Chapter V.

Factorial design of serum protein profiling by LC-MS

Manuscript in preparation

1. Introduction

Biomarker discovery requires complex procedures that involve close collaboration between medical, analytical and computational sciences (Figure 1). Biofluids that are easily accessible, such as serum, are generally used for biomarker discovery, but they are extremely complex. During the process of biomarker discovery doctors working in medical sciences are responsible for diagnosis and patient classification, sample collection and storage as well as documentation. Scientists working in analytical science are responsible for sample preparation and analysis (e.g. by LC-MS), and must strive to minimize the variation introduced by these analytical steps. Scientists working in computational science process and analyze the generated data by extracting information related to the abundance of the various proteins and peptides in different samples and by using sophisticated statistical methods to find significant abundance differences between preclassified samples. Due to the complexity of the overall process, it is possible that certain experimental parameters may have an effect on the abundance of the measured protein and peptides thus affecting the outcome of the statistical analysis with the risk of creating false positives with respect to biomarker candidates or of missing potentially relevant candidates. It is thus important to determine the various factors, which affect the measured peptide profiles, in the case of trypsin-digested serum. This will allow the experimentalist to keep close control over the most important factors and to relax the stringency of the experimental protocol for those factors, which have little effect on the outcome of the
analyses. There are preanalytical factors that cannot be changed retrospectively, for example, when working with already existing biobanks (e.g. the hemolysis level) but that may well affect the protein composition. These factors may be used as entry criteria and can help to remove samples that do not meet these criteria prior to analysis with the added benefit that non-class-specific variance will be reduced.

To assess the importance of a selected number of factors, we used a Factorial Design approach on serum samples that were depleted of the 6 most abundant proteins. As our analytical procedure by cap-LC-MS is rather time-consuming, we prioritized seven factors that were analyzed at two levels each. The analyzed factors were:

- type of blood collection tube (BD 367784, BD 368430, abrv.: Blo)
- hemolysis level (low and high, abrv.: Hem)
- clotting time (2 and 6 hours, abrv.: Clo)
- number of freeze-thaw cycles (1 and 3 cycles, abrv.: Fre)
- trypsin to protein ratio (1:20 and 1:100, abrv.: Try)
- stopping of trypsin digestion with acid (yes and no, abrv.: Sto)
- residence time in the autosampler (0 and 30 days, abrv.: Sta) at 4°C

A full factorial design would require $2^7=128$ analyses and would give information on the main effect of each factor and on all interactions between different factors from the second till the seventh order. In order to reduce the total analysis time, we have chosen to perform a two-level fractional factorial design with resolution VI (a so-called $2^{7-3}_{VI}$ design). This design requires only 16 analyses while resolution VI means that only the main effect of each factor can be evaluated directly, since these effects are not confounded by other effects or by two-factor interactions. Two-factor interactions are confounded in triads and thus in this design it is not possible to distinguish which of them is responsible for the measured effect. Higher order interactions between factors are generally negligible and thus confounding them with main and two-factor interactions will not affect the overall evaluation of our design. Table 1 contains the main effect and two-factor confounding pattern of the $2^{7-3}_{VI}$ design as used in this study. In order to estimate the data variances not related to the studied factors we have performed 3 repetitions of one analysis leading to total of 19 analyses [1, 2].

Automatic processing of ion-trap LC-MS data resulted in 8 000 - 10 000 aligned features or measured variables. The evaluation of this multivariate data set was done by eliminating one factor at a time to produce the average highest variance in the dataset.
Figure 1. Schematic view of the collaborative environment of a biomarker discovery project. The final result may be affected by variable factors, as exemplified in red.

Table 1. Main effect and two-factors confounding pattern of a $2^{-3}_{7}$ fractional factorial design as used in this study.

