3Na]^{3+}$, $[S_1F_1H_5N_4 + 3Na]^{3+}$, $[S_3H_6N_5 + 3Na]^{3+}$ and $[S_3F_1H_6N_5 + 3Na]^{3+}$ glycan ions are indicated.

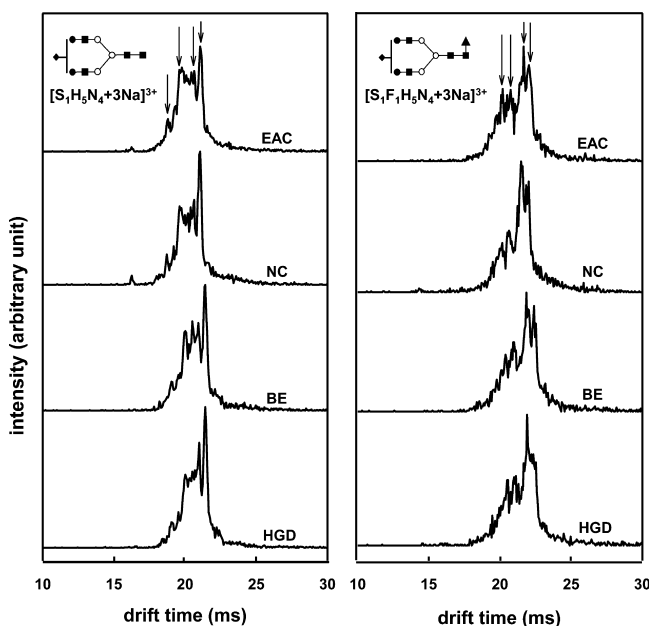


Figure 2. Ion mobility distributions of N-linked glycans from the four phenotypes [esophageal adenocarcinoma (EAC), normal control (NC), Barrett's esophagus (BE) and high-grade dysplasia (HGD)]. Each phenotype is represented by a single individual. Left and right panels show distributions for $[S_1H_5N_4 + 3Na]^{3+}$ and $[S_1F_1H_5N_4 + 3Na]^{3+}$ glycan ions respectively. Main reproducible features of the mobility distributions are indicated by arrows. Glycan structures are shown as insets.

range of 17.5 to 27.5 ms and from $m/z = 700$ to 1400. The most intense feature is at $m/z = 946.7$ which corresponds to the

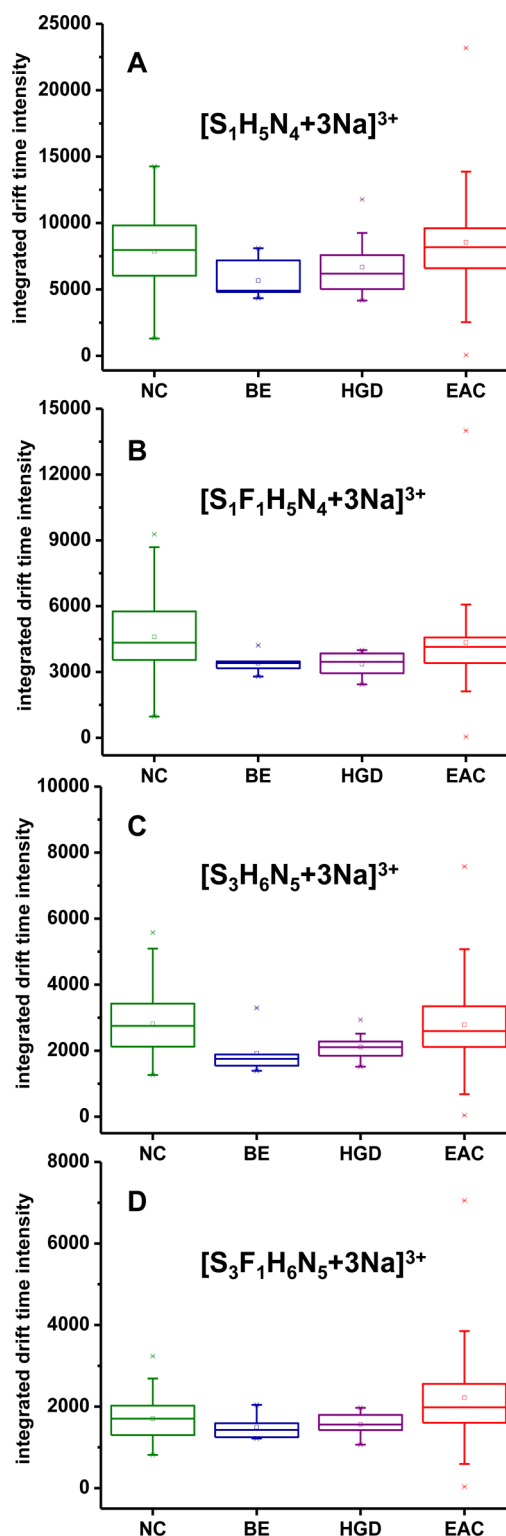


Figure 3. Box plots of integrated raw drift time intensities for (A) $[S_1H_5N_4 + 3Na]^{3+}$, (B) $[S_1F_1H_5N_4 + 3Na]^{3+}$, (C) $[S_3H_6N_5 + 3Na]^{3+}$ and (D) $[S_3F_1H_6N_5 + 3Na]^{3+}$ for all samples analyzed in this study. The data are organized according to disease phenotypes: green, blue, purple and red symbols corresponding to normal control (NC), Barrett's esophagus (BE), high-grade dysplasia (HGD) and esophageal adenocarcinoma (EAC) samples, respectively. The number of samples analyzed are respectively 61, 7, 12 and 56.

