Biomarker Discovery for Arsenic Exposure Using Functional Data Analysis and Feature Learning of Mass Spectrometry Proteomic Data

Oct 30, 2007

Jaroslaw Harezlak[1,⋆], Michael C. Wu[2,⋆], Mike Wang[3], Armin Schwartzman[2,4], David C. Christiani[3], Xihong Lin[2]

[1] Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, 46202

[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

[3] Department of Environmental Health, Harvard School of Public Health, Boston, MA 02115

[4] Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, MA 02115

⋆ - Email: harezlak@iupui.edu and mwu@hsph.harvard.edu. These authors contributed equally to this work.

Abstract

Plasma biomarkers of exposure to environmental contaminants play an important role in early detection of disease. The emerging field of protemics presents an attractive opportunity for candidate biomarker discovery, as it simultaneously measures and analyzes a large number of proteins. This article presents a case study for measuring arsenic concentrations in a population residing in an As-endemic region of Bangladesh using plasma protein expressions measured by SELDI-TOF mass spectrometry. We analyze the data using a unified statistical method based on functional learning to preprocess mass spectra and extract MS features and to associate the selected mass spectrometry features with arsenic exposure measurements. The task is challenging due to several factors, high dimensionality of mass spectrometry data, complicated error structures, and a multiple comparison problem. We use nonparametric functional regression techniques for MS modeling, peak detection based on the significant zero-downcrossing method and peak alignment using a

warping algorithm. Our results show significant associations of arsenic exposure to either under- or over-expressions of 20 proteins.

1 Introduction

Arsenic, classified as a human carcinogen by International Agency for Research on Cancer (IARC), is one of the most serious environmental health hazards, with chronic arsenic exposure occurring mainly through drinking water(Gebel [2000], Hughes [2002], IARC [2000]). In Bangladesh, it is estimated that 35 to 77 million people are at risk of drinking arsenic contaminated water, the largest mass poisoning of a population in history (Guha [1998]). Pre-malignant skin lesions, including hyperpigmentation and hyperkeratosis, are hallmarks of chronic arsenic ingestion by humans, which may eventually lead to the development of skin cancer(Haque [2003], Council [2001]). In addition, population-based epidemiologic studies have associated chronic arsenic exposure with numerous adverse health outcomes, including internal organ cancers, neurological effects, hypertension, cardiovascular disease, pulmonary disease, peripheral vascular disease, and diabetes mellitus (Buchet [1981]).

With high affinity to the sulfhydryl groups in keratin, arsenic is usually found at the highest levels in the hair, skin, and nails, and is isolated from further metabolic processes once it is deposited in keratin matrix Chen [1999]. Thus, arsenic detected in the keratin will reflect only those conditions that occurred during its deposition, making it an ideal biomarker of chronic arsenic exposure. Epidemiologic studies have correlated nail arsenic levels with water exposure (Karagas [1996], Bonassi [2002]). Nail concentrations of arsenic are good biomarkers of chronic body burden of arsenic exposure from drinking, representing an internal dose exposure 9-12 months before sample collection.

Research on the biomarkers of the effects caused by chronic arsenic exposure, such as premalignant skin lesions, is very limited. Without valid biomarkers of early effects, it is more difficult to intervene for prevention and control of environmental disease. Biomarkers can be singular measures of a protein, enzyme activity or a small molecule associated with health, disease or toxicity (Wetmore [2004]). For many complex diseases and toxicities,

it is unlikely that any single biomarker may be a sufficient indicator. Instead, multiple markers that function as the signature patterns/profiles are necessary to achieve higher sensitivity, and enable the early discovery of acute toxicity onset or the molecular signatures of long-term toxicant exposure and disease (Bischoff [2004]).

Mass spectrometry proteomics provides such a tool by an attempt to identify and quantify relative abundance of a large number of proteins using high-throughput technologies. We concentrate in this paper on the surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) mass spectrometry (MS) which was developed by Ciphergen Biosystem (Hutchens and Yip [1993]) for profiling protein (peptide) biomarkers from complex biological samples.

In the present study, we investigate the plasma "proteome" profiles using SELDI TOF mass spectrometry in a population with well-characterized arsenic exposure. The main objective is to find the plasma proteomic profiles associated with arsenic-induced effects, and to identify specific proteins that can be used as biomarkers of early diagnoses in high-risk populations and hopefully used to find new treatment targets. In order to achieve these goals, we propose a new statistical method based on functional learning for the analysis of mass spectrometry proteomic data.

Several characteristics of proteomic spectra mandate pre-processing of the data. It has been well recognized that appropriate preprocessing of MS data is critical to ensure meaningful analysis of the association between proteins and disease or exposure in second-stage analysis (Baggerly *et al*. [2003] and Listgarten and Emili [2005]). Standard pre-processing steps often involve several sequential ad hoc steps including baseline subtraction, normalization, peak detection and peak alignment. Some of these steps are implemented in the Ciphergen software and the R package PROcess in Bioconductor (Li *et al*. [2005]). Specifically, baseline subtraction is performed to remove the elevation due to the presence of the energy absorbing molecule contamination. Total area normalization is used to normalize the intensities from different spectra. Peak detection is most frequently done by threshold-

ing the signal over noise ratio using non-optimal cutoffs, and peak matching is performed by binning the etected peaks on different spectra by location proximity. Correlation of protein intensity measures over the spectra is often ignored.

We develop in this paper a unified statistical method that simultaneously takes into account different sources of variation that are present in mass spectrometry measurements. Specifically, we propose a functional learning method to process the raw MS data, and use the resulting dimensional-reduced measures of protein abundance to relate to the arsenic exposure.

In particular, in the first stage, we decompose the MS raw spectrum into four components: baseline, signal, instrumental noise component and random noise. Our primary interest is in the signal component which is used to further define the peaks in the spectrum corresponding to proteins detected in plasma. Characterization of the baseline and two noise components is necessary though, since they influence the quantification and existence of peaks respectively. We pose minimal assumption on the statistical models used, and estimate individual baseline and signal components nonparametrically using kernel methods. The instrumental noise component is modeled as a harmonic function with two frequencies detected in the preliminary analysis, and an exponential decay component corresponding to the inverse relationship between the mass over charge ratio and the noise level. We perform peak detection based on the significant zero-downcrossing method, and peak alignment using a warping algorithm. At the second stage, the identified protein expression features, e.g., peak intensities are related to arsenic exposure using two-sample comparison and the dose response relationships are studied using linear regression. False discovery rates are used to account for multiple comparison. Super-proteins are identified using principal component analysis.

