Research Article

Journal of MASS SPECTROMETRY

Received: 10 April 2008 Accepted: 9 October 2008 Published online in Wiley Interscience: 8 December 2008

(www.interscience.com) DOI 10.1002/jms.1527

A strategy for the prior processing of high-resolution mass spectral data obtained from high-dimensional combined fractional diagonal chromatography

Dirk Valkenborg,[a,b]* Grégoire Thomas,[c] Luc Krols,[d] Koen Kas[c] and Tomasz Burzykowski[b]

Combined fractional diagonal chromatography (COFRADIC) is a novel suite of gel-free technologies for the identification of biomarkers in complex peptide mixtures. For this purpose, reversed-phase high performance liquid chromatography (HPLC) technology and, in this case, matrix assisted laser desorption /ionization- time of flight (MALDI-TOF) mass spectrometers are extensively used. The particular characteristic of COFRADIC mass spectrometry data is the high number of chromatographic fractions, over which a peptide can be scattered. This can obstruct the quantification of the peptide abundance in the biological sample, which is required for statistical analysis. On the other hand, because of the superior peptide sorting properties of the methodology, the mass spectra become less crowded. Consequently, each peptide appears in a mass spectrum as a series of peaks with peak heights proportional to the probability of occurrence of the isotopic variants of the peptide. In this manuscript, we propose an analysis strategy concerned with the preprocessing of COFRADIC mass spectra prior to a downstream statistical analysis. The preprocessing algorithm produces for each mass spectrum a peptide list by exploiting the characteristic features that should be associated with peaks corresponding to an isotopically resolved cluster of peptide peaks. This reduction step is necessary to facilitate the clustering used in a next step to assemble the validated monoisotopic peptide peaks found over several fractions into a single peptide abundance. To assess the performance of the algorithm, two technical experiments were conducted. The proposed strategy is memory and computationally efficient.

Supporting information may be found in the online version of this article.

Keywords: mass spectral signal processing; *N*-terminal COFRADIC; gel-free label-free quantification; high-resolution mass spectrometry; bioinformatics

Introduction

In the search for new biomarkers, surrogate endpoints, or markers for classification of diseases, shotgun proteomic techniques are often used to rapidly visualize and compare the protein content in complex mixtures. These shotgun proteomic techniques mostly incorporate multidimensional liquid chromatography (LC), mass spectrometry (MS), and database-searching algorithms, and are characterized by the enormous amount of generated data. We focus particularly on combined fractional diagonal chromatography (COFRADIC), introduced by Gevaert *et al*.,[1] as an example of a high-dimensional LC technique for the separation of dense protein mixtures such as serum. To provide a better understanding of the complexity of the data generated from the COFRADIC methodology, the procedure and typical characteristics of this separation technique will be briefly explained. For more details on the COFRADIC methodology and the MS settings, we refer to the paper of Sandra *et al*.,[2] which describes the procedures quite extensively.

The basic strategy of the *N*-terminal COFRADIC, which focuses on amino-terminal peptides for reducing and separating complex protein mixtures, can be summarized in five consecutive steps. First, proteins from a biological sample are acetylated. This means that an acetyl group is introduced to the amino acid lysine (K) and to the *N*-terminus of the protein. In the second step, the protein is digested (in this case, by trypsin) to peptides. Trypsin cleaves only the *C*-terminus of arginine (R), since lysine is acetylated. There are now two kinds of peptides in the mixture: those that first formed the original protein *N*-terminus, and those resulting from the proteolytic cleavage with trypsin. The former carry acetylated amines, while the latter present a free amine (i.e. a free new *N*-terminus). In the third step, the complex peptide mixture is separated during primary fractionation on a reversed-phase high-pressure liquid chromatography (RP-HPLC) column, which separates the sample on the basis of its hydrophobicity.

