PROTEOMICS METABOLOMICS GENOMICS INFORMATICS

GLYILEVALCYSGLUGLNALASERLEUASPARG
CYSVALLYSPROLYSPHETYRTHRLEUHISLYS

A Highly Accurate Mass Profiling Approach to Protein Biomarker Discovery Using HPLC-Chip/ MS-Enabled ESI-TOF MS

Authors

Christine Miller and
Bryan Miller

Agilent Technologies,
Santa Clara, US

Abstract

LC/MS-based workflows play an important role in the field of biomarker discovery and validation. Classical shotgun proteomics approaches relying on data-dependent acquisition are generally acknowledged to provide confident identification of only a subset of the actual proteins present in the sample. This application note explores a protein profiling approach combined with differential analysis to highlight and identify putative biomarkers. The protein profiling approach comprises two steps: (1) Rapid differential expression analysis of samples using accurate mass ESI-TOF data, followed by (2) Profile-directed identification from MS/MS data of differentially-expressed putative markers. Agilent's 1200 Series HPLC-Chip/6210 TOF LC/MS system demonstrates the mass accuracy and resolution necessary for profile-directed biomarker discovery, as well as the ability to deal with the identification of low-abundance proteins in the presence of much higher-abundance proteins. The results of this study demonstrate that a profile-directed approach using the TOF profiling system is a powerful method for identifying low-level, differentially-expressed components within complex samples. The profiling approach allows for the efficient, targeted analysis of only differentially expressed features and thus makes this method a more effective, sensitive, and reliable alternative to traditional data-dependent MS/MS for biomarker discovery.

Introduction

Liquid chromatography and mass spectrometry (LC/MS) have become core technologies for protein identification and quantification, and many LC/MS workflows have been applied to proteomics research. While earlier proteomics research often was aimed at basic proteomic characterization, an increasing number of investigations are focused on protein biomarker discovery and validation. At present, no single LC/MS workflow has been adopted by the scientific community as the gold standard for protein biomarker discovery and validation. The common data-dependent strategy of biomarker identification by MS/MS (based on the most intense ions eluting over a specific time in full scan MS mode) typically provides data representing only a subset of the actual proteins present in a sample. Extensive fractionation of complex samples may be necessary to identify more proteins, significantly increasing the number of analyses required for an individual sample. These approaches have been somewhat effective; however, a significant concern with this approach is that lower level proteins and potential biomarkers may frequently be missed.

This application note explores a protein profiling approach combined with differential analysis to highlight and identify putative biomarkers. Agilent's 1200 Series HPLC-Chip/6210 TOF LC/MS system demonstrates the mass accuracy and resolution necessary for profile-directed biomarker discovery, as well as the ability to deal with challenges such as high sample throughput and identification of low-abundance proteins in the presence of much higher-abundance proteins; challenges encountered daily in many laboratories. The performance of new informatics tools for data extraction and differential analysis of the complex data sets inherent in proteomic profiling is illustrated in this context.

Experimental

Sample Preparation

For these experiments, *E. coli* lysate was used as a model representing a fairly complex sample. In order to mimic up- and down-regulation in complex samples, equivalent amounts of lysate were spiked with varying amounts of bovine proteins (bovine serum albumin and serotransferrin) for detection using the biomarker profile-directed approach. *E. coli* lysate (BioRad), bovine serum albumin and serotransferrin (Sigma) were digested with trypsin using a protocol based on 2,2,2-trifluoroethanol as the denaturant. The digests were aliquoted, dried, and stored frozen until use. *E. coli* digests were spiked with the two bovine protein digests at different levels. The samples were prepared so that a 2-µL injection would result in the amounts on-column shown below in Table 1.

Sample	*E. coli* lysate (ng total protein)	BSA (fmol)	Serotransferrin (fmol)
Control	400	100	200
Sample A	400	200	100
Sample B	400	400	50

Table 1. Amount injected on-column for the samples in the model set.