2 Materials and Methods

Study Population

All study participants were selected from a large Arsenic Case-Control Study of Skin Disease in Bangladesh as described previously (Kile [2005]). Briefly, from 2000 to 2003, 900 pairs of skin lesion cases and controls were recruited from the Pabma district of Bangladesh, located north of Dhaka on the Padma (Ganges) River, which is a region considered to be moderately affected by arsenic contamination in drinking water. Blood samples were collected from each study participant, and plasma were separated by centrifugation and frozen at -80^oC. Hair and toenails samples were also collected from each study participant, and toenail levels of arsenic were analyzed as described previously (Kile [2005]). In addition, questionnaires regarding exposure history, diet and lifestyle factors were collected, as well as a drinking/tube well water sample. To investigate plasma proteomic profiles of arsenic exposure, samples were selected by first sorting toenail arsenic levels in 900 healthy controls, and then selecting 100 high toenail arsenic and 100 low toenail arsenic samples with matched age (± 3 years), gender and living location.

SELDI-TOF/MS

Surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) mass spectrometry (MS) was developed by Ciphergen Biosystem (Hutchens and Yip, 1993) for profiling protein (peptide) biomarkers from complex biological samples. This technology has been applied to different body fluids, including serum, urine and nipple aspirate fluid, and has been employed successfully in discovery of protein profiles of several diseases. For example, protein profiles were used to distinguish ovarian (Petricoin *et al*., 2002), prostate (Adam *et al*., 2002) and breast cancer (Li *et al*., 2002) patients from healthy controls.

Serum sample collection

We used IMAC (Immobilized Metal Affinity Chromatography) ProteinChip array (Ciphergen Biosystem Inc., Fremont, CA) in this study. Before analysis, ProteinChips were

washed with 50% acetonitrile (HPLC grade; Aldrich, Milwaukee, WI, USA) for 2 x 5 min, dried for 1 h at room temperature, loaded onto a 192-well bioprocessor (Ciphergen), and equilibrated with 10% acetonitrile/0.1% trifluoroacetic acid (Fisher Scientific International, Hampton, NH, USA). Plasma samples were thawed at 4^oC, centrifuged at 10,000 $\times$ g at 4^oC for 10 min to remove any precipitates, and then aliquots of each sample were subjected to Ciphergen fractionation or Sigma multiple-removal column. After mixing with sample buffer (8 M urea, 2% 3-w (3-cholamidopropyl) dimethylammoniox-1-propansulfonate, pH 7.4) in a volume ratio of 2:3, the following procedure was carried out using a fully automated liquid-handling robotic system (Biomek FX, Beckman Coulter, Fullerton, CA, USA). Ten microliter sample mix was dispensed onto array spot, incubated for 1 h, washed, and air dried according to manufacturer's instruction. After applying energy absorbing matrix (EAM) molecule, sinapinic acid (SPA; Fluka, Buchs, Switzerland), mass spectrometry was carried out with the Protein Biology System II SELDI-TOF mass spectrometer reader (Ciphergen). The reader was externally calibrated with 8 different calibrants (Ciphergen) with molecular weights ranging from 1296.5 to 43,240 Da. Time-of-flight spectra were derived at two different laser settings: one low-energy protocol, which is most suitable for detection of peptides and proteins less than 10,000 Da; and a high-energy protocol, which is optimal for capturing proteins between 10,000 and 40,000 Da, as recommended by the manufacturer.

Statistical Model for Preprocessing MS data

We model the MS spectra as continuous functions using functional learning methods and focus on feature (peak) extraction, alignment and quantification for the second-stage analysis. All the analyses are done on the time scale, and the final results are presented using mass over charge (m/z) ratios corresponding to the instrument observation times. We propose a model consisting of baseline, normalization, signal, and noise components. We further decompose the noise structure into a harmonic instrument-related noise and a random noise. Our main interest is nonparametric estimation of the signal portion of each individual MS, and peak detection using significant zero down-crossings estimated by cal-

culating the confidence intervals of the derivatives of the nonparametric signal curves. The detected MS features (peaks) are then aligned using a warping algorithm. The resulting MS features, e.g., peak intensities, are related to the arsenic exposure level in the second stage that is described in detail in the next section.

We first describe our proposed generic functional model for MS spectra,

$$Z\{w(t)\} = B(t) + NS(t) + \epsilon(t),$$

where $Z(t)$ is the MS intensity at time t, $w(t)$ is a warping (peak-alignment) function, $B(t)$ is a baseline function, N is a normalizing constant, $S(t)$ is the signal of interest and $\epsilon(t)$ is the noise process. We assume that $B(t)$ and $S(t)$ vary smoothly and estimate them using kernel smoothing. Specifically, $B(t)$ is estimated by fitting a slowly varying smooth function to a low quantile of the individual spectra. The normalizing constant N is calculated as the area under the baseline-subtracted portion of the spectrum. The signal curve $S(t)$ is estimated using local polynomial kernel smoothing with the bandwidth selected by a plug-in method. Plug-in method of bandwidth selection has been shown to perform well in practice (Wand and Jones [1995]). It was chosen in our application as it is a computationally efficient method and can be easily modified to incorporate a correlation structure of the error terms. Estimation of the optimal bandwidth requires good estimates of the variance and the covariance of the errors. To understand the error structure, we performed exploratory analysis of the errors using fast Fourier transform in conjunction with kernel smoothing. This analysis suggested to further decompose the error term into the instrument-related noise and the random noise as

$$\begin{aligned} \epsilon(t) &= H(t) + e(t), \\ H(t) &= \exp\{-kt\} * \{a_{1s}\sin(\omega_1 t) + a_{1c}\cos(\omega_1 t) + a_{2s}\sin(\omega_2 t) + a_{2c}\cos(\omega_2 t)\}, \end{aligned} \quad (1)$$

where k is the decay coefficient, a_{1s}, a_{1c}, a_{2s}, a_{2c} are the coefficients of the harmonic oscil-

lations due to the instrument, ω_1 and ω_2 are the two periods of the harmonic component, and $e(t)$ is the random noise assumed to follow a distribution with mean 0, variance $\sigma^2(t)$ and autoregressive correlation structure.

Features (peak locations) are identified for individual spectra and the mean spectrum based on the zero-downcrossing idea of Chaudhuri and Marron [1999]. Specifically, the zero downcrossing T is defined as the point where the first derivative estimate of the curve is zero and the confidence bands are above zero for the values $t < T$ and below zero for $t > T$. The estimated zero-downcrossings are regarded as peak locations. We estimate the peak locations for each individual and the mean spectrum of all the individuals. To align the peaks, we use the mean spectrum peaks as empirical landmarks and individual curves are aligned to the landmarks using piecewise linear warping function $w(t)$ and partial matching of individual and mean curve peaks.