* *Correspondence to: Dirk Valkenborg, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Agoralaan 1, B3590, Diepenbeek, Belgium. E-mail: dirk.valkenborg@uhasselt.be*

a *Interuniversity Institute for Biostatistics and statistical Bioinformatics, Katholieke Universiteit Leuven, Leuven, Belgium*

b *Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Diepenbeek, Belgium*

c *Pronota, VIB Bio-Incubator, Zwijnaarde, Belgium*

d *LifeScience Lighthouse, Brasschaat, Belgium*

Journal of MASS SPECTROMETRY

In the fourth step, each collected fraction undergoes a chemical modification of a target subset of peptides to alter the column retention properties. In this case, the free amines, generated during the tryptic digest, are modified by 2,4,6-trinitrobenzene sulfonic acid (TNBS), such that the altered peptide becomes more hydrophobic and shifts to higher retention times. The *N*-terminal peptides of the proteins, which were acetylated prior to digestion, remain unaltered during the TNBS modification. During the fifth step, each of the fractionated and modified sample is further separated on a RP-HPLC column under identical conditions as the primary fractionation. During this refractionation, the *N*-terminal peptides elute from the column at approximately the same time as in the primary fractionation. On the other hand, internal peptides, modified by TNBS, shift out of the original collection interval and are separated from the *N*-terminal peptides. Hence, theoretically, a protein is represented by its acetylated *N*-terminal peptide, while the internal peptides are filtered out from the sample. This procedure maximally reduces the complexity of the peptide mixture, as only one peptide per protein is retained in the sample. Because the *N*-terminal peptide is used as a signature to link to the parental protein, it is of paramount importance that this peptide is observed in the mass spectrum, and that it gives rise to a positive and unique identification on tandem MS or, equivalently, MS2; otherwise, information about the protein is lost from the sample. It should be noted that the complex process to link the selected *N*-terminal peptide to its parent protein is not discussed in this paper. Note that each obtained COFRADIC fraction is spotted onto a MALDI plate. Thus, each fraction gives rise to a single mass spectrum. When processing the *N*-terminal COFRADIC fractions on a high-resolution mass spectrometer, one expects less crowded and better resolved mass spectra.

The particular characteristics of COFRADIC MS data are as follows:

- For a single biological sample, a large number of mass spectra are obtained, corresponding to the number of fractions generated by the primary and secondary chromatographic fractionation, which results in an enormous amount of data (in the considered case approximately 3.88 GB per biological sample).
- A peptide can be present in the spectra from several fractions; to obtain its overall abundance, information from multiple spectra needs to be combined.
- The data quality can differ between consecutive MALDI-TOF mass spectra.

Of course, these characteristics are inherent to most high-throughput gel-free proteomic experiments and therefore the proposed strategy should also be applicable to other LC-MS settings. The only requirement is that the spectra are generated from high-mass-accuracy, high-resolution LC-MS systems, i.e., the individual isotopic variants of a peptide should be discernable in the mass spectrum. However, in this paper we focus on the COFRADIC setting.

Clearly, a major obstacle of peptide-based protein identification is the huge number of tandem MS analyses required. The number of peptides selected for identification is data-dependent, i.e., mostly highly abundant peptides are selected for sequencing. As indicated by Li *et al*.,[3] such an approach can lead to serious undersampling of the available information in the biological sample. Also, because mostly abundant peptides are selected for identification on MS2, this can cause a poor dynamic range. However, when screening and comparing whole proteomes between different biological conditions, the main interest is not necessarily the identification of all proteins, but rather the identification of differentially expressed proteins. Hence, instead of pursuing an MS2 interrogation on 'every' detected (and abundant) peptide peak, we may target selected sets of 'interesting' peptides from MS1 by using well-known statistical analysis methods. Of course, this requires a stable preprocessing algorithm to handle the amount of *N*-terminal COFRADIC MS data. Therefore, our aim is to introduce a strategy concerned with the prior processing of MS1 mass spectral data generated from a COFRADIC setting, to which the techniques for a quick, automated, and yet sensitive analysis, which we develop in this paper, can be applied. The proposed preprocessing strategy is able to translate the massive amount of data into proteomic profiles, similar to microarray data, such that classical statistical techniques such as, e.g., SAM[4] can be applied in the analysis of MS data, so that proteins that are found differentially expressed can be further investigated on MS2. The prior processing allows discrimination and scoring of a series of peaks in an MS1 scan that are possibly related to a peptide from those generated by error. This is achieved by exploiting characteristic features that should be associated with peaks corresponding to an isotopically resolved group of peptide peaks. Valid peptide peaks, which are separated over successive fractions, are assembled in a unique way, such that they represent the relative abundance of a peptide in a mixture.