Instrumentation and Software

The protein profiling approach comprised two steps:

- Rapid differential expression analysis of samples using accurate mass ESI-TOF data, followed by
- Profile-directed identification from MS/MS data of differentially-expressed putative markers

All profiling experiments were performed on an Agilent 1200 Series HPLC-Chip/MS system interfaced to an Agilent 6210 TOF mass spectrometer. The HPLC-Chip configuration included a 40-nL enrichment column and 150 mm x 75 µm analytical column packed with Zorbax 300SB-C18, 5 µm material. A 100-minute long gradient method was used. The solvents employed were: A. 0.1% formic acid in water and B. 90% acetonitrile with 0.1% formic acid. After initial loading at 3% B, the gradient stepped to 8% B at 0.5 minutes, then 45% B at 85 minutes, 80% B from 90 to 92 minutes and back to 3% B at 92.01 minutes. The analytical flow rate was 300 nL/min and the sample was loaded at 4 µL/min.

The LC system consisted of a nanoflow pump, a capillary pump for sample loading and a microwell-plate autosampler with cooler. Complete system control was accomplished using Agilent TOF LC/MS software. The software includes internal reference mass correction (IRMC), which automatically corrects mass assignments as mass spectra are written to disk, thus simplifying and speeding later data processing. During data acquisition, code associated with the TOF device driver locates reference ion(s) in each mass spectrum, and uses the known mass positions to determine new A and t_0 coefficients ("IRM coefficients") for the base equation used for time-to-mass conversion. A moving average of these new coefficients is saved to disk together with the raw

time/abundance data. When spectra are read from raw data by the data analysis code, if present the IRM coefficients are used instead of base coefficients to convert from time to mass.

Accurate mass LC/MS data were extracted and evaluated using a specialized molecular feature extraction algorithm and Mass Profiler software. Targeted LC/MS/MS analyses were performed using an HPLC-Chip/MS system interfaced to an Agilent 6330 Ion Trap mass spectrometer. Peptides were identified using Spectrum Mill MS Proteomics Workbench software with the SwissProt protein database.

Results and Discussion

Sample complexity and low-level peptides

Biological samples are frequently very complex, and protein levels can vary greatly. To address this complexity, we employed a combination of high chromatographic performance and extreme sensitivity by using an HPLC-Chip/6210 TOF MS system to resolve and detect proteins. Data from one sample showed *E. coli* lysate spiked with BSA and serotransferrin generated a total ion chromatogram (TIC) that demonstrated a highly complex sample (Figure 1).

Further analysis was performed on a single peptide (m/z = 504.2506) by generating an extracted ion chromatogram (EIC) using a narrow mass window of +/– 1.9 ppm. The EIC showed multiple peaks, further indication of the complexity of the sample (Figure 2). A single peak at 9.2 min was identified from the EIC and the full mass spectrum was obtained for this time point (Figure 3).

The spectrum showed that the peptide at m/z = 504.2506 is relatively low in abundance compared to other ions. These data illustrate a situation that can occur when searching for biomarkers. The combination of the HPLC-Chip with the 6210 TOF LC/MS system was able to detect a low-level peptide within a highly complex sample. More importantly, with typical data-dependent acquisition this particular low-abundance peptide would likely not be selected for MS/MS analysis because of the presence of more abundant ions.

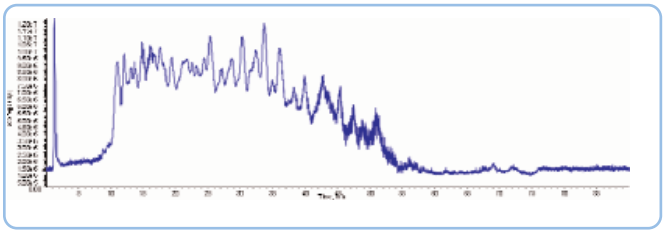

Figure 1. Total ion chromatogram (TIC) for Sample B.

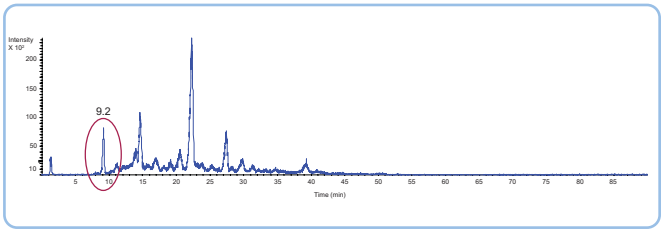

Figure 2. Extracted ion chromatogram (EIC) for m/z 504.2507 +/– 1.9 ppm.

