Anal. Chem. **2007**, *79,* 7014−7026

Pretreatment of Mass Spectral Profiles: Application to Proteomic Data

Reidar Arneberg,[†,‡] Tarja Rajalahti,[§,∥] Kristian Flikka,[⊥,#,◊] Frode S. Berven,[♦] Ann C. Kroksveen,[♦,@] Magnus Berle,[♦] Kjell-Morten Myhr,[§,∥,&] Christian A. Vedeler,[§,∥] Rune J. Ulvik,[♦,$] and Olav M. Kvalheim[*,%]

Center for Integrated Petroleum Research, Department of Clinical Medicine, Proteomics Unit (PROBE), Department of Informatics, Institute of Medicine, Institute of Molecular Biology, and Department of Chemistry, University of Bergen, Bergen, Norway, Pattern Recognition Systems AS, Bergen, Norway, Department of Neurology, The National Competence Centre for Multiple Sclerosis, and Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen, Norway, and Computational Biology Unit, Bergen Centre for Computational Science, Bergen, Norway

Mass spectral profiles are influenced by several factors that have no relation to compositional differences between samples: baseline effects, shifts in mass-to-charge ratio (*m/z*) (synchronization/alignment problem), structured noise (heteroscedasticity), and, differences in signal intensities (normalization problem). Different procedures for pretreatment of whole mass spectral profiles described by almost 50 000 *m/z* values are investigated in order to find optimal approaches with respect to revealing the information content in the data. In order to quantitatively assess the impact of different procedures for pretreatment of mass spectral profiles, we use factorial designs with the ratio between intergroup and intragroup (replicate) variance as response. We have examined the influence of smoothing, binning, alignment/synchronization, noise pattern, and normalization on data interpretation. Our analysis shows that the spectral profiles have to be corrected for heteroscedastic noise prior to normalization. An *n*th root transform, where *n* is a small, positive integer, is used to create a homoscedastic noise structure without destroying the linear correlation structures describing individual components when using whole mass spectral profiles. The choice of *n* is decided by a simple graphic procedure using replicate information. Log transform is shown to change the heteroscedastic noise structure from being dominant in high-intensity regions, to produce the largest noise in the low-intensity regions. In addition, log transform has a negative effect on the collinearity in the profiles. Factorial designs reveal strong interactions between several of the pretreatment steps, e.g., noise structure and normalization. This underlines the limited usability of looking at the different pretreatment steps in isolation. Binning turns out to be able to substitute smoothing of spectra by, for example, moving average or Savitsky−Golay, while, at the same time, reducing the data point description of the profiles by 1 order of magnitude. Thus, if the sampling density is high, binning seems to be an attractive option for data reduction without the risk of losing information accompanying the integration of profiles into peaks. In the absence of smoothing, binning should be executed prior to alignment. If binning is not performed, the order of pretreatment should be smoothing, alignment, *n*th root transform, and normalization.

Mass spectrometry has become a widely used method for analyzing samples of biological origin. Especially in proteomics and metabolomics, mass spectrometry has become the dominant instrumental technique. The main reason is that mass spectrometry compared to, for example, nuclear magnetic resonance, can detect components at a very low concentration level. In a clinical context, the appearance of disease-specific peptide and protein patterns in body fluids like plasma and cerebrospinal fluid (CSF) may provide information about presence, status, or progress of a particular disease.[1−5] Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry is one of the most used methods.[6] The biomolecules (proteins or peptides) are caught within a crystalline structure, referred to as a spot, and bombarded with laser pulses. The matrix vaporizes and the biomolecules ionize

[*] Correponding author. E-mail: Olav.Kvalheim@kj.uib.no. Tel: +47 55583366. Fax: +47 55589490.

[†] Center for Integrated Petroleum Research, University of Bergen.

[‡] Pattern Recognition Systems AS.

[§] Department of Clinical Medicine, University of Bergen.

[∥] Department of Neurology, Haukeland University Hospital.

[⊥] Bergen Centre for Computational Science.

[#] Proteomics Unit at University of Bergen (PROBE).

[◊] Department of Informatics, University of Bergen.

[♦] Institute of Medicine, University of Bergen.

[@] Institute of Molecular Biology, University of Bergen.

[&] The National Competence Centre for Multiple Sclerosis, Haukeland University Hospital.

[$] Laboratory of Clinical Biochemistry, Haukeland University Hospital.

[%] Department of Chemistry, University of Bergen.

(1) Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C.; Liotta, L. A. *Lancet* **2002**, *359* (9306), 572−577.

(2) Kozak, K. R.; Su, F.; Whitelegge, J. P.; Faull, K.; Reddy, S.; FArias-Eisner, R. *Proteomics* **2005**, *5*, 4589−4596.

(3) Laronga, C.; Becker, S.; Watson, P.; Gregory, B.; Cazares, L.; Lynch, H.; Perry, R. R.; Wright, G. L.; Drake, R. R.; Semmes, O. J. *Dis. Markers* **2003**, *19*, 229−238.

(4) Li, J.; White, N.; Zhang, Z.; Rosenzweig, J.; Mangold, L. A.; Partin, A. W.; Chan, D. W. *J. Urol.* **2004**, *171*, 1782−1787.

(5) Li, L.; Tang, H.; Wu, Z.; Gong, J.; Gruidl, M.; Zou, J.; Tockman, M.; Clark, R. A. *Artif. Intell. Med.* **2004**, *32*, 71−83.

(6) Karas, M.; Hillenkamp, F. *Anal. Chem.* **1988**, *60*, 2299−2301.

10.1021/ac070946s CCC: $37.00
Published on Web 08/21/2007

and are accelerated in an electric field. The mass of each molecule is calculated based on time of flight (TOF) to the detector. The ensemble of molecules produces a multicomponent spectral profile of intensities (ion counts at the detector) for the measured mass-to-charge ratios (m/z). Usually the procedure is repeated several times at different positions in a spot and the final spectral profile calculated as an average. This reduces the variations in intensity that may be observed even between spots originating from the same sample.

In addition to differences due to heterogeneity in the sample, mass spectral data are influenced by several factors that have no relation to compositional differences between samples. Disease-associated proteomic patterns can only be extracted if the influence of these experimental and instrumental factors is minimized. Baseline effects, shifts in m/z values, structured noise (heteroscedasticity), and differences in signal intensities caused by analytical workup and the instrumental technique represent some of the major factors. Baseline level in MALDI spectra decays almost exponentially with decreasing m/z value. Most instrument manufacturers have implemented their own procedure for baseline correction. These algorithms are not perfect and some produce negative intensities in the spectral profiles.[7] Negative spectral intensities can be corrected by assigning zero intensity to the lowest signal in each spectrum. The absolute value of the lowest number is added to the intensities at each m/z number throughout the entire profile. This is done independently for each spectral profile.

The problem of m/z shift between corresponding molecules in different spots, commonly called the alignment or synchronization problem, has attracted a lot of attention because it represents an obstacle to the comparison of whole spectral profiles. A common solution has been to reduce the spectral profile to peaks, but this approach leads to loss of information in complicated spectra. Overlapping peaks need to be resolved in order to minimize the information loss, but this is not a straightforward procedure and may create other problems. Therefore, techniques for adjusting profiles to maximize cross-correlation between a set of spectral profiles have been developed. An early attempt to solve the alignment problem for digitized instrumental profiles is attributed to Andersson and Hämäläinen.[8] Some distinct peaks that were present in all the profiles were selected and used as targets for adjusting the profiles. Simplex optimization was used to maximize the cross-correlation with the profiles of the target peaks for all the samples, and the profiles were adjusted piecewise and independently around each target peak for each sample. Later on, many refined procedures based on the same principle of maximizing the cross-correlation were developed.[9−12] Recently, Wong et al.[11,12] developed a fast Fourier transform cross-correlation method implemented in a software called SpecAlign.[13] The method is fast, provides excellent results, and is relatively simple to use. It is therefore used in this work to cope with the alignment problem.

Structured noise, i.e., noise increasing with signal size (heteroscedasticity), represents a major problem in the comparison of mass spectral profiles. This creates a problem when normalizing profiles to correct for multiplicative effects due to differences in sample size.[14,15] Heteroscedastic noise gives false negative correlations between major peaks and also impacts the minor peaks by imposing false positive correlations between them. Various procedures have been proposed to overcome this problem, e.g., the nth root or logarithmic transform to provide a homoscedastic noise pattern.[14] The logarithmic transform destroys linear correlations in the profiles. This is a problem when using whole spectral profiles since one component is described by many linearly correlated m/z numbers. The nth root transform preserves perfect linear correlations but reduces correlations in regions with only partial correlation. Furthermore, the nth root transform reduces the intensity of major peaks compared to minor peaks, but this may turn out to be an advantage since the information content in minor peaks in many cases may be higher than in major peaks.[16] In this work, we use the nth root transform on instrumental replicates to find the value of n providing homoscedastic noise. Replicate information is thus crucial to assess the severity of heteroscedasticity in the data. As far we can see, there is no alternative to the use of such a data-driven nonparametric approach to correct for heteroscadisticity.