The main reason for the construction of the proposed analysis strategy is twofold. First, available software packages, such as, e.g., SpecArray,[3] PEPPeR,[5] and SuperHirn,[6] which can be used to analyze, combine, and interpret data from high-dimensional LC and MS as an automated procedure, are based on data-driven methods or image processing methods to extract the peptide features from LC-MS. We present an approach that does not operate on the LC-MS image and we argue that the prior processing of LC-MS data can be split into two parts. In the first part, the MS-scans are processed separately, in order to extract peptide features on the basis of prior biological knowledge about the peptide's isotopic distribution, and not on a data-driven method. In the second part, the extracted peptide features are combined across the LC-dimension, reconstructing the LC-profile. This paper gives a detailed description of how prior processing can transform high-dimensional LC-MS data into a simple protein list. Note that the proposed strategy operates directly on the unprocessed, raw ASCII data files, which is supported by most mass analyzer instruments.

Second, the accuracy of quantification gains a lot from working on a clear mass spectral signal. Peaks which are due to noise, nonpeptidic contaminants, complex baselines, or co-eluting peptides make the detection of differently expressed proteins much more difficult. This is a general problem when analyzing complex samples that generate complex LC-MS profiles, as stated by Schmidt and Aebersold.[7] De-isotoping algorithms used in commercially available software often lead to the detection of many spurious peaks, which do not correspond to genuine peptide peaks. For instance, in the part below 1000 mass-to-charge (m/z), the software can detect a number of peaks, while this region is known for its chemical noise and matrix peaks. A possibility to circumvent this problem is to use information (a score) that would indicate how well the found peaks correspond to a series of bonafide peptide peaks. However, most available software packages do not provide this information.

Table 1. Peptides found in bovine cytochrome C tryptic digest and internal standards with information about the coefficient of variation (CV) before and after total ion count normalization. The CV is based on 384 measurements

Bovine cytochrome c (CC)				
			CV	
nr.	Sequence	Mass (*M*)	Before	After
CC1	IFVQK	633.38	–	–
CC2	YIPGTK	677.37	–	–
CC3	MIFAGIK	778.44	0.2928	0.1071
CC4	KYIPGTK	805.46	–	–
CC5	EDLIAYLK	963.52	0.2810	0.0874
CC6	TGPNLHGLFGR	1167.61	0.2181	0.0256
CC7	GEREDLIAYLKK	1433.78	0.2398	0.0399
CC8	TGQAPGFSYTDANK	1455.66	0.2709	0.0801
CC9	KTGQAPGFSYTDANK	1583.75	0.2662	0.0767
CC10	IFVQKCAQCHTVEK	1632.81	0.2401	0.0816
CC11	GITWGEETLMEYLENPK	2008.94	0.2275	0.0843
CC12	GITWGEETLMEYL ENPKK	2137.03	0.2184	0.0863
Internal standards (IS)				
			CV	
nr.	Sequence	Mass (*M*)	Before	After
IS1	RPPGF	572.30	–	–
IS2	DRVYIHPF	1045.53	0.2435	0.0537
IS3	ZLYENKPRRPYIL	1671.90	0.2072	0.0412
IS4	RPVKVYPNGAED ESAEAFPLEF	2464.19	0.2096	0.0725
IS5	FVNQHLCGSHLVEALYL VCGERGFFYTPKA	3493.67	–	–

Materials

Two technical experiments were specially designed to evaluate the proposed analysis strategy concerned with the prior processing of mass spectra. Further, the experiments were used to answer some technical questions regarding the reproducibility and variability of the used COFRADIC setting.

Bovine cytochrome C mass spectra

A peptide mixture of tryptic-digested bovine cytochrome C was purchased from LC Packings and mixed with five internal standards from Laser BioLabs used for the calibration of the mass spectrometer. According to the data sheets of the suppliers, the bovine cytochrome C tryptic digest and internal standard mixture contains 17 protein fragments. The amino acid sequence and theoretical monoisotopic masses (*M*) of these fragments are presented in Table 1.

Note that the molecules are protonated by the MALDI-procedure (MH^+). Therefore, the monoisotopic mass, as reported in Table 1, should be corrected by adding 1.00783 Da. Further, it should be noted that the third internal standard 'ZLYENKPRRPYIL' has a 'Z' in the sequence, indicating uncertainty between pyroGlu ('Q') and pyroGln ('E') at this location. The two forms have the same mass, however; so this should not affect the detection of the calibrant in a mass spectrum.

