BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–8

Quantitative quality-assessment techniques to compare fractionation and depletion methods in SELDI-TOF mass spectrometry experiments

Jaroslaw Harezlak[a] [*], Mike Wang[b], David Christiani[b], Xihong Lin[a]

[a] Department of Biostatistics, [b] Department of Environmental Health, Harvard School of Public Health 655 Huntington Ave. Boston, MA 02115

ABSTRACT

Motivation: Mass-spectrometry (MS), such as the surface-enhanced laser desorption and ionization time-of-flight (SELDI-TOF) MS, provides a potentially promising proteomic technology for biomarker discovery. An important matter for such a technology to be used routinely is its reproducibility. It is of significant interest to develop quantitative measures to evaluate the quality and reliability of different experimental methods.

Results: We compare the quality of SELDI-TOF MS data using unfractionated, fractionated plasma samples and abundant-protein depletion methods in terms of the numbers of detected peaks and reliability. Several statistical quality-control and quality-assessment techniques are proposed, including the Graeco-Latin square design for the sample allocation on a Protein chip, the use of the pairwise Pearson correlation coefficient as the similarity measure between the spectra in conjunction with multi-dimensional scaling (MDS) for graphically evaluating similarity of replicates and assessing outlier samples; and the use of the reliability ratio for evaluating reproducibility. Our results show that the number of peaks detected is similar among the three sample preparation technologies, and the use of the Sigma multi-removal kit does not improve peak detection. Fractionation of plasma samples introduces more experimental variability. The peaks detected using the unfractionated plasma samples have the highest reproducibility as determined by the reliability ratio.

Availability: Our algorithm for assessment of SELDI-TOF experiment quality is available at *http://www.biostat.harvard.edu/˜xlin*.

Contact: jharezla@hsph.harvard.edu

1 INTRODUCTION

The field of proteomics attempts to identify and quantify relative abundance of a large number of proteins using high-throughput technologies. The most common profiling technologies are based on liquid chromatography and/or mass spectrometry. The spread of these tools has attracted an increasing attention to the need of quantitative assessment of the quality of different sample preparation methods. The main purpose of this paper is to provide several statistical quality-control and quality-assessment methods to reduce experimental variability and bias and to evaluate the reproducibility of different sample preparation techniques.

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. 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.

In recent years, researchers have emphasized more and more on the importance of reliability and reproducibility of a mass spectrometry technology in protein profiling (Diamandis, 2004). Aivado *et al*., 2005 report that increase in the number of technical replicates significantly increases the reliability of the protein profiles. They also advocate the use of a robotics system to lower experimental variation. Baggerly *et al*., 2005 discuss reproducibility of findings in three experiments performed on ovarian cancer and normal tissues. They conclude that the differences in the proteomic profiles uncovered in the previous experiment (Petricoin *et al*., 2002) are due to sample processing and not the underlying biology of cancer. These findings demonstrate that it is important to develop statistical methods for quality-control in sample processing and for quantitatively assessing the reliability of different sample preparation methods.

This article provides such quantitative quality-assessment techniques by comparing three different plasma preparation methods, including whole plasma, fractionated plasma, and a abundant-protein depletion method using Sigma Kit FH. The major goals of this paper are as follows. First, we propose the use of a Graeco-Latin square design for sample allocation on ProteinChips to control chip/spot variability between different subjects and replicates. Second, we propose to use the pairwise Pearson correlation coefficient as the similarity measure between the spectra in conjunction with multi-dimensional scaling (MDS) and pairwise correlation plots to qualitatively assess the similarity of technological replicates and to check for the outlier spectra. Third, we compare the number of peaks detected between the three plasma preparation methods and propose to use the reliability ratio to assess the reproducibility of peak detection for each method.

[*] to whom correspondence should be addressed

2 EXPERIMENTAL SETTING AND DATA GENERATION

2.1 Plasma Preparation Methods

One of the major obstacles of proteomic technologies, such as SELDI-TOF MS, in plasma biomarker discovery is that plasma proteome is characterized by a large variation in individual protein expression intensities. Indeed, about 60-80% of total plasma proteins are the most abundant proteins, including albumin or immunoglobulins. Highly expressed proteins can decrease the number of peaks detected by MS technologies. Currently, there are two common strategies applied to biological samples prior to proteomic analysis using the SELDI-TOF MS bioprocessor. They include expanding the range of protein measurements and removing most abundant proteins from the plasma.