Several procedures for normalization exist, the most common being normalization to constant sum or constant length. The so-called TIC normalization as implemented by Conrad et al.[16] differs only by a scalar from the common procedure of normalization to constant sum. Thus, TIC normalization provides the same correlation patterns as the well-known normalization to constant sum. Normalization to unit length (the norm) provides larger weight to the regions with high intensity in the spectra. It is therefore slightly more vulnerable to heteroscedastic noise than normalization to constant sum. As long as the profiles are properly corrected for heteroscedastic noise prior to normalization, there should from theoretical considerations be no large difference between the two procedures. We have chosen to use the norm for normalization in this work, but calculations confirm that there are no real differences between the two procedures.

The use of whole spectral profiles leads to large quantities of data that need to be processed. In this work, the raw profiles are described by 40−50 000 m/z numbers. Several methods exist for reducing profiles. Karstang and Eastgate[17] proposed the use of a maximum entropy criterion whereby the profiles were reduced by adding adjacent data points in the mean spectral profile until the summed intensities exceeded a precalculated limit depending on the number of points used to describe the reduced profiles. In

[7] Williams, B.; Cornett, S.; Crecelius, A.; Caprioli, R.; Dawant, B.; Bodenheimer, B. *Proceedings of the 43rd Annual ACM Southeast Regional Conference*; Kennesaw, GA, 2005; Vol. 1, pp 137−142.

[8] Andersson, R.; Hämäläinen, M. D. *Chemom. Intell. Lab. Syst.* **1994**, *22*, 49−61.

[9] Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard J. *J. Chromatogr., A* **1998**, *805*, 17−35.

[10] Tomasi, G.; van den Berg, F.; Andersson, C. *J. Chemom.* **2004**, *18*, 231−241.

[11] Wong, J. W. H.; Durante, C.; Cartwright, H. M. *Anal. Chem.* **2005**, *77*, 5655−5661.

[12] Wong, J. W. H.; Cagney, G.; Cartwright, H. M. *Bioinf. Appl. Note* **2005**, *21*, 2088−2090.

[13] *SpecAlign Documentation Version 2.3.* Jason W. H. Wong Cartwright Group, Physical & Theoretical Chemistry Lab, University of Oxford.

[14] Kvalheim, O. M.; Brakstad, F.; Liang, Y.-Z. *Anal. Chem.* **1994**, *66*, 43−51.

[15] Rietjens, M. *Anal. Chim. Acta* **1995**, *316*, 205−215.

[16] Conrad, T. O. F.; Leichtle, A.; Hagehülsmann, A.; Diederichs, E.; Baumann, S.; Thiery, J.; Schütte, C. In *Computational Life Sciences II, Proceedings*; Berthold, M. R., Glen, R., Fischer I., Eds.; Springer-Verlag: Berlin, 2006; *4216*, pp 119−128.

[17] Karstang, T.; Eastgate, R. J. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 209−219.

this way, high-intensity regions were given higher data point density than low-intensity regions, i.e., more emphasis on large peaks compared to small peaks. Binning represents another attractive way of reducing the profiles without losing major information in the profiles. Binning adds a fixed number of adjacent data points into one combined variable. Both maximum entropy and binning provides a smoothing of the profiles, which may help to reduce the alignment problem, assuming of course that the reduction is executed before alignment. The essential point in the reduction is that all important features are retained in the reduced profile. Since maximum entropy leads to a varying number of m/z values being added; the results are more difficult to interpret than binning. Furthermore, binning provides a better description of minor peaks than the maximum entropy approach. Therefore, we have chosen to use binning for data reduction and investigated its effect on the information content in the data. We collect data in bins so that each point in the reduced profile represents ~1 Da for most of the analyzed data sets.

The aim of this work is to provide a scheme for pretreatment of mass spectral profiles that eliminates differences in profiles resulting from experimental and instrumental procedures, but at the same time preserves the compositional information. As pointed out by Baggerly et al.,[18] we may expect interaction between the different preprocessing steps. Therefore, they should not be considered in isolation. In order to incorporate the possibility of interaction and assess the relative importance of the different preprocessing steps, we use factorial design with different pretreatment steps as design variables and the inter- to intragroup variation between replicates as response variable to decide on the optimal procedure. This is done for three sample sets of different origin. From the resulting models, we are able to propose optimal schemes for pretreatment of mass spectral data using MALDI and related techniques.

THEORY

Removal of signals related to noncompositional information from spectral profiles is important since they are superimposed on the compositional correlation patterns. Thus, these disturbing factors make it difficult to compare profiles from different populations of samples, e.g., profiles from a group of controls with profiles from a group with a particular diagnosis. Often such comparisons are performed using a latent variable projection method[19] of which the most common is principal component analysis (PCA):

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

Here, **X** represents the matrix of spectral profiles with each row representing the profile of one sample and each column representing a particular m/z number in all samples. Decomposing on principal components produces a matrix of scores **T** and loadings **P**. The superscript T implies the transpose. The matrix **E** contains the residuals, i.e., the difference between measured and modeled values.

Mass spectral data are influenced by many experimental and instrumental effects that need to be removed before a correlation analysis by PCA or other correlation methods looking for patterns in spectral profiles is performed.

Baseline Correction. The baseline is a mass-to-charge ratio-dependent offset that is characterized as a near-exponential decay.[20] The baseline needs to be removed without removing compositional information in the spectrum simultaneously. Several approaches have been proposed to remove the baseline.[7,20,21] Most instrument manufacturers have developed their own software for baseline correction. In this work, we have used the FlexAnalysis software from Bruker Daltonics for baseline correction. Since baseline effects differ a lot between different mass spectroscopic techniques, and we want to keep our results general, we decided not to include different procedures for baseline correction as a variable in our investigation of effects of pretreatment.

Smoothing. Mass spectra are quite noisy, and this may give rise to problems in the alignment or synchronization of spectra along the m/z axis using cross-correlation algorithms (see below). Smoothing may help to reduce these problems. Many methods for smoothing have been proposed, the most used being moving average and Savitsky−Golay. In this work, we have used moving average with a window size of 10 for all the data sets except one where the manufacturer's software was used (FlexAnalysis software version 2.4 from Bruker Daltonics).

Alignment/Synchronization. When comparing spectral profiles in, for example, PCA, it is assumed that each column of the data matrix represents the same variable in all samples. As pointed out by Andersson and Hämäläinen,[8] violation of this assumption seriously limits the possibility of extracting information from instrumental profiles. Even small shifts in a series of profiles may cause large inconsistencies in the data matrix and thereby seriously reduce the cross-correlation between chemically similar profiles. Wong et al.[11,12] used fast Fourier transform to align spectra to a reference profile. They aligned segments of the profiles and optimized cross-correlation between reference and profiles to be aligned for each segment. Their method, called FFT/Peak matching combined method implemented in SpecAlign,[13] is fast when working on complex profiles with high data point density. We have therefore chosen to use this method in our work presented here.

Binning. The use of whole spectral profiles may lead to data processing problems due to the large number of points describing a profile. This is one of the reasons why many researchers reduce profiles to peaks. Reduction of a profile to peaks inherently leads to loss of information. Minor peaks close to noise level and shoulders on larger peaks due to overlapping neighboring peaks frequently disappear when integrating profiles to peaks. This can to a certain extent be reduced by using a proper procedure for curve resolution, but this leads to additional work and is not straightforward for spectral profiles since information about number of overlapping peaks in a cluster can only be assessed by comparing many similar samples in a mathematical procedure.[22] Maximum entropy reduction[17] and binning represent two methods for compromising between full spectral profiling and peak integration. Maximum entropy adds intensity of adjacent data

[18] Baggerly, K. A.; Morris, J. S.; Wang, J.; Gold, D.; Xiao, L.-C.; Coombes, K. R. *Proteomics* **2003**, *3*, 1667−1672.

[19] Lee, K. R.; Lin, X.; Park, D. C.; Eslava, S. *Proteomics* **2003**, *3*, 1680−1686.

[20] Wagner, M.; Naik, D.; Pothen, A. *Proteomics* **2003**, *3*, 1692−1698.

[21] Liu, Q.; Krishnapuram, B.; Pratapa, P.; Liao, X.; Hartemink, A.; Carin, L. *Asilomar Conference on Signals, Systems and Computers*, November 2003.

