Analyst

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c6an02697d

Delineation of disease phenotypes associated with esophageal adenocarcinoma by MALDI-IMS-MS analysis of serum N-linked glycans†

M. M. Gaye, *‡[a] T. Ding,[b] H. Shion,[c] A. Hussein,[d] Y. HU,[d] S. Zhou,[d] Z. T. Hammoud,[e] B. K. Lavine,[b] Y. Mechref,[d] J. C. Gebler[c] and D. E. Clemmer*[a]

Received 20th December 2016,
Accepted 22nd March 2017

DOI: 10.1039/c6an02697d

rsc.li/analyst

N-Linked glycans, extracted from patient sera and healthy control individuals, are analyzed by Matrix-assisted laser desorption ionization (MALDI) in combination with ion mobility spectrometry (IMS), mass spectrometry (MS) and pattern recognition methods. MALDI-IMS-MS data were collected in duplicate for 58 serum samples obtained from individuals diagnosed with Barrett's esophagus (BE, 14 patients), high-grade dysplasia (HGD, 7 patients), esophageal adenocarcinoma (EAC, 20 patients) and disease-free control (NC, 17 individuals). A combined mobility distribution of 9 N-linked glycans is established for 90 MALDI-IMS-MS spectra (training set) and analyzed using a genetic algorithm for feature selection and classification. Two models for phenotype delineation are subsequently developed and as a result, the four phenotypes (BE, HGD, EAC and NC) are unequivocally differentiated. Next, the two models are tested against 26 blind measurements. Interestingly, these models allowed for the correct phenotype prediction of as many as 20 blinds. Although applied to a limited number of blind samples, this methodology appears promising as a means of discovering molecules from serum that may have capabilities as markers of disease.

Introduction

Glycosylation is the most common posttranslational modification of proteins (70% of the human proteome)[1,2] and it has now been established that a correlation exists between aberrant glycosylation and the occurrence of cancer.[2,3] MALDI-TOF-MS in combination with informatics and multivariate data analysis has been successfully used to analyze the human serum glycome and to generate glycan profiles that can delineate healthy individuals from an individual diagnosed with a given disease.[3–9] This methodology has been used to differentiate patients diagnosed with hepatocellular carcinoma or chronic liver disease from normal controls.[4] Other studies have reported specific changes in the serum glycome associated with disease phenotypes. For example, the occurrence of breast cancer has been correlated to an increase in the sialylation and fucosylation of N-linked glycans (branched carbohydrate attached on a protein to the nitrogen atom of an asparagine residue within the sequence Asn-X-Ser-Thr, where X can be any amino acid).[5] Unlike in the case of breast cancer, a decrease in fucosylation and no change in sialylation of N-linked glycans were observed when serum from patients diagnosed with esophageal adenocarcinoma (EAC) was analyzed by MALDI-TOF-MS.[6] More recent studies were focused on improving the reproducibility and high-throughput capabilities of MALDI-TOF-MS experiments. To this end, heterogeneous crystallization was limited by using matrix-prespotted plates,[7] the release of glycans from serum samples, and their subsequent permethylation was enhanced by using 96-well plate platforms;[8] and linkage-specific enzymatic release of glycans was implemented.[9] Although MALDI-TOF-MS has been successfully used to establish serum glycan profiles, structural information cannot be obtained without the use of endo- and exo-glycosidase as well as tandem mass spectrometry (MS/MS).

Because ion mobility spectrometry (IMS) allows the separation of molecules according to both their mass-to-charge ratio (m/z) and their shape, isomeric and/or conformeric separation is obtained in favorable cases. This is a considerable advantage for the analysis of glycans as a single N-linked

[a]Department of Chemistry, Indiana University, Bloomington, IN 47405, USA. E-mail: mgaye@indiana.edu, clemmer@indiana.edu
[b]Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, USA
[c]Waters Corporation, Pharmaceutical Business Operations, Milford, MA 01757, USA
[d]Department of Chemistry & Biochemistry, Texas Tech University, Lubbock, TX 79409, USA
[e]Department of Surgery, Henry Ford Hospital, Detroit, Michigan 48202, USA

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c6an02697d

‡Present address: Department of Chemistry & Chemical Biology, IUPUI, Indianapolis, IN 46202.

glycan can exist as numerous isomers due to the differences in branching patterns and isobaric monosaccharide composition.[10] In 1995, Von Helden *et al.*[11] successfully coupled a MALDI source with an IMS-MS instrument and applied this new technique to the structural characterization of polyethylene glycol polymers. Following this groundwork, MALDI-IMS-MS studies were performed on protein digests,[12–14] ganglioside structural isomers[15] and N-linked glycans released from standard glycoproteins.[14,16] When compared to MALDI-TOF-MS, MALDI-IMS-MS was reported to reduce chemical noise, increase sequence coverage (up to 70% sequence coverage for a cytochrome C tryptic digest) and allow high-throughput separation of a complex mixture.[12–14] There are many reports in the literature on the use of MALDI-TOF-MS[17–19] and MALDI-IMS-(MS/MS)[20–23] both in combination with MS imaging for *in situ* detection of cancer biomarkers (peptides, proteins and lipids). However, to the best of our knowledge, MALDI-IMS-MS has not been used for the discovery of biomarkers within the serum glycome.

We have previously demonstrated the capabilities of IMS-MS techniques used in combination with principal component analysis (PCA) for discrimination of disease phenotypes.[24–26] With this in mind, 58 serum samples from patients diagnosed with Barrett's esophagus (BE), high-grade dysplasia (HGD), EAC, and normal control (NC) individuals were analyzed in duplicate by MALDI-IMS-MS. In order to assess the contribution of sample preparation to this analysis, each MALDI spot is treated as an individual sample leading to a total of 116 MALDI-IMS-MS measurements. A composite IMS distribution of nine N-linked glycans is created and a subset of features identified by a genetic algorithm for variable selection is used to develop a classifier, which (in turn) is validated using 26 blind measurements.

Experimental section

Materials

Peptide-*N*-glycosidase F (PNGase F, EC 3.5.1.52; Sigma), ammonium bicarbonate (≥99.0% purity), sodium hydroxide beads (97% purity), methyl iodide (99% purity), dithiothreitol (DTT, ≥98% purity) and iodoacetamide were purchased from Sigma (St Louis, MO). Chloroform (99.8% purity), trifluoroacetic acid (TFA, 99% purity) were obtained from Aldrich (Milwaukee, WI). Dimethyl sulfoxide (DMSO, 99.9% purity), micro-spin columns and C18 Sep-Pak cartridges were purchased from J. T. Baker (Phillipsburg, NJ), Harvard Apparatus (Holliston, MA) and Waters (Milford, MA) respectively. Finally, β-*N*-acetylglucosaminidase (Endo-M) was obtained from TCI (Portland, OR).

Serum samples and release of N-glycans from human blood serum

Serum samples from patients with documented phenotypes (BE, HGD, and EAC) and disease-free volunteers (NC) were obtained from the Henry Ford Health clinic (Detroit, MI) with all the needed institutional review board (IRB) approvals for sample collection. Blood serum samples from disease free individuals and patients diagnosed with different esophagus diseases were randomized and treated with a mixture of Endo-M and PNGase F.[27] Briefly, 10 μL of human serum plasma was suspended in 200 μL of 100 mM ammonium bicarbonate buffer solution, to which 5 μL of 10 mM DTT was added, while the mixture was incubated at 56 °C for 45 min. After cooling, 200 μL of 55 mM iodoacetamide prepared in 100 mM ammonium bicarbonate buffer solution was added to the mixture prior to incubation at room temperature for 30 min in the dark. The pH is adjusted to be 7.5 to be optimum for both enzymes using 100 mM phosphate buffer. The N-glycans were then enzymatically released from the tryptically digested samples using PNGase F, isolated from *Escherichia coli* expressing the gene for PNGase F from *Chryseobacterium (Flavobacterium) meningosepticum*, and Endo-M. A 5 mU aliquot of both enzymes was added, and the reaction mixture was then incubated overnight (18–22 h) at 37 °C.