<table>
<thead>
<tr>
<th>Term</th>
<th>Confounded with</th>
<th>Confounded with</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blo</td>
<td>Clo*Try</td>
<td>Sto*Sta</td>
</tr>
<tr>
<td>Hem</td>
<td>Clo*Try</td>
<td>Sto*Sta</td>
</tr>
<tr>
<td>Clo</td>
<td>Clo*Try</td>
<td>Sto*Sta</td>
</tr>
<tr>
<td>Fre</td>
<td>Clo*Sta</td>
<td>Try*Sto</td>
</tr>
<tr>
<td>Try</td>
<td>Clo*Sta</td>
<td>Try*Sto</td>
</tr>
<tr>
<td>Sto</td>
<td>Clo*Sta</td>
<td>Try*Sto</td>
</tr>
<tr>
<td>Sta</td>
<td>Clo*Sta</td>
<td>Try*Sto</td>
</tr>
<tr>
<td>Blo*Hem</td>
<td>Hem*Try</td>
<td>Fre*Sta</td>
</tr>
<tr>
<td>Blo*Clo</td>
<td>Hem*Try</td>
<td>Fre*Sta</td>
</tr>
<tr>
<td>Blo*Fre</td>
<td>Clo*Sta</td>
<td>Try*Sto</td>
</tr>
<tr>
<td>Blo*Try</td>
<td>Hem*Clo</td>
<td>Fre*Sto</td>
</tr>
<tr>
<td>Blo*Sto</td>
<td>Hem*Sta</td>
<td>Fre*Try</td>
</tr>
<tr>
<td>Blo*Sta</td>
<td>Hem*Sta</td>
<td>Clo*Fre</td>
</tr>
<tr>
<td>Hem*Fre</td>
<td>Clo*Sta</td>
<td>Try*Sta</td>
</tr>
</tbody>
</table>
2. Methods

2.1. Factorial design

Different handling of serum samples can influence the resulting LC-MS data reflecting the peptide/protein composition of the samples. Based on previous experience with the analysis of serum samples we prioritized the following seven factors and defined the levels to be investigated: tubes for blood collection, high/low hemolysis level, different clotting times (2 or 6h), freeze-thaw cycles (1 or 3), digestion with trypsin at different enzyme to substrate ratios (1:20 or 1:100), stopping of trypsin digestion with acid or not, stability of the samples at 4°C prior to LC-MS analysis. Samples were analyzed according to the factorial design scheme given in Table 1.

Table 2. $2^{7-3}$ Fractional design pattern.