$[S_2H_5N_4 + 3Na]^{3+}$ glycan ion. Its ion mobility distribution spans the range from 19.5 to 25.1 ms. In all 136 samples, the 11 N-

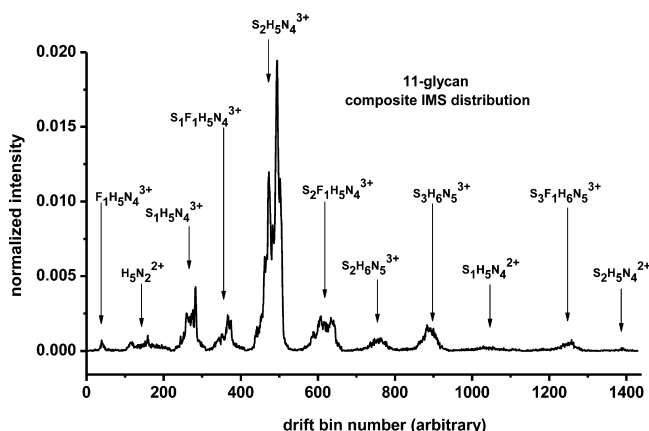


Figure 4. Eleven-glycan composite IMS distribution. Ion mobility distributions from individual glycan ions are sequentially spliced together to provide the distribution for each sample. The ion mobility distribution intensity of each glycan ion has been normalized using the sum of intensities over the entire concatenated distribution of a single sample. Data set features corresponding to the 11 glycans (Table 1) are labeled.

linked glycans reported in Table 1 have been identified by comparing the experimentally measured m/z values to theoretical values that are calculated from known atomic masses. The relatively broad features (in drift time) observed for all glycans could correspond to the presence of multiple isomers, or broad features could arise from multiple gas-phase ion conformations that exist for a single isomeric form.

Comparison of Glycan Ion Mobility Profiles

In order to assess whether or not glycans can be used to distinguish the different phenotypes, it is instructive to compare the mobility profiles of specific glycan ions for samples from each group. Figure 2 illustrates ion mobility distributions for the $[S_1H_5N_4 + 3Na]^{3+}$ and $[S_1F_1H_5N_4 + 3Na]^{3+}$ ions across the four different phenotypes (NC, BE, HGD, and EAC). In Figure 2, each group is represented by a data set from a unique sample (obtained from one single individual), and both glycan ions are extracted from the same data set. The glycans $S_1H_5N_4$ and $S_1F_1H_5N_4$ are of interest because each has been previously associated with a disease state [a congenital disorder of glycosylation referred to as a COG8 (conserved oligomeric Golgi complex) mutation]⁴⁷ and with cancer,¹ respectively. Prior work^{47,49} indicates that $S_1H_5N_4$ exists as two positional isomers with the sialic acid residue being on the α -1,3 or α -1,6 branch and these appear to be resolvable based on differences in mobilities.¹ In the complete data set for all samples, the ion mobility distributions obtained for the $[S_1H_5N_4 + 3Na]^{3+}$ ions display four features (designated by the arrows in Figure 2 showing the positions of peak maxima that are observed across all data sets)—one sharp peak at 21.4 ms and three others at 18.8, 19.6, and 20.9 ms. We note that visual examination of the data shows that the ion mobility distributions for this ion have relatively similar profiles for the different phenotypes.

It is useful to quantify the relative intensities of the features in Figure 2 for all data sets. This can be done by comparing peak intensity ratios [normalized intensities for the two higher-mobility features (at peak maxima) are summed and divided by the sum of normalized intensities for the two lower-mobility features] across the four different samples. For example, the ratio of intensities of the two higher-mobility (18.8 and 19.6 ms) features to both lower-mobility (20.9 and 21.4 ms) features is

determined to be 0.86 for the NC sample. The ratios for the EAC, BE, and HGD distributions are 0.93, 0.78 and 0.41, respectively. Therefore, for these samples, the ion mobility distributions for the $[S_1H_5N_4 + 3Na]^{3+}$ species appears to vary significantly from one phenotype to the next.

The fucosylated equivalent of the $[S_1H_5N_4 + 3Na]^{3+}$ species, the $[S_1F_1H_5N_4 + 3Na]^{3+}$ ions, also exhibits four features at 20.3, 20.8, 21.7, and 22.1 ms (Figure 2, right panel). Although less resolved than the features observed for the $[S_1H_5N_4 + 3Na]^{3+}$ ions, two sets of features are observed: high-mobility features (20.3 and 20.8 ms) and low-mobility features (21.7 and 22.1 ms). Again, the intensity ratio of the high-mobility features to the low-mobility features is determined to be 0.82 for the NC group. The ratios for the EAC, BE, and HGD distributions are 0.93, 0.80 and 0.61, respectively. Again, the relative intensities of the data set features appear to vary from one group to another. The degree to which such variations can be used to distinguish the various phenotypes is discussed below.