To expand protein measurements, fractionation technology is used to divide plasma proteome into subproteomes with minimal protein loss. Ciphergen provides a standard anion exchange fractionation procedure that is most frequently used. This procedure produces six plasma fractions with different pH and organic content: Fraction 1 (pH9 + flowthrough), Fraction 2 (pH7), Fraction 3 (pH5), Fraction 4 (pH4), Fraction 5 (pH3), and Fraction 6 (organic wash). A previous study of analyzing fractions by bicinchoninic acid assay revealed that protein recovery is around 75% of total protein amount (Solassol *et al*., 2005). Furthermore, plasma albumin is detected mainly in Fraction 4 and the majority of immunoglobulins is detected in Fraction 1. Fractionation by strong anion exchange chromatography has been shown to increase the number of peaks detected in plasma by SELDI (Fung *et al*., 2002, Linke *et al*., 2004, Adam *et al*., 2002), as well as to improve the detection of low-abundance proteins (Koopman *et al*., 2004, Solassol *et al*., 2005). The drawbacks of this strategy are that it significantly increases the cost and introduces additional sources of variability (Koopman *et al*., 2004). In addition, there is potential of a loss of low-abundant proteins which bind to albumin, as a result of lower albumin recovery in anion fractionation (Mehta *et al*., 2003-2004).

To remove highly abundant proteins from the plasma, a variety of depletion methods for specific removal of highly abundant proteins from plasma have been developed. The immunoaffinity columns based on specific antibodies to the most abundant proteins, isoelectric trapping, dye-ligand chromatography, peptide affinity chromatography, and ultrafiltration. (Echan *et al*., 2005; Bjorhall *et al*., 2005). Among them, the immunoaffinity method is often preferred, as it provides the most efficient depletion of targeted proteins with less nonspecific binding to other proteins. More recently, immunoaffinity columns have been developed for removal of multiple abundant proteins. Several comparative studies demonstrated that the multi-removal columns could provide increased resolution and improved intensity of low-abundant proteins in a reproducible fashion (Govorukhina *et al*., 2003, Steel *et al*., 2003, Echan *et al*., 2005, Bjorhall *et al*., 2005). The selection of immunoaffinity columns from widely available commercial resources is mainly based on the cost of columns, loading capacity, and the ease of integration with subsequent proteomic MS analysis. Taking into consideration the sample dilution and the aforementioned requirements, we selected ”Sigma kit FH” which required only 2-3 fold dilution of original sample by elution buffer and elution could be applied directly to SELDI-TOF analysis. Although the application of multi-removal columns was reported in the proteomic studies using two-dimensional electrophoresis with subsequent MALDI-TOF mass spectrometry, LC-MS/MS, and ESI mass spectrometry, there is no report of applying this technology to SELDI-TOF mass spectrometry.

In this paper, we compare three plasma preparation methods: whole plasma, fractionated plasma and whole plasma with the Sigma-removal kit FH. We compare the number of detected peaks of each of the three methods and the reliability of peak detection of each method.

2.2 Data generation

We applied our comparative procedure to a sample of 4 lung cancer patients who were selected from a large case-control study at Massachusetts General Hospital (MGH) in Boston (Liu *et al*., 2004) Selected subjects were males, former smokers with ages around the median age of the study participants (62 years). We considered four experimental conditions: whole plasma (“W”), Ciphergen fraction 1 (“1”), Ciphergen fraction 4 (“4”) and multi-removal column using Sigma kit FH (“S”). We used Ciphergen fractions 1 and 4 to minimize the overlap of chemically separated fractions and to keep the number of experimental conditions managable. We used a 16-spot Ciphergen ProteinChip on a 192-well bioprocessor. Each sample was divided into four parts, i.e., four replications, and loaded via a programmed robotic instrument on a ProteinChip. In summary, we considered 4 subjects, 4 experimental conditions, and 4 replicates per sample. These gave a total of 64 spectra. The details of the sample preparation are presented in Section 2.2.1 and the randomization of the sample placement is described in Section 2.2.2.

2.2.1 Technological design We used cationic (CM10) 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 × 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.

Table 1. The Graeco-Latin square design of sample allocations on four ProteinChips

Spot	Layout			
1	B-S	D-W	C-1	D-S
2	B-S	D-W	C-1	D-S
3	D-4	B-1	A-S	B-1
4	D-4	B-1	A-S	B-1
5	C-1	A-4	D-4	C-W
6	C-1	A-4	D-4	C-W
7	A-W	C-S	B-W	A-4
8	A-W	C-S	B-W	A-4
9	C-4	A-1	B-4	A-W
10	C-4	A-1	B-4	A-W
11	A-S	C-W	D-W	C-4
12	A-S	C-W	D-W	C-4
13	B-W	D-S	A-1	B-S
14	B-W	D-S	A-1	B-S
15	D-1	B-4	C-S	D-1
16	D-1	B-4	C-S	D-1

2.2.2 Experimental design We employed a randomization according to a Graeco-Latin square design (Bose, 1947) for sample allocation on the ProteinChips to control chip/spot variability between subjects and replicates obtained from plasma preparation methods. The Graeco-Latin design is an extension to a triple grouping of a Latin square design which is used for double grouping by eliminating any systematic differences among ProteinChips and spots. This design makes it possible to have a balanced design among subjects, replicates, preparation technologies and their combinations so that between-spot/chip variability can be controlled for. Consequently, comparison of spectra protein intensities between different preparation technologies would not be confounded by spot and chip effects.