From the first stage functional modeling of the spectra, we get the estimates of the signal functions, their derivatives, and standard errors which are used to arrive at the confidence bands for the first derivative of the signal functions and the locations of the peaks and their intensities.

Specifically, we give a simplified description of the algorithm used to recover the protein expression features (peaks) here, and provide more details in the Appendix. First, for each subject $i = 1, \ldots, n$, we estimate the baseline function $B_i(t)$ using a local polynomial kernel smoothing technique with a large bandwidth fit to the low quantile points in the large neighborhood around values of t. Second, we subtract this estimate from the raw values $Z_i(t)$, calculate the area under the curve N_i and obtain the baseline-subtracted normalized values $Z_i^\star(t) = [Z_i(t) - B_i(t)]/N_i$. Third, we use the backfitting procedure to estimate the signal curve $S_i(t)$ using local linear kernel regression and the harmonic component $H_i(t)$ using nonlinear least squares of the noise process. Fourth, we obtain the locations of the peaks on the individuals curves using the significant zero-downcrossing method and denote them as $\widehat{T}_{i1}, \ldots \widehat{T}_{iQ_i}$, where Q_i is the number of peak locations for subject i.

Fifth, we estimate the average curve $\overline{S}(t)$ using the average baseline-subtracted spectrum $\bar{Z}^{\star}(t) = n^{-1}\sum_{i=1}^{n} Z_i^{\star}(t)$, where the average is a cross-sectional averaging at each t over all the subjects. Sixth, we obtain the locations of the peaks on the average curve via the significant zero-downcrossing method and denote them by $\widehat{T}_1, \ldots \widehat{T}_Q$, where Q is the number of peak locations on the average curve. Seventh, we associate the subject-specific peaks with the peaks detected on the average curve and warp the subject-specific signal estimates $S_i(t)$ to locally align the subject-specific peaks to the peaks on the average curve. If an individual does not have a peak in the neighborhood of any population peak, this individual peak is not aligned. We go back to the fifth step, where we use the warped individual curves to go through detection of average peaks and alignment till convergence. We define the stopping criterion to be the distance between the estimates of the average curve to be small and the number of peaks detected on an average curve to be not changing. The individual peak intensities at the population peak locations $T_1, \ldots, T_Q$ are regarded as the dimension reduced individual features from the first stage analysis.

Second Stage Analysis

To identify protein peaks associated with high arsenic exposure, we first treat the exposure status as a dichotomous variable (0 = low/1 = high) and for each ascertained peak, the log-transformed protein expression value was regressed on exposure status. Wald p-values testing for association between exposure status and protein expression level were computed for all protein peaks. To account for multiple hypothesis tests, false discovery rates (FDRs) are calculated and the corresponding q-values are reported (Benjamini and Hochberg [1995], Storey [2002]). To account for confounding of covariates, we also perform confounder adjusted analysis by including covariates - age, body mass index (BMI), sex, smoking, smoking environment, chewing tobacco, and chewing betel nuts - in the model and regressing the log-transformed protein expression values at each peak on all variables. The Wald p-values and FDRs for testing for an association between exposure status and protein expression are calculated while adjusting for potential confounders.

Dose Response Analysis

We next perform a dose response analysis to examine the continuous relationship between arsenic concentration and protein expressions. Specifically, we fit a linear regression model of log-transformed protein expression at each peak on continuous log toe-nail arsenic concentration. Both unadjusted analysis and confounder adjusted analysis are performed and the Wald p-values and FDR q-values are calculated.

"Super-Protein" Based Analysis

In proteomic profiling, some of the protein peak intensities are often highly correlated, possibly as a result of doubly charged molecules or biological mechanism, such as protein interactions. As such, the individual peak analyses described above, such as FDR estimates, tend to be conservative as the number of comparisons is inflated. To overcome this problem, we conduct a "super-protein" analysis. We first perform an agglomeration hierarchical clustering analysis to all the peak intensities with the pairwise correlation between peaks used as the distance metric. The cluster dendrogram can be cut to generate clusters which are then combined to form the membership of a "super-protein". The expression of each super-protein is calculated as the first principal component of its constituent protein peaks' expressions. The unadjusted and confounder adjusted analyses as described above are performed.

3 Results

We provide in Table 1 descriptive statistics of the covariates in the Bangladesh study. The covariates are similar in the high and low arsenic exposure groups except that slightly more subjects in the high arsenic exposure group chewed tobacco and betal nuts. No statistically significant difference in the covariates between the two arsenic exposure groups was found.

We next applied the proposed method to the study of the arsenic exposure effects on protein expressions in the Bangladesh data described in Section 2. We restrict our analysis to the range of mass over charge (m/z) values of 3000Da and 20,000Da. All analyses were performed on the equally spaced time scale and the final results are presented on the m/z scale. Fast Fourier analysis using the original baseline-subtracted data suggested an AR(1) error structure for the noise component $\epsilon_i(t)$, with a preliminary estimate of the correlation parameter equal to $\rho = 0.6$. This estimate was used to obtain preliminary estimates of the individual signal curves $S_i(t)$, which were used in turn to get estimates of the harmonic component of the noise $H_i(t)$ (see Figure 1). For the harmonic noise component $H_i(t)$, the estimated frequencies ω_1 and ω_2 were approximately 32.0 MHz and 93.8 MHz (see Figure 2). The estimates were almost identical for all the subjects indicating that the harmonic noise is due to the instrument itself, not due to the plasma samples. The removal of the harmonic component $H_i(t)$ resulted in the errors $e_i(t)$ showing little or no correlation in the final iteration with the correlation coefficient ρ not statistically significant from 0 (mean=0.075, sd=0.052). The variance $\sigma^2(t)$ of the noise component $e(t)$ was estimated using difference-based estimator and was used in the construction of confidence bands for the derivatives of the signal curves $S_i(t)$.

The iterative warping procedure took four steps to converge. We detected a total of $Q = 77$ protein peaks on the average curve $\overline{S}(t)$. The number of peaks on the individual signal curves ranged from 73 to 161. We used the 77 population peak locations as population landmarks and aligned the individual curves by warping them locally using the population landmarks. To illustrate the results, we present in Figure 3 the signal estimates in the range 8.8kDa and 9.0kDa for 10 randomly selected subjects before and after the alignment. The vertical line denotes the peak detected at 8,911Da. The results after alignment show good agreement in the peak locations. The individual intensities from the estimated individual curves at the 77 population peak locations are used for the second stage analysis.