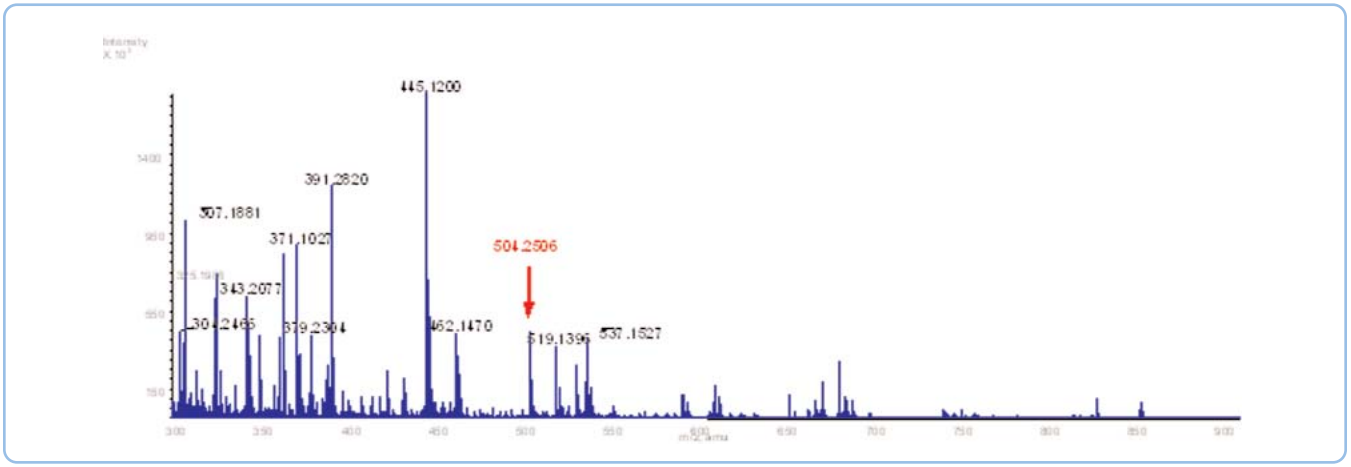

Figure 3. Full MS at 9.2 minutes. Peptide with m/z 504.25 would most likely have not been chosen for MS/MS using regular data-dependent strategies because more intense peaks are present at this time point.

Feature extraction

To identify all components—including low-level components—within each sample, TOF LC/MS accurate mass data were extracted and evaluated using a specialized molecular feature extraction algorithm. The algorithm located the groups of co-variant ions in a chromatogram. Each of these groups represented a unique compound. Thus, the algorithm identified all the components in a chromatogram, instead of just identifying chromatographic peaks, which may conceal multiple components. This facilitated very effective removal of chemical background data. Next, peaks were clustered in retention time (RT) and *m/z* to form 3-D peaks. The 3-D peaks were centroided and a peak volume determined for each peak. Related 3-D peaks (isotopes, adducts, dimers, trimers, multiple charge states) were combined and assigned a neutral mass and total volume.

The molecular feature extraction algorithm effectively removed noise (demonstrated as streaks in the contour plot, Figure 4A) and extracted even low-level peptides (Figure 4B). For this study, the algorithm was set to extract only multiply-charged features with signal-to-noise ratio ≥4. The extracted information for all replicates of each sample were then combined and analyzed for differential features against the control sample.

Differential expression analysis

In biomarker discovery, statistical analysis can be employed to aid in the segregation of experimental and within-population variations from cross-population expression level changes. For statistical analysis in these experiments, Mass Profiler software was used, which enables retention time and *m/z* alignment for features across samples, response normalization, and t-test statistics for identification of significant differences between samples.

After feature extraction, all multiply-charged features that occurred in at least one data file (out of 20 files: 10 replicates for control, 10 replicates for Sample A or Sample B) were plotted. Results for Sample B and control sample replicates are shown in Figure 5.