Purification and permethylation of released N-glycans

As described previously, the enzymatically released oligosaccharides were preconcentrated by applying the digestion mixture to C18 Sep-Pak cartridges preconditioned with ethanol and deionized water.[28,29] The collected eluent containing released N-glycans was further purified by passing through a home-packed activated carbon microspin column. Prior to the application of samples, the columns were preconditioned with acetonitrile and equilibrated with 0.1% TFA aqueous solution. After applying an aliquot of the diluted sample, the activated carbon microspin columns were washed with 0.1% TFA aqueous solution, while the glycans were eluted with 50% acetonitrile, and 0.1% TFA aqueous solution. The samples were dried under vacuum and permethylated using a previously published procedure.[28] All N-linked glycans were permethylated with microspin columns packed with sodium hydroxide beads. To protect the packing material from moisture, sodium hydroxide beads were immediately suspended in acetonitrile. The packing was accomplished pneumatically, while the columns were conditioned with DMSO prior to use. Typically, a sample was resuspended in 90 μL of DMSO, before the addition of 2.7 μL of water and 33.6 μL of methyl iodide. Next, samples were infused through the sodium hydroxide-packed column by centrifugation at 2000 rpm for two minutes and collected into microtubes. Lastly, permethylated samples were extracted with chloroform and washed several times with water before drying under vacuum.

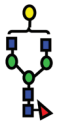
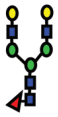
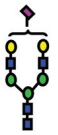
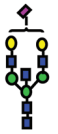
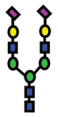
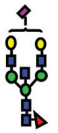
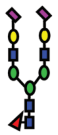
MALDI-IMS-MS measurement and data extraction

Data are acquired using a Waters Synapt G2-S (Waters Corporation, Manchester, UK) travelling wave ion mobility mass spectrometer (TWIMS) coupled with a MALDI source and operated in positive mode. Each sample was dissolved in 2 μL water : methanol (1 : 1, v : v) and mixed with 2 μL of the MALDI matrix (2,5-dihydroxybenzoic acid) prepared at 10 mg mL^{-1} in water : methanol (1 : 1, v : v) with 2 mM sodium acetate.

Sample/matrix mixtures were spotted in duplicate (2 μL each, one immediately following the other) on two 96-well MALDI plates (referred to as plate 1 and plate 2) and dextran was spotted after every ten samples as a control. Samples from different disease groups were randomized prior to spotting on the MALDI plates. The laser used was a frequency-tripled Nd:YAG laser (355 nm), firing at a rate of 1000 Hz at the energy lever of 450 (a laser energy level of 500 corresponding to approximately 100 μJ) and in a reverse-spiral pattern, as carbohydrates are known to be preferentially localized at the edges of a MALDI spot. The trap collision cell voltage was set to 6 eV in MS mode, while the transfer voltage was kept at 4 eV. A peak height voltage of 40 V and a T-wave velocity of 350 m s^{-1} were applied to the mobility cell. An external calibration was performed using a MassPREP™ calibration mix containing polyethylene glycol (Waters Corporation, Milford, MA) and mass spectra were acquired from 1000 to 5000 *m*/*z* for three minutes. All data are recorded within a 24-hour window.

For each sample, data included in a diagonal selection across the drift bin (*m*/*z*) two-dimensional spectrum (2D-plot) containing the N-linked glycans were extracted using Driftscope software (Waters Corporation, Manchester, UK). A box selection was performed across a specific drift bin and *m*/*z* range corresponding to a single N-linked glycan ion $[M + Na]^+$. Although 12 sodiated N-linked glycans up to 3775 *m*/*z* were detected, nine glycan ions until 3000 *m*/*z* had sufficient S/N across all samples and are further examined in this work. The box selection was repeated for these nine glycan ions (see Table 1). Data in the box selection were exported to MassLynx software (Waters Corporation, Manchester, UK) with the *retain drift time* function enabled in order to obtain mobility distributions (also referred to as arrival time distributions) for the glycan ions. As described in our previous studies,[24–26] A composite IMS distribution is obtained by sequentially splicing together the mobility distributions of the selected N-linked glycans across an arbitrary drift bin axis. Individual mobility distribution intensities are scaled uniformly prior to being sequentially spliced together. These 9-glycan composite ion mobility distributions were obtained for each of the 116 MALDI-IMS-MS measurements and were used as an input for pattern recognition analysis of the data.

Table 1 N-Linked glycans used for the statistical evaluation of the dataset

Glycan composition[a]	*m*/*z*[b] $[M + Na]^+$	Structure
$F_1H_3N_4$	1835.9	
$F_1H_4N_4$	2040.0	
$F_1H_5N_4$	2244.1	
$S_1H_5N_4$	2431.2	
$S_1F_1H_5N_4$	2605.3	
$S_1H_5N_5$	2676.3	
$S_2H_5N_4$	2792.4	
$S_1F_1H_5N_5$	2850.4	
$S_2F_1H_5N_4$	2966.5	

[a] F represents fucose (red triangle), H represents hexose (mannose green circle, galactose yellow circle), N represents *N*-acetyl glucosamine (blue square) and S represents sialic acid (purple diamond). [b] Permethylated glycans with free reducing end.

Multivariate analysis of the dataset

Potential N-linked glycan markers for EAC and associated phenotypes were identified utilizing a pattern recognition approach based on identifying the smallest subset of features within the dataset that optimize the separation of the sample classes in a plot of the two or three largest principal components (PCs) of the data. Because PCs maximize variance, the bulk of the information encoded by these variables is about differences between the classes in the data set. This approach to variable selection avoids overly complicated solutions that do not perform as well on a prediction set because of overfitting, which is a serious problem with many wrapper based variable selection methods.[30] Although filters[31] that select variables by ranking them are preferred by many workers because of their computational and statistical scalability, variables that are selected by filters are usually not optimal for a prediction because they score variables individually and are independent of each other. As such filters cannot determine feature combinations that give the best classification results.