<table>
<thead>
<tr>
<th>Experiment name</th>
<th>Run order</th>
<th>Blood collection tube</th>
<th>Hemolysis</th>
<th>Clotting time</th>
<th>Freeze-thaw cycles</th>
<th>Trypsin digestion</th>
<th>Stopping trypsin</th>
<th>Stability sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>11</td>
<td>BD368430</td>
<td>Low</td>
<td>2 hours</td>
<td>1 cycle</td>
<td>1:20</td>
<td>Yes</td>
<td>0 days</td>
</tr>
<tr>
<td>N2</td>
<td>3</td>
<td>BD367784</td>
<td>Low</td>
<td>2 hours</td>
<td>1 cycle</td>
<td>1:100</td>
<td>Yes</td>
<td>30 days</td>
</tr>
<tr>
<td>N3</td>
<td>9</td>
<td>BD368430</td>
<td>High</td>
<td>2 hours</td>
<td>1 cycle</td>
<td>1:100</td>
<td>No</td>
<td>0 days</td>
</tr>
<tr>
<td>N4</td>
<td>17</td>
<td>BD367784</td>
<td>High</td>
<td>2 hours</td>
<td>1 cycle</td>
<td>1:20</td>
<td>No</td>
<td>30 days</td>
</tr>
<tr>
<td>N5</td>
<td>18</td>
<td>BD368430</td>
<td>Low</td>
<td>6 hours</td>
<td>1 cycle</td>
<td>1:100</td>
<td>No</td>
<td>30 days</td>
</tr>
<tr>
<td>N6</td>
<td>2</td>
<td>BD367784</td>
<td>Low</td>
<td>6 hours</td>
<td>1 cycle</td>
<td>1:20</td>
<td>No</td>
<td>0 days</td>
</tr>
<tr>
<td>N7</td>
<td>10</td>
<td>BD368430</td>
<td>High</td>
<td>6 hours</td>
<td>1 cycle</td>
<td>1:20</td>
<td>Yes</td>
<td>30 days</td>
</tr>
<tr>
<td>N8</td>
<td>16</td>
<td>BD367784</td>
<td>High</td>
<td>6 hours</td>
<td>1 cycle</td>
<td>1:100</td>
<td>Yes</td>
<td>0 days</td>
</tr>
<tr>
<td>N9</td>
<td>8</td>
<td>BD368430</td>
<td>Low</td>
<td>2 hours</td>
<td>3 cycles</td>
<td>1:20</td>
<td>No</td>
<td>30 days</td>
</tr>
<tr>
<td>N10</td>
<td>15</td>
<td>BD367784</td>
<td>Low</td>
<td>2 hours</td>
<td>3 cycles</td>
<td>1:100</td>
<td>No</td>
<td>0 days</td>
</tr>
<tr>
<td>N11</td>
<td>13</td>
<td>BD368430</td>
<td>High</td>
<td>2 hours</td>
<td>3 cycles</td>
<td>1:100</td>
<td>Yes</td>
<td>30 days</td>
</tr>
<tr>
<td>N12</td>
<td>7</td>
<td>BD367784</td>
<td>High</td>
<td>2 hours</td>
<td>3 cycles</td>
<td>1:20</td>
<td>Yes</td>
<td>0 days</td>
</tr>
<tr>
<td>N13</td>
<td>12</td>
<td>BD368430</td>
<td>Low</td>
<td>6 hours</td>
<td>3 cycles</td>
<td>1:100</td>
<td>Yes</td>
<td>0 days</td>
</tr>
<tr>
<td>N14</td>
<td>6</td>
<td>BD367784</td>
<td>Low</td>
<td>6 hours</td>
<td>3 cycles</td>
<td>1:20</td>
<td>Yes</td>
<td>30 days</td>
</tr>
<tr>
<td>N15</td>
<td>5</td>
<td>BD368430</td>
<td>High</td>
<td>6 hours</td>
<td>3 cycles</td>
<td>1:20</td>
<td>No</td>
<td>0 days</td>
</tr>
<tr>
<td>N16</td>
<td>19</td>
<td>BD367784</td>
<td>High</td>
<td>6 hours</td>
<td>3 cycles</td>
<td>1:100</td>
<td>No</td>
<td>30 days</td>
</tr>
<tr>
<td>N17</td>
<td>14</td>
<td>BD368430</td>
<td>Low</td>
<td>2 hours</td>
<td>1 cycle</td>
<td>1:20</td>
<td>Yes</td>
<td>0 days</td>
</tr>
<tr>
<td>N18</td>
<td>1</td>
<td>BD368430</td>
<td>Low</td>
<td>2 hours</td>
<td>1 cycle</td>
<td>1:20</td>
<td>Yes</td>
<td>0 days</td>
</tr>
<tr>
<td>N19</td>
<td>4</td>
<td>BD368430</td>
<td>Low</td>
<td>2 hours</td>
<td>1 cycle</td>
<td>1:20</td>
<td>Yes</td>
<td>0 days</td>
</tr>
</tbody>
</table>

The samples were injected into the LC-MS according to the order shown in column 2 “Run order”.

2.1.1. Description of samples

Serum samples were obtained from 2 healthy volunteers (Fact. Design #1, female and Fact. Design #2, male) and provided by the Department of Gynecological Oncology (University Medical Center Groningen, The Netherlands). They were stored at -80°C in aliquots until analysis.
2.1.2. Blood collection tubes

Two kinds of tubes were used for the collection of blood: BD368430 (a “red stopper clotting tube”, which is a glass tube with a siliconized inner wall used to avoid retention of red blood cell on the walls of the tube) and BD367784 (a “gel tube”, which is a glass tube with a separation gel and micronized silica to accelerate clotting). During centrifugation, the polymer gel moves up the inner wall of this tube forming a barrier between the supernatant (serum) and sediment (blood clot and cells)). These were selected, since they were both used at the University Medical Center Groningen, The Netherlands to acquire the available serum biobank from cervical cancer patients.

2.1.3. Variation of clotting time

Serum samples were allowed to clot for 2 or 6 hours at room temperature before centrifugation to obtain serum.