The tryptic-digested bovine cytochrome C and internal standards were mixed with the matrix molecules and automatically spotted 384 times by a robot on a stainless steel MALDI-plate with 384 spots. Note that the same mixture was spotted on the MALDI-plate, in order to evaluate the inter-spot variability. The plate was processed on a 4800 MALDI-TOF/TOF analyzer (Applied Biosystems) mass spectrometer, which resulted in 384 mass spectra. No tandem MS information was available to us. These mass spectra are primarily used for the evaluation of the peptide validation method described in this paper.

COFRADIC mass spectra

To assess the performance of the peak-assembling algorithm for COFRADIC described in this paper, we used three technical COFRADIC replicates of a complex biological mixture: that is, human blood serum from a healthy volunteer was processed three times according to the COFRADIC methodology. Note that only one biological sample is used. This was done to evaluate the variability introduced by the COFRADIC procedure.

The retention time dimension ranged from 50 to 170 min during the primary RP-HPLC fractionation and was represented by 30 so-called primary fractions. This means that one primary fraction was equivalent to a 4-min collection interval. After modification of the primary amines by TNBS, the secondary fractionation was performed under conditions identical to the first separation, except that the secondary fractions were collected and separated over a slightly extended interval. In practice, the secondary fractionation of a primary fraction commences at the elution time of that primary fraction minus 1 min and stops at the end of that primary fraction plus 1 min. Therefore the primary fraction, collected in the primary fractionation during 4 min, is further separated in the secondary fractionation during 6 min in 48 fractions. This means that one secondary fraction is equivalent to a collection interval of 7.5 s. As a result, the biological sample is split into 1440 (secondary) fractions, which, ideally, cover the complete amount of *N*-terminal peptides from the primary fractionation. The m/z dimension ranges from 500 to 4000. Because the peptides detected with MALDI-TOFMS are mostly singly charged, we use the mass measure dalton (Da) interchangeably with the term m/z.

Each COFRADIC replicate, i.e., the 1440 fractions, were automatically spotted together with matrix molecules and five internal standards from Laser BioLabs over four stainless steel MALDI plates by the same robot. The plates were processed on the same 4800 MALDI-TOF/TOF analyzer (Applied Biosystems). This resulted in 1440 mass spectra per COFRADIC replicate. One spectrum was represented as a 150000×2 data matrix, and took approximately 2.8 MB in raw ASCII format. Thus, one COFRADIC replicate used approximately 3.88 GB. For technical details about the COFRADIC methodology and the mass spectrometer settings used to generate the data, we refer the reader to the paper of Sandra *et al.*[2]

Methods

A typical feature of COFRADIC data, or more generally LC-MS data, is that a peptide signal $s(t, m/z)$ is present in two dimensions, where t represents the retention time dimension and m/z represents the mass-to-charge dimension. Because of complexity issues, we consider the peptide signal $s(t, m/z) = s_1(t) \cdot s_2(m/z)$ as a combination of two independent signals, with $s_1(t)$ denoting the elution profile and $s_2(m/z)$ denoting the mass signal, as proposed