A genetic algorithm (GA) is employed in this study to implement this approach to feature selection. PCA which is incorporated into the fitness function of the pattern recognition GA serves as an information filter significantly reducing the size of the search space as it restricts the search to variables whose PC plots show clustering on the basis of their sample class membership. To evaluate and compare different chromosomes (*i.e.*, variable subsets), a fitness function called PCKaNN[32,33] that quantifies the fitness of the different variable subsets was formulated utilizing both PCA and K-NN (K-nearest neighbor) to score each variable subset (*i.e.*, chromosome) in a population of potential solutions. For each chromosome in the population, PCA is applied to assess the information content of variable subsets by projecting the data onto a plot defined by the two or three largest PCs of the data. K-NN is then used to characterize the degree of class separation achieved by the variables in the PC plot. During each generation, class and sample weights are computed as shown by eqn (1) and (2) respectively, where CW(c) is the weight of class c, and SW(s) is the weight of sample s in class c. The sum of the sample weights for spectra assigned to a particular class is equal to the class weight, and the sum of all class weights in the data set is equal to 100.

$$\mathrm{CW(c)} = 100\frac{\mathrm{CW(c)}}{\sum_{\mathrm{c}}\mathrm{CW(c)}} \tag{1}$$

$$\mathrm{SW(s)} = \mathrm{CW(c)}\frac{\mathrm{SW(s)}}{\sum_{\mathrm{s}\in\mathrm{c}}\mathrm{SW(s)}} \tag{2}$$

For a given data point in the PC plot (*i.e.*, sample), Euclidean distances are computed between this point and every other point in the PC plot. These distances are arranged from the smallest to the largest. A poll is then taken of the point's K_c-nearest neighbors. (K_c is set by the user, and for the most rigorous classification of the data, K_c equals the number of samples in the class to which the point belongs.) The number of K_c-nearest neighbors with the same class label as the sample point in question, called the sample hit count (SHC), is computed ($0 \leq \mathrm{SHC(s)} \leq K_c$). It is then a simple matter to score a principal component plot (see eqn (3)).

$$F(d) = \sum_{\mathrm{c}}\sum_{\mathrm{s}\in\mathrm{c}}\frac{1}{K_c} \times \mathrm{SHC(s)} \times \mathrm{SW(s)} \tag{3}$$

To better understand the scoring of the PC plots, consider a data set comprised of two classes, with each assigned equal class weights. One class has 10 samples, and the other has 20 samples. At generation 0, all classes will have the same class weight and all samples in a given class will have the same sample weight. Thus, each sample in class 1 (10 samples) has a sample weight of 5, whereas each sample in class 2 (20 samples) has a weight of 2.5. Suppose a sample from class 1 has 7 samples from class 1 as its nearest neighbors. For this sample, $\mathrm{SHC}/K = 0.7$, and $(\mathrm{SHC}/K) \times \mathrm{SW} = 0.7 \times 5$, which equals 3.5. By summing $(\mathrm{SHC}/K_c) \times \mathrm{SW}$ for all samples, each PC plot is scored. A PC plot with a higher score indicates greater separation among the classes in the variable subset from which the plot was generated.

PCKaNN is able to focus on those samples (*i.e.* spectra) and classes (*i.e.* disease state) that are difficult to classify by boosting their sample and class weights over successive generations. In order to boost these factors, it is necessary to compute both the sample-hit rate (SHR), which is the mean value of SHC/K_c for all feature subsets produced in a particular generation (see eqn (4)), and the class-hit rate (CHR), which is the mean sample hit rate of all samples in a class (see eqn (5)). The variable ϕ, in eqn (4), is the number of chromosomes in the population, whereas ∀ and AVG in eqn (5) refer to all samples in the class and the average or mean value. During each generation, class and sample weights are adjusted using a perceptron (see eqn (6) and (7)) with the momentum, P, set by the user. ($g + 1$ is the current generation, whereas g is the previous generation.) Classes with a lower class hit rate are boosted more heavily than classes that score well.

$$\mathrm{SHR(s)} = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{\mathrm{SHC}_i\mathrm{(s)}}{K_c} \tag{4}$$

$$\mathrm{CHR}_g\mathrm{(c)} = \mathrm{AVG}(\mathrm{SHR}_g\mathrm{(s)}:\ \forall_{\mathrm{s}\in\mathrm{c}}) \tag{5}$$

$$\mathrm{CW}_{g+1}\mathrm{(s)} = \mathrm{CW}_g\mathrm{(s)} + \mathrm{P}(1 - \mathrm{CHR}_g\mathrm{(s)}) \tag{6}$$

$$\mathrm{SW}_{g+1}\mathrm{(s)} = \mathrm{SW}_g\mathrm{(s)} + \mathrm{P}(1 - \mathrm{SHR}_g\mathrm{(s)}) \tag{7}$$

Boosting is crucial for the successful operation of the pattern recognition GA using PCKaNN as its fitness function because it modifies the fitness landscape by adjusting the values of the class and sample weights which are an integral part of the fitness function. This helps to minimize the problem of convergence to a local optimum because the fitness function of the pattern recognition GA is changing as the population is evolving towards a solution. Boosting obviates the potential problem of a deceptive fitness landscape.

In this study a second fitness function was also utilized to select variables.[34] This function employs the Hopkins statistic to assess sample clustering.[35–37] By coupling the Hopkins statistic to PCKaNN, features are selected to optimize clustering in the PC plot using all of the data points (both the training set and the blind samples *via* the Hopkins statistic) while simultaneously seeking to identify features that create class separation using only the labeled data points (training set samples *via* PCKaNN). The advantage of using this compound fitness function to select variables is that transductive learning is used not only to predict future data, but also to identify truly informative features in the data set, thereby ensuring a reliable classification of the data. By varying the contribution of PCKaNN and the Hopkins statistic to the scoring of the chromosomes (*i.e.*, feature subsets) obtained during each generation, it is possible to tune the fitness function of the pattern recognition GA, enabling it to explore the structure of a large data set and to uncover hidden relationships in the

data. Using this approach, variable selection, classification, and prediction can be performed in a single step.

Results and discussion

Mobility distributions associated with different phenotypes

For each one of the 116 MALDI-IMS-MS spectra, nine glycan ions (depicted in Table 1) are chosen and combined into a single mobility profile along an arbitrary drift bin axis. Unlike our previous studies,[24–26] serum samples were processed using a mixture of two enzymes: Endo-M cleaves glycans after the *N*-acetylglucosamine residue attaches to Asn on the glycoprotein; and is inactive in the presence of core fucosylation and highly branched glycans; PNGase F cleaves glycans after the Asn residue on the glycoprotein.[27] As a result, a 9-glycan composite IMS distribution dominated by fucosylated species (6 out of the 9 ions depicted in Table 1) and including three N-linked glycans not observed in our previous IMS experiments is obtained. Six of the N-linked glycans represented in Table 1 have been characterized in our previous studies[24–26] as doubly- and/or triply-charged sodiated ions ($F_1H_5N_4$, $S_1H_5N_4$, $S_1F_1H_5N_4$, $S_1H_5N_5$, $S_2H_5N_4$ and $S_2F_1H_5N_4$); three of the nine selected glycan ions are characterized for the first time by IMS ($F_1H_3N_4$, $F_1H_4N_4$ and $S_1F_1H_5N_5$). We have shown previously that variations in mobility profiles can be correlated to disease phenotypes.[24–26] With this in mind, because variations in sialylation and fucosylation of N-linked glycans have been associated with the occurrence of cancer,[2] the mobility profiles of $S_1H_5N_4$, $F_1H_5N_4$ and $S_1F_1H_5N_4$ (Fig. 1, left, middle and right