[22] Hämäläinen, M. D.; Liang, Y.-Z.; Kvalheim, O. M.; Andersson, R. *Anal. Chim. Acta* **1993**, *271*, 101−114.

points in the mean spectrum until a preset limit is reached. In the ideal case, this should lead to an almost flat mean spectrum after data reduction. The method adds many adjacent data points in regions with low signal intensity compared to regions with high intensity. Thus, the method assumes that the information content is higher in regions with high intensity. This may not be a valid assumption in proteomic applications since we may expect that, in many cases, the low-intensity regions may be more information-rich than the high-intensity regions and should therefore not be reduced more than the high-intensity regions. By the binning procedure, adjacent m/z numbers with a fixed window size are added throughout the spectrum. The window size is chosen to retain a "good" description of the features in the profiles, and the choice of window size is performed according to the number of points needed to describe a typical peak in the profile. For instance, if a peak is originally described by 100 m/z numbers, a window size of 10 (corresponding to approximately 1 and 5 Da for low and medium mass range, respectively; see Experimental Section) would in most cases be appropriate to balance the time needed for data processing with the need of retaining the spectral features. Both binning and maximum entropy has an additional effect of smoothing the spectra and may make other commonly used smoothing procedures, such as Savitsky−Golay or moving average, superfluous.

Normalization. Without internal standards, it is a common procedure to normalize mass spectral data to create a profile of relative intensities. A variety of procedures exists for this step. The most common procedure is normalization to constant sum:

$$\mathbf{z}_i^{\mathrm{T}} = \mathbf{x}_i / \sum_{j=1}^{M} \mathrm{x}_{ij} \quad i = 1, 2, ...N \tag{2a}$$

In eq 2a, $\mathbf{x}_i$ and $\mathbf{z}_i$ represent the profile for sample i before and after normalization, respectively. M is the number of points describing the profile, and N is the number of instrumental profiles acquired. In bypassing, we note that the so-called TIC normalization[13,16] deviates from normalization to constant sum only by a scalar and thus provides the same correlation patterns and information as obtained by normalization to constant sum.

Instrumental profiles can also be normalized to unit length:

$$\mathbf{z}_i^{\mathrm{T}} = \mathbf{x}_i / ||\mathbf{x}_i|| \quad i = 1, 2, ...N \tag{2b}$$

Normalization to unit length implies larger impact of the more intense parts of the profiles in the normalization procedure, but the effect of heteroscedastic noise on this normalization should else be closely similar to its effect on normalization to constant sum.

Heteroscedasticity. It is well-known that normalization is influenced by structured noise. Increasing noise level with increasing signal induces false positive correlations between spectral regions with low signal intensity and negative correlations between regions with high signal intensity.[14−15] This may give rise to false biomarker candidates. Several procedures have been proposed to reduce this effect. Selective normalization, whereby a subset of medium-sized peaks with relatively small variation is used as internal standards, was proposed by Johanson et al.[23] This procedure reduces the effect of heteroscedasticity on normalization, but a successful selection of internal standards needs replicated samples or prior knowledge about components that are only varying with sample size. Furthermore, the method is subjective in the selection step. Another possibility may be to remove regions with major peaks before normalization, but this may lead to loss of information if the components present in these regions are of discriminatory importance. Kvalheim et al.[14] pointed out that the nth root transform might be used to reduce heteroscedasticity in spectral profiles. For instance, if there is a linear relationship between mean signal and the variance of signal, a square root transform will provide homoscedastic noise. In addition to taking care of structured noise, the nth root transform has a major advantage compared to the log transform: It does not influence perfect correlations in profiles and destroys partial correlations to a much lesser extent than the log transform. This is important since one component in a mass spectral profile is described by several correlated m/z numbers. The conservation of this correlation is crucial for a successful correlation analysis using, for example, PCA on the preprocessed profiles. The value of n is chosen by plotting variance as a function of mean signal intensity for different choices of n for replicated profiles. The lowest n that provides an approximate homoscedastic noise structure is chosen.

The nth root transform reduces the signal intensity. This influences to a larger extent the most intense regions in the profiles than those with lower intensity. This may be a drawback if regions with high intensity contain much more information than regions with low intensity. For tasks such as searching for biomarkers to be used for early diagnosis of diseases, one may expect that regions with low intensity might be at least as important with respect to information as the high-intensity regions. Thus, a relative reduction of the intense regions compared to the low-intensity regions may actually be an advantage. The reduction in signal intensity accompanying the nth root transform also impacts regions with intensity close to noise level, and this may lead to loss of information for such regions. In such cases, it may be convenient to do a normalization and local PCA just on these regions without the nth root transform since without the presence of high-intensity regions, normalization of data is less influenced by heteroscedastic noise.[24]

Inter- to Intragroup Variation. The total variation V_{total} in a collection of spectral profiles can be partitioned into variation of spectral profiles within groups of replicated samples (or spots), V_{intra}, and variation between these groups, V_{inter}:

$$V_{\mathrm{total}} = V_{\mathit{int}\mathrm{ra}} + V_{\mathrm{inter}} \tag{3}$$

The purpose of the pretreatment of data is to enhance the information in the spectral profiles. This is obtained when the ratio R of between-group variation to intragroup variation increases:

$$R = V_{\mathrm{inter}} / V_{\mathrm{intra}} \tag{4}$$

In order to decide on the optimal pretreatment procedure, we therefore use the ratio of inter- to intragroup variation R as

[23] Johansson, E.; Wold, S.; Sjödin, K. *Anal. Chem.* **1984**, *56*, 1685−1688.

[24] Sletten, E.; Kvalheim, O. M.; Kruse, S.; Farstad, M.; Søreide, O. *Eur. J. Cancer* **1990**, *26*, 615−618.

response in factorial designs covering the different pretreatment options needed. The response R is the summed inter- to intragroup variation over each m/z number. In our work here, this means summing as many as 44 403 contributions. If whole profiles are converted into peaks, the number of contributions to R is the number of peaks. The number R increases when executing pretreatment procedures that correct for experimental effects of analytical workup and instrumental conditions. Furthermore, the use of factorial designs enables the possibility to assess the importance of the different preprocessing steps.

EXPERIMENTAL SECTION

Sampling. Cerebrospinal fluid (CSF) samples were taken early in the morning by a standard lumbar puncture procedure from patients at the Department of Neurology, Haukeland University Hospital. About 10 mL of CSF was immediately placed on ice before freezing at −80 °C. The study was approved by The Regional Committee for Medical Research Ethics of Western Norway and included a written, informed consent from the selected patients.

Data Set 1. A CSF pool was created by mixing CSF from five different patients. Six aliquots of 0.5 mL were prepared from the CSF pool. Three of these were spiked with 1600 pM peptide standard (Bruker). The two sets of triplicates were fractionated and spotted three times each on the MALDI target and analyzed in the low molecular weight range (1−9 kDa) as described by Berven et al.[25] This provided a data set consisting of 18 spectral profiles, each being described by 44 403 m/z values (see paragraph on mass spectrometry).

Data Set 2. Fresh CSF was obtained from a 30-year-old male with mild hemiparesis and sensory symptoms in his right arm and leg. The CSF sample contained some traces of blood that leaked into the CSF upon collection. The sample was split and stored under 15 different storage conditions as described by Berven et al.[25] Each of the 15 0.5-mL aliquots from the patient was fractionated through individual 20-kDa molecular weight cutoff (MWCO) filters to create one low MW fraction and one fraction of proteins that could pass through the filter after treatment with 6 M guanidinium hydrochloride (designated the guanidinium fraction). Only the guanidinium fraction in the mass range 6−17 kDa was analyzed in this work. The samples were analyzed in triplicates to provide 45 spectral profiles, each described by 16 598 m/z values.

Data Set 3. A CSF pool was created by mixing CSF from 10 different patients. Five replicates were fractionated and spotted three times each on the MALDI target plate and analyzed in the low molecular mass range (1−9 kDa) as described by Berven et al.[25] Each profile is described by 44 403 m/z values.

Mass Spectrometry. All samples were analyzed using an AutoFlex (Bruker Daltonics) mass spectrometer in a positive linear mode. Data were acquired in two different ranges, for data sets 1 and 3 in the 1−9 kDa defined as the low mass range, and for data set 2 in the 6−17 kDa defined as the medium mass range (for details, see Berven et al.[25]). Each spectral profile acquired using the low mass range (data sets 1 and 3) was described by intensities at 44 403 m/z numbers, starting at 740.04 Da and increasing in steps of 0.08 to 8999.84 Da. For the medium mass range (data set 2), each spectral profile was described by intensities at 16 598 m/z numbers, starting at 6000.04 Da and increasing in steps of 0.07 to 16 442.62 Da.