A standardized procedure established in the DF/HCC Cancer Proteomics Core (Boston, MA) puts one sample on 2 consecutive spots. With this restriction, our design had a full factorial layout from 4 subjects and 4 fractions on 2 ProteinChips (32 spots). In order to evaluate between-chip variability, we repeated the experiment on 2 additional ProteinChips resulting in 4 technical replicates (2 replicates per chip) per subject-fraction combination. The Graceo-Lation square sample layout in presented in Table 1, where the letters A, B, C and D denote the subjects' IDs, and the characters W, 1, 4 and S denote the experimental conditions: whole plasma, Ciphergen fraction 1, Ciphergen fraction 4 and Sigma kit respectively. We used a Graeco-Latin square design to balance the placement of the samples coming from four subjects and prepared using four techniques within four ProteinChips. Each ProteinChip has eight rows, not counting the duplicates. In order to accomodate the instrument's requirement of 8 spots by 4 ProteinChips, we combined two different 4×4 Graeco-Latin square designs. We opted for the complete balance within each set of 2 ProteinChips, so on each of the four spots we would have samples corresponding to four subjects and four techniques. Our goal was to have a balanced design within ProteinChips, since the variation between the chips is greater than within the chip.

3 STATISTICAL ANALYSIS

3.1 Preprocessing of the MS data

In further analysis, we use spectra obtained using low laser energy protocol with m/z values below 20,000Da. All spectra were preprocessed using PROcess package from Bioconductor (Li *et al.*, 2005). PROcess is an R package designed for extraction of features (peaks) from raw SELDI-TOF data. Ideally a spectrum should rest on the zero horizontal line. The raw spectra exhibit an elevated baseline caused by the chemical noise in the EAM and ion overload. Baseline subtraction was performed by fitting a smooth curve to the local minima of each spectrum. The spectra were truncated at m/z equal to 3000 Da as the beginning m/z region is mainly contaminated by large chemical noise from the systematic EAM signature.

3.2 Spectra similarity and outlier detection

It is important to perform exploratory analysis to examine for outlier and problematic spectra before formal statistical analysis is carried out. Given high-dimensional spectra data, traditional exploratory data analysis is difficult. We use the pairwise Pearson correlation coefficient as the similarity measure between the spectra. We then use the multidimensional scaling (MDS) technique (Cox and Cox, 2001) and the heatmap of similarity matrix constructed using the pairwise Pearson correlation coefficients to visually explore the data and examine for outlier and problematic samples. We discuss in Sections 4 and 5 an automatic way of outlier detection using the multiple outlier test based on the extreme studentized deviate using the MDS and the Pearson correlation coefficients of individual spectra against a reference spectrum. Specifically, multidimensional scaling is a method for visualizing the similarity of a set of objects, especially for high-dimensional data, where each object is characterized by a collection of traits. MDS transforms similarities between objects into Euclidean distances. This allows one to easily identify outliers for high-dimensional data. It also provides a convenient tool to visually examine spectra variabilility between different groups, e.g., the four plasma preparation methods, "W", "1", "4" and "S".

A typical object in MDS is a numeric vector in a space R^p containing data used to explore the relationship among the objects. In our case a numeric vector is a mass spectrum obtained from a plasma sample of a patient using the SELDI-TOF MS, where p is the number of mass over charge ratios. Each vector of relative abundance of proteins from a proteomic experiment can be represented by a point in R^p. Since p in general is very large and can be in the order of thousands, MDS may be used to achieve dimension reduction and to display the objects in a lower dimensional space R^d , $d << p$. We used one minus the pairwise Pearson correlation coefficient as the entries in the dissimilarity matrix, and a classical MDS using Euclidean distances to reduce the dimension to $d = 1, 2,$ or 3 which correspond to the eigenvectors of the first three largest eigenvalues of the dissimilarity matrix respectively. We visually compared the clustering of the spectra among subjects, technologies and replicates.

Another way to visually assess the similarity of the spectra is to plot the correlation matrix of the biological replicates (Subjects A, B, C, and D in our case) against one another for each technological condition. Ideally, given all the spectra are from the same population (lung cancer cases), the pairwise correlations should be close to one (or dissimilarities close to zero). To contruct our heatmap of the pairwise spectrum correlations (Figure 4), we first obtained the average spectra of the four replicates for each subject (A, B, C, D), and then calculated the pairwise Pearson correlation coefficients between the spectra.

3.3 Peak detection

We are interested in comparing the number of peaks detected by each of the four sample preparation technologies (whole plasma ("W"), fraction 1 ("1"), fraction 4 ("4"), and whole plasma with sigma-depletion ("S")).