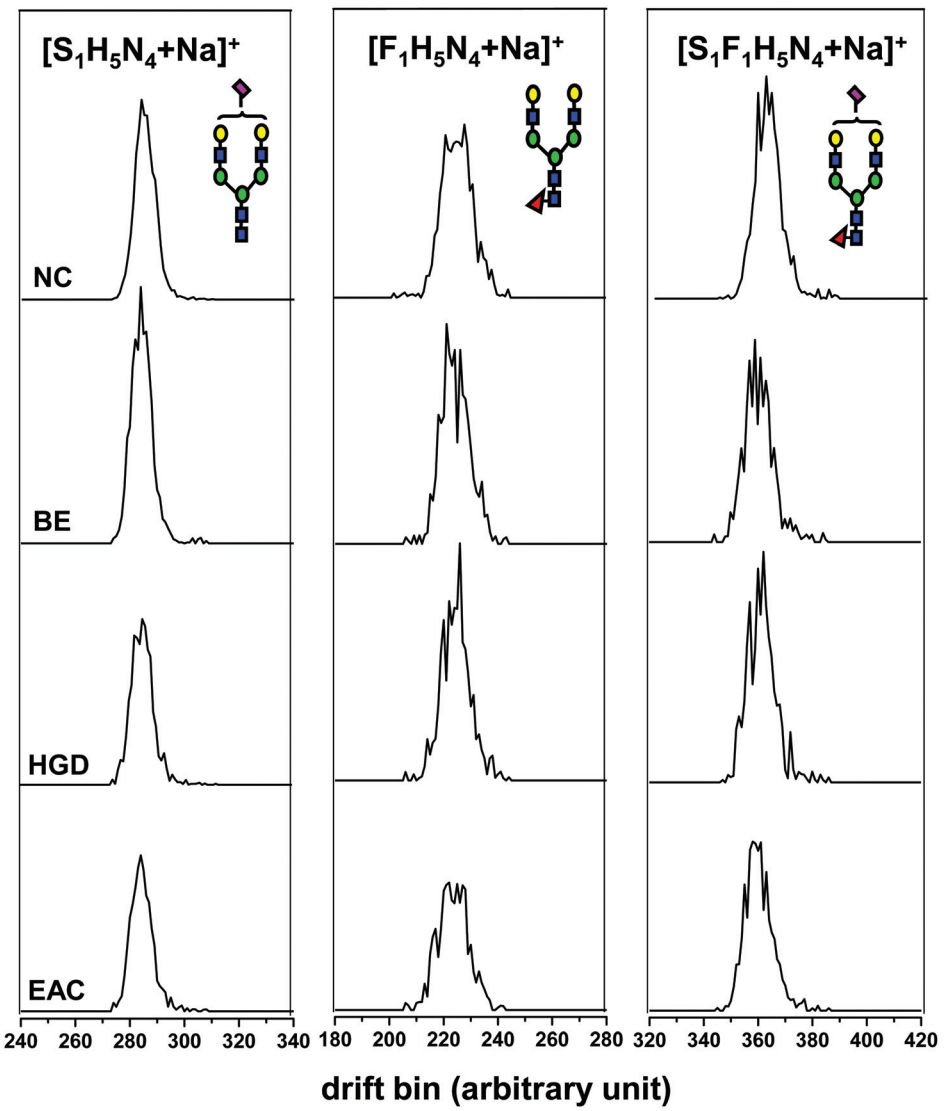


Fig. 1 Ion mobility distribution of the N-linked glycan ions $[S_1H_5N_4 + Na]^+$, $[F_1H_5N_4 + Na]^+$ and $[S_1F_1H_5N_4 + Na]^+$. Esophageal adenocarcinoma (EAC), high grade dysplasia (HGD), Barrett's esophagus (BE) and normal control (NC) phenotypes are represented by a single individual each. Glycan structures are shown as insets: F represents fucose (red triangle), H represents hexose (mannose green circle, galactose yellow circle), N represents *N*-acetylglucosamine (blue square) and S represents sialic acid (purple diamond).

panel respectively) are scrutinized. For each glycan ion, the four phenotypes NC, BE, HGD and EAC are represented by a single individual (Fig. 1) chosen in the center of the group clusters obtained after pattern recognition analysis (Fig. 4). Mobility distributions for three additional individuals are represented in Fig. S-1† (glycan ion $[S_1H_5N_4 + Na]^+$), Fig. S-2† (glycan ion $[F_1H_5N_4 + Na]^+$) and Fig. S-3† (glycan ion $[S_1F_1H_5N_4 + Na]^+$).

Mobility distributions of the glycan ion $[S_1H_5N_4 + Na]^+$ (Fig. 1 and S-1†) for all phenotypes are dominated by a single feature at 284 drift bin number (dbn). Some individual profiles for NC and BE phenotypes display a peak shoulder at 286 dbn with an additional unresolved feature (282 dbn) observed for the BE phenotype only. The mobility distribution associated with the HGD phenotype shows two additional features: a peak shoulder (277 dbn) and a more elongated unresolved structure (293 dbn). Interestingly enough, unlike in the case of NC and BE phenotypes, the most abundant peak lies at 285 dbn with an unresolved feature at 283 dbn. With the occurrence of EAC, the same peak as in the mobility distributions of NC and BE phenotypes is observed (284 dbn) and in some cases the peak shoulder at 286 dbn is visible but the unresolved structure at 282 dbn and the peak shoulder is not present. Additionally, the unresolved structure at 293 dbn in the HGD mobility profile appears as a peak shoulder in the EAC distribution. Although upon visual inspection these differences are not dramatic, these variations in the mobility profiles for different disease phenotypes are the elements which are taken into account by the GA for pattern recognition and compared across all MALDI-IMS-MS measurements.

Mobility distributions for $[F_1H_5N_4 + Na]^+$ and $[S_1F_1H_5N_4 + Na]^+$ glycan ions (Fig. 1) display more features than for $[S_1H_5N_4 + Na]^+$, which could increase the probability of finding elements capable of delineating phenotypes within $F_1H_5N_4$ and $S_1F_1H_5N_4$ distributions. The mobility distribution for $[F_1H_5N_4 + Na]^+$ (Fig. 1, Fig. S-2 and enlarged view of Fig. S-7†) for the NC phenotype is a broad peak with three unresolved features at drift bin numbers of 221, 225 and 228 respectively and a minor more elongated feature at 234 dbn. Interestingly enough, for some individuals within the BE group, features at 221 dbn and 225 dbn are partially resolved with a peak shoulder at 228 dbn. In the case of the HGD phenotype, the mobility distribution is dominated by the feature at 226 dbn, which is partially resolved from the more compact, lower intensity structure at 221 dbn. In a similar manner to the NC mobility distribution, the EAC phenotype displays a broad distribution with three unresolved features at 222, 227 and 228 dbn respectively but with an additional partially resolved peak at 217 dbn. Overall, variations in the mobility profiles across the disease phenotypes of the glycan ion $[F_1H_5N_4 + Na]^+$ are greater than for its sialylated counterpart $[S_1H_5N_4 + Na]^+$.

Lastly, mobility distributions for $[S_1F_1H_5N_4 + Na]^+$ are also depicted in Fig. 1 and S-3.† The NC phenotype displays two main, partially resolved features (358 and 361 dbn); the more elongated feature is separated into an additional unresolved feature (363 dbn). Looking at the BE phenotype, in some cases, the two main features are comprised of an additional unresolved feature (357 and 359 dbn for the compact feature, 361 and 363 dbn for the more elongated one); and additionally, peak shoulders are observed at 354 and 366 dbn. At the HGD stage, the mobility distribution for $[S_1F_1H_5N_4 + Na]^+$ is dominated by the more elongated structure (unresolved peaks at 360 and 362 dbn) and the feature corresponding to a drift bin number of 357 is a single peak as in the case of the NC phenotype. Similar to the mobility distribution for BE phenotype, shoulders at 353 and 367 dbn are present but in different proportions. Interestingly enough, an elongated feature (372 dbn) also displayed for NC and BE phenotypes as a minor feature is for the HGD phenotype a partially resolved structure ~1/4 of the main peak abundance. Finally, the mobility distribution associated with EAC phenotype for $[S_1F_1H_5N_4 + Na]^+$ displays a main broad structure between 358 and 361 dbn and in some cases two peak shoulders at higher (355 dbn) and lower mobilities (363 dbn).