It is instructive to consider whether or not the raw intensities of the glycan ions are statistically different for the different data sets. Box plots of integrated drift time intensities for $[S_1H_5N_4 + 3Na]^{3+}$, $[S_1F_1H_5N_4 + 3Na]^{3+}$, $[S_3H_6N_5 + 3Na]^{3+}$ and $[S_3F_1H_6N_5 + 3Na]^{3+}$ ions from 61 NC, 56 EAC, 7 BE, and 12 HGD samples are shown in Figure 3. The mean integrated drift time intensities for the $[S_1H_5N_4 + 3Na]^{3+}$ glycan ion are 7866 ± 3192 , 5665 ± 1412 , 6678 ± 2182 and 8546 ± 3903 for NC, BE, HGD and EAC groups, respectively. As observed in Figure 3, the variability in intensities for the $[S_1H_5N_4 + 3Na]^{3+}$ glycan ion can be quite significant (5 to 10-fold) for the NC and EAC samples. The smaller standard variation (reflected in the smaller standard deviation) observed for the HGD and BE groups are most likely due to the fact that fewer samples have been analyzed compared to the NC and EAC groups (4 to 5 times fewer samples). For the EAC and NC samples of greater numbers, the mean intensities are similar; however, it is noteworthy that the lowest and highest intensities are observed within the EAC group. The observed trends are also true for the $[S_1F_1H_5N_4 + 3Na]^{3+}$, $[S_3H_6N_5 + 3Na]^{3+}$ and $[S_3F_1H_6N_5 + 3Na]^{3+}$ glycan ions.

Pattern Recognition Analysis of IMS-MS Data

The data shown in Figure 3 suggest that comparisons based on the integrated raw intensities of different glycan ions would be inconclusive. That is, the glycan ion intensities observed in a mass spectrum would not contain sufficient information to distinguish the various phenotypes. In part this may result from the relatively small number of samples used in the study. However, it is also possible that although the overall intensities of specific glycans do not vary significantly across the sample phenotypes, the intensities at specific positions across the mobility distribution would contain more information. The increased number of data set features requires the use of a more rigorous computational approach.

To test the effect of these species in distinguishing phenotype, PCA has been performed using the ion mobility distributions of the different glycan ions. Figure 4 illustrates the profile of one control sample where ion mobility distributions from 11 N-linked individual glycan ions have been sequentially spliced together. This 11-glycan composite IMS distribution demonstrates the relative intensities of the various glycan ions. The most intense feature belongs to the glycan ion $[S_2H_5N_4 + 3Na]^{3+}$ and the least intense to $[S_1H_5N_4 + 2Na]^{2+}$. The data show that some glycan ions, such as $[S_2F_1H_5N_4 + 3Na]^{3+}$ and $[S_3H_6N_5 + 3Na]^{3+}$, exist as broad mobility distributions exhibiting many peaks;