Table 1. Design Matrix and Results for the Response Variable for Reference and Spiked Sample (Data Set 1)

exp	smooth/ binning	alignment	noise	normalization	inter/intra-variation
1	−1	−1	−1	−1	0.22
2	1	−1	−1	−1	0.22
3	−1	1	−1	−1	0.22
4	1	1	−1	−1	0.22
5	−1	−1	1	−1	0.27
6	1	−1	1	−1	0.27
7	−1	1	1	−1	0.27
8	1	1	1	−1	0.27
9	−1	−1	−1	1	0.76
10	1	−1	−1	1	0.76
11	−1	1	−1	1	1.21
12	1	1	−1	1	0.95
13	−1	−1	1	1	1.02
14	1	−1	1	1	1.02
15	−1	1	1	1	1.08
16	1	1	1	1	1.27
17	1	1	fourth root	1	1.27
18	1	1	1	constant sum	1.26
19	1	1	fourth root	constant sum	1.26

Table 2. Design Matrix and Results for the Response Variable for Storage Data (Data Set 2)[a]

exp	alignment/ shift	binning/ reduction	noise	normalization	R_1	R_2
1	−1	−1	−1	−1	6,67	6,67
2	1	−1	−1	−1	6,76	6,76
3	−1	1	−1	−1	6,70	6,70
4	1	1	−1	−1	6,78	6,72
5	−1	−1	1	−1	7,35	7,35
6	1	−1	1	−1	7,33	7,33
7	−1	1	1	−1	7,42	7,42
8	1	1	1	−1	7,39	7,42
9	−1	−1	−1	1	13,30	13,30
10	1	−1	−1	1	14,05	14,05
11	−1	1	−1	1	13,84	13,84
12	1	1	−1	1	14,43	13,69
13	−1	−1	1	1	16,52	16,52
14	1	−1	1	1	16,49	16,49
15	−1	1	1	1	18,87	18,87
16	1	1	1	1	18,30	19,69

[a] R_1 and R_2 are inter- to intragroup variation with alignment prior to binning and binning prior to alignment, respectively.

Design/Pretreatment. All three data sets were baseline corrected using the FlexAnalysis software from Bruker Daltonics. The baseline correction produced some regions with negative intensities. To provide profiles without negative intensities, the profiles were independently shifted by the absolute value of the largest negative intensity in each profile prior to further pretreatment. For data set 2, smoothing was also performed using FlexAnalysis prior to executing the factorial designs. For all three data sets, a 2^4 factorial design was performed. For data sets 1 and 3, smoothing (−1) versus binning (+1), no alignment (−1) versus alignment (+1), heteroscedastic (−1) versus homoscedastic noise (+1), and non-normalized (−1) versus normalized (+1) profiles constituted the 16 initial runs (Tables 1 and 3). A

[25] Berven, F. S.; Kroksveen, A. C.; Berle, M.; Rajalahti, T.; Flikka, K.; Arneberg, R.; Myhr, K.-M.; Vedeler, C.; Kvalheim, O. M.; Ulvik, R. J. *Proteomics Clin. Appl.* In press.

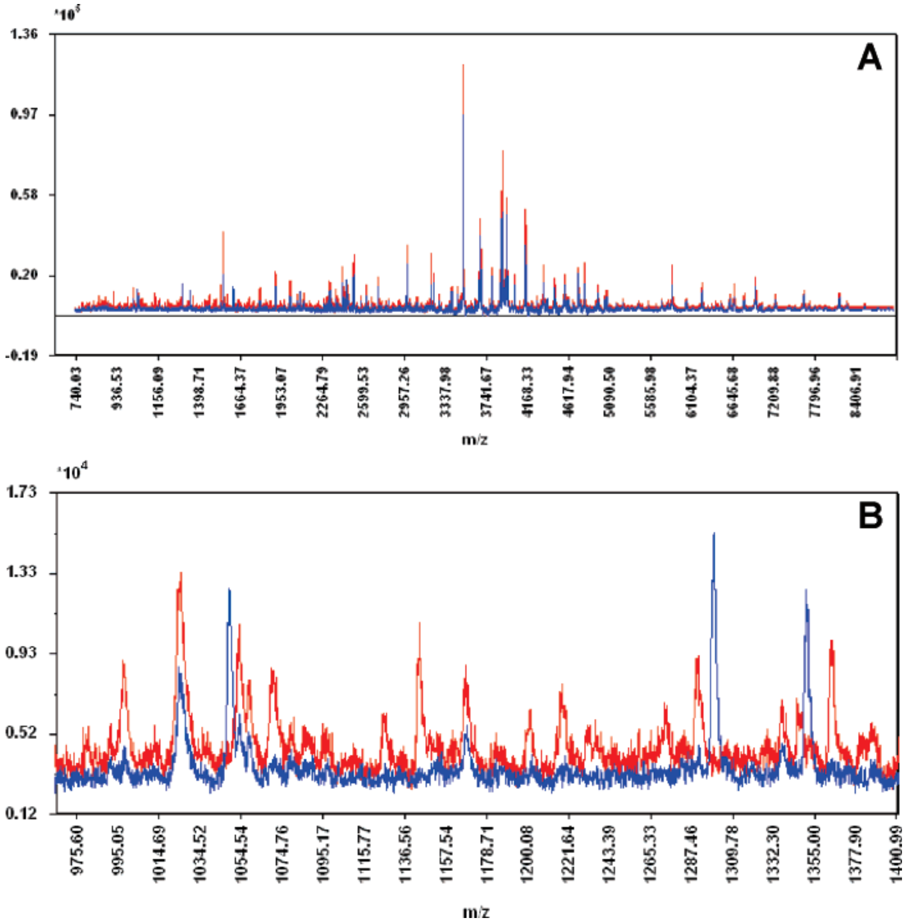

Figure 1. Baseline-corrected (A) and zoomed baseline-corrected (B) raw data, one profile from reference sample SP0 (red) and one profile from spiked sample SP3 (blue).

Table 3. Design Matrix and the Results for the Response Variable for Replicated Sample (Data Set 3)

exp	smooth/ binning	alignment	noise	normalization	inter/intra-variation
1	−1	−1	−1	−1	0.54
2	1	−1	−1	−1	0.54
3	−1	1	−1	−1	0.37
4	1	1	−1	−1	0.41
5	−1	−1	1	−1	0.49
6	1	−1	1	−1	0.49
7	−1	1	1	−1	0.44
8	1	1	1	−1	0.45
9	−1	−1	−1	1	2.35
10	1	−1	−1	1	2.36
11	−1	1	−1	1	1.34
12	1	1	−1	1	1.32
13	−1	−1	1	1	1.60
14	1	−1	1	1	1.60
15	−1	1	1	1	0.88
16	1	1	1	1	0.99
17	1	1	1	constant sum	0.99

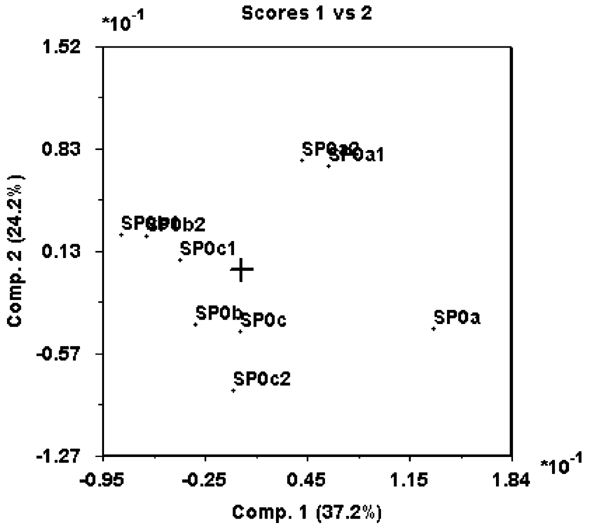

Figure 2. Scores on principal component 1 vs principal component 2 for reference sample SP0 (three replicates spotted three times). Baseline-corrected and normalized data.

few additional runs were executed for data set 1 as explained in the Results and Discussion section. For data set 2, two designs were executed. For both designs, the four variables were as follows: no binning (−1) versus binning (+1), no alignment (−1) versus alignment (+1), heteroscedastic (−1) versus homoscedastic noise (+1), and non-normalized (−1) versus normalized (+1). The 16 runs are listed in Table 2. The difference between the two designs for data set 2 is the order of alignment and binning. The ratios R_1 and R_2 refer to inter- to intragroup variation with alignment prior to binning and vice versa, respectively. Spectra were pretreated according to the setting of the variables in the design. Smoothing was performed by moving average using a 10-point window. Binning was performed by adding 10 points, reducing the data point resolution to ~1 Da. Alignment was executed by SpecAlign using a window size of 20. Transformation