It is interesting to note that the mobility distributions for the sialylated and fucosylated N-linked glycan ion $[S_1F_1H_5N_4 + Na]^+$ display more variations across disease phenotypes than the fucosylated or sialylated related ions. These observations are the result of a visual examination of mobility profiles from a single individual in each phenotype group. Because of the number of samples and features in the dataset, a more systematic analysis of the data using a GA for pattern recognition analysis has been implemented to correlate N-linked glycan mobility distributions to disease phenotypes.

PCA of all spectral features

For each one of the 90 spectra (MALDI plates 1 and 2) of the training set, the 9-glycan composite mobility distribution contained 1791 drift bins. Because many of the 1791 initial drift bins are zeros (present before and after each individual glycan) or have similar intensities, only 404 drift bins are considered for pattern recognition analysis. The two largest PCs of the 404 spectral features are depicted in Fig. 2A. Each training set measurement is represented as a point in the PC plot. Data points for the four phenotypes (28 NC, 20 BE, 10 HGD and 32 EAC measurements) lie between −15 and +15 along the PC1 axis; and between −15 and +10 along the PC2 axis. EAC samples are partially resolved from the other three phenotypes but more noticeably, three outliers are present (two NC and one HGD). A visual examination of the data reveals that these three outliers are represented by spectra that are markedly different from the other spectra in the training set. For this reason, the three outliers are removed and PCA is again performed on the truncated training set. A plot of the two largest PCs of this analysis is shown in Fig. 2B. The training set can be divided into two groups. A dashed line at +5 along the PC2 axis and parallel to the PC1 axis delineating the separation of spectra in the training set is shown in Fig. 2B. There were no variations in instrumental parameters within one MALDI plate or between the two plates as the instrument used was continually tuned using an external standard. Examining the origin of each measurement, we observed that all samples above the

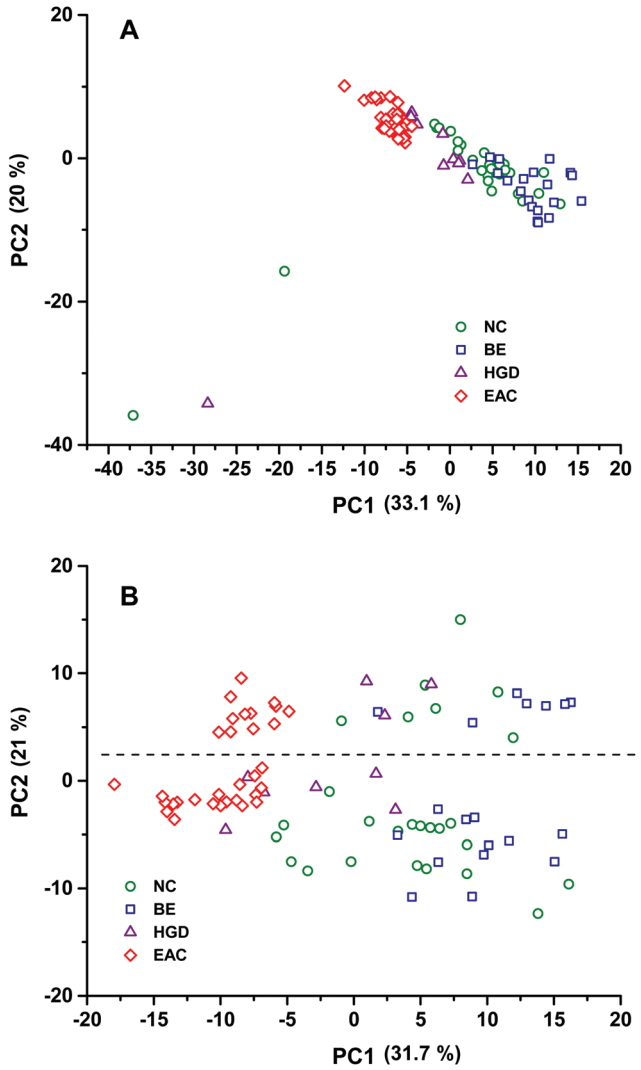


Fig. 2 (A) Representation of the two largest principal components (PC) for the 45 samples of the training set (90 measurements) obtained by PCA of all spectral features. Normal control (NC, 28 measurements, green circle), Barrett's esophagus (BE, 20 measurements, blue square), high grade dysplasia (HGD, 10 measurements, purple triangle) and esophageal adenocarcinoma (EAC, 32 measurements, red diamond) phenotypes are represented. (B) Plot of the two largest PC for all spectral features after removal of the three outlier measurements depicted in the PC plot (A) (2 NC and 1 HGD). A dashed line parallel to PC1 axis showing a separation in the dataset is depicted.

dashed line (Fig. 2B) are from the first MALDI spot for each sample within MALDI plate 1 (duplicates were spotted back to back on the MALDI plate, the first spot labeled a and the second b). All samples from the second spot on plate 1 and both spots on plate 2 lie below the dashed line in Fig. 2B. The observed clustering in the PC plot is probably due to the quality of the sample spotting technique, which improved during the course of the experiment. Following this line of investigation, the first set of spectra from plate 1 (denoted as plate 1a samples) and the second set of spectra from plate 1 combined with both sets of spectra from plate 2 (denoted as plate 1b, plate 2a and 2b samples) were analyzed separately using PCA. Separation of the different phenotype groups is observed in both PC plots (figures not shown). Although sample spotting is a source of variation in the data, differences between phenotypes are a larger source of variation. Because of this, data from all plates were analyzed by the GA for pattern recognition. This analysis, which allowed an evaluation of the training set with respect to a possible bias introduced by sample preparation, is valuable as it helps ensure that observed differences are due to phenotype rather than due to experimental conditions.

PCA of features selected by the genetic algorithm for pattern recognition

A pattern recognition GA is used to identify informative features in the MALDI-IMS-MS dataset correlated to the physiological state of the patient by sampling key feature subsets (chromosomes) and scoring their PC plots. In the present work, the capability of two different sets of chromosomes for delineation between NC, BE, HGD and EAC phenotypes is assessed. For each feature subset, the two largest PCs for the best set of chromosomes (obtained after 200 generations) are depicted in Fig. 3 (model 1) and in Fig. 4 (model 2) respectively. The first model is based on 38 features identified by the GA after the removal of seven samples (outliers) from the training set (Fig. 3A). Remarkably, the four phenotypes are delineated. This is an improvement from our previous study where only NC and EAC phenotypes were distinguished.[26] The spectra for NC and EAC phenotypes constitute tighter clusters than for BE and HGD phenotypes (Fig. 3A). A possible explanation is that NC and EAC are the two extreme phenotypes of this training set for whom the clinical diagnosis is less subject to errors than for BE or HGD. Two spectra from the NC group (at ~−3 along PC1 and at ~1 along PC2, Fig. 3A) are closer to the BE cluster than to the other NC samples. This could be an indication of a NC subject not yet identified as BE or of a NC subject with a glycan profile different from the other individuals within this phenotype. In Fig. 3B, 26 blind measurements are projected onto the PC plot comprising the first model. Almost all of the unknown samples fall into a phenotype cluster but an ambiguity remains for three samples (Fig. 3B). For the second model, 24 features were selected by the GA after the removal of eight outliers from the training set (Fig. 4A). Seven of the eight samples are the same as those removed from the training set for model 1. Again, the four phenotypes are unequivocally distinguished and in addition, the clusters are tighter (only one sample lies outside of a cluster) than those from model 1. The observation of tighter phenotype clusters suggests greater prediction power for the second model. This is the case when examining the projection of the 26 blind measurements with the PC plot (Fig. 4B) comprising model 2. Unlike model 1, all blind samples fall within a given phenotype. Glycans contributing to phenotype differentiation are identified by examining the position along the arbitrary drift bin axis of the features selected by the pattern recognition GA. 68% of the features for the first model and 44% of