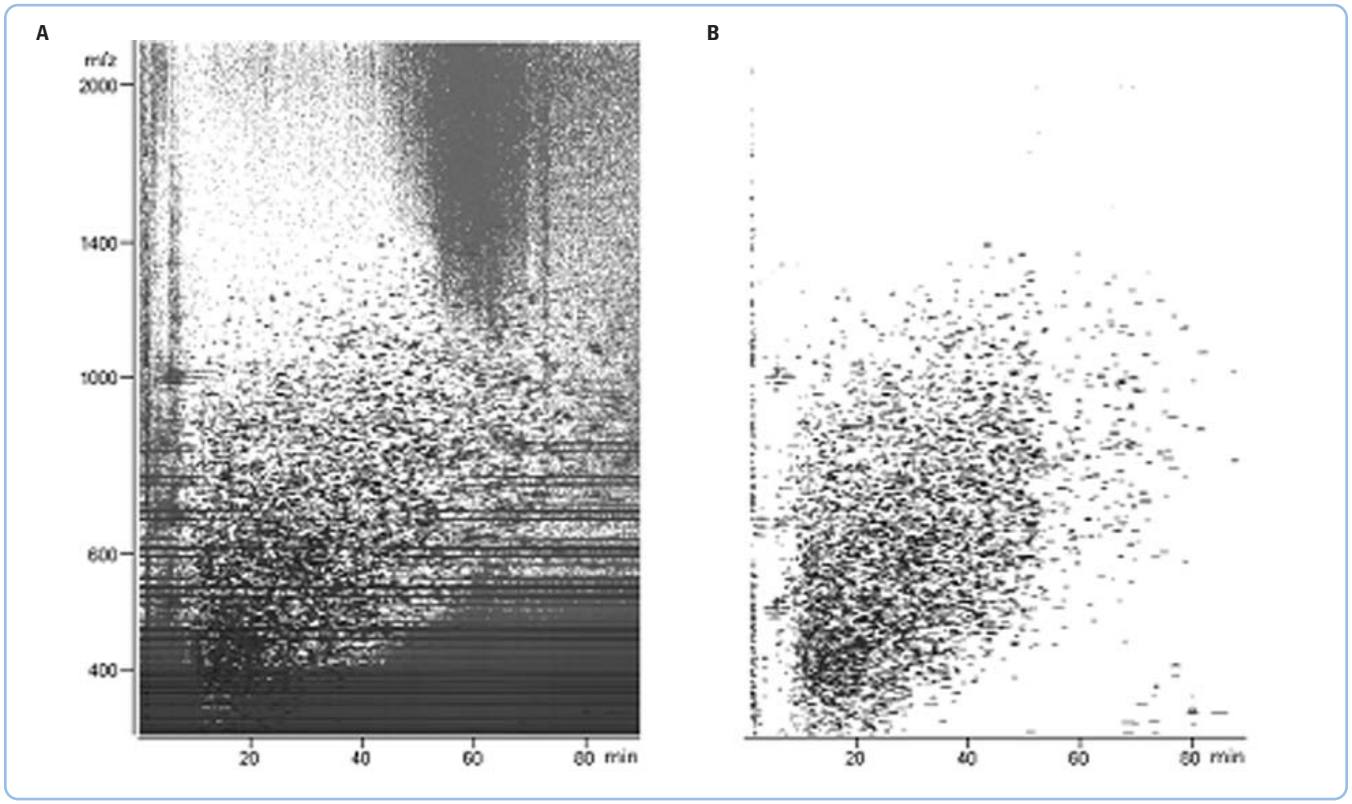

Figure 4. Contour plots for one replicate (out of 10) of Sample B showing the effectiveness of feature extraction. Figure 4A shows data before feature extraction. Figure 4B shows data after feature extraction.