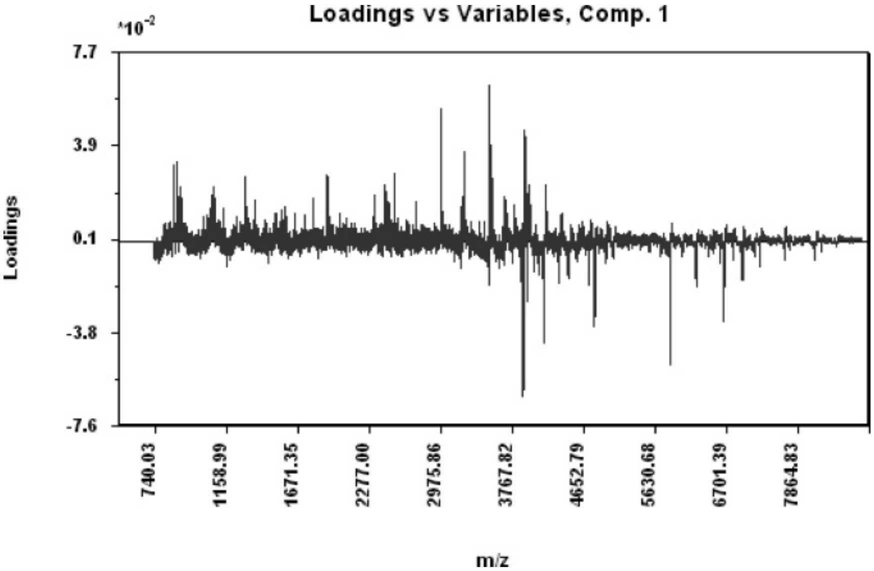

Figure 3. Loadings on the first principal component for reference sample (SP0). Baseline-corrected and normalized data.

from heteroscedastic to homoscedastic noise was carried out by a third root transform except for a few additional runs for data set 1 where the fourth root was used (see Table 1). The decision to use the third root was based on noise structure in replicated spots. Normalization was performed to unit length (eq 2b) except for a few additional runs for data sets 1 and 3, where normalization to constant sum was performed.

The ratio R of between-group variation to intragroup variation (eq 4) was calculated by partitioning the samples in the following way: For data set 1, profiles were divided into two groups, reference samples and spiked samples, respectively. For data set 2, profiles were divided into 15 groups of triplicates, each triplicate corresponding to a particular storage condition.[25] For data set 3, the profiles were divided into five groups, each group being replicates spotted three times. For data sets 1 and 2, the optimal pretreatment procedures should reduce the intraclass variation more than the between-group variation, implying that R should increase during preprocessing. For data set 3, the best pretreatment procedures are those that remove the undesired effects without increasing R.

Software. The MALDI-TOF sampling and preanalysis was performed using FlexAnalysis. This analysis also includes the baseline correction procedure. SpecAlign version 2.3 from Cartwright Group PTCL, University of Oxford, was used for peak alignment. The ratio of inter- to intragroup variation was programmed and calculated by using MATLAB version 6.5 from Mathworks Inc. Sirius version 7.0 from Pattern Recognition Systems was used for all additional analysis including smoothing, binning, pretreatment, and PCA.

RESULTS AND DISCUSSION

In this section, we assess the influence of the different pretreatment procedures for three data sets of spectral profiles. All the spectral profiles were baseline corrected using the instrument vendor's software. The baseline correction produces some regions with negative intensities. Prior to further preprocessing, the entire profiles were therefore independently adjusted by the absolute value of the largest negative intensity so that the lowest intensity became zero. For spectral alignment, we have used the software SpecAlign[13] developed by Wong et al.[11,12] For smoothing, we use moving average with a 10-point window, and for binning, we use a 10-point window that corresponds approximately to merging m/z numbers into units of 1 Da. For all three data sets discussed below, we use the intensities of the whole spectral profiles after chosen pretreatment to calculate the response $R = V_{\mathrm{inter}}/V_{\mathrm{intra}}$ and as input to PCA. Note that PCA is only used for illustrative purposes in this work. For quantification of different procedures for pretreatment, the reader is referred to the response R.

Profiles from Reference and Spiked Sample (Data Set 1). Figure 1A shows the spectral profiles of two samples from the data set with three reference samples each spotted three times and three samples spiked with a standard and also spotted three times each. The zoomed part in Figure 1B shows differences between the profile from the reference sample and the profile from the spiked sample. Before further analysis, the profiles were examined for outliers. Figure 2 shows the PCA score plot of the profiles acquired for the reference sample. Prior to PCA, the profiles were normalized to unit length to make them comparable. Heteroscedastic noise influences the normalization, but for the purpose of revealing outliers in similar samples, this is of little concern. We observe that the spot from the profile labeled SP0a represents a strong outlier on PC1. The two other spots from the same sample (SP0a1 and SP0a2) are located as an isolated cluster in the score plot. The loading plot on PC1 (Figure 3) shows differences in regions with high intensity between the three spots from the outlying sample and the profiles from the two other reference samples. The three profiles (from sample SP0a) were therefore removed from the data set. The spiked profiles formed a homogeneous group in the PCA score plot. The further analysis is therefore based on 15 spectral profiles, 6 profiles from the reference sample (SP0b and SP0c) and 9 from the spiked sample (SP3a, SP3b, and, SP3c).

Three homogeneous replicated profiles from one sample were selected and analyzed for heteroscedasticity. Panels A–E in Figure

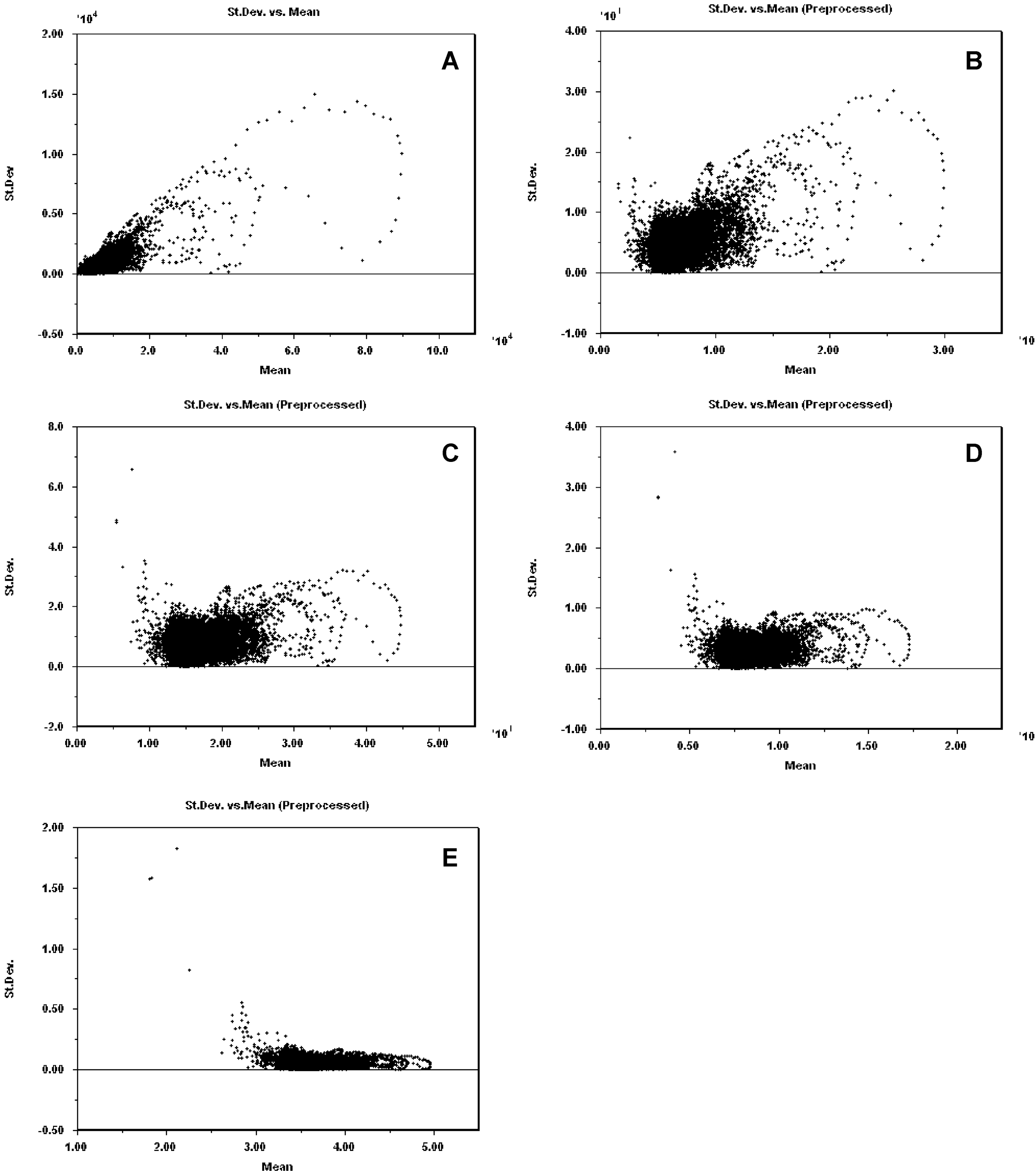

Figure 4. Standard deviation plotted vs mean intensities for a baseline-corrected reference sample spotted three times. (A) Raw data; (B) square root transform; (C) third root transform; (D) fourth root transform; (E) log transform.