At the second stage analysis, we associated the intensity of each of 77 protein peaks with

arsenic exposure. We first treated the exposure status as a dichotomous variable (0 = low/1 = high) and regressed the log-transformed protein expression value on the exposure status. We performed both crude analysis and covariate-adjusted analysis. The covariates in the covariate-adjusted analysis included age, body mass index (BMI), sex (female = 0/male = 1), smoking (no = 0/yes = 1), smoking environment (no = 0/yes = 1), chews tobacco (no = 0/yes = 1), and chews betel nuts (no = 0/yes = 1) (Kile [2005]) The results are presented in table 2.

Setting the FDR at 5%, for the crude analysis, 28 protein peaks were found to be significantly associated with the high arsenic exposure status. Among these, 16 were over-expressed for subjects with high exposure. After controlling for covariates, a total of 24 protein peaks were found to be significantly associated with the high exposure status. Among these, 14 protein peaks were over-expressed for subjects with high exposure. These peaks are a subset of the 28 peaks identified by the unadjusted analysis. We present in Table 3 the dose-response analysis results. Without the covariate adjustment, we identified 23 significant peaks at the FDR = 0.05 level. As expected, these dose response results were quite consistent with the earlier dichotomous exposure analysis with a substantial overlap of significant peaks. Adjusting for covariates yielded identical results to the crude analysis.

Although these individual protein analyses help in identifying promising candidate single biomarkers for arsenic exposure, some protein peaks are highly correlated, as can be seen from the left panel of figure 4, a heatmap of the correlation structure of the peaks. Therefore, we applied the super-proteins based analysis to increase the statistical power. Following complete aglommeration hierarchical clustering, the cluster dendrogram was then cut a height of 1.5. This left a final group of 46 super-proteins. The pairwise correlations of the super-proteins are shown in the right panel of figure 4. Although a few super-proteins are still highly correlated, overall, the blocks of strongly correlated markers have been removed.

Using the super-proteins, both the unadjusted and covariate-adjusted analyses were per-

formed. At the FDR = 0.05 level, 17 super-proteins were determined to be significantly associated with exposure status. Results are given in table 4. Although only 17 super-proteins are significant, in contrast to the 28 protein peaks identified in the initial crude analysis, these super-proteins actually correspond to 31 protein peaks. Out of these 31 constituent protein peaks, 25 were identified in the earlier analysis which also identified 3 peaks not found here. However, the super-protein based analysis identified 6 additional proteins not previously found. The covariate-adjusted analysis identified 14 super-proteins as significant, but these super-proteins correspond to 24 proteins. Although using the covariate-adjusted super-protein based analysis did not identify more protein peaks than the original covariate-adjusted analysis, the list of interesting protein peaks were not the same: 19 protein peaks were identified by both analyses, while each analysis found 5 more proteins peaks that the other did not.

4 Conclusions

Mass spectrometry shows significant potentials to identify biomarkers for early detection of a disease. Pre-processing of the MS data is critical in proteomic research. However, the existing pre-procesing methods are often ad-hoc and lack sound statistical justifications and raise concerns of the validity and reliability of the identified peaks. We propose in this paper a unified statistical approach based on functional learning to pre-process protein expressions in plasma as measured by the SELDI-TOF-MS instrument and study the association of the identified protein peaks and arsenic exposure in the population of Bangladesh. This unified framework allows us to flexibly model the mass spectra using advanced nonparametric regression techniques such as error structures, peak alignment, and peak detection. In the second stage analysis, we perform both individual protein analysis and super-protein analysis. Although our method is applied to the SELDI-TOF-MS data, this functional learning method is applicable for pre-processing of mass spectrometry data collected from other platforms, such as LC-MS-MS and MALDI-TOF-TOF.

Acknowledgments: This research was supported in part by NIH grants: R01ES011622 , R01ES015533, ES00002 and R37CA76404.

References

Gebel, T., Confounding variables in the environmental toxicology of arsenic. *Toxicology* 2000, **144, (1-3)**, 155-62.

Hughes, M. F., Arsenic toxicity and potential mechanisms of action. *Toxicol Lett* 2002, **133, (1)**, 1-16.

IARC, In Overall Evaluations of Carcinogenicity to Humans.

Guha Mazumder, D. N.; Haque, R.; Ghosh, N.; De, B. K.; Santra, A.; Chakraborty, D.; Smith, A. H., Arsenic levels in drinking water and the prevalence of skin lesions in West Bengal, India. *Int J Epidemiol* 1998, **27, (5)**, 871-7.

Haque, R.; Mazumder, D. N.; Samanta, S.; Ghosh, N.; Kalman, D.; Smith, M. M.; Mitra, S.; Santra, A.; Lahiri, S.; Das, S.; De, B. K.; Smith, A. H., Arsenic in drinking water and skin lesions: dose-response data from West Bengal, India. *Epidemiology* 2003, **14, (2)**, 174-82.

Council, N. R. Arsenic in Drinking Water: 2001 Update; National Academy Press: Washington, DC, 2001.

Buchet, J. P.; Lauwerys, R.; Roels, H., Urinary excretion of inorganic arsenic and its metabolites after repeated ingestion of sodium metaarsenite by volunteers. *Int Arch Occup Environ Health* 1981, **48, (2)**, 111-8.

Chen, K. L.; Amarasiriwardena, C. J.; Christiani, D. C., Determination of total arsenic concentrations in nails by inductively coupled plasma mass spectrometry. *Biol Trace Elem Res* 1999, **67, (2)**, 109-25.

Karagas, M. R.; Morris, J. S.; Weiss, J. E.; Spate, V.; Baskett, C.; Greenberg, E. R., Toenail samples as an indicator of drinking water arsenic exposure. *Cancer Epidemiol Biomarkers Prev* 1996, **5, (10)**, 849-52.

Bonassi, S.; Au, W. W., Biomarkers in molecular epidemiology studies for health risk prediction. *Mutat Res* 2002, **511, (1)**, 73-86.

Wetmore, B. A.; Merrick, B. A., Toxicoproteomics: proteomics applied to toxicology and pathology. *Toxicol Pathol* 2004, **32, (6)**, 619-42.

Bischoff, R.; Luider, T. M., Methodological advances in the discovery of protein and peptide disease markers. *J Chromatogr B Analyt Technol Biomed Life Sci* 2004, **803, (1)**, 27-40.

Kile, M. L.; Houseman, E. A.; Rodrigues, E.; Smith, T. J.; Quamruzzaman, Q.; Rahman, M.; Mahiuddin, G.; Su, L.; Christiani, D. C., Toenail arsenic concentrations, GSTT1 gene polymorphisms, and arsenic exposure from drinking water. *Cancer Epidemiol Biomarkers Prev* 2005, **14, (10)**, 2419-26.