Let $y_{ijk}(t_l)$ be the baseline-subtracted signal for the kth replicate ($k = 1, 2, 3, 4$) of subject i ($i = A, B, C, D$), using the sample preparation technology j ($j = W, 1, 4, S$) measured at the l^{th} m/z value t_l. To proceed with peak detection, we first calculate the average of the four replicates for each person as $\bar{y}_{ij\cdot}(t_l)$. We then estimate the smoothed signals of $y_{ij\cdot}(t)$ using a moving-average estimator, which is a particular nonparametric kernel smoothing estimator (Wand and Jones, 1995). We estimate the smoothed noise levels $\sigma_{ij\cdot}(t)$ using a moving-mean absolute deviation estimator, which provides a robust estimator of the noise level (Percival and Walden, 2000). Specifically, our estimators can be written as

$$\begin{aligned}\widehat{y}_{ij\cdot}(t_l) &= \frac{\sum_{m=-h}^{h} \bar{y}_{ij\cdot}(t_{l+m})}{2h+1},\\ \widehat{\sigma}_{ij\cdot}(t_l) &= \frac{\sum_{m=-h_\sigma}^{h_\sigma} |\bar{y}_{ij\cdot}(t_{l+m}) - \bar{y}^M_{ij\cdot}(t_l)|}{(2h_\sigma+1)0.6745},\end{aligned}$$

where $\bar{y}^M_{ij\cdot}(t_l)$ is the median of $\bar{y}_{ij\cdot}(t_{l+m}), m \in \{-h, \dots, h\}$. We estimate bandwidth h using the plug-in method (Wand and Jones, 1995) for the optimal bandwidth estimation and $h_\sigma = 3h$ is a function of h where constant 3 was chosen to stabilize the standard deviation estimation.

Define the adjusted signal-to-noise (SN) ratio as

$$SN_{ij\cdot}(t_l) = \frac{\widehat{y}_{ij\cdot}(t_l)}{\widehat{\sigma}_{ij\cdot}(t_l) + c_j},$$

where c_j is a small constant that is added to numerically stabilize the *SN* estimators in order to avoid numerical explosion of the *SN* ratios when the $\widehat{\sigma}_{ij\cdot}(t_l)$'s are very small, especially in flat regions when the values of $\widehat{y}_{ij\cdot}(t_l)$ are small. A similar variance stablization strategy is employed in microarray analysis (Efron *et al.*, 2001). For each subject i and technology j, we estimate the $\widehat{\sigma}_{ij\cdot}(t_l)$ at each t_l-value. For a given technology j, we calculate the percentiles of the distribution of $\widehat{\sigma}_{ij\cdot}(t_l)$ $(i = 1, \dots, n; l = 1, \dots, L)$, and use the first quartile of the empirical distribution of the $\widehat{\sigma}_{ij}(t_l)$ to estimate c_j. We define a location t_l as a peak if both of the following two conditions are satisfied:

$$\begin{aligned} SN_{ij\cdot}(t_l) &> d && (1)\\ \widehat{y}_{ij\cdot}(t_l) &> \widehat{y}_{ij\cdot}(t_{l+r}), \text{ for } r \in \{-R, \dots, R\}\backslash\{0\} && (2)\end{aligned}$$

where d is a hard-threshold constant and R defines a neighborhood around t_l used to check whether t_l is a local maxima. In our analysis, we set the hard-threshold $d = 3$ and the neighorhood range $R = 45$. The condition (1) guarantees that the intensity at t_l exceeds the noise at least d-fold, and the condition (2) ensures the spectrum has a local maximum at t_l. We apply this peak detection method to the four sample preparation technologies $j = W, 1, 4, S$.

3.4 Reliability ratio as a reproducibility measure

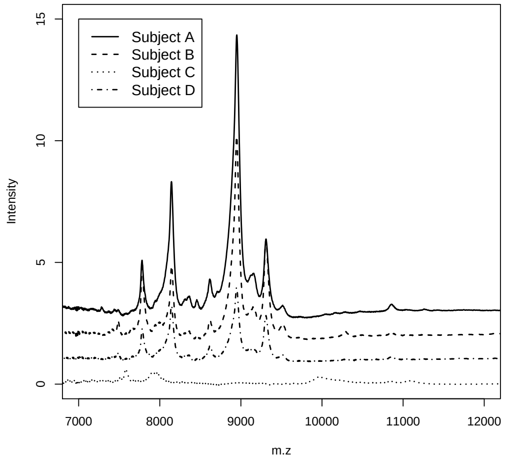
The number of peaks detected using the method in 3.3 quantifies the performance of a sample preparation technology in terms of its ability to detect abundance of proteins. To properly assess the performance of a technology, it is equally important to study the reproducibility of peak detection of the technique. For example, if the same peaks are detected and the same peak intensities are measured on all replicates for a subject using a given sample preparation technique, we would say that the technique has a good reproducibility. On the other hand, if the detection of the peaks and their intensities have large variability within a subject, the reproducibility is poor. In order to quantify the reproducibility of a technique, we propose to use the reliability ratio measure, which is commonly used in the measurement error literature (Carroll *et al.*, 2006)