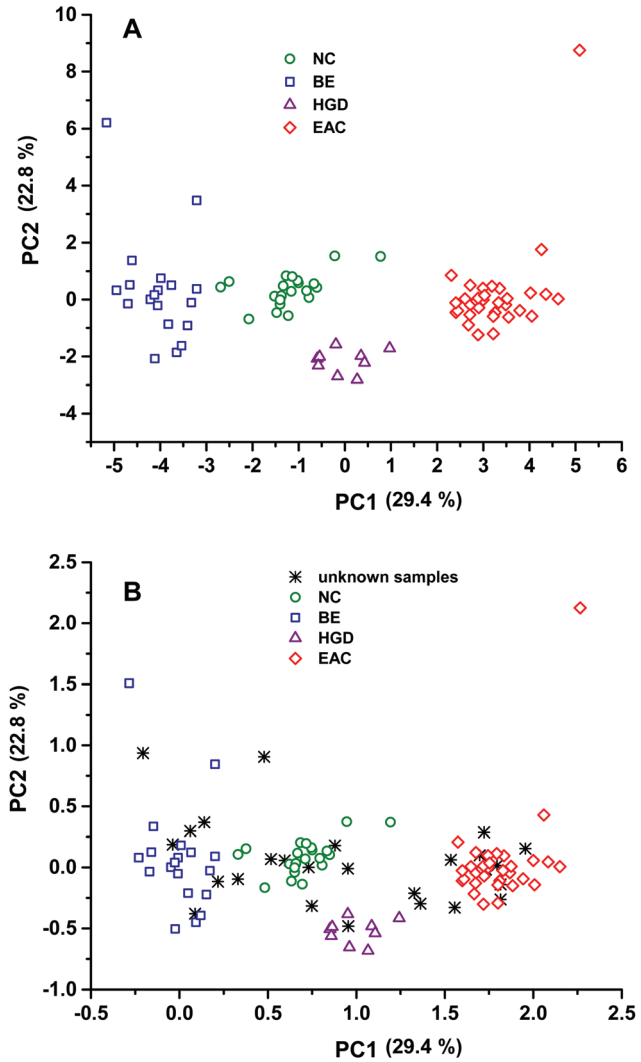


Fig. 3 (A) PC plot of the 38 features selected after 200 generations of the genetic algorithm for pattern recognition and removal of 7 outliers for 83 measurements (chromosome 5000, model 1). All four phenotypes are illustrated: NC (23 measurements), BE (19 measurements), HGD (9 measurements) and EAC (32 measurements). (B) Representation of the two largest PC for model 1 overlaid with a PC plot of the 26 blind measurements (black star).

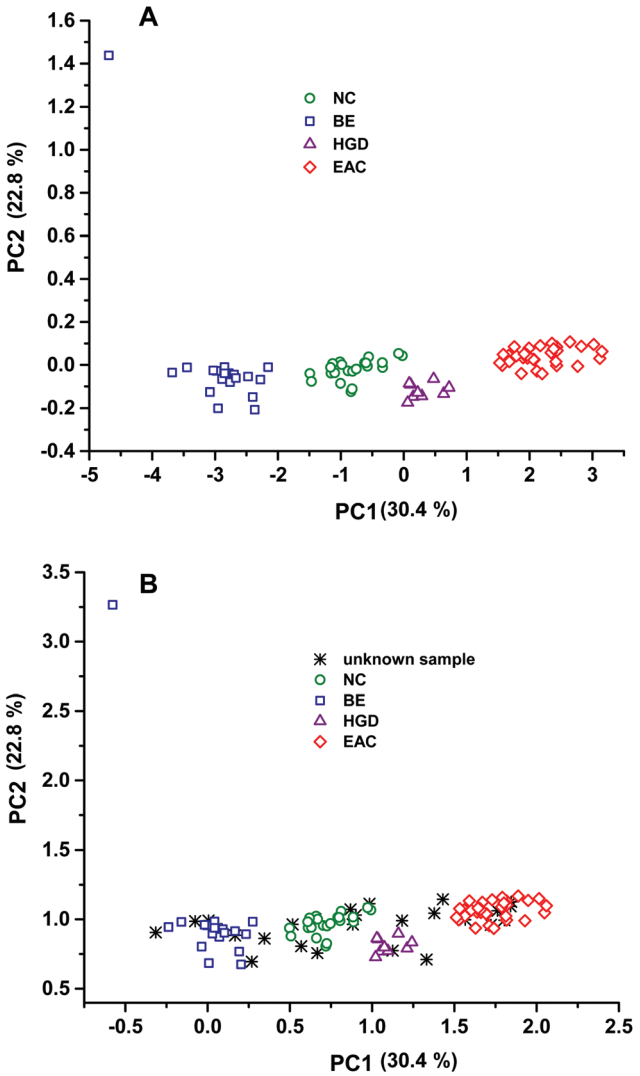


Fig. 4 (A) Plot of the two largest PC of the 24 features obtained after 200 generations of the genetic algorithm for pattern recognition and removal of 8 outliers for 82 measurements (chromosome 10 000, model 2). (B) PC plot for model 2 overlaid with the PC plot of the 26 blind measurements (black star). NC (22 measurements), BE (19 measurements), HGD (9 measurements) and EAC (32 measurements) phenotypes are depicted.

the features for the second model are localized on the mobility distribution of four and five glycans respectively. $S_1F_1H_5N_4$, $F_1H_4N_4$ and $F_1H_5N_4$ N-linked glycans mainly contribute to the delineation between phenotypes for both proposed models, while $S_1H_5N_5$ is specific to the first model; $S_1H_5N_4$ and $S_2H_5N_4$ are specific to the second model. It is interesting to note that $S_1H_5N_4$ and $S_2H_5N_4$ were also identified as the main contributors in phenotype differentiation in our previous study.[26]

Evaluation of the phenotype prediction

After a visual inspection of the PC plots obtained for model 1 (Fig. 3) and model 2 (Fig. 4), the second model appears more fit for delineating between disease states. The prediction made for each one of the 26 blinds are shown in Table 2 and summarized in a bar diagram (Fig. 5). Among the 26 blinds, 17 phenotypes are correctly predicted by model 1 and 20 by model 2. Using the first model, the prediction for 3 blinds led to an ambiguity. That is, in two cases the model cannot predict between HGD and EAC, the correct phenotype being HGD; and in one case the ambiguity is between NC and HGD, the correct phenotype being NC (Table 2). HGD and EAC are both disease phenotypes whereas an ambiguity between NC and HGD has more serious consequences in terms of diagnosis. Nevertheless, in both cases, the correct phenotype is based on the prediction. Interestingly enough, the second model did not yield any ambiguity. Both models led to the same number and same identification of false positives and false negatives (2 and 4 respectively, Table 2 and Fig. 5). Within the false

Table 2 Phenotype prediction based on the analysis of 9-glycan composite ion mobility distributions

Blind measurement ID	Phenotype	Prediction using	
		Model 1	Model 2
U_1	BE	BE	BE
U_2	BE	NC	NC
U_3$^{\alpha}$	BE	BE	BE
U_4$^{\beta}$	EAC	EAC	EAC
U_5$^{\gamma}$	EAC	EAC	EAC
U_6	EAC	EAC	EAC
U_7	HGD	HGD or EAC	HGD
U_8$^{\delta}$	HGD	HGD or EAC	HGD
U_9	NC	HGD	HGD
U_10	NC	NC	NC
U_11	NC	EAC	EAC
U_12$^{\gamma}$	EAC	EAC	EAC
U_13	BE	BE	BE
U_14	BE	BE	BE
U_15	EAC	EAC	EAC
U_16	HGD	NC	NC
U_17	NC	NC	NC
U_18$^{\beta}$	EAC	EAC	EAC
U_19$^{\delta}$	HGD	NC	NC
U_20$^{\alpha}$	BE	BE	BE
U_21$^{\varepsilon}$	BE	NC	NC
U_22$^{\zeta}$	EAC	EAC	EAC
U_23$^{\eta}$	NC	NC or HGD	NC
U_24$^{\varepsilon}$	BE	BE	BE
U_25$^{\zeta}$	EAC	EAC	EAC
U_26$^{\eta}$	NC	NC	NC

$^{\alpha-\eta}$Indicates duplicate samples (two separate MALDI spots, IMS-MS measurements and data processing).