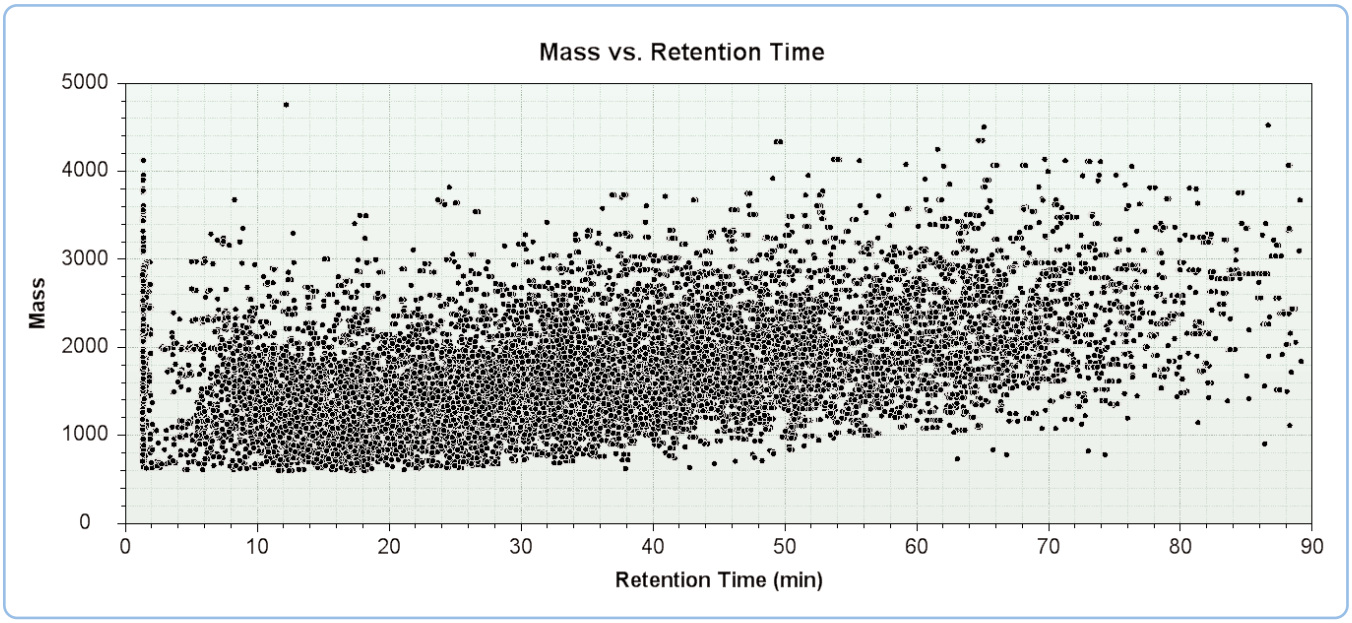

Figure 5. All multiply-charged features that occurred in at least one data file (out of 20) with no differential filters applied produced 22,011 features for Sample B and control sample replicates.

Results filters were then used to reduce the number of differential features (peptides in this case) to be investigated. The comparison of Sample A with control replicates shows all multiply-charged differential features found in 100% of the replicates with a minimum differential score of 85 and a minimum $\log_2$ ratio of 0.8 (indicating a roughly 2-fold difference in abundance). These criteria identified 22 differential features in Sample A at the expected 2x and 0.5x ratios for BSA and serotransferrin, respectively (Figure 6). For the comparison of Sample B with control replicates, the criteria were all multiply-charged differential features found in 80% of the replicates with a minimum differential score of 85 and a minimum $\log_2$ ratio of 1.5 (indicating a roughly 4-fold difference in abundance). A total of 19 differential features were found in Sample B at the expected 4x and 0.25x ratios for BSA and serotransferrin, respectively (Figure 7, see next page). These results strongly demonstrate the power of the mass profiling approach for discovering potential biomarkers and determining their relative abundances within complex samples.