Denote by $T_{j1}, \dots, T_{jL_j}$ the estimated pooled peak locations across subjects using technology j $(j = W, 1, 4, S)$. Denote by y_{ijkl} the intensity for subject i, repetition k using the technique j at the lth peak location T_l. For each peak location T_l $(l = 1, \cdots, T_{L_j})$, we assume that y_{ijkl} follows a mixed effects model (Laird and Ware, 1982)

$$y_{ijkl} = \mu_{jl} + b_{ijl} + \epsilon_{ijkl}, \qquad (3)$$

where μ_{jl} denotes the group mean over the subjects using the technique j at the lth peak T_l, $b_{ijl} \sim N(0, \theta^2_{jl})$ is a random subject effect and the variance component θ^2_{jl} measures between-subject variability for a given technique j at peak T_{jl}, and $\epsilon_{ijkl} \sim N(0, \sigma^2_{jl})$ is a random replicate error and the variance component σ^2_{jl} measures the within-subject variablility for a given technique j at peak T_{jl}. We fit this model using mixed effects models by estimating μ_{jl} using maximum likelihood and the variance components θ^2_{jl} and σ^2_{jl} using restricted maximum likelihood (REML) to account for small sample sizes (Harville, 1977).

Fig. 1. Baseline-subtracted spectra from one technical repetition of the experiment for all four subjects (A, B, C and D) using 'whole plasma' ("W"). To visually compare different spectra, different constant shifts are added artificially to the four spectra. Note that the original baseline-substracted data are used in the analysis.



The reliability ratio λ_{jl} for technique j at a given peak location T_{jl} is defined as the ratio of the within-subject variance divided by the total variance, the sum of the between-subject variance and the within-subject variance and can be written as

$$\lambda_{jl} = \frac{\theta^2_{jl}}{\sigma^2_{jl} + \theta^2_{jl}}. \qquad (4)$$

One can easily see that the reliability ratio statisfies $0 \leq \lambda_{jl} \leq 1$ and can be interpreted as a percentage of reproducibility of a technique. A large λ means a technique has a high reproducibility, while a small λ means it has a low reproducibility. For example, if $\lambda_{jl} = 1$ then a technique generates all the same replicates for a given subject, i.e., there is no between-replicate variability and all the variability comes from between-subject variability. Hence the technique is 100% reproducible. On the other hand, if $\lambda_{jl} = 0$ then all variability comes from the between-replicate variability and there is no between-subject variability. Hence the technique is 0% reproducible. The intuition behind the reliability ratio use is that if a procedure is repeated many times on the same biological sample and the values do not change much between replicates when compared with the biological variation across different biological samples, then the reproducibility is high.

4 RESULTS

For illustration, we present the baseline-subtracted data between 7000Da-12,000Da in Figures 1 and 2. Specifically, Figure 1 presents data from one spectrum using whole plasma for each of the four subjects, while Figure 2 shows the data from subject A using the four technologies ("W","1","4" and "S"). Figure 1 shows that the spectrum for subject C is problematic. This is further demonstrated using multi-dimensional scaling analysis.

Fig. 2. Spectra from one technical repetition for subject A using four different technologies (“W”,”1”,”4” and “S”). To visually compare different spectra, different constant shifts are added artificially to the four spectra. Note that the original baseline-substracted data are used in the analysis.

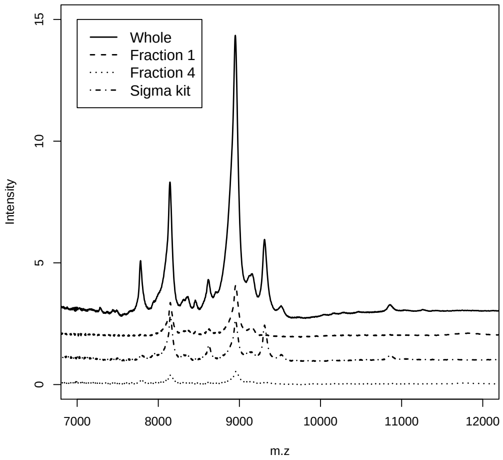


Figure 3 presents a 1-dimensional multi-dimensional scaling plot using the first eigenvector corresponding to the largest eigenvalue of the dissimilarity matrix obtained from the pre-processed spectrum data by subjects (A, B, C and D) and fractionation technologies (different panels). We can see clearly that subject C is an outlier, since all of his spectra are clustered separately in a far distance from the other three subjects for most fractionation techniques, while the data of the three subjects (A, B, and D) are clustered together. This is particularly apparent using the whole plasma and the sigma-depletion kit. Figure 3 also shows that samples using the whole plasma (with/without sigma-depletion) show smaller between-replicate variability, while fractionated samples have higher between-replicate variablity. This suggests that whole plasma is likely to have better reproducibility than fractionated plasma.