Hutchens, T. W. and Yip, T. T. (1993) New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Commun. Mass. Spectrom.*, **7**, 567-580.

Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GLJr., (2002) Serum protein Fingerprinting coupled with a pattern matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, *Cancer Res.* **62**: 3609-3614.

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Bofelli,F.*et al.*, (2002) Use of proteomic patterns in serum to identify ovarian cancer, *Lancet*, **359**, 572-577.

Li J, Zhang Z, Rosenzweig J, Wang YY, and Chan DW (2002) Proteomics and Bioinformatics Approaches for Identification of Serum Biomarkers to Detect Breast Cancer. *Clinical Chemistry* **48**, 1296-1304

Chaudhuri, P. and Marron, J. S. (1999), SiZer for Exploration of Structures in Curves, *Journal of the American Statistical Association*, **94**, 807-823

Storey, JD (2002) A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B*, **64**, 479-498.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*, **57**, 289-300.

Baggerly KA, Morris JS, Wang J, Gold D, Xiao LC, and Coombes KR (2003) A comprehensive approach to the analysis of MALDI-TOF proteomics spectra from serum samples, *Proteomics*, :1667-1682

Listgarten J and Emili A (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry, *Molecular & Cellular Proteomics*, **4**:419–434

Diamandis E. P. (2004) Mass spectrometry as a diagnostic and cancer biomarker discovery tool. *Molecular and Cellular Proteomics*, **3**, 367-378.

Fung ET., Enderwick C. (2002) ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques*, **(Suppl. 3)**: 34-38, 40-41.

Koopmann J, Zhang Z, White N, Rose nzweig J, Fedarko N, Jagannath S, Canto MI, Yeo CJ, Chan DW, Goggins M. (2004) S erum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorpt ion and ionization mass spectrometry. *Clin Cancer Res.*, **10(3)**, 860-8.

Li X., Gentleman R., Lu X., Shi Q., Igleha rt J.D., Harris L. and Miron A., (2005) SELDI-TOF Mass Spectrometry Protein Data . In Gentleman, R. et al: Bioinformatics and Computational Biology Solutions Usi ng R and Bioconductor, Springer.

Wand M.P. and Jones M.C. (1995) Kernel smoothing, Chapman and Hall.

5 Appendix

Notation

Let $B_i(t)$ be a subject-specific baseline function, $S_i(t)$ be a signal function for subject i, and $H_i(t)$ be a harmonic part of the noise process. The observed MS $Z_i(t)$ is decomposed as

$$\begin{aligned} Z_i\{w_i(t)\} &= B_i(t) + N_i S_i(t) + \epsilon_i(t), \quad (2) \\ \epsilon_i(t) &= H_i(t) + e_i(t), \\ H_i(t) &= exp\{-k_i t\} * [a_{1s}\sin(\omega_1 t) + a_{1c}\cos(\omega_1 t) + a_{2s}\sin(\omega_2 t) + a_{2c}\cos(\omega_2 t)]. \quad (3) \end{aligned}$$

We provide here a detailed description of the functional learning algorithm for estimation of the population peak locations $T_1, \ldots, T_Q$ and subject-specific peak intensities at these locations.

1. For a subject $i, i = 1, \ldots, n$, estimate the baseline-subtracted signal curve using a kernel smoother with a large bandwidth fitted to the low quantiles of the raw data. Denote it by $\hat{B}_i(t)$ and calculate $Z_i^\star(t) = Z_i(t) - \hat{B}_i(t)$.

2. Calculate the area under the $Z_i^\star(t)$ curve and denote it by $\widehat{N}_i$. Calculate $Z_i^{N\star}(t) = \text{median}(\widehat{N}_i) Z_i^\star(t)/N_i$. For the following steps, we work with the so-called "baseline-subtracted and normalized" curve and drop the $N\star$ from the Z_i to simplify the notation.

3. Use a local polynomial kernel smoother with a location-varying bandwidth $\widehat{h}_i(t)$ obtained using the plug-in method with starting values for the variance $\sigma^2(t)$ and correlation coefficient ρ to get a preliminary estimate of $S_i(t)$. Denote the residuals of this fit by $\epsilon_i^S(t) = Z_i(t) - S_i(t)$.

4. Fit a parametric function $H_i(t)$ to the $\epsilon_i^S(t)$, and remove the instrument-related harmonic noise from the data as $Z_i^S(t) = Z_i(t) - S_i(t)$.

5. For subject i, set the initial warping function $w_i(t) = t$.

6. Calculate the average spectrum $\overline{\mathbf{z}} = \{\overline{Z}(t_1), \ldots, \overline{Z}(t_m)\}^\top$, and obtain the estimates of the signal $S(t)$ and harmonic part of the noise $H(t)$ using the same steps as those for the individual spectra $Z_i(t)$.

7. Estimate the peak locations on the average curve: $T_1, \ldots, T_Q$ using the zero-downcrossing method, Specifically, T_q is declared to be a peak location if

   (a) $S^{(1)}(T_q) = 0$ and $S^{(1)}(T_q^-) > 0,\ S^{(1)}(T_q^+) < 0$

   (b) T_q is a significant zero-downcrossing,

   where $S^{(1)}(t)$ denotes the first derivative of $S(t)$

8. Estimate the significant zero-downcrossings on the individual curves $i = 1, \ldots, m$ similar to step 8. Denote the elements of the set of individual downcrossing locations as T_{iq_i} where $i = 1, \ldots, m$ and $q_i = 1, \ldots, Q_i$.

9. For each subject, estimate the individual time-warping function $w_i(t)$ using a piecewise linear function by registering each individual peak locally using the population peaks $T_1, \cdots, T_Q$ as landmarks. Register the curves using the estimated warping functions.

10. Go back to step 6. Repeat until convergence defined as:

$$\begin{aligned} ||\overline{Z}^{(b+1)} - \overline{Z}^{(b)}|| &< \varepsilon, \\ \text{and} \qquad Q^{(b+1)} &= Q^{(b)}, \end{aligned}$$

    where b is the iteration number.

11. At convergence, calculate the individual peak intensities F_{iq} for all Q peaks and all curves $i = 1, \ldots, n$. The peak intensities F_{iq} ($q = 1, \cdots, Q$) at the peak locations $T_1, \cdots, T_Q$ are used in the second stage analysis.