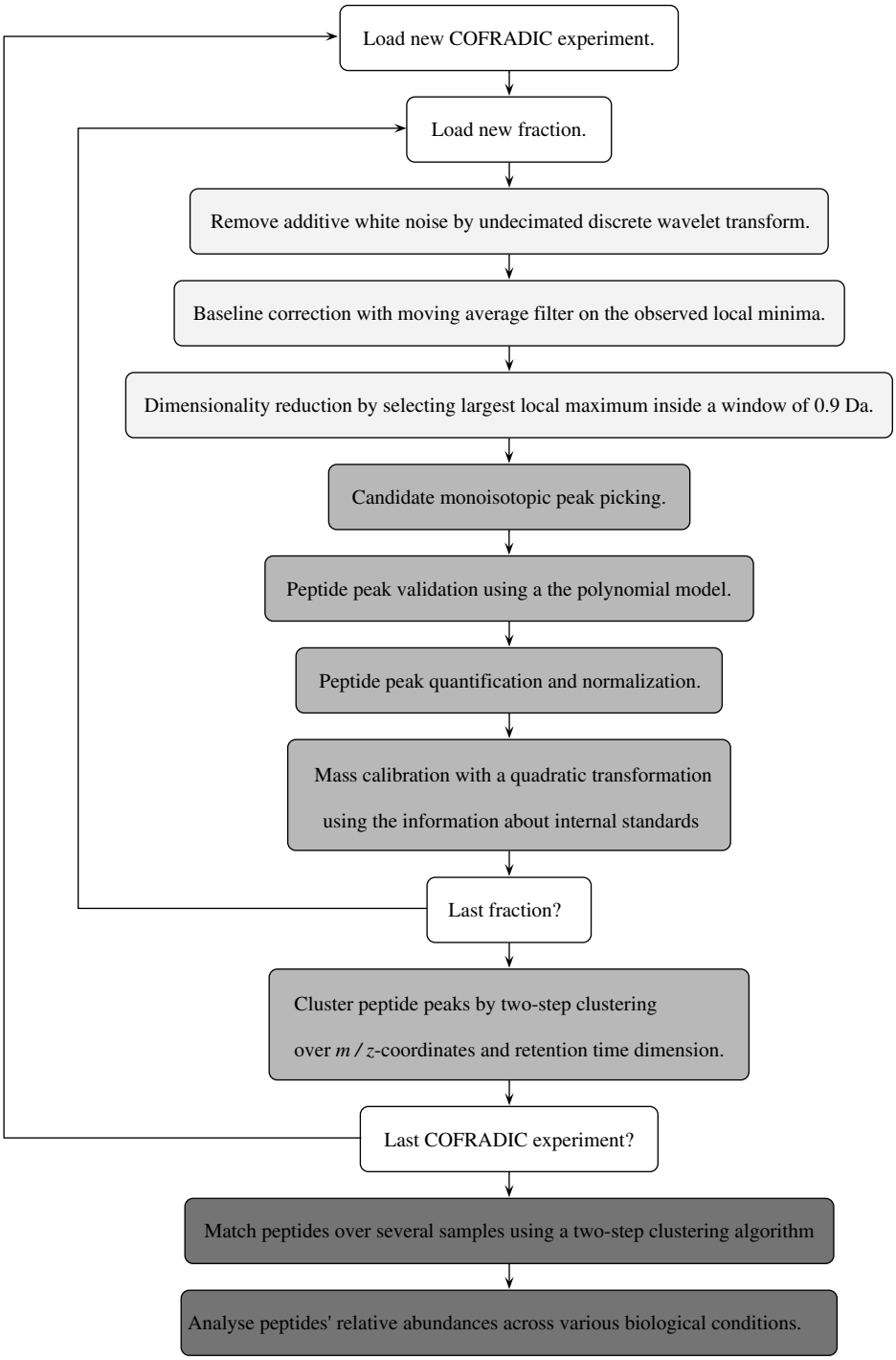

Figure 1. Outline of the proposed preprocessing algorithm.

by Hilario *et al*.[8] Under this assumption, we can first process each mass spectrum independently over the mass dimension, for which peptide peaks are selected from the mass spectrum. In the next step, the selected peptides are assembled over the retention time (fractions) in order to study the elution profile in the LC dimension. This approach is especially beneficial for a high-throughput proteomics framework, considering the massive amount of generated data, where it is difficult to directly operate on the complete data set owing to limitations of computing power. Nevertheless, there are methods that consider the complete LC-MS map, e.g., the method proposed by Schulz-Trieglaff *et al*.[9] They require, however, an intermediate step, in which peptide-containing regions of a manageable size are selected from the LC-MS map before further analysis can be applied. We argue that this step is unnecessary and time consuming, as it can be avoided by the proposed algorithm. The general outline of the proposed analysis strategy is depicted in Fig. 1. The light gray labels indicate the low-level preprocessing steps, the middle gray labels represent the mid-level processing, and the dark gray labels correspond to the high-level analysis.

An unprocessed MS1 mass spectrum typically consists of numerous intensity measurements, not always representing valuable information. Therefore, the low-level processing of mass spectral data is a critical step, because it provides a way to reduce the dimensionality of the data. For this purpose, we propose a model for the noise sources of MALDI-TOFMS data. Further, we suggest signal processing approaches to eliminate the noise, such that the potential for the detection of valid peptide peaks is increased.