2.1.4. Stability of trypsin-digested serum samples in the autosampler prior to LC-MS analysis

Stability of the trypsin-digested serum samples was evaluated by keeping them for 30 days at 4ºC in the autosampler. Samples labeled “0 days samples” were injected directly after thawing (storage at -80ºC).

2.1.5. Level of hemolysis

To simulate a high level of hemolysis, a lysate of red blood cells was added to the serum prior to depletion. Red blood cells were collected according to the following protocol: 0.5 mL lysis buffer (NH₄Cl 155 mmol/L, EDTA 0.1mmol/L) was added to 0.5 mL fresh blood and centrifuged for 20 minutes at 2000 rpm. 4 mL of lysis buffer was added to the pellet and incubated overnight at 4ºC. The next day the lysate was filtered through spin filters (0.22µm; # 5185-5990, Agilent) at 13000 rpm. Aliquots of the filtrate were stored at -80ºC.

The amount of lysed red blood cells that should be added to serum to mimic a high level of hemolysis was determined by the addition of different amounts (1, 3, 5, 7 and 10 µL) of red blood cell lysate to 20 µL serum that was immediately diluted with ice-cold water to a total volume of 60 µL followed by centrifugation at 13000 rpm for 30 min at 4ºC. Another 15 µL of ice-cold water were added to the supernatant and the absorbance was measured at 340, 380, 415 and 450 nm (Biowave S2100 UV/Vis Diode Array Spectrophotometer (Biochrom Ltd, Cambridge, UK)). A calibration line with respect to hemoglobin (Hb) was obtained using the formula: Hb [g/L] = (167.2 * A₄₁₅ – 83.6 * A₄₅₀/380 – 83.6 * A₄₅₀) / 1000 [4]. The absorbance of a serum sample from the serum bank that appeared to be very red, indicating a high level of hemolysis, was measured and the amount (in µL) of red blood cells to reach the corresponding Hb level was calculated from the calibration curve. To determine the concentration of hemoglobin that corresponded to a given volume of lysed red
blood cells, a calibration curve of hemoglobin (Hemoglobin human, Sigma, #9008-02-0) was made. It was calculated that 4 µL red blood cell (containing 6.68 µg hemoglobin) should be added before depletion to the serum to simulate a high level of hemolysis as observed in clinical samples from the biobank.

2.1.6. Preparation of serum samples

80 µL (80% of the total amount (20 µL of crude serum mixed with 80 µL of buffer A (Agilent)) of diluted crude serum was injected on a Multiple Affinity Removal column (Agilent, 4.6 x 50 mm, # 5185-5984) after filtration through 0.22 µm spin filters (# 5185-5990) at 13000g and 4°C for 10 min to remove particulates. The removal of abundant proteins was performed on a LaChrom HPLC System (Merck Hitachi, www.merck.com) with detection at 280 nm using the following timetable: 0-9 min, 100% buffer A (0.25 mL/min); 9.0-9.1 min, linear gradient 0-100 B % (1 mL/min), 9.1-12.5 min, 100% buffer B (1 mL/min); 12.5-12.6 min, linear gradient 100-0% buffer B (1 mL/min); 12.6-20 min, 100% buffer A (1 mL/min). The flow-through fraction (depleted serum collected between 2-6 min) of a total volume of appr. 1 mL was collected [3].

Protein concentrations were determined with the Micro BCA™ Protein assay reagent kit (www.piercenet.com) and calculated for an average protein molecular weight of 50 kDa. BSA was used as the calibration standard.

2.1.7. Digestion of serum samples

Trypsin (sequencing grade modified trypsin, Promega, # V5111, USA) in ratios 1:20 or 1:100 wt/wt (enzyme to total protein in depleted serum) was used for digestion at 37ºC at 450rpm (Eppendorf thermomixer, overnight).

2.1.8. Stopping of trypsin digestion (according to Table 1)

To stop the reaction with trypsin, formic acid was added after overnight digestion to reach a final concentration of 0.5% (v/v).

2.1.9. Freeze-thaw cycles

One or three freeze-thaw cycles (-80°C/room temperature) were include for comparative analysis (according to Table 1).