Table 1: Descriptive Statistics of the 214 Subjects in the Bangladesh Proteomic Study of the Arsenic Exposure Effect

Continuous Characteristics				
Characteristic		Overall Median (Range)	High Exposure Median (Range)	Low Exposure Median (Range)
Toenail Arsenic (ugg)		4.06 (0.12-44.74)	7.31 (2.8 0-44.74)	0.59 (0.12-0.79)
Age		28.5 (16-78)	29.5 (16-78)	27.5 (16-75)
BMI		19.4 (14.7-33.8)	19.1 (14.7-20.8)	19.7 (15. 4-33.8)
Discrete Characteristics				
Characteristic	Value	Overall Number (%)	Number Among High Exposure (%)	Number Among Low Exposure (%)
Overall		214 (100)	116 (54.3)	98 (45.8)
Sex	Male	116 (54.2)	67 (57.8)	55 (56.1)
	Female	92(43.0)	49 (42.2)	43 (43.9)
Ever Smoked	Yes	58 (27.1)	31 (26.7)	27 (27.6)
	No	154 (72.0)	84 (72.4)	70 (71.4)
	Unknown	2 (0.9)	1 (0.9)	1 (1.0)
Smoking Environment	Yes	161 (75.2)	90 (77.6)	71 (72.4)
	No	51 (23.8)	24 (20.7)	27 (27.6)
	Unknown	2(0.9)	2 (1.7)	0 (0.0)
Chew Tobacco	Yes	32 (15.0)	21 (18.1)	11 (11.2)
	No	177 (82.7)	90 (77.6)	87 (88.8)
	Uknown	5 (2.3)	5 (4.3)	0 (0.0)
Chew Betel Nuts	Yes	45 (21.0)	29 (25.0)	16 (16.3)
	No	167 (78.0)	86 (74.1)	81 (82.7)
	Unknown	2 (0.9)	1 (0.9)	1 (1.0)

Table 2: Regression coefficient estimates and p-values of the identified proteins that are significantly associated with the arsenic exposure status (high/low) at the FDR = 0.05 level for both the unadjusted analysis and the covariate-adjusted analysis for the Bangladesh study. The $\uparrow$ sign indicates that the protein is over-expressed in the high exposed group, while $\downarrow$ indicates that the protein is under-expressed in the exposed group.

	Unadjusted Analysis			Adjusted Analysis			Trend in
Protein Intensity (m/z)	Estimate	p-value	q-value	Estimate	p-value	q-value	Exposed Subjects
3255	-0.319	0.015	0.044	-0.288	0.034	0.090	$\downarrow$
3428.6	0.145	< 0.001	0.002	0.131	0.001	0.007	$\uparrow$
3669.7	-0.224	0.001	0.005	-0.230	0.001	0.007	$\downarrow$
4124.7	-0.192	< 0.001	0.004	-0.179	0.002	0.011	$\downarrow$
4277	-0.401	0.002	0.009	-0.388	0.003	0.017	$\downarrow$
4566.9	0.124	0.009	0.028	0.122	0.013	0.043	$\uparrow$
4640.1	0.145	0.004	0.019	0.133	0.012	0.043	$\uparrow$
5248.1	0.314	0.001	0.005	0.291	0.002	0.011	$\uparrow$
5577.5	-0.147	0.001	0.005	-0.150	0.001	0.007	$\downarrow$
5800.9	-0.272	< 0.001	0.004	-0.272	0.001	0.007	$\downarrow$
5905.1	0.320	0.008	0.028	0.263	0.035	0.090	$\uparrow$
6630.9	0.107	0.014	0.041	0.116	0.011	0.043	$\uparrow$
6849.2	-0.330	< 0.001	0.002	-0.330	< 0.001	0.005	$\downarrow$
7052.7	-0.158	0.001	0.006	-0.165	0.001	0.007	$\downarrow$
7426	0.105	< 0.001	0.002	0.111	< 0.001	0.002	$\uparrow$
7465	0.068	0.003	0.012	0.077	0.001	0.007	$\uparrow$
7558.8	0.152	< 0.001	< 0.001	0.153	< 0.001	0.001	$\uparrow$
7754.5	0.179	< 0.001	0.001	0.170	< 0.001	0.004	$\uparrow$
8114.1	-0.107	0.018	0.048	-0.104	0.021	0.062	$\downarrow$
8324.8	-0.098	0.007	0.027	-0.096	0.010	0.041	$\downarrow$
8584.6	0.150	0.009	0.028	0.133	0.023	0.065	$\uparrow$
8911	0.147	0.006	0.023	0.139	0.012	0.043	$\uparrow$
10034.7	-0.215	< 0.001	< 0.001	-0.204	< 0.001	0.001	$\downarrow$
10487.2	0.076	0.003	0.014	0.074	0.006	0.026	$\uparrow$
12367.5	-0.104	0.001	0.006	-0.091	0.006	0.026	$\downarrow$
14576.6	0.130	0.002	0.009	0.142	0.001	0.007	$\uparrow$
15140.1	0.164	0.013	0.041	0.171	0.013	0.043	$\uparrow$
15745.3	0.208	0.006	0.024	0.222	0.006	0.026	$\uparrow$

Table 3: Regression coefficient estimates and p-values of the identified protein peaks that have significant dose responses for the Bangladesh data . Both unadjusted and adjusted results are provided. False discovery rates were set 0.05 and linear dose response models were fit.

	Unadjusted Analysis			Adjusted Analysis		
Protein Peaks (m/z)	Estimate	p-value	q-value	Estimate	p-value	q-value
3428.6	0.044	0.001	0.008	0.041	0.003	0.016
3669.7	-0.074	0.001	0.008	-0.080	0.001	0.008
4124.7	-0.060	0.002	0.010	-0.057	0.005	0.024
4277	-0.122	0.006	0.025	-0.124	0.007	0.033
4566.9	0.041	0.013	0.043	0.039	0.023	0.069
4640.1	0.044	0.013	0.043	0.040	0.032	0.087
5248.1	0.096	0.003	0.016	0.087	0.008	0.033
5577.5	-0.048	0.002	0.010	-0.051	0.001	0.011
5800.9	-0.089	0.001	0.007	-0.093	0.001	0.008
6849.2	-0.109	< 0.001	0.006	-0.112	< 0.001	0.007
7052.7	-0.056	0.001	0.007	-0.058	0.001	0.008
7426	0.033	< 0.001	0.006	0.035	< 0.001	0.007
7465	0.021	0.009	0.032	0.022	0.006	0.029
7558.8	0.053	< 0.001	< 0.001	0.053	< 0.001	0.001
7754.5	0.065	< 0.001	< 0.001	0.061	< 0.001	0.004
8911	0.053	0.004	0.017	0.049	0.011	0.043
10034.7	-0.073	< 0.001	< 0.001	-0.069	< 0.001	0.001
10487.2	0.029	0.001	0.008	0.028	0.003	0.016
12367.5	-0.032	0.004	0.017	-0.029	0.013	0.048
14576.6	0.049	0.001	0.007	0.050	0.001	0.008
15745.3	0.074	0.005	0.020	0.084	0.003	0.016
18011	0.030	0.008	0.032	0.032	0.008	0.033
19985.8	-0.020	0.003	0.016	-0.022	0.003	0.016

Table 4: Regression coefficient estimates, p-values, and q-values of super-proteins significantly associated with exposure status at the FDR = 0.05 level for both the crude analysis and the covariate adjusted analysis. The "+" indicates the combination of protein peaks into a super-protein.