The different terms contributing to the intensity $s_2(m/z)$ measured at a particular m/z location in a MALDI-TOF mass spectrum can be presented as follows:

$$s_2(m/z) = \delta(m/z) \otimes P(ac, ab, ms) + B(m/z) + \delta(m/z) \otimes M(ac, ab, ms) + c(m/z) + \varepsilon \quad (1)$$

Note that it is not the intention to directly fit this model to the obtained mass spectra, but rather to use it to illustrate the components contributing to a mass spectrum. In model (1), $\delta(m/z)$ is a Dirac impulse that represents the monoisotopic mass of a biochemical molecule. The Dirac impulse is convoluted (symbol: $\otimes$) with a predefined shape $P(ac, ab, ms)$, where the shape depends on the atomic composition (ac),[10] abundance (ab) of the peptide, and mass spectrometer specifications (ms). Only those peaks that really represent the presence of some biological mass are considered as the signals of interest. In what follows, other terms contributing to the mass spectrum are considered as noise. Peaks $M(ac, ab, ms)$, introduced by chemical compounds such as matrix or solvent, have a shape similar to biochemical compounds, but do not contain information about the *N*-terminal peptides in the fraction. Further, the baseline $B(m/z)$ and a low- intense oscillation $c(m/z)$, often referred to as chemical noise, are visible in the spectrum. In the lower mass regions, mainly below 1500 m/z, the chemical noise $c(m/z)$ appears as a structured periodic signal. In the higher mass regions, the chemical noise has an increased frequency and decreased intensity, while at the end of the mass range (in this case, near 2500 Da) the chemical noise appears as low-intense correlated noise. Owing to the low frequency of this structured noise at the lower mass region, it is difficult to filter it out from the spectrum without affecting the peptide peaks $P(ac, ab, ms)$. The last term of the MS signal in Eqn (1) is the normal random noise ε, and has a minor effect on the signal. Nevertheless, during the low-level processing (light gray labels in Fig. 1), the additive noise is removed from the mass spectrum by using the undecimated discrete wavelet transform proposed by Baggerly *et al.*[11] The baseline is removed to improve the reproducibility. To decrease the complexity of a mass spectrum, the data are reduced without loss of relevant information, i.e., the information about the isotopic distribution is retained. These findings are based on the observations of a controlled experiment, in which the content of a sample was known, and will be explained later in this paper. There exist several strategies to filter out information about the isotopic distribution from the mass spectra. These are usually based on the peak shape or on the accumulated intensity measures composing the peak. However, we argue that information about the peak height (stick representation) should be adequate.

During the mid-level processing (middle gray labels in Fig. 1), valid peptide peaks are selected from the mass spectrum. After mass calibration, the peptide peaks, possibly spread over several subsequent fractions, are aggregated and normalized such that they represent the relative abundance of the peptide in a sample.

The high-level analysis (dark gray labels in Fig. 1) focuses on the statistical comparison of peptide abundances from different biological conditions. In the next sections each of the prior processing steps are discussed in more detail.

Low-level processing

Baseline correction

Let us define the ion count / as the sum of the intensity measurements for the data points in a mass spectrum corresponding to the isotopic distribution of a peptide. The ion count is used as a measure for the relative abundance of a peptide in a fraction. Therefore, it is mandatory to subtract the baseline from the spectrum before calculating the ion count, so that the baseline variability does not influence the measure of abundance. Another reason favoring baseline correction is that, in our approach, for selecting peptide-related peaks in a mass spectrum, the height of the peaks is used. The baseline would influence the height of the observed peaks and, consequently, would complicate the assessment of valid peptide peaks. Baseline is found by applying a moving median filter (greedy baseline correction) or a moving minimum filter (stingy baseline correction) on the observed local minima in a mass spectrum. For our purpose, the smoothing with a moving median filter and a window width of 10 Da was found to be optimal. The dashed line in Fig. 2a represents the baseline. After interpolating and subtracting the baseline from the MS signal, all negative values are truncated at zero. The result of the baseline correction can be seen in Fig. 2b. Baseline correction based on a moving median filter is fast and yields results comparable to, e.g., locally weighted scatterplot smoother (LOWESS), but is less computationally involved.

Data reduction

The data from one mass spectrum contain approximately 150 000 data points. However, we are interested in finding only the group of peaks corresponding to the isotopic distribution of a peptide. Therefore, we apply a data reduction technique that reduces the number of intensity measurements to approximately 3500 without loss of information about the isotopic distribution. The data reduction is achieved by, first, selecting all the local maxima from the original spectrum. All other measurements are set to zero. Second, the largest local maximum inside a window of size less than 1.0015 Da is selected. In our case, a window of 0.95 Da is used, because isotopic peaks in a MALDI-TOF spectrum are separated by approximately 1 Da. The outcome of the data reduction step can be observed in Fig. 2a, where the selected peaks are indicated by a cross.