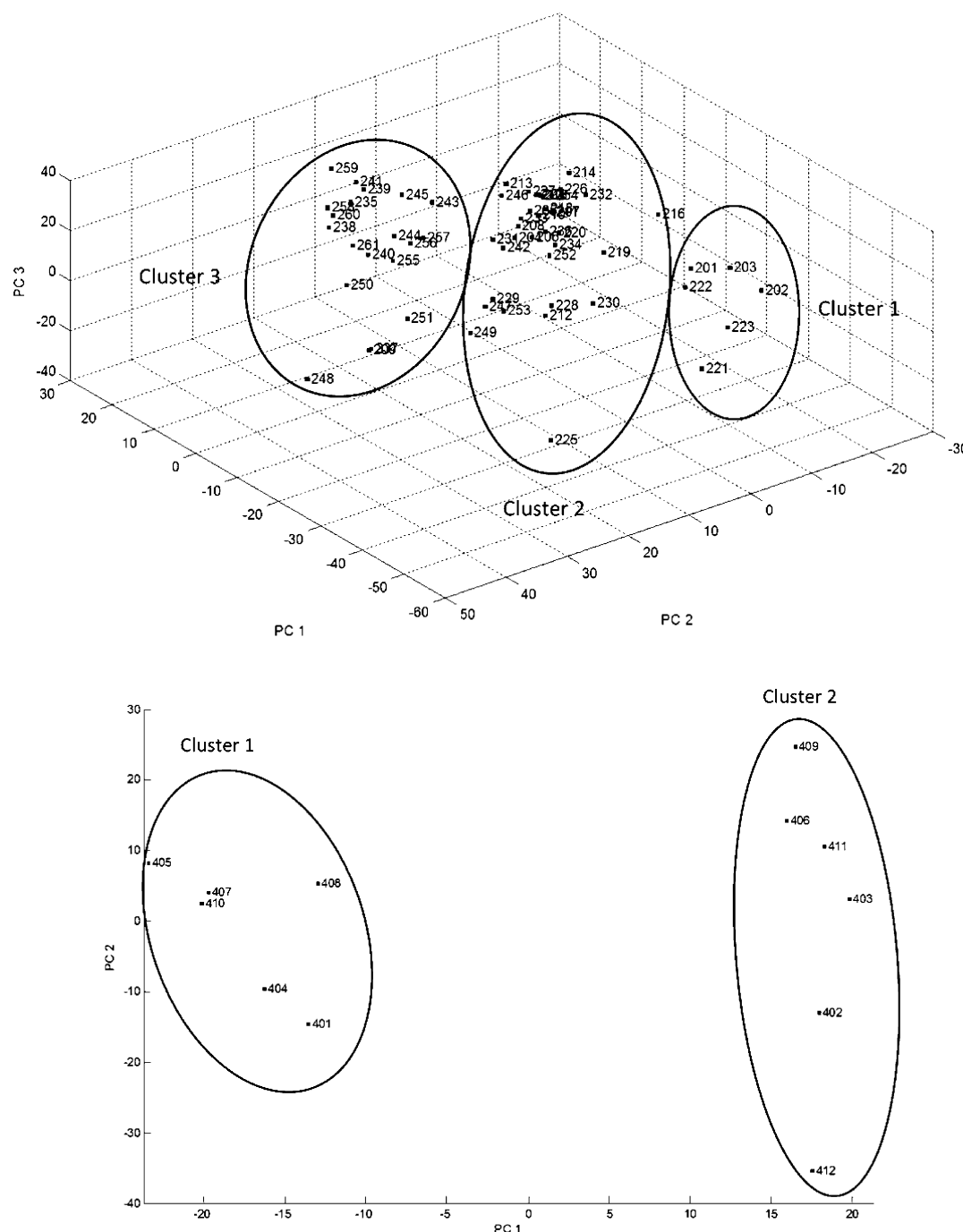


Figure 5. PC plot of the 1431-point IMS spectra of the 61 NC samples (Top) and PC plot of the 1431-point IMS spectra of the 12 HGD samples (Bottom). Each sample is represented by its id number in the PC plot of the data.

others are narrower (e.g., $[\text{S}_2\text{H}_5\text{N}_4 + 3\text{Na}]^{3+}$ or $[\text{S}_1\text{H}_5\text{N}_4 + 3\text{Na}]^{3+}$).

Because outliers have the potential to adversely affect the performance of statistical and pattern recognition methods, outlier analysis was performed on each sample class in the IMS-MS data sets. As a result, seven EAC samples were identified as outliers. Three of these samples were dirty and a visual analysis of the IMS spectra of the other four samples indicated the presence of experimental artifacts in the spectra. As these seven samples belonged to the same lot, this suggested that problems occurred when processing these samples. For the NC samples, a PC plot of the 1431-point spectra (Figure 5, top) indicated the presence of three distinct clusters in the data which we also attributed to

differences in the sample work up. Nine samples were identified as outliers from the generalized distance test and the sample leverage. HGD also showed similar clustering (Figure 5, bottom) with cluster 1 containing the outliers (6 samples) based on a visual analysis of their IMS spectra. Using the same methodology, one outlier was identified within the BE group. The 23 outliers were not taken into account for the pattern recognition analysis.

Figure 6(top) shows a plot of the two largest principal components of the 46 spectral features identified by the pattern recognition GA (see Experimental Section) to differentiate IMS-MS spectra of EAC patients from NC. The pattern recognition GA identified these 46 features by sampling key feature subsets, scoring their principal component plots, and tracking samples