	Unadjusted Analysis			Adjusted Analysis		
Super-Protein Peaks (m/z)	Estimate	p-value	q-value	Estimate	p-value	q-value
3005.4+3255+3763+4277	0.643	0.009	0.028	0.586	0.021	0.065
3428.6+5336.9+5905.1	-0.662	0.002	0.009	-0.551	0.012	0.042
4124.7+7177.6	0.557	0.002	0.009	0.529	0.005	0.032
4464.8+4566.9	-0.431	0.018	0.049	-0.396	0.037	0.09
4640.1	0.388	0.004	0.018	0.357	0.012	0.042
4676.3+5248.1	-0.543	0.004	0.017	-0.488	0.011	0.042
5577.5+5800.9+6849.2	0.838	< 0.001	0.002	0.843	< 0.001	0.002
7052.7	-0.453	0.001	0.008	-0.471	0.001	0.007
7426+7465	-0.664	< 0.001	0.002	-0.721	< 0.001	0.001
7558.8+7754.5	-0.811	< 0.001	< 0.001	-0.792	< 0.001	0.001
8114.1+8324.8	0.488	0.010	0.029	0.476	0.013	0.042
8584.6	0.359	0.009	0.028	0.318	0.023	0.065
8911	0.378	0.006	0.021	0.357	0.012	0.042
10034.7	-0.660	< 0.001	< 0.001	-0.625	< 0.001	< 0.001
12367.5	-0.444	0.001	0.008	-0.386	0.006	0.035
14576.6	0.426	0.002	0.009	0.463	0.001	0.007
15140.1+15745.3	0.503	0.008	0.028	0.531	0.008	0.040

Figure 1: Estimate of the harmonic component of the noise $H_i(t)$ for one subject in the range of 'm/z' between 7,800 Da and 8,000 Da.

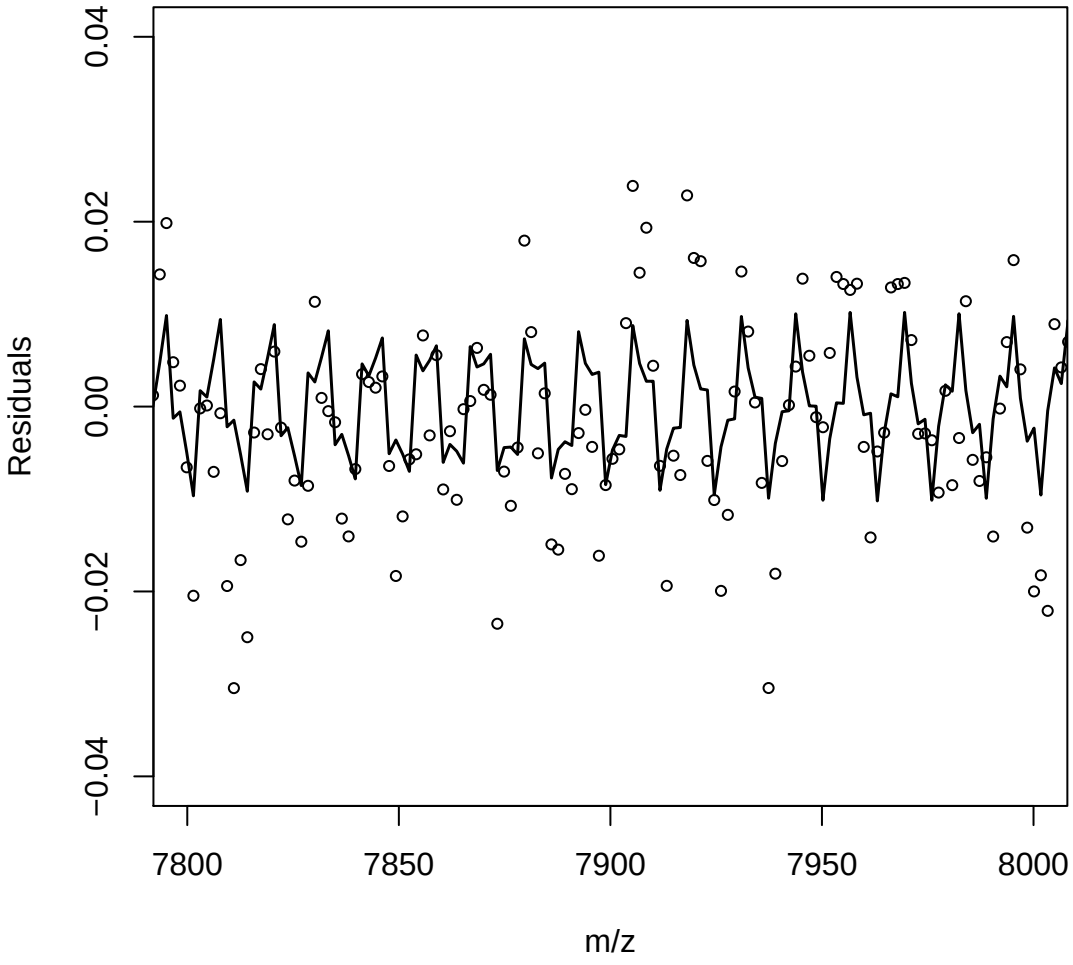

Figure 2: Estimate of the frequency of the harmonic component of the noise $H_i(t)$ for one subject in the range of 'm/z' between 1,700 Da and 3,700 Da.