We used an outlier test based on the extreme studentized deviate (ESD) (see Rosner, 1975) on the first dimension of the MDS solution corresponding to the largest eigenvalue of the dissimilarity matrix (see Figure 3) with subjects as grouping factors. The median values of the first coordinate for the four subjects were: -0.403 (A), -0.361 (B), 0.620 (C), and -0.368 (D) (see Figure A in the online supplement). If the variation as summarized by MDS for all subjects depended only on the fractionation technique, the medians for all subjects should be similar, since we employed a balanced design where each subject had 4 repetitions for 4 fractionation techniques. We found that the value for a subject C was an outlier (p-value $= .0013$).

Figure 4 shows the heatmap of the pairwise Pearson correlation coefficients of the subject-specific spectra averaged over the four replicates for each fractionation technique. It confirms the finding from the MDS plot (Figure 3). Subjects A, B and D show a good correlation agreement within each panel (fractionation), while subject C is a clear outlier. This conclusion is especially pronounced for the whole plasma and Sigma abundant protein removal technologies. A detailed examination of subject C’s plasma sample found that it had undergone substantial hemolysis and should be excluded. Based on the above results, the quantitative comparisons in peak detection and reliability between different techniques in Section 4.1 and 4.2) are based on the data collected on subjects A, B and D.

Fig. 3. Multi-dimensional scaling plot of the first dimension of the eigenvector corresponding to the largest eigenvalue of the dissimilarity matrix based on the (1-correlation) distance metric. Panels correspond to the technologies (”W”, ”1”, ”2”, ”S”), and the subjects (A,B,C,D) are represented on the x-axis

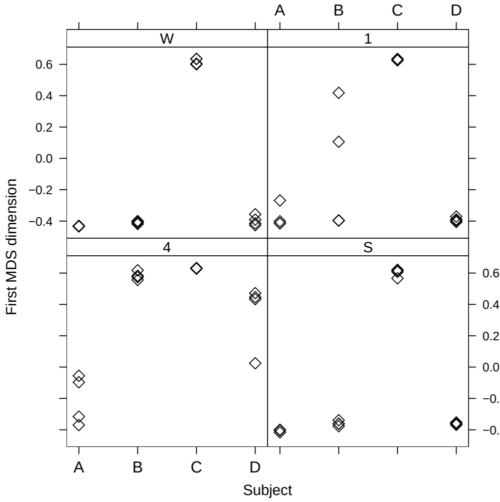
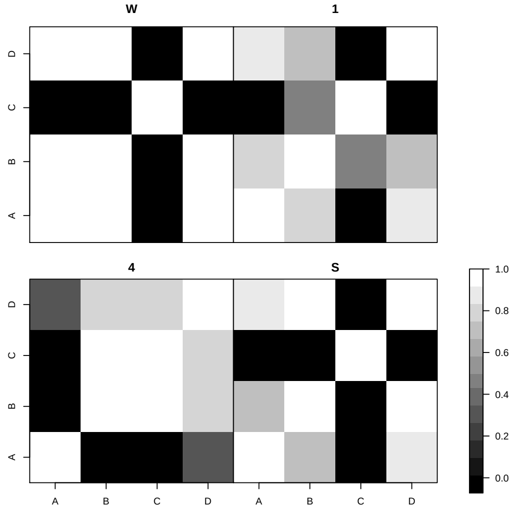


Fig. 4. Correlations for the 4 subjects and 4 fractionation technologies averaged over the repetitions, where white represents perfect agreement (correlation equal to 1), and black represents disagreement (correlation equal to 0)



4.1 Comparison of the number of detected peaks between different techniques

To compare the number of detected peaks between the four techniques (j=”W”, ”1”, ”4”, ”S”), we averaged the four spectra for each subject, and estimated the number of peaks in each spectrum for each technique using the method described in Section 3.1. The variance stablizing constant c_j was estimated to be approximately the same for each technique. We hence used $c_1 = \ldots = c_4 = c = 0.008$. We first compare the number of peaks

Fig. 5. Comparison of the smoothed probabilities of peak detection as a function of peak m/z locations between the four techniques using smoothed logistic regression: W-whole plasma, 1-fraction 1, 4-fraction 4, S-whole plasma with Sigma depletion kit