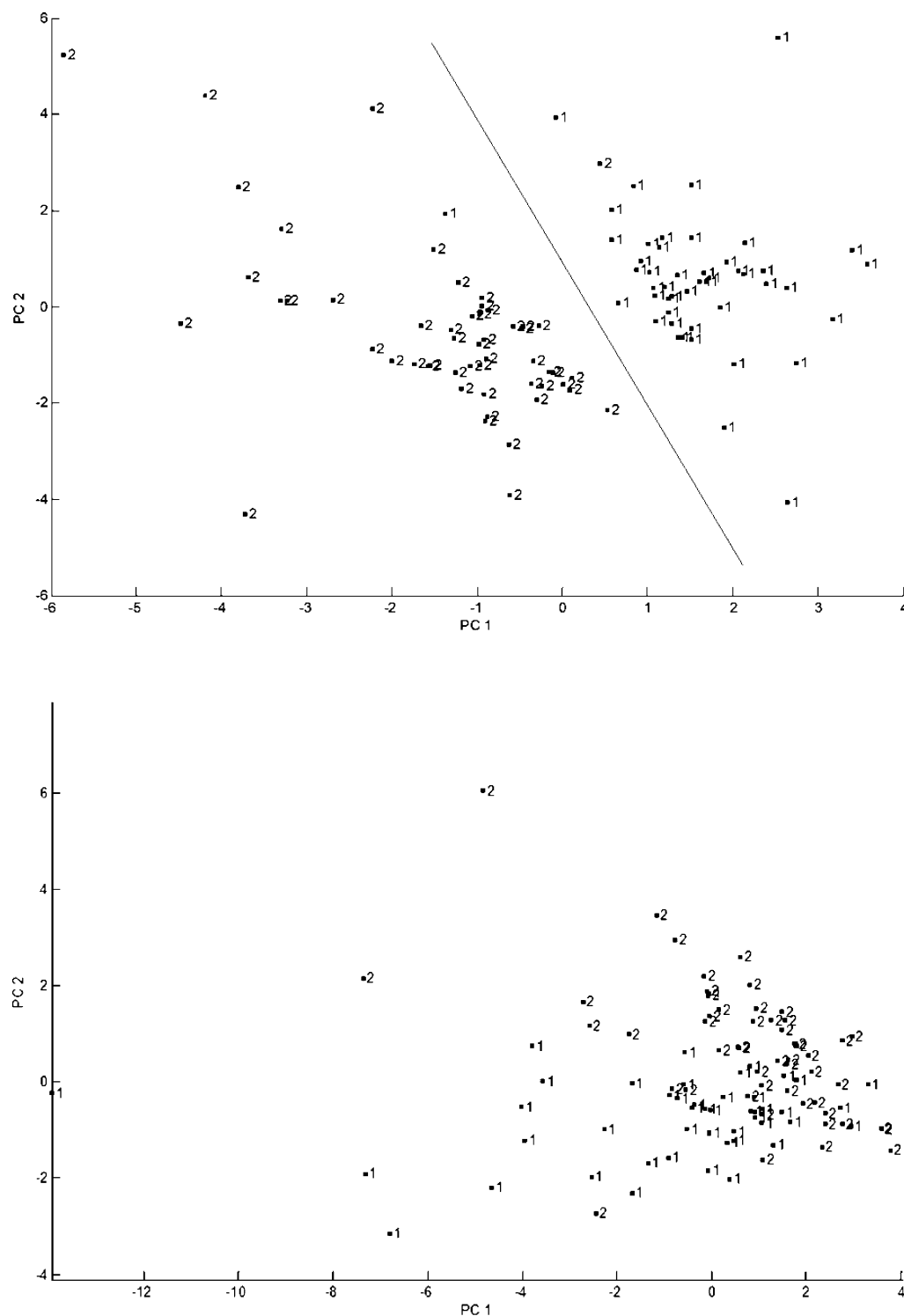


Figure 6. (Top) Plot of the two largest principal components of the 46 spectral features identified by the pattern recognition GA to differentiate IMS-MS spectra of EAC patients from NC. 1 = EAC and 2 = NC. (Bottom) Plot of the two largest principal components of the 12 spectral features identified by the pattern recognition GA to differentiate MALDI-TOF spectra of EAC patients from NC. 1 = EAC and 2 = NC.

that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. The degree of separation of EAC patients from NC individuals in the PC plot of the 46 IMS spectral features indicates separation between these two sample classes.

Additionally, the positions of the 46 spectral features within the combined mobility distribution of the 11 N-linked glycans (Figure 4) were examined in order to identify glycan ions

contributing to phenotype differentiation. Interestingly, 60% of the 46 features identified by the pattern recognition GA match only four glycan ions out of the 11 considered here. The triply charged glycan ion $S_3F_1H_6N_5$ (Figure 4, at ~1250 drift bin number) is the main contributor to group distinction, as ten features have been identified by the GA across its distribution. The glycan ions $[S_1H_5N_4 + 2Na]^{2+}$ (Figure 4, at ~1050 drift bin number), $[S_2H_6N_5 + 3Na]^{3+}$ (Figure 4, at ~750 drift bin

number) and $[S_2H_3N_4 + 2Na]^{2+}$ (Figure 4, at ~ 1390 drift bin number) also appear to play a major role in the delineation of EAC versus NC samples. Respectively, seven, six and five features across the mobility distributions of these glycan ions were selected by the GA.

One disadvantage of selecting features across the mobility distribution in this manner is that it is a point by point approach. That is, no information can be obtained about the contribution (in terms of abundance) to disease delineation of the different gas-phase conformers and (or) isomers present underneath each peak. To address this question, a future study will include the use of a set of Gaussian functions to represent the mobility distributions, where each function represents a population of isomers or gas-phase conformers. This approach was recently used to successfully probe the populations of solution state conformations of ubiquitin by IMS-MS.⁴⁸ The integrated area underneath each Gaussian function could then be used to perform PCA and the resulting separation of sample groups would be specific to single populations of ions (isomers or gas-phase conformers). To obtain the required number of Gaussian functions is a significant task. This would require examining the fits for all 11 glycans across all samples. That said, the high-throughput IMS-MS approach is uniquely suited for this type of analysis as the mobility distributions for multiple glycans from multiple samples can be generated on a very short time scale.

Despite the limitations of the PCA approach described in this study, it is worth noting the advantage of the technique. For example, when a similar study was performed using MALDI-TOF data for the same set of samples, the two classes appear indistinguishable as there was a significant degree of overlap between the same two classes in a PC plot of the spectral features selected by the pattern recognition GA to separate EAC samples from NC ones (Figure 6, bottom). For these samples, IMS-MS appears to perform better than MALDI-TOF for the delineation of disease states using the glycan content of patient sera.