4 show the standard deviation as a function of mean intensity for raw, second, third, fourth root, and log transform, respectively, for these three replicated profiles. The loop structure observed in the plots is a consequence of using whole mass spectral profiles and the heteroscedastic noise structure. Peak clusters with varying intensities and level of noise show up as loops in these plots. Strong heteroscedastic noise is observed in the raw data (Figure 4A). The square root transform (Figure 4B) removes most of the heteroscedasticity, and the third (Figure 4C) and fourth root transform (Figure 4D) provides a uniform noise distribution. The log transform (Figure 4E) is too strong and provides a noise distribution where noise decreases with mean intensity of the replicates. Since *n*th root transform reduces both signal size and partial linear correlations in signal, we propose to be conservative and use either the square or third root transform with results like those in Figure 4. In our further analysis, we use the third root transform, but the investigator has to make his choice based on the characteristics of his own data.

A 2^4 factorial design was executed on the baseline-corrected data. Four pretreatment procedures on the spectral profiles were investigated: smoothing using moving average or binning, both with a window size of 10, alignment (SpecAlign), noise removal using the third root transform, and normalization to unit length (eq 2b). For all possible combinations, the inter- to intragroup variation was calculated according to eq 4. The design matrix and the results for the response variable are shown in Table 1.

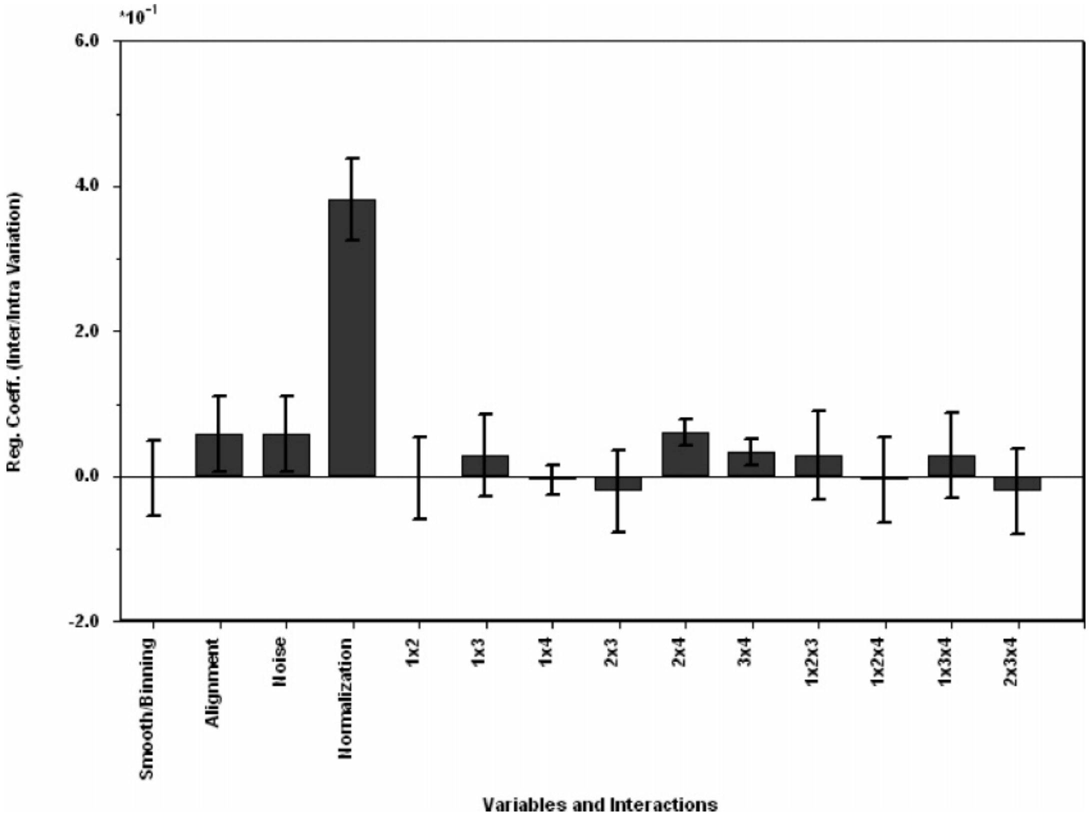

Figure 5. Regression coefficients with uncertainty estimate for main effects and interactions between the different pretreatment steps for data set 1.

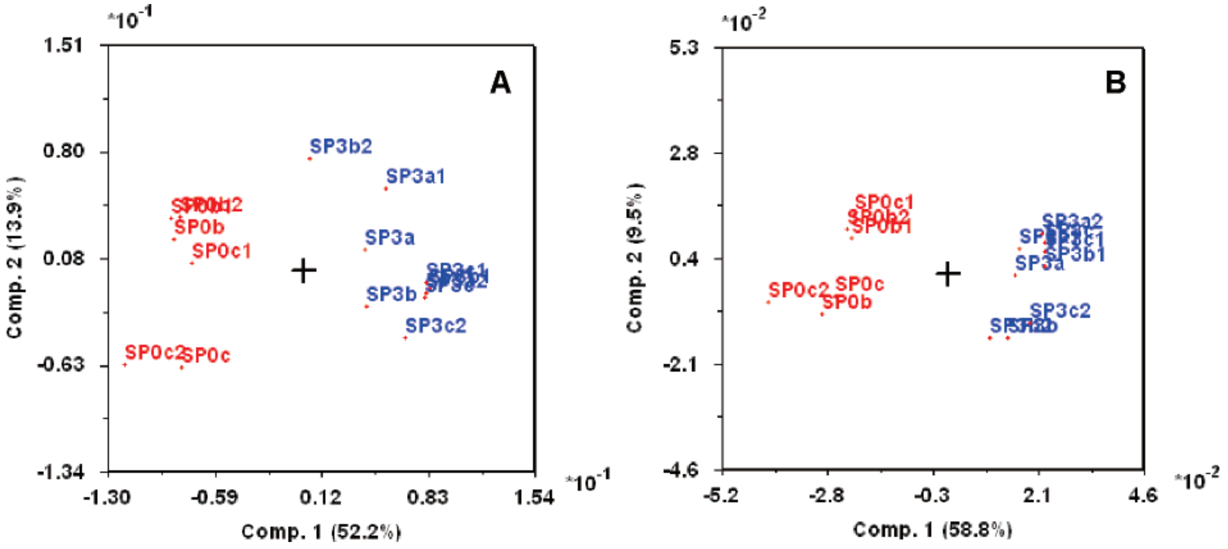

Figure 6. Score plot on first and second principal components for reference and spiked samples (A) for run 12 in Table 1 (without heteroscedastic noise correction) and (B) for run 16 in Table 1 (with heteroscedastic noise correction).