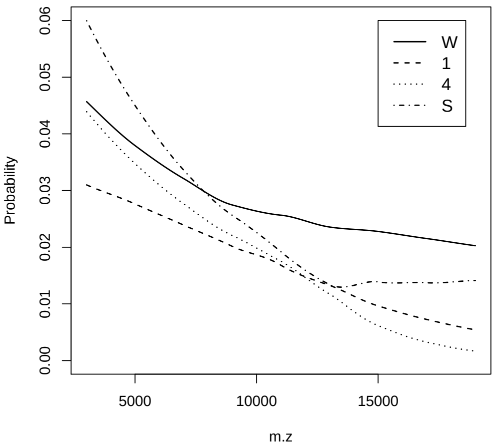
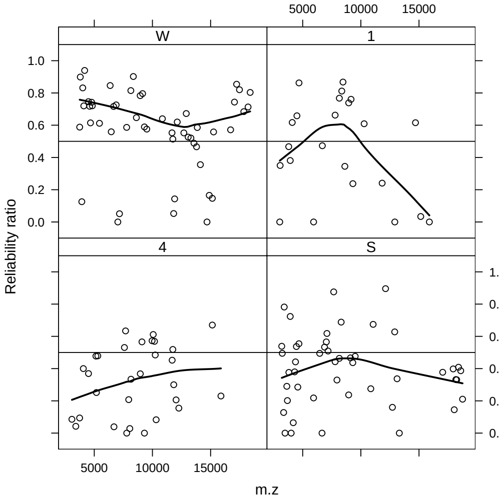


Fig. 6. Scatterplot of the reliability ratios for the four technologies as a function of "m/z" location (fitted smooth curves overlay the scatterplots)



per subject and per technique. Table 2 gives the total number of detected peaks over the whole spectra for each of the three subjects and each of the four techniques.

Table 2. The total number of detected peaks over the whole spectra for each of three subjects (A, B, D) and each of four techniques (W-whole plasma, 1-fraction 1, 4-fraction 4, S-whole plasma with Sigma depletion kit

	Fraction				
Subject	W	1	4	S	1+4
A	43	29	22	35	39
B	50	28	34	42	51
D	41	25	31	43	50
Average	45	27	29	40	47

Table 2 shows that the Sigma abundant-protein depletion kit producs fewer peaks than using the unfractionated whole plasma. Each fraction ("1" and "4") yields a much smaller number of peaks compared to the whole plasma (with/without Sigma depletion). In order to compare the peak detection capacity of jointly using fractions 1 and 4, we combined the distinct peaks detected by fractions 1 and 4 and created a super-fraction "1+4" which contained non-overlapping peaks from separate fractions. We defined the peaks to be distinct when their locations differed by more than 0.5% of the mass over charge ratio. If the peaks for fraction 1 and fraction 4 were in the distance smaller than 0.5% of the m/z value we defined the peaks to come from the same protein/peptide. The number of peaks detected in this combined fraction ("1+4") is comparable to those using unfractionated samples. An interesting question is how many additional peaks that are detected in the fractionated samples but not in the whole plasma sample and vice versa. We found on average 11 peaks that were detected from the super-fraction ("1+4") but not from the whole plasma, while on average 9 peaks that were detected from the whole plasma but not by the fractionated samples.

We are also interested in comparing the peak detection ability of each technique as a function of m/z values. Figure 5 plots the probability of a peak being detected as a smooth function of m/z values for each technique. These curves show that majority of the detected peaks are in the mass range of 5000-10,000Da. There are not many expressed proteins for large masses. One can see that the Sigma-kit is more likely to detect a peak in the early m/z region but detects fewer peaks than the whole plasma for larger m/z values. This suggests that Sigma kit enhances the peak detection for small to moderate protein masses, but might overly remove proteins with larger masses.

4.2 Assessment of reproducibility of the four technologies

We compared the reproducibility between the four sample preparation techniques using the reliability ratio. By applying the linear mixed model method described in Section 3.4, we estimated the reliability ratio for each technique at each peak m/z location, T_{jl} $(j = 1, \ldots 4; l = 1, \cdots, L_j)$. Figure 6 plots for each technique the estimated reliability ratios against the peak locations and super-imposes a smooth curve using loess smoothing. Table 3 gives the summary statistics of the reliability ratio across peak locations for each technique. These results show that the whole plasma technique has the highest reliability ratio and is most stable across the spectrum mass. The fractionated plasma have on average the lowest reliability. Fraction 1 in particular has fair but quite variable reliability in the beginning region ($< 10,000$Da) but poor reliabiilty in the region with large m/z values. Fraction 4 has poor reproducibility throughout the experimental range. The Sigma multi-removal column has moderate reliability in the beginning region but performed poorly at m/z values above 15 kDa where the reliability ratio is below 0.5 for all locations.

5 DISCUSSION AND CONCLUSIONS

Quality assessment of advanced mass spectrometry proteomic techniques is important in the field of proteomics. We provide in this paper both graphical and quantitative methods for comparison of reproducibility of different sample preparation techniques for the mass spectra generated using SELDI-TOF. Specifically, we propose to use the Graeco-Latin square design for the experimental design on ProteinChips to balance spot/chip variation in different samples. The multi-dimensional scaling method provides a convenient way to visualize high-dimensional mass-spectra data and is useful

Table 3. Summary of the reliability ratios for each of the four techniques (W-whole plasma, 1-fraction 1, 4-fraction 4, S-whole plasma with Sigma depletion kit) including the mean, standard deviation (SD), median and interquartile range (IQR).