SUMMARY AND CONCLUSIONS

N-linked glycans extracted from 136 serum samples for individuals that were diagnosed with Barrett's esophagus, high-grade dysplasia, and esophageal adenocarcinoma, as well as a normal control group were analyzed by a combination of IMS-MS and PCA techniques. In the data sets that were produced it was typically possible to assign ~ 20 to 30 glycan ions based on MS information. These ions often display multiple features when separated by IMS, indicating the possibility of multiple isomeric forms. In some cases, such as the $[S_1H_5N_4 + 3Na]^{3+}$ and $[S_1F_1H_5N_4 + 3Na]^{3+}$ glycan ions, multiple reproducible features which vary in intensities depending upon the phenotype of the individual are observed. As a means of understanding the ability of IMS data to differentiate between phenotypes, ion mobility profiles of 11 ions were combined for each sample of the data set. From a statistical analysis, 23 outliers were identified and removed. An important aspect of the present work is the application of an iterative genetic algorithm to the data set followed by a PCA on selected features. This approach determined that 46 features within the combined distributions were statistically capable of unequivocally distinguishing esophageal adenocarcinoma samples from those from the normal control group. Interestingly the majority of the selected 46 features correspond to four of the 11 glycan ions chosen for this analysis ($[S_3F_1H_6N_5 + 3Na]^{3+}$, $[S_1H_5N_4 + 2Na]^{2+}$, $[S_2H_6N_5 + 3Na]^{3+}$ and $[S_2H_3N_4 + 2Na]^{2+}$). Overall, the application of high-throughput IMS-MS analysis combined with PCA appears to be a

promising means of identifying species that may be used to distinguish between these phenotypes. Current efforts are underway to improve the separation efficiency of the IMS so that individual isomer contributions can be directly observed.

AUTHOR INFORMATION

Corresponding Author

*E-mail: clemmer@indiana.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the National Institute of Health (NIH-SR01GM93322-2).

REFERENCES

- (1) Isailovic, D.; Kurulugama, R. T.; Plasencia, M. D.; Stokes, S. T.; Kyselova, Z.; Goldman, R.; Mechref, Y.; Novotny, M. V.; Clemmer, D. E. Profiling of human serum glycans associated with liver cancer and cirrhosis by IMS-MS. *J. Proteome Res.* **2008**, *7*, 1109–1117.
- (2) Isailovic, D.; Plasencia, M. D.; Gaye, M. M.; Stokes, S. T.; Kurulugama, R. T.; Pungpapong, V.; Zhang, M.; Kyselova, Z.; Goldman, R.; Mechref, Y.; Novotny, M. V.; Clemmer, D. E. Delineating diseases by IMS-MS profiling of serum N-linked glycans. *J. Proteome Res.* **2012**, *11* (2), 576–585.
- (3) Phillips, W. A.; Lord, R. V.; Nancarrow, D. J.; Watson, D. I.; Whiteman, D. C. Barrett's esophagus. *J. Gastroenterol. Hepatol.* **2011**, *26* (4), 639–648.
- (4) Mechref, Y.; Hussein, A.; Bekesova, S.; Pungpapong, V.; Zhang, M.; Dobrolecki, L. E.; Hickey, R. J.; Hammond, Z. T.; Novotny, M. V. Quantitative serum glycomics of esophageal adenocarcinoma, and other esophageal disease onsets. *J. Proteome Res.* **2009**, *8*, 2656–2666.
- (5) Bereman, M. S.; Williams, T. I.; Muddiman, D. C. Development of a nanoLC LTQ Orbitrap mass spectrometric method for profiling glycans derived from plasma from healthy, benign tumor control, and epithelial ovarian cancer patients. *Anal. Chem.* **2009**, *81* (3), 1130–1136.
- (6) Williams, T.; Toups, K.; Saggese, D.; Kalli, K.; Cliby, W.; Muddiman, D. C. Epithelial ovarian cancer: Disease etiology, treatment, detection, and investigational gene, metabolite, and protein biomarkers. *J. Proteome Res.* **2007**, *6* (8), 2936–2962.
- (7) Bereman, M. S.; Young, D. D.; Deiters, A.; Muddiman, D. C. Development of a robust and high throughput method for profiling N-linked glycans derived from plasma glycoproteins by nanoLC-FTICR mass spectrometry. *J. Proteome Res.* **2009**, *8* (7), 3764–3770.
- (8) Kyselova, Z.; Mechref, Y.; Al Bataineh, M. M.; Dobrolecki, L. E.; Hickey, R. J.; Vinson, J.; Sweeney, C. J.; Novotny, M. V. Alterations in the serum glycome due to metastatic prostate cancer. *J. Proteome Res.* **2007**, *6*, 1822–1832.
- (9) Saldova, R.; Royle, L.; Radcliffe, C. M.; Abd Hamid, U. M.; Evans, R.; Arnold, J. N.; Banks, R. E.; Hutson, R.; Harvey, D. J.; Antrobus, R.; Petrescu, S. M.; Dwek, R. A.; Rudd, P. M. Ovarian cancer is associated with changes in glycosylation in both acute phase proteins and IgG. *Glycobiology* **2007**, *17* (12), 1344–1356.
- (10) Goldman, R.; Ransom, H. W.; Varghese, R. S.; Goldman, L.; Bascug, G.; Loffredo, C. A.; Abdel-Hamid, M.; Gouda, I.; Ezzat, S.; Kyselova, Z.; Mechref, Y.; Novotny, M. V. Detection of hepatocellular carcinoma using glycomic analysis. *Clin. Cancer Res.* **2009**, *15*, 1808–1813.
- (11) Arnold, J. N.; Saldova, R.; Hamid, U. M. A.; Rudd, P. M. Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation. *Proteomics* **2008**, *8*, 3284–3293.
- (12) Arnold, J. N.; Saldova, R.; Galligan, M. C.; Murphy, T. B.; Mimura-Kimura, Y.; Telford, J. E.; Godwin, A. K.; Rudd, P. M. Novel glycan biomarkers for the detection of lung cancer. *J. Proteome Res.* **2011**, *10* (4), 1755–1764.