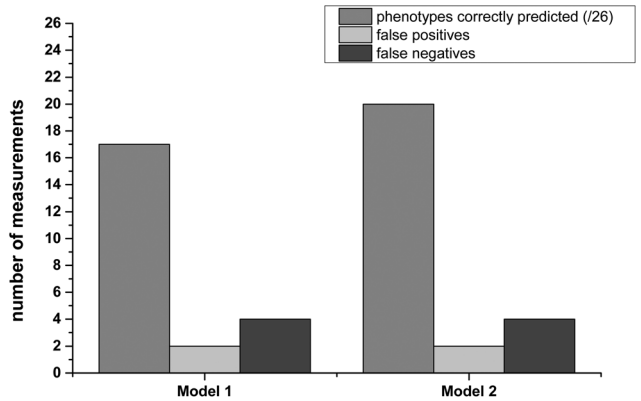


Fig. 5 Bar diagram summarizing the phenotype prediction based on the two models developed using the pattern recognition GA.

negative predictions, two measurements are predicted as NC instead of BE and two as NC instead of HGD; and within the false positive predictions, one EAC prediction and one HGD prediction are incorrectly made. Furthermore, with the exception of two samples, phenotype predictions for duplicate measurements are in agreement (Table 2). Overall, the second model has a better ability to predict a disease state correctly (80% sensitivity *versus* 70% for model 1) and to exclude a disease state correctly (66% specificity *versus* 50% for model 1). Although a larger blind sample set is necessary for a true clinical evaluation of the sensitivity and specificity of the two proposed models, this methodology is promising for disease phenotype delineation.

Overfitting occurs when a classifier describes random error or noise instead of the underlying relationships in the data. The results of overfitting will be high classification success rates for the training set but low classification success rates for the blinds. To avoid the problem of overfitting, principal component analysis was used in this study to develop discriminants. Although a PC plot is not a sharp knife edge for discrimination, if we have a PC plot that shows clustering, then our experience is that we will be able to predict robustly using this set of measurements. Predicting 20 of the 26 blind samples successfully from our principal component models is consistent with our previous experience. Because principal component analysis displays the variability between large numbers of samples and shows the major clustering trends present in data sets, the presence of confounding relationships in the data is uncovered, thereby gaining insight into how the decision for the classification has been made. The problems associated with chance or spurious classification, which is always of concern when using any variable selection technique, is mitigated because of the stringent criterion used for variable selection based on the PC score plots to assess the information content of discriminating features.

Conclusions

Serum N-linked glycans extracted from patients diagnosed with BE, HGD, EAC and NC are analyzed by MALDI-IMS-MS. The study design used here serves as a test to determine whether information characteristic of the disease state of the subject can be extracted from the mass spectral data, while ensuring that the potential impact of variations under the experimental conditions is monitored. A close examination of mobility profiles for the glycan ions $[S_1H_5N_4 + Na]^+$ $[F_1H_5N_4 + Na]^+$ and $[S_1F_1H_5N_4 + Na]^+$ revealed that in some cases, variations across different phenotypes are immediately noticeable. Because of the number of samples and ions examined within each sample, a pattern recognition based methodology utilizing variable selection is implemented in order to assess the capability of the dataset for disease phenotype delineation. To perform this task, mobility distributions for nine N-linked glycan ions ($F_1H_3N_4$, $F_1H_4N_4$, $F_1H_5N_4$, $S_1H_5N_4$, $S_1F_1H_5N_4$, $S_1H_5N_5$, $S_2H_5N_4$, $S_1F_1H_5N_5$ and $S_2F_1H_5N_4$) are extracted from the dataset and combined into a composite IMS distribution. A total of 58 serum samples are examined in duplicate (45 samples for the training set and 13 blind samples) yielding 116 MALDI-IMS-MS spectra. As a result, two models, differing by the nature of their feature subsets, are developed. Noticeably, NC, BE, HGD and EAC phenotypes are unambiguously distinguished for both proposed models. Interestingly enough, the second model displayed tighter clustering and appears overall more fit as it correctly predicted the phenotype

for 20 of the 26 blind measurements. Among the nine N-linked glycan ions selected for this analysis, three are the major contributors for distinguishing phenotypes for both proposed models ($S_1F_1H_5N_4$, $F_1H_4N_4$ and $F_1H_5N_4$); $S_1H_5N_5$ on the one hand and $S_1H_5N_4$ and $S_2H_5N_4$ on the other hand are unique to the first and second models respectively. Overall, this study demonstrates the capability of the combination of MALDI-IMS-MS and pattern recognition techniques for disease phenotype delineation.

Acknowledgements

Partial support for this work was provided by grants from the National Institutes of Health (NIH-5R01GM93322-2) and the Waters Corporation through the Indiana University Waters Center of Innovation.

References

1 R. Apweiler, H. Hermjakob and N. Sharon, On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database, *Biochim. Biophys. Acta*, 1999, **1473**(1), 4–8.

2 *Essentials of Glycobiology*, ed. A. Varki, R. Cummings, J. Esko, H. Freeze, G. Hart and J. Marth, CSHL Press, 2nd edn, 2009.

3 Z. Lin and D. M. Lubman, Permethylated N-glycan analysis with mass spectrometry, *Methods Mol. Biol.*, 2013, **1007**, 289–300.

4 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 Res.*, 2008, **7**, 603–610.

5 Z. Kyselova, Y. Mechref, P. Kang, J. A. Goetz, L. E. Dobrolecki, G. W. Sledge, L. Schnaper, R. J. Hickey, L. H. Malkas and M. V. Novotny, Breast cancer diagnosis and prognosis through quantitative measurements of serum glycan profiles, *Clin. Chem.*, 2008, **54**, 1166–1175.

6 Y. Mechref, S. Bekesova, V. Pungpapong, M. Zhang, L. E. Dobrolecki, R. J. Hickey, Z. T. Hammond and M. V. Novotny, Quantitative Serum Glycomics of Esophageal Adenocarcinoma and Other Esophageal Disease Onsets, *J. Proteome Res.*, 2009, **8**, 2656–2666.

7 C. W. Park, Y. Jo and E. J. Jo, Enhancement of ovarian tumor classification by improved reproducibility in matrix-assisted laser desorption ionization time-of-flight mass spectrometry of serum glycans, *Anal. Biochem.*, 2013, **443**, 58–65.

8 H. J. Jeong, Y. G. Kim, Y. H. Yang and B. G. Kim, High-Throughput Quantitative Analysis of Total N-Glycans by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry, *Anal. Chem.*, 2012, **7**, 3453–3460.