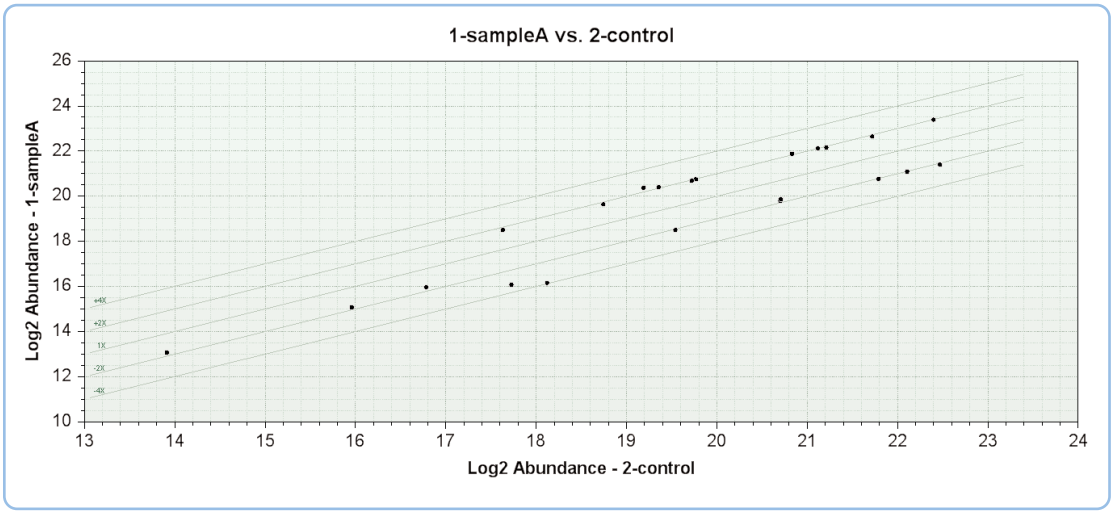

Figure 6. Mass Profiler software identified Sample A features with relative 2-fold in abundance compared to control sample.

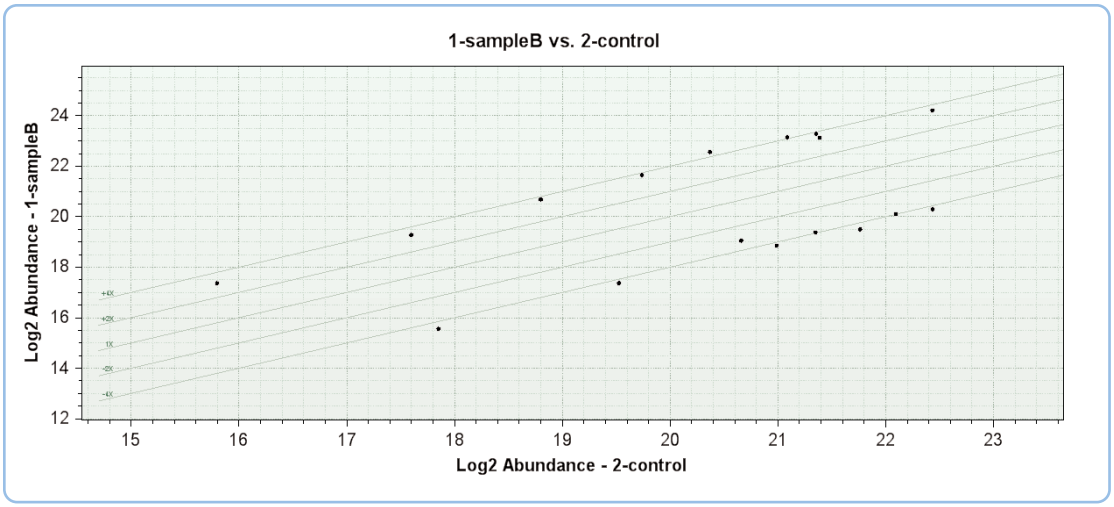

Figure 7. Mass Profiler identified Sample B features with relative 4-fold differences in abundance compared to control sample.

To further demonstrate the effectiveness of the mass profiling analysis, data generated by the Mass Profiler software were exported to Agilent's GeneSpring GX software, which offers advanced visualization and statistical tools for differential analysis of multiple samples. A principal component analysis (PCA) was performed, resulting in well-defined clustering of the 5 technical replicates of each sample for one day's experiment (Figure 8).

Protein identification

As the second step in this approach for the identification of biomarkers, a targeted identification of the differential features was performed by importing the TOF differential features' mass data as an "include" mass list into the instrument control software of an Agilent 6330 ion trap LC/MS, and then reanalyzing Samples A and B using the same LC conditions. Peptides were identified using Spectrum Mill software and SwissProt as the protein database. The MS/MS data from this targeted identification method resulted in the correct identification of the BSA and serotransferrin peptides in both Samples A and B (data not shown). These results demonstrate the efficiency and effectiveness of using a TOF mass profiling approach to direct MS/MS data acquisition in order to target acquisition of spectra for specific differentially expressed peptides within a complex sample.

Figure 8. GeneSpring GX principal component analysis showing clustering of five technical replicates for Sample A (red), Sample B (yellow), and control (blue).