Multiple linear regression provides the regression coefficients in Figure 5. Positive bar means that the inter- to intragroup variation ratio R is increasing when a variable is changed from -1 to $+1$ level, while negative bar means that a change from -1 to $+1$ level reduces R. For data set 1, increasing R is implying that profiles from one group become more similar without destroying the compositional correlation pattern. We observe that normalization is by far the most important factor for enhancing differences between spiked and reference samples and reducing replicate variation. Alignment and removal of heteroscedastic noise have a comparable positive effect, while there is no difference between smoothing and binning on increasing or decreasing differences between profiles. This latter observation is positive since it implies that by binning it is possible to reduce the number of m/z values with 1 order of magnitude without losing information. We further observe a positive interaction between alignment and normalization, proving the point that looking at one pretreatment step in isolation from the others may be of limited value. The score plots without and with reduction of heteroscedastic noise, Figure 6A and B, respectively, show the importance of noise

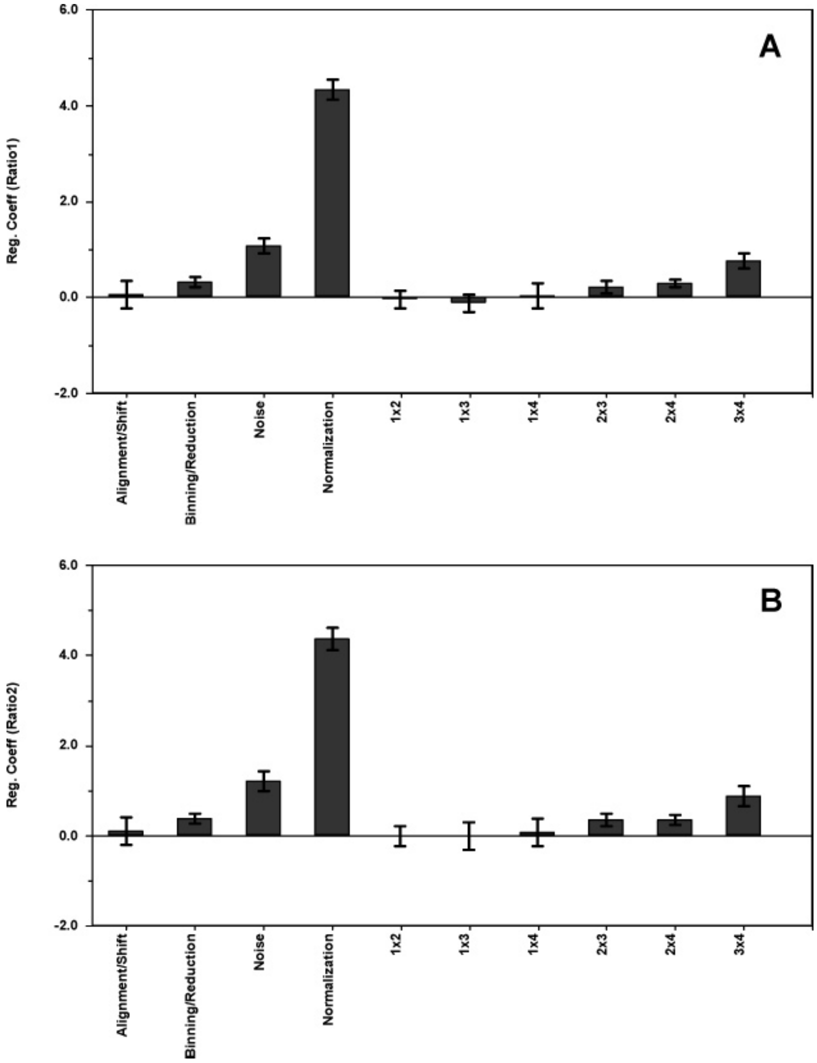

Figure 7. Regression coefficients with uncertainty estimate for main effects and interactions between the different pretreatment steps for data set 2. (A) Alignment prior to binning; (B) binning prior to alignment.

reduction prior to normalization. The increased inter- to intragroup variation after third root transform is reflected in the larger similarity between replicates. Furthermore, explained variance on PC1 is increased after the third root transform, proving that linear correlations have survived the transform. Three additional pretreatment procedures were investigated for these data: using fourth root transform in run 17, changing normalization to constant sum in run 18, and combining fourth root transform with normalization to constant sum in run 19 (Table 1). These three additional procedures provided the same inter- to intragroup variation to the second decimal place as run 16 showing that our result is robust to these changes in pretreatment procedures.

Profiles from Storage Study (Data Set 2). These profiles from the medium molecular weight fraction were both baseline corrected and smoothed using the instrument vendor's software. A 2^4 factorial design was performed with alignment, binning, third root transform, and normalization to unit length as design variables. We further investigated the effect on R of changing the order of alignment and binning, i.e., binning before or after alignment. The reason for this is that binning act as a smoothing agent, which can be of importance for a successful alignment if smoothing by moving average or Savitsky−Golay is dropped. The results are shown in Table 2. Regression models for the design with alignment before binning, and vice versa, are shown in Figure 7A and B. The two models are very similar: Normalization is most important, followed by correction for heteroscedastic noise. We notice a strong interaction between noise pattern and normalization, again underlining the danger of looking at the different pretreatment steps in isolation. Binning has a small positive effect on enhancing differences between different profiles and reducing differences between replicated profiles. This implies that it is possible to reduce the m/z number by 1 order of magnitude without losing information. Surprisingly, alignment had no overall positive effect on the profiles. This may imply that the smoothing routine used in FlexAnalysis also affects positively the synchronization of profiles. Further analysis of these spectral profiles from a study assessing effects of storage on CSF samples can be found in ref 25.

Profiles from One Replicated Sample (Data Set 3). The profiles were acquired on five replicated samples fractionated independently. Each replicate was spotted three times. Figure 8A (whole) and Figure 8B (zoomed) show spectra from two different wells after baseline correction, smoothing, and alignment. For these data, a successful pretreatment should not increase separa-

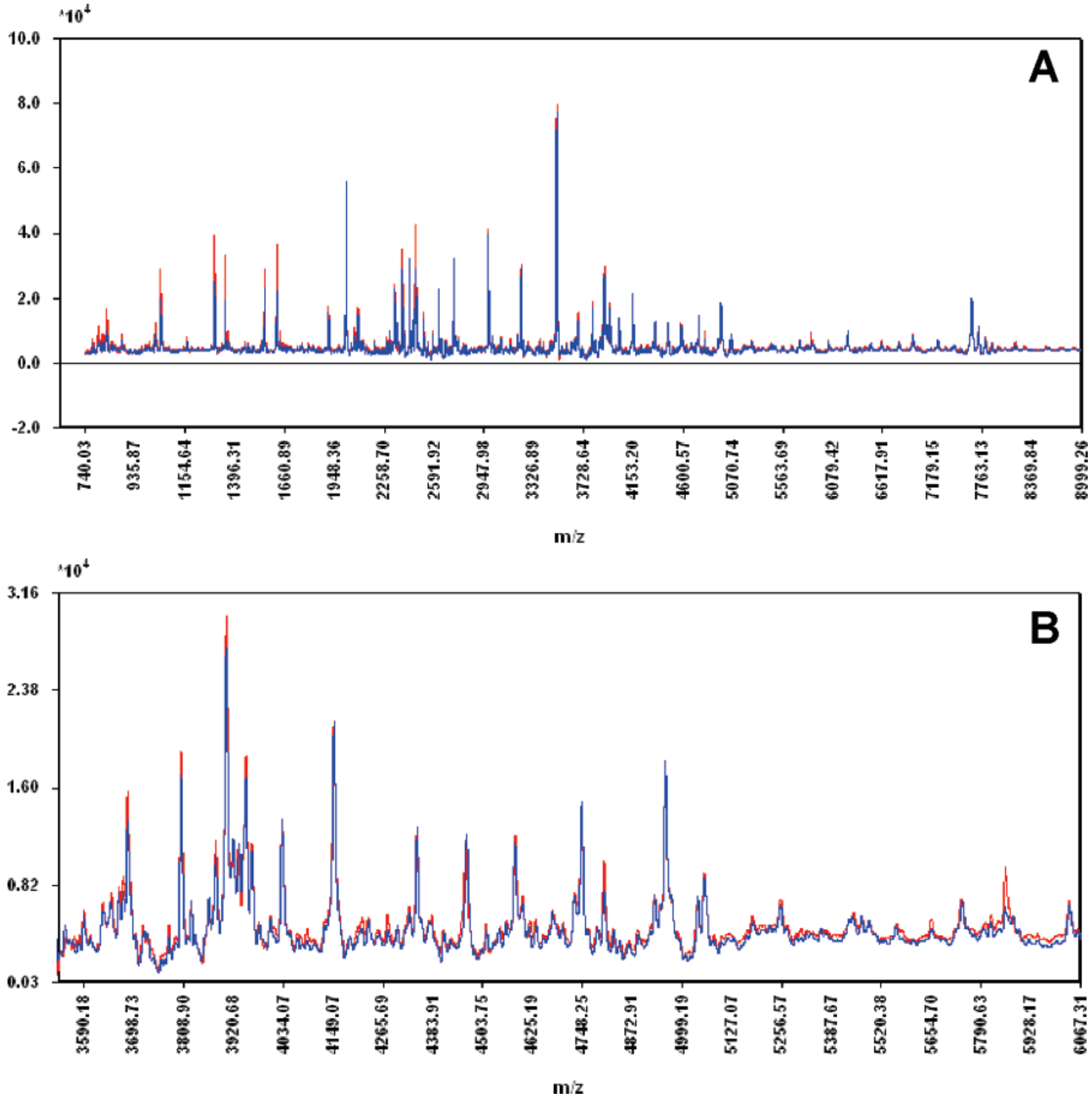

Figure 8. Baseline-corrected, smoothed, and aligned profiles (A) and zoomed baseline-corrected, smoothed, and aligned profiles (B) from replicated samples 1a (red) and 4a (blue).