- (13) Robbe-Masselot, C.; Herrmann, A.; Maes, E.; Carlstedt, I.; Michalski, J. C.; Capon, C. Expression of a Core 3 Disialyl-Le(x) hexasaccharide in human colorectal cancers: a potential marker of malignant transformation in colon. *J. Proteome Res.* **2009**, *8*, 702–711.
- (14) Varki, A.; Cummings, R.; Esko, J.; Freeze, H.; Hart, G.; Marth, J., Eds. *Essentials of Glycobiology*, 2nd ed.; CSHL Press: Cold Spring Harbor, NY, 2009.
- (15) Cummings, R. D.; Pierce, J. M., Eds. *Handbook of Glycomics*; Academic Press: San Diego, 2009.
- (16) Sansom, C.; Markman, O. *Glycobiology*; Scion Publishing Ltd: Oxford, 2007.
- (17) Lebrilla, C. B.; An, H. J. The prospects of glycan biomarkers for the diagnosis of diseases. *Mol. BioSyst.* **2009**, *5*, 17–20.
- (18) Zhu, M.; Bendiak, B.; Clowers, B.; Hill, H. H., Jr. Ion mobility-mass spectrometry analysis of isomeric carbohydrate precursor ions. *Anal. Bioanal. Chem.* **2009**, *394* (7), 1853–1867.
- (19) Plasencia, M. D.; Isailovic, D.; Merenbloom, S. I.; Mechref, Y.; Clemmer, D. E. Resolving and assigning N-linked glycan structural isomers from ovalbumin by IMS-MS. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (11), 1706–1715.
- (20) Williams, J. P.; Grabenauer, M.; Holland, R. J.; Carpenter, C. J.; Wormald, M. R.; Giles, K.; Harvey, D. J.; Bateman, R. H.; Scrivens, J. H.; Bowers, M. T. Characterization of simple isomeric oligosaccharides and the rapid separation of glycan mixtures by ion mobility mass spectrometry. *Int. J. Mass Spectrom.* **2010**, *298* (1–3), 119–127.
- (21) Kyselova, Z.; Mechref, Y.; Al Bataineh, M. M.; Dobrolecki, L. E.; Hickey, R. J.; Vinson, J.; Sweeney, C. J.; Novotny, M. V. Alterations in the serum glycome due to metastatic prostate cancer. *J. Proteome Res.* **2007**, *6*, 1822–1832.
- (22) Kang, P.; Mechref, Y.; et al. High-throughput solid-phase permethylation of glycans prior to mass spectrometry. *Rapid Commun. Mass Spectrom.* **2008**, *22* (5), 721–734.
- (23) Kyselova, Z.; Mechref, Y.; Kang, P.; Goetz, J. A.; Dobrolecki, L. E.; Hickey, R. J.; Malkas, L. H.; Novotny, M. V. Breast cancer diagnosis and prognosis through quantitative measurements of serum glycan profiles. *Clin. Chem.* **2008**, *54*, 1166–1175.
- (24) Kang, P.; Mechref, Y.; et al. Solid-phase permethylation of glycans for mass spectrometric analysis. *Rapid Commun. Mass Spectrom.* **2005**, *19* (23), 3421–3428.
- (25) Clemmer, D. E.; Jarrold, M. F. Ion Mobility Measurements and their Applications to Clusters and Biomolecules. *J. Mass Spectrom.* **1997**, *32*, 577–592.
- (26) Collins, D. C.; Lee, M. L. Developments in ion mobility spectrometry-mass spectrometry. *Anal. Bioanal. Chem.* **2002**, *372* (1), 66–73.
- (27) Valentine, S. J. Developing liquid chromatography ion mobility mass spectrometry techniques. *Expert Rev. Proteomics* **2005**, 553–565.
- (28) Koeniger, S. L.; Merenbloom, S. I.; Valentine, S. J.; Jarrold, M. F.; Udseth, H.; Smith, R.; Clemmer, D. E. An IMS-IMS analogue of MS-MS. *Anal. Chem.* **2006**, *78*, 4161–4174.
- (29) Hoaglund-Hyzer, C. S.; Clemmer, D. E. Ion trap/ion mobility/quadrupole/time-of-flight mass spectrometry for peptide mixture analysis. *Anal. Chem.* **2000**, *73* (2), 177–184.
- (30) Merenbloom, S. I.; Koeniger, S. L.; Bohrer, B. C.; Valentine, S. J.; Clemmer, D. E. Improving the efficiency of IMS-IMS by a combing technique. *Anal. Chem.* **2008**, *80* (6), 1918–1927.
- (31) Hoaglund, C. S.; Valentine, S. J.; Sporleder, C. R.; Reilly, J. P.; Clemmer, D. E. Three-dimensional ion mobility/TOFMS analysis of electrosprayed biomolecules. *Anal. Chem.* **1998**, *70*, 2236–2242.