Technology	Mean	SD	Median	IQR
W	0.58	0.25	0.61	0.22
1	0.46	0.30	0.47	0.46
4	0.33	0.22	0.35	0.40
S	0.40	0.22	0.40	0.28

for checking outliers and problematic samples. The reliability ratio provides an objective comparison of the reproducibility of different sample preparation techniques.

Our results show that the Ciphergen fractionation method and the Sigma abundant protein removal method provide overall similar numbers of detected peaks compared to that using the whole plasma, while the Sigma kit detects more peaks in the <10,000Da region but fewer peaks in the higher m/z region. This suggests the Sigma kit might overly remove proteins of larger mass. The whole plasma gives the best and the most stable reproducibility of deteced peaks over the whole range of the m/z values, while fractionation and the Sigma-removal kit have lower reproducibility.

Based on our results in Section 4, we propose to use Pearson correlation coefficients calculated for each spectrum against the reference spectrum (e.g. average spectrum or group-specific average spectrum) as a general recommendation for outlier detection. Spectra exhibiting low correlation coefficients are declared to be outliers. Our proposal uses multiple outlier test based on the extreme studentized deviate (ESD) as studied in (Rosner, 1975) among others.

We recommend the use of the reliability ratio as a quantitative assessment of reproducibility of a mass spectrometry technique. Such an assessment should be an important routine practice before a MS technique is used for processing a large number of samples. A high reliability ratio is needed in order for a MS technique to be used to generate reproducible results.

ACKNOWLEDGEMENT

This research was supported in part by the NIH grants R01 CA76404 and R01 CA074386.

REFERENCES

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.

Aivado M, Spentzos D, Alterovitz G, Otu HH, Grall F, Giagounidis AA, Wells M, Cho JY, Germing U, Czibere A, Prall WC, Porter C, Ramoni MF, Libermann TA. (2005) Optimization and evaluation of surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) with reversed-phase protein arrays for protein profiling. *Clin Chem Lab Med.* **43**: 133-140.

Baggerly KA, Morris JS, Edmonson S, and Coombes KR. (2005) Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *Journal of the National Cancer Institute* **97**:307-309

Bjorhall K, Miliotis T, Davidsson P. (2005) Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* **5(1)**:307-17.

Bose, R.C. (1947). Mathematical Theory of the Symmetrical Factorial Design, *Sankhya*, **8**, 107-166

Carroll R.J., Ruppert D., Stefanski L.A. and Crainiceanu C. (2006) Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition, CRC

Cox, T. F. and Cox, M. A. (2001). Multidimensional Scaling. Chapman and Hall/CRC.

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

Echan LA, Tang HY, Ali-Khan N, Lee K, Speicher DW. Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. *Proteomics* **5(13)**:3292-303.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

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

Govorukhina NI, Keizer-Gunnink A, van der Zee AG, de Jong S, de Bruijn HW, Bischoff R. (2003) Sample preparation of human serum for the analysis of tumor markers. Comparison of different approaches for albumin and gamma-globulin depletion. *J Chromatogr A.* **1009(1-2)**:171-8.

Harville, D. A. (1977), Maximum Likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320-340.

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.

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

Kruskal, W. H., and Wallis, W. A. (1952), Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**, 583-621.

Laird, N. M., and Ware, J. H. (1982), Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

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

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

Linke T, Ross AC, Harrison EH, Profiling of rat plasma by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry, a novel tool for biomarker discovery in nutrition research, *J. Chromatogr.* **A 1043**: 65-71.

Liu G, Zhou W, Park S, Wang LI, Miller DP, Wain JC, Lynch TJ, Su L, Christiani DC. (2004) The SOD2 Val/Val genotype enhances the risk of nonsmall cell lung carcinoma by p53 and XRCC1 polymorphisms. *Cancer*, **101**, 2802-2808.

Mehta AI, Ross S, Lowenthal MS, Fusaro V, Fishman DA, Petricoin EF 3rd, Liotta LA. (2003-2004) Biomarker amplification by serum carrier protein binding. *Dis Markers.* **19**, 1-10.

Percival, D. B. and Walden, A. T. (2000) Wavelet Methods for Time Series Analysis, Cambridge University Press

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.

Rosner, B, (1975) On the detection of many outliers,*Technometrics*, **17**, 221-227.

Solassol J, Marin P, Demettre E, Rouanet P, Bockaert J, Maudelonde T, Mange A. (2005) Proteomic detection of prostate-specific antigen using a serum fractionation procedure: potential implication for new low-abundance cancer biomarkers detection. *Anal Biochem.* **338(1)**:26-31.

Steel LF, Trotter MG, Nakajima PB, Mattu TS, Gonye G, Block T. (2003) Efficient and specific removal of albumin from human serum samples. *Mol Cell Proteomics.* **2(4)**:262-70.

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