tion to a large extent since any difference between samples and spots should, by definition, be random. A score plot of the raw data (Figure 9A) shows some grouping of spots from the same replicate. A score plot of normalized data (Figure 9B) shows even stronger grouping of spots from the same replicate, underlining the potential of heteroscedastic noise to induce false correlations in data. Furthermore, the variation between replicates shows that the experimental workup of the low molecular weight fraction introduces systematic differences that are larger than those produced through spotting and mass spectral analysis. Table 3 shows the results of a 2^4 factorial design with smoothing/binning, alignment, third root transform, and normalization as design variables, Runs 9 and 10 show the strong impact of heteroscedastic noise on normalization. Compared to the runs without normalization (runs 1−8), the inter- to intragroup variation has increased by a factor of 5−6. The regression model (Figure 10) shows no differences between smoothing and binning, but strong effects of normalization, alignment, noise, and their interactions. Alignment and third root transform reduce the differences between spots from different samples as is desired, while normalization increases the differences. There are, however, strong interactions between alignment and normalization and between noise pattern and normalization. Both these interactions reduce *R*. Again the results show the necessity to assess the impact of pretreatment procedures jointly. By comparing experiments 15 and 16 with experiments 9 and 10 in Table 3, we note that alignment and third root transform prior to normalization make replicates more similar by a factor of 2−3 compared to normalization alone. Thus, normalization has a clear effect of increasing differences between the replicates instead of making them more similar, as desired. This reflects analytical errors inherent to the analytical workup. Compositional heterogeneities can arise from, for example, sample fractionation, the spotting of sample and matrix onto the MALDI target plate or during data acquisition on the mass spectrometer. Normalization amplifies such induced compositional differences, and sadly, there is no way to remove these undesired effects mathematically.

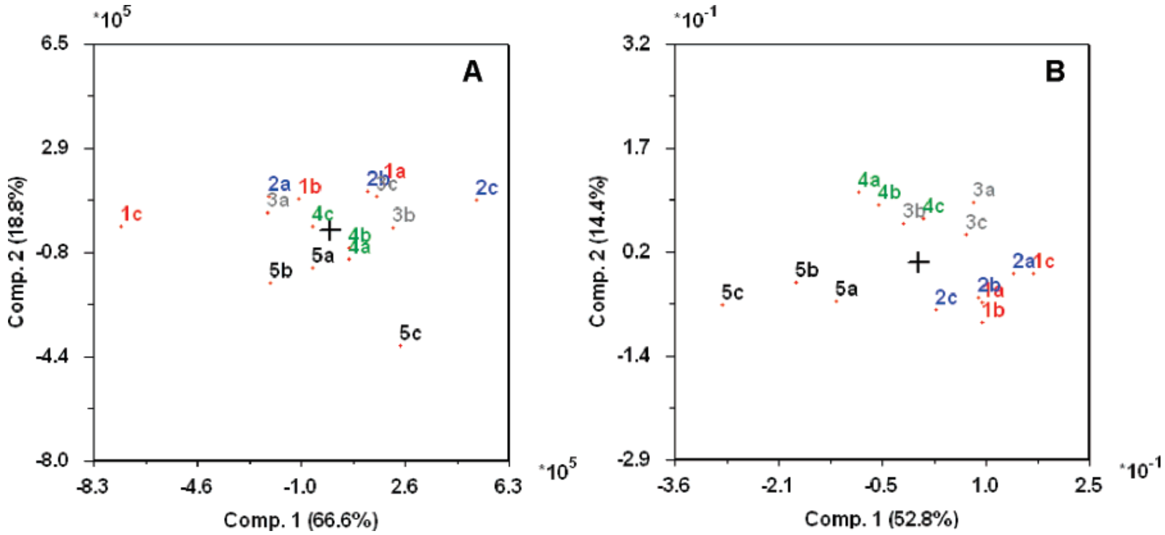

Figure 9. Score plot on first and second principal components for 5 replicates spotted three times (a, b, c). (A) Baseline corrected. (B) Baseline corrected and normalized.

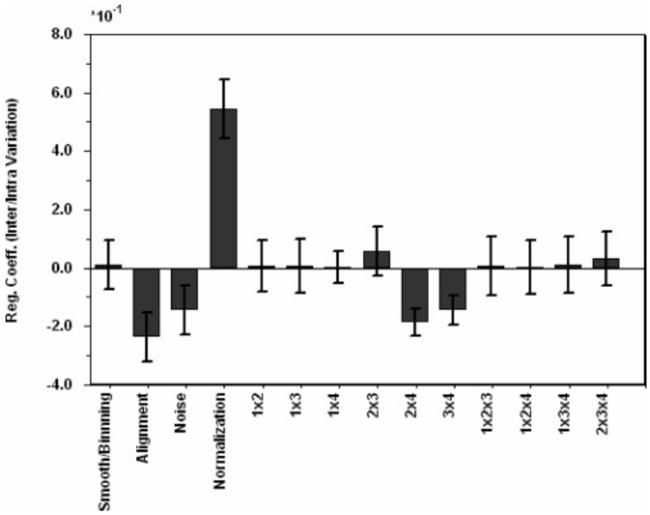

Figure 10. Regression coefficients with uncertainty estimate for main effects and interactions between the different pretreatment steps for data set 3.

CONCLUSIONS

We have investigated several pretreatment steps for making mass spectral profiles from, for example, MALDI experiments amenable to correlation analysis. Such analysis is crucial for finding discriminating spectral regions that may contain information about differences in protein signatures between, for example, controls and disease-impacted persons. The aim of correlation analysis in such cases is to be able to use the spectral profiles for early medical diagnosis of diseases that are otherwise difficult to diagnose clinically at their early stage. In order to make a quantitative assessment of the impact of the different pretreatment procedures on the mass spectral profiles, we used factorial designs with the ratio between intergroup and intragroup variance as response.

Normalization to make spectral signatures comparable in the absence of internal standards has been shown to be vulnerable to heteroscedastic noise. Normalization without prior adjustment of the noise structure may give rise to false biomarker candidates. The remedy for this problem is to perform the *n*th root transform with *n* chosen so that the standard deviation becomes uniform with increasing signal for the regions investigated for a set of homogeneous instrumental replicates. Our analysis has shown that the procedure for removing heteroscedastic noise is reasonable robust to the choice of *n*, i.e., second, third, or fourth root transform provides almost the same result on the data set analyzed in this work. The choice of *n* can therefore, to some extent, be made to balance the contributions of small and large peaks in the analysis. The log transform is too rough for mass spectral profiles. It gives rise to a decreasing standard deviation with increased signal and, in addition, destroys linear correlations in the spectral profiles. The deteriorative effect of the log transform can be easily demonstrated by selecting a m/z region describing a single peak and perform PCA on the log-transformed profiles. The use of normalization to constant sum (equivalent to TIC normalization) or to unit length provides no differences for the data sets analyzed in this work. Binning was shown to be able to reduce the number of data points in the profiles with 1 order of magnitude without loss of information. For profiles acquired with high sampling density, this may be a useful option to reduce memory resources and speed up data analysis. Furthermore, smoothing seems to be unnecessary when binning is performed. This is not surprising since binning will act as a smoothing agent. Alignment of profiles is necessary in most cases and should be routinely performed.

What about the order of the different pretreatment steps? There are no principal objections against executing the *n*th root transform prior to alignment, but it may have a negative effect on the alignment of minor peaks since these will be reduced in size. In the absence of smoothing, binning should be executed prior to alignment. If binning is not performed, the order of pretreatment should be smoothing, alignment, *n*th root transform, and normalization.

Our analysis shows that the different pretreatment steps cannot be assessed in isolation. Strong interactions are observed, especially between heteroscedastic noise and normalization, and normalization and alignment. Also, our analysis has shown the need for replicate analysis; both to be able to reduce heteroscedastic noise and to assess experimental errors.

ACKNOWLEDGMENT

This work was supported by grants from the aid of EXTRA funds from the Norwegian Foundation for Health and Rehabilitation, the Kjell Almes Legacy, the Bergen MS Society, the Odd Fellow, the Meltzer foundation, the Norwegian Society of Multiple Sclerosis, and Fritz and Ingrid Nilsen's Legacy for Research in Multiple Sclerosis, Norway. The authors are partly supported by the National Programme for Research in Functional Genomics (FUGE) funded by the Norwegian Research Council, and the Western Norway Regional Health Authority. Pattern Recognition Systems AS is thanked for partial financing of R.A. and for providing the Sirius software free of charge.

Received for review May 9, 2007. Accepted July 3, 2007.

AC070946S