Data reproducibility

Effective biomarker detection and identification requires the ability of a system to reproducibly generate reliable data. To determine the quality of data generated by HPLC-Chip/TOF MS profiling, cross-sample response relative standard deviation (RSD) values were calculated. As shown in Table 2 for peptides identified for serotransferrin in Sample B, retention time RSD values for 19–20 replicates were approximately 0.3%. The most variability was exhibited at the RT of 9.13 min (corresponding to m/z = 504.2506); this particular peak was located at the beginning of the elution of a large collection of peptides and therefore would be more likely to show such variability (see Figure 2 for the TIC).

Mass RSD values were also calculated (Table 2). Impressively, the standard deviation (SD) values for all masses were ≤2.0 mDa for 19–20 replicates. These values translate to SD values of 0.99–1.42 ppm, demonstrating the exceptional mass accuracy ability of the 6210 TOF LC/MS.

Using Mass Profiler data, the reproducibility of serotransferrin peptide abundances within Sample B and control samples was also determined (Table 3). Within the control, most RSD values for the average abundances were ≤10%. For Sample B, most values were ≤18%.

These results show that the mass profiling software effectively compensates for cross-analysis variations in RT, measured masses, and abundances to enable effective biomarker detection.

Number of replicates	RT (min)	% RSD of RT	Mass	SD of Mass (mDa)	SD of Mass (ppm)
20	9.13	1.02	1509.7273	1.6	1.05
19	20.10	0.32	2016.9039	2.0	0.99
19	20.52	0.33	1310.6421	1.3	0.99
19	27.74	0.29	1388.6666	1.6	1.15
20	28.35	0.31	918.5474	1.3	1.42

Table 2. RT and mass reproducibility for Sample B and corresponding control samples.

		Control		Sample B	
Identified peptide	RT	Average abundance	% RSD	Average Abundance	% RSD
(R)KPVTDAENCHLAR	9.22	1655910	24.03	814822*	19.04
DQTVIQNTDGNNEQWQK	20.09	2080001	5.33	467082	9.83
ELPDPQESIQR	20.52	4471566	9.20	1129758	12.73
TSDANINWNNKL	27.74	2671411	7.52	682671	17.45
GYLAVAVVK	28.35	5674326	7.85	1290613	6.06

Table 3. Reproducibility of response for Sample B and control samples.

*** The peptide (R)KPVTDAENCHLAR at m/z 504.25 showed a clear difference in the Day 1 versus Day 2 samples due to coeluting component(s) in Day 1 replicates. Therefore, for Sample B m/z 504.25 results, only the Day 2 replicates were used.**

Conclusions

Effective biomarker detection and identification requires sensitivity, accuracy, and reproducibility. In this study, a two-step profiling approach is described that demonstrates all these attributes. The Agilent HPLC-Chip/6210 TOF MS system provides sensitivity and high mass accuracy for low-level peak detection and reliable results. In addition, the system enables highly reproducible mass profiling, which is a prerequisite for reliable and comparable sample studies. Methods developed on the Agilent 1200 Series HPLC-Chip/6210 TOF MS profiling system are directly transferable to the HPLC-Chip/6330 Ion Trap MS system for targeted protein identification using Spectrum Mill protein identification software. The results of this study demonstrate that a profile-directed approach using the TOF profiling system is a powerful method for identifying low-level, differentially-expressed components within complex samples. The profiling approach allows for the efficient, targeted analysis of only differentially expressed features and thus makes this method a more effective, sensitive, and reliable alternative to traditional data-dependent MS/MS for biomarker discovery.

About Agilent's Integrated Biology Solutions

Agilent Technologies is a leading supplier of life science research systems that enable scientists to understand complex biological processes, determine disease mechanisms, and speed drug discovery. Engineered for sensitivity, reproducibility, and workflow productivity, Agilent's integrated biology solutions include instrumentation, microfluidics, software, microarrays, consumables, and services for genomics, proteomics, and metabolomics applications.

For more information

Learn more: www.agilent.com/chem/lcms

Buy online: www.agilent.com/chem/store

Find an Agilent customer center in your country:
www.agilent.com/chem/contactus

U.S. and Canada
1-800-227-9770
agilent_inquiries@agilent.com

Europe
info_agilent@agilent.com

Asia Pacific
adinquiry_aplsca@agilent.com

Printed in the U.S.A. April 19, 2006
5989-5083EN