2.2. Data processing

For processing and multivariate statistical analysis of the original Bruker Daltonics HPLC-MS data, the files were converted into ASCII-format with the Bruker data analysis software LC/MSD Trap, version 3.3 (build 146) (Bruker Daltonics, Bremen, Germany) and saved in centroid mode. The time alignment algorithm was written in C++ using Microsoft Visual Studio [ver. 8.0.50727.762 (SP.050-727-7600), Redmont, WA, USA]. For further data analysis, Matlab [version 7.4.0.287 (R2007a), Mathworks, Natick, Massachusetts, USA] was used.
Centroid data were smoothed and reduced using a normalized two-dimensional Gaussian filter, with rounding the nominal m/z ratios to 1 m/z (the original data had a resolution of 0.1 m/z). In the retention time dimension no data reduction was performed. This meshing procedure reduced the number of available data points by roughly a factor 10 and corrected for shifting m/z values as a result of different loadings of the ion-trap during elution of abundant peptides, a phenomenon that is common for ion-trap mass spectrometers [5,6]. After meshing the data files of all chromatograms, they were time-aligned (warped) to a reference data file using Correlation Optimized Warping (COW) [7] based on peak lists obtained from the chromatograms. This time alignment algorithm optimizes time shifts using the overlapping between the extracted peaks using 2-dimensional extent of the peaks (retention time and m/z value) in the two chromatograms. Thus all chromatograms were aligned to each other in a pair-wise manner starting with randomly selected chromatogram as reference. The accuracy of correcting retention time shifts by time alignment was manually checked by visualization the 4 most intensive peaks for 10 equal time segments between 60-155 min before and after time alignment.

A modified M-N rule was applied for peak detection on meshed data with data reduction 1:10 in m/z resulting rounded integer m/z in the mass spectra by first calculating a median local baseline using the sliding window technique separately for each rounded m/z trace. A median window size of 1200 data points, corresponding to 20.17 min, was used with a moving rate of 10 points and a minimum median value of 200 counts. According to the M-N rule, a threshold of M-times the local baseline was used and a peak was assigned if, within one m/z trace, the signal exceeded this threshold for at least N consecutive points [8]. For each detected peak the m/z value, the mean retention times of the three highest measured intensities within the same peak reduced by the local baseline were stored in a peak list created for every chromatogram.

We used a similar approach as Radulovic et al. [8] to obtain optimal settings for M and N. Different values for M (1.5-4) and N (4-8) were applied to two blank LC-MS runs and two LC-MS runs of depleted, trypsin-digested serum samples. Settings were used at which the ratio between the number of peaks between 60 and 155 min in the samples relative to the blank chromatograms was highest and at which a minimal number of peaks was extracted from the noise in the blank chromatogram (M = 2 and N = 5 in our case).

In order to combine the peak lists from different samples into a common peak matrix, one-dimensional peak matching was performed by using the sliding window technique, in which the same m/z traces were evaluated for peaks that are proximate in time (step size 0.1 min; search window 1.0 min; maximal accepted standard deviation for all retention times within a group of
matched peaks 0.75 min). Missing peak allocation was performed by extracting the background subtracted local signal of the given m/z trace at the average retention time corresponded to the chromatograms where that peak was present. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak(row)-sample(column)-intensity(value) matrix. This peak matrix was used for analysis. All data preprocessing work was done on a personal computer equipped with a dual core +3800 MHz AMD 64 X2 processor equipped with 4 GB of RAM.

2.3. Statistical analysis

Multiple Linear Regression was performed using the MODDE software (ver 7.0.0.1) from Umetrics (Kinnelon, NJ, USA) (Umetrics INC, USA, ver 7.0.0.4). The evaluation of aligned peak matrices was performed using a stepwise method. First the aligned peak matrix was mean centered and normalized with respect to the standard variation for each peak (row) and thus relative standard deviation was obtained (RSD). Because of the small sample size and the high number of peaks (variables), occurrence of small standard deviations were avoided by adding 0.001. At each step the main factor effects were removed from the peak’s intensity by subtracting the corresponding mean factor-level . At a given step the factor resulting in the highest decrease in overall sum of the RSD (SRSD) was retained and eliminated from the remaining iterations. For the remaining steps the peak matrix corrected with the selected factor was used. Pearson correlation plots produced in Matlab were used to visualize the results.
3. Results

Specific effect of the selected seven preanalytical parameters on LC-MS data was evaluated by integrating the peak areas of peptides belonging to high-abundant proteins (Apolipoprotein A, α-2-Macroglobulin precursor, human Complement C3 precursor with masses of 756.7, 909.6, 682.5, 753.5, 619.2, 694.4). These protein-derived peptides have been shown by us to be rather stable in quantity in serum samples [3] and may thus be considered to reflect major effects of preanalytical factors, if affected.