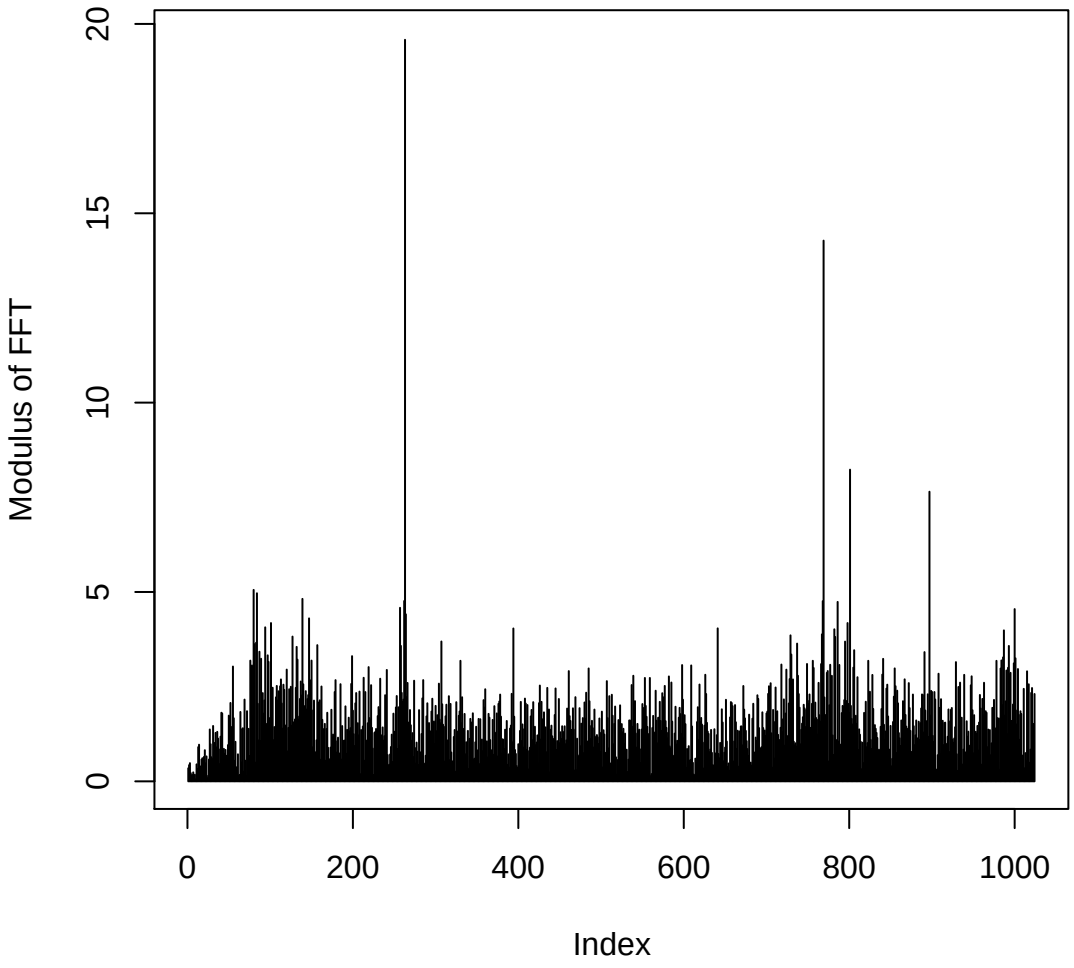

Figure 3: Estimates of the signal curves $S_i(t)$ in the range of 'm/z' between 8,800 Da and 9,000 Da for 10 randomly selected subjects: before alignment (left) and after alignment (right).

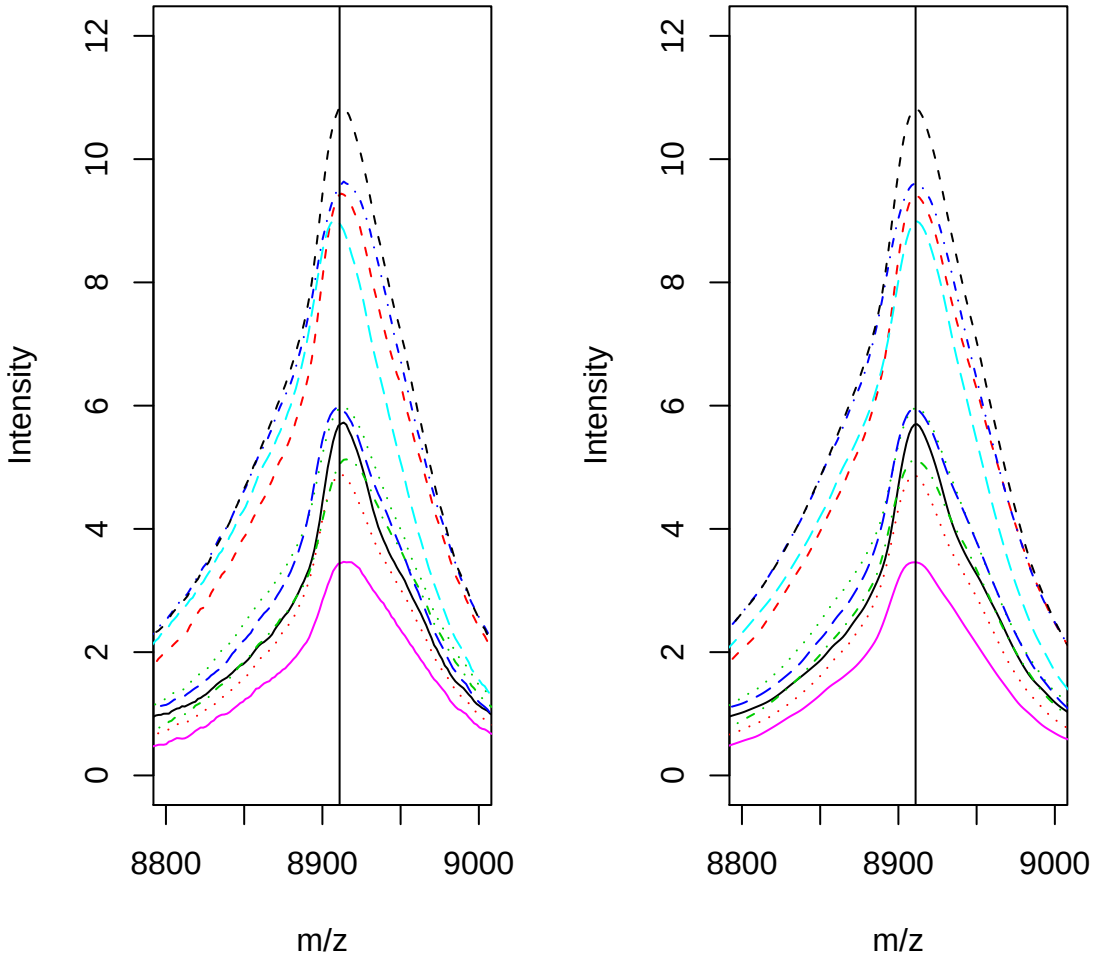

Figure 4: Heatmap of the correlation structure of the initial 77 protein peaks (left) and the 46 super-protein peaks (right). Highly correlated blocks of protein peaks are present in the protein peak heatmap and not as apparent among the super-proteins.