9 K. R. Reiding, D. Blank, D. M. Kuijper, A. M. Deelder and M. Wuhrer, High-Throughput Profiling of Protein N-Glycosylation by MALDI-TOF-MS Employing Linkage-Specific Sialic Acid Esterification, *Anal. Chem.*, 2014, **86**, 5784–5793.

10 W. R. Alley and M. V. Novotny, Structural Glycomic Analyses at High Sensitivity: A Decade of Progress, *Annu. Rev. Anal. Chem.*, 2013, **6**, 237–265.

11 G. Von Helden, T. Wyttenbach and M. T. Bowers, Inclusion of a MALDI Ion Source in the Ion Chromatography Technique: Conformational Information on Polymer and Biomolecular Ions, *Int. J. Mass Spectrom. Ion Processes*, 1995, **146**, 349–364.

12 K. J. Gillig, B. Ruotolo, E. G. Stone, D. H. Russell, K. Fuhrer, M. Gonin and A. J. Schultz, Coupling high-pressure MALDI with ion mobility/orthogonal time-of flight mass spectrometry, *Anal. Chem.*, 2000, **72**, 3965–3971.

13 B. T. Ruotolo, K. J. Gillig, E. G. Stone, D. H. Russell, K. Fuhrer, M. Gonin and J. A. Schultz, Analysis of protein mixtures by matrix-assisted laser desorption ionization-ion mobility-orthogonal-time-of-flight mass spectrometry, *Int. J. Mass Spectrom.*, 2002, **219**, 253–267.

14 L. S. Fenn and J. A. McLean, Simultaneous glycoproteomics on the basis of structure using ion mobility-mass spectrometry, *Mol. BioSyst.*, 2009, **5**, 1298–1302.

15 S. N. Jackson, B. Colsch, T. Egan, E. K. Lewis, J. A. Schultz and A. S. Woods, Gangliosides' analysis by MALDI-ion mobility MS, *Analyst*, 2011, **136**, 463–466.

16 D. J. Harvey, C. A. Scarff, M. Crispin, C. N. Scanlan, C. Bonomelli and J. H. Scrivens, MALDI-MS/MS with Traveling Wave Ion Mobility for the Structural Analysis of N-Linked Glycans, *J. Am. Soc. Mass Spectrom.*, 2012, **23**, 1955–1966.

17 K. Schwamborn, Imaging mass spectrometry in biomarker discovery and validation, *J. Proteomics*, 2012, **75**, 4990–4998.

18 H. Bateson, S. Saleem, P. M. Loadman and C. W. Sutton, Use of matrix-assisted laser desorption/ionisation mass spectrometry in cancer research, *J. Pharmacol. Toxicol. Methods*, 2011, **64**, 197–206.

19 Y. T. Cho, Y. Y. Chiang, J. Shiea and M. F. Hou, Combining MALDI-TOF and molecular imaging with principal component analysis for biomarker discovery and clinical diagnosis of cancer, *Genomic Med., Biomarkers, Health Sci.*, 2012, **4**, 3–6.

20 J. Stauber, L. MacAleese, J. Franck, E. Claude, M. Snel, B. K. Kaletas, I. M. V. D. Wiel, M. Wisztorski, I. Fournier and R. M. A. Heeren, On-Tissue Protein Identification and Imaging by MALDI-Ion Mobility Mass Spectrometry, *J. Am. Soc. Mass Spectrom.*, 2010, **21**, 338–347.

21 M. C. Djidja, E. Claude, M. F. Snel, P. Scriven, S. Francese, V. Carolan and M. R. Clench, MALDI-Ion Mobility Separation-Mass Spectrometry Imaging of Glucose-Regulated Protein 78 kDa (Grp78) in Human Formalin-Fixed, Paraffin-Embedded Pancreatic Adenocarcinoma Tissue Sections, *J. Proteome Res.*, 2009, **8**, 4876–4884.

22 L. M. Cole, M. C. Djidja, J. Bluff, E. Claude, V. A. Carolan, M. Paley, G. M. Tozer and M. R. Clench, Investigation of protein induction in tumour vascular targeted strategies by MALDI MSI, *Methods*, 2011, **54**, 442–453.

23 K. Chughtai, L. Jiang, T. R. Greenwood, K. Glunde and R. M. A. Heeren, Mass spectrometry images acylcarnitines, phosphatidylcholines, and sphingomyelin in MDA-MB-231 breast tumor models, *J. Lipid Res.*, 2013, **54**, 333–344.

24 D. Isailovic, R. T. Kurulugama, M. D. Plasencia, S. T. Stokes, Z. Kyselova, R. Goldman, Y. Mechref, M. V. Novotny and D. E. Clemmer, Profiling of Human Serum Glycans Associated with Liver Cancer and Cirrhosis by IMS–MS, *J. Proteome Res.*, 2008, **7**, 1109–1117.

25 D. Isailovic, M. Plasencia, M. Gaye, S. Stokes, R. Kurulugama, V. Pungpapong, M. Zhang, Z. Kyselova, R. Goldman, Y. Mechref, M. V. Novotny and D. E. Clemmer, Delineating Diseases by IMS–MS Profiling of Serum N–linked Glycans, *J. Proteome Res.*, 2012, **11**, 576–585.

26 M. M. Gaye, S. J. Valentine, Y. Hu, N. Mirjankar, Z. T. Hammoud, Y. Mechref, B. K. Lavine and D. E. Clemmer, Ion Mobility-Mass Spectrometry Analysis of Serum N-Linked Glycans from Esophageal Adenocarcinoma Phenotypes, *J. Proteome Res.*, 2012, **11**, 6102–6110.

27 Z. M. Segu, A. Hussein, M. V. Novotny and Y. Mechref, Assigning N-Glycosylation Sites of Glycoproteins using LC/MSMS in Conjunction with Endo-M/Exoglycosidase, *J. Proteome Res.*, 2010, **9**, 3598–3607.

28 P. Kang, Y. Mechref, I. Klouckova and M. V. Novotny, Solid-phase permethylation of glycans for mass spectrometric analysis, *Rapid Commun. Mass Spectrom.*, 2005, **19**, 3421–3428.

29 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, *J. Proteome Res.*, 2007, **6**(5), 1822–1832.

30 I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.*, 2003, **3**, 1157–1182.

31 I. Guyon, S. Gunn, M. Nikravesh and L. Zadeh, *Feature extraction: foundations and applications*, Springer Verlag, Berlin, 2006.

32 B. K. Lavine, N. Mirjankar and S. Delwiche, Classification of the waxy condition of durum wheat by near infrared reflectance spectroscopy using wavelets and a genetic algorithm, *Microchem. J.*, 2014, **117**, 178–182.

33 B. K. Lavine, N. Mirjankar, R. LeBouf and A. Rossner, Prediction of mold contamination from microbial volatile organic compound profiles using head space gas chromatography/mass spectrometry, *Microchem. J.*, 2012, **103**, 119–124.

34 B. K. Lavine, K. Nuguru and N. Mirjankar, One stop shopping - feature selection, classification, and prediction in a single step, *J. Chemom.*, 2011, **25**, 111–129.

35 A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice Hall, 1988.

36 R. G. Lawson and P. C. Jurs, New index for clustering tendency and its application to chemical problems, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 36–41.

37 V. Centner, D. L. Massart and O. E. de Noord, Detection of inhomogeneities in sets of NIR spectra, *Anal. Chim. Acta*, 1996, **330**, 1–17.