In Figure 2 peak areas for these peptides are shown for serum from female (2a) and male blood (2b). The deviations of the mean retention time (RT) of endogenous standards were less than 0.8% confirming reproducibility of the LC-MS analyses. Variation in peak area ranged from approximately 40% a the peak derived from Apolipoprotein A (doubly charged peptide DLATVYVVDVLK, m/z 619.2, accession number P02647) up to 130% for a peak derived from Alpha-2-macroglobulin precursor (doubly charged peptide AAQVTIQSSGTFFSK, accession number P01023); (m/z 753.5) for factorial design #1 (female serum) (Figure 2a) showing that the chosen preanalytical factors do have a significant effect even on peptides derived from high-abundance proteins.

![Figure 2](image1.png)

Figure 2. Six m/z values derived from high-abundance proteins remaining after depletion were analyzed to assess the overall effect of the selected seven preanalytical factors (see Table 2 for the factorial design) based on changes in peak areas of extracted ion chromatograms. a) female serum; b) male serum. Peptides with the same m/z values: 619,2; 694,4; 753,5 were eluted twice at different retention time. Standard deviations are calculated based on the peak area’s of the Selected Ion Chromatogram of ±0.5 m/z values.

In male samples (factorial design #2) the deviation of the mean RT of the same peptides was also less than 0.8%. Standard deviations of peak areas varied from approximately 75% for a peak m/z 756.7 up to 135% for the peak with m/z 753.5 from Alpha-2-macroglobulin precursor.

In order to study the effects of the seven selected factors on these peptides, we applied a multiple linear regression model (Figure 3). The model showed that only one peptide was significantly affected by the studied factors in female serum (m/z 694.4, shown in red).
Investigation: LC-MS Fac (MLR)

Summary of Fit

N=19         Cond. no.=1.5811
DF=11        Y-miss=0

R2          Q2
Model Validity  Reproducibility

Figure 3. Peak areas of peptides derived from high-abundance proteins (see Table 3) were analyzed using a Multiple Linear Regression model. Data were obtained from two fractional factorial design studies (one in male and one in female serum). In the figure R2 refer to Residual Standard Deviation and is the fraction of variation of the response explained by model. Q2 refer to the predictive power of the model (Prediction Residual Sum of Squares) and is the fraction of variation of the response that can be predicted by the model.

The peak area of the peptide derived from Apolipoprotein A (doubly charged ion of peptide VSFLSALEEYTK, accession number P02647, m/z 694.4) in female serum showed a linear correlation between the expected (as predicted by the model) and observed peak areas (Figure 4a). Analysis of the contribution of different factors showed that clotting time and the trypsin-to-substrate ratio affected the measured peak area significantly (significance level of 95%) (Figure 4b). In male dataset the peak areas of the same peptide were less well predicted by the model, although a similar pattern was observed (data not shown).

In order to perform a global analysis of all detected features and how they respond to the variation of a given factor, it is necessary to generate a common peak matrix from all 19 LC-MS runs. This requires correction of retention time shifts between the different LC-MS analyses. However, Correlation Optimized Warping using the Total Ion Chromatogram (COW-TIC) was not successful, due to the major differences in chromatographic patterns when changing certain factors (Figure 5a). Poor time alignment of chromatograms lead to mismatched peaks and resulted in misinterpretations. The failure of the COW-TIC algorithm to align the chromatograms is due to the fact that some factors have a large effect on the measured peak profiles leading to very high analytical variability and to rather diverse TICs in spite of the fact that the serum sample was taken from the same patient (no biological variance included in the fractional factorial design experiment). Applying the COW algorithm not to the TIC but to a 2-dimensional extent of all detected peaks in the retention time and m/z dimensions (2D-COW), thus taking the 3-dimensional structure of the data into account, resulted in proper alignment of most peaks (see Figure 5b for an example).
Investigation: LC-MS Fac (MLR)
area2 694.4 female