- (32) Henderson, S. C.; Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. ESI/ion trap/ion mobility/time-of-flight mass spectrometry for rapid and sensitive analysis of biomolecular mixtures. *Anal. Chem.* **1999**, *71*, 291–301.
- (33) Shaffer, S. A.; Prior, D. C.; Anderson, G. A.; Udseth, H. R.; Smith, R. D. An ion funnel interface for improved ion focusing and sensitivity using electrospray ionization mass spectrometry. *Anal. Chem.* **1998**, *70*, 4111–4119.
- (34) Tang, K.; Shvartsburg, A. A.; Lee, H.; Prior, D. C.; Buschbach, M. A.; Li, F.; Tomachev, A.; Anderson, G. A.; Smith, R. D. High sensitivity ion mobility spectrometry/mass spectrometry using electrodynamic ion funnel interfaces. *Anal. Chem.* **2005**, *77*, 3330–3339.
- (35) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
- (36) Lavine, B. K.; Brown, S. D. Winning at chemometrics. *Managing Modern Lab.* **1998**, *3* (1), 9–14.
- (37) Stapanian, M. A.; Garner, F. C.; Fitzgerald, K. E.; Flatman, G. T.; J. M. Nocerino, J. Finding suspected causes of measurement error in multivariate environmental data. *Chemom.* **1993**, *7*, 165–176.
- (38) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; Wiley Series in Probability and Statistics: New York, 2003.
- (39) Schwager, S. J.; Margolin, B. H. Detection of multivariate normal outliers. *Annu. Stat.* **1982**, *10*, 943–953.
- (40) Otto, M. *Chemometrics – Statistics and Computer Application in Analytical Chemistry*; Wiley-VCH: New York, 1999; p 209.
- (41) Lavine, B. K.; Ritter, J.; Moores, A. J.; Wilson, M.; Faruque, A.; Mayfield, H. T. Genetic algorithms for deciphering the complex chemosensory code of social insects. *Anal. Chem.* **2000**, *72* (2), 423–431.
- (42) Lavine, B. K.; Davidson, C. E.; Moores, A. J.; Griffiths, P. R. Raman Spectroscopy and Genetic Algorithms for the Classification of Wood Types. *Appl. Spectrosc.* **2001**, *55* (8), 960–966.
- (43) Lavine, B. K.; Davidson, C. E.; Moores, A. J. Innovative genetic algorithms for chemoinformatics. *Chemom. Intell. Lab. Instrum.* **2002**, *60* (1), 161–171.
- (44) Lavine, B. K.; Davidson, C. E.; Breneman, C.; Katt, W. Electronic Van der Waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases. *J. Chem. Inf. Sci.* **2003**, *43*, 1890–1905.
- (45) Karasinski, J.; Andreescu, S.; Sadik, O. A.; Lavine, B.; Vora, M. N. Multisensor sensors with pattern recognition for the detection, classification, and differentiation of bacteria at subspecies and strain levels. *Anal. Chem.* **2005**, *77* (24), 7941–7949.
- (46) Eiceman, G. A.; Wang, M.; Prasad, S.; Schmidt, H.; Tadjimukhamedov, F. K.; Lavine, B. K.; Mirjankar, N. Pattern recognition analysis of differential mobility spectra with classification by chemical family. *Anal. Chim. Acta* **2006**, *579* (1), 1–10.
- (47) Kranz, C.; Ng, B. G.; Sun, L. W.; Sharma, V.; Eklund, E. A.; Miura, Y.; Ungar, D.; Lupashin, V.; Winkel, R. D.; Cipollo, J. F.; Costello, C. E.; Loh, E.; Hong, W.; Freeze, H. H. COG8 deficiency causes new congenital disorder of glycosylation type II. *Hum. Mol. Genet.* **2007**, *16*, 731–741.
- (48) Shi, H.; Pierson, N.; Valentine, S.; Clemmer, D. E. Conformation types of ubiquitin $[M + 8H]^{8+}$ ions from water: methanol solutions: evidence for the N and A states in aqueous solution. *J. Phys. Chem. B* **2012**, *116* (10), 3344–3352.
- (49) Consortium for functional glycomics and Nature publishing group, 2006 Functional Glycomics Gateway, available at www.functionalglycomics.org.