N=19         R2=0.905     R2 Adj.=0.844
DF=11        Q2=0.682     RSD=1281191,7500

Conf. lev.=0.95

Investigation: LC-MS Fac (MLR)
Scaled & Centered Coefficients for P2 694.4 female (Extended)

Figure 4. Multiple Linear Regression model of the peptide derived from Apolipoprotein A (doubly charged ion of peptide VSFLSALEEYTK, accession number P02647, m/z 694.4). A) Correlation of the predicted with the observed peak area (R^2 = 0.905; N=19 experiments), b) coefficient of factors obtained in the MLR model.

Fig. 5. Time alignment of the 19 chromatograms of the fractional factorial design study (the example of one peptide peak at appr. 107 min retention time (m/z 619) is shown) using the COW-TIC algorithm (a) or COW in conjunction with a newly developed algorithm that aligns peaks in 2-dimensions (retention time and m/z) (b).

This allowed automatic processing of thousands of peaks in the datasets and the subsequent evaluation of the effect of the selected factors on the overall pattern. Correlation plots with hierarchical clustering of the aligned peak
matrixes from the study of the different factors showed that the level of hemolysis is the most important factor affecting the overall correlation between LC-MS datasets (Figure 6).

Figure 6 Correlation plots with hierarchical clustering of two fractional factorial design studies (see Table 2) for (a) male serum, and (b) female serum. Note that samples with low or high hemolysis levels, respectively, cluster together.

Figure 7. Sum of relative standard deviation (SRSD) after correction with the mean of corresponding factors for each level for all 7 factors (a, male) and (b, female).

Further analysis of the contribution of each factor to the overall sum of relative standard deviation (SRSD) between the datasets was carried out by correcting the contribution of each factor by the mean of each level. The factor showing the lowest SRSD has the greatest influence on peak intensity for largest number of peaks, and may thus be considered to be the most important preanalytical factor in the design with respect to the analysis of depleted from the 6 most abundant protein and trypsin-digested serum samples (Figure 7). Again the level of hemolysis stands out as being highly relevant. In order to
assess the relative importance of the remaining factors, the most important factor was discarded from the list of factors and the aligned peak matrix corrected with the most important factor was retained to determine the second most important factor. These iterations resulted in a list of factors ordered by their importance (Figure 8). The most important factors were: hemolysis level > trypsin-to-protein ratio > stopping the trypsin digestion with acid. Factors such as clotting time, type of blood collection tube, stability of sample in the autosampler or freez-thaw cycles had rather small effects on the resulting LC-MS data. The results concerning clotting time corroborate our findings described in Chapter IV.

Figure 8. Differences of the SRSD (log scale; log SRSD) after recursive correction with the mean of factor effect having the highest influence on SRSD using all peptide peaks of the 2D-COW aligned-peak matrices for male (a) or female serum (b).

4. Conclusions

The main result of this study is that the hemolysis level, which is a parameter that needs to be controlled during blood collection (responsibility of medical sciences; see Figure 1), makes the largest contribution to variability in the datasets. This implies that sera with high hemolysis levels have to be discarded from further analysis because this level cannot be set a priori but is largely influential on the expected differences in concentrations of potential biomarkers. Determining this relationship will be a matter of future studies. It is, however, wise to take special care to avoid hemolysis during blood collection and the preparation of serum. The other two most important factors, trypsin-to-protein ratio and stopping digestion with acid, can be easily controlled by the analytical chemist.

Our results describe only a partial evaluation of the multivariate data obtained from the fractional factorial design study of depleted and trypsin-digested serum samples. We only evaluated the main effect of each factor on the final data and more detailed analyses of the data need to apply multivariate statistical methods (e.g. Partial Least Squares methods) to obtain additional
information about possible interactions between two factors. Beside the factor and interaction order we would like to assess the significance level of the main factors and multiple factor interactions. As the design of the experiments has a resolution of IV, meaning that at least 3 out of two factor-interactions are confounded, a full factorial design, of the 3 main factors should be performed. This work is presently ongoing.
References