A possible disadvantage of this method is that it captures information only about the height of peaks in the mass spectrum. Information about the shape of the peaks is removed during this process, as it is not relevant for the selection of candidate monoisotopic peaks. The lack of information about the peak shape can be problematic if we want to detect overlapping peaks. However, owing to the peptide sorting properties of the COFRADIC methodology, it is less likely that peptides with similar nominal mass are introduced at the same time by the LC-column.[12] In our case study data, described in the Section on COFRADIC Mass Spectra, only a few overlapping peptide peaks were observed. Exact quantification of the amount of overlapping peptides is difficult, as it requires visual inspection of the data.

Journal of MASS SPECTROMETRY

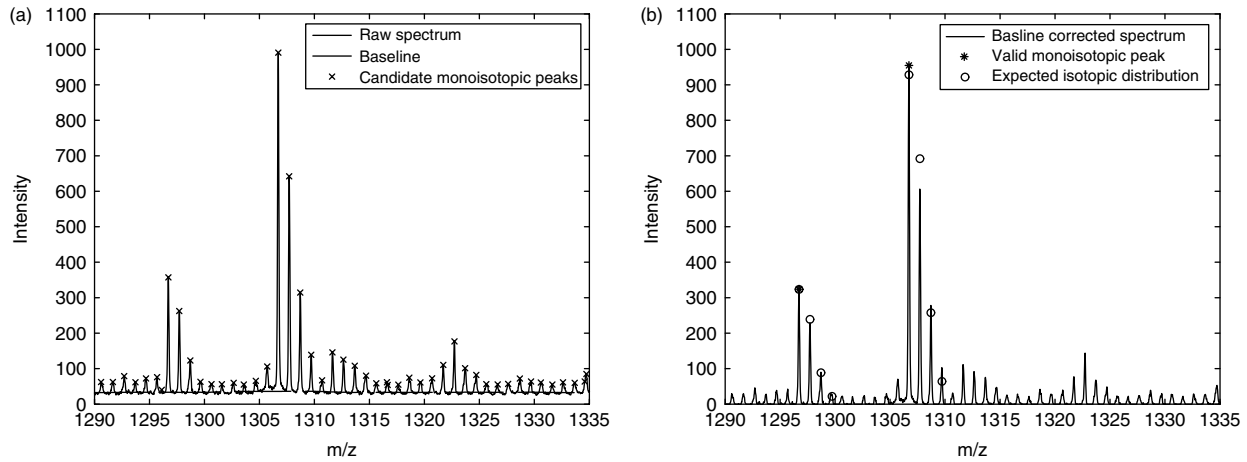

Figure 2. Mass spectrum from bovine cytochrome C mixture. (a) Unprocessed MALDI-TOF spectrum (solid line) with baseline (dotted line) and reduced mass spectrum (cross). (b) Baseline-corrected MALDI-TOF spectrum (solid line) with two valid monoisotopic peptide peaks (star). The expected isotopic distributions are indicated by circles.

Undoubtedly, variable data quality can result in a loss of detection power. Peaks are selected from the data, but this does not prevent us from selecting peaks generated by noise. In the next step, noise peaks are removed from the list of local maxima peaks by using information about peptide's characteristics.

Mid-level processing

Candidate monoisotopic peak picking

From the list of local maxima obtained by the method described in the Section on Low-level Processing, we select candidate monoisotopic peaks. This is done by selecting the peaks fulfilling the following criteria:

1. There are at least four 1.0015 Da separated consecutive peaks present in the spectrum. This condition is imposed because the peaks corresponding to an isotopic distribution of a peptide should appear as peaks separated by approximately 1 Da. Depending on machine precision, a margin of 100 ppm is allowed on the position of the subsequent peaks. The set of four peaks is called an *isotopic cluster*.
2. The signal-to-noise ratio of the isotopic cluster is defined as the intensity of the largest peak in the cluster divided by the noise. This value should be equal to or larger than 1.5. The noise is estimated locally by the maximum value of intensity measurements within a small region of approximately 1 Da surrounding the corresponding isotopic cluster.

If both conditions are satisfied, then the peak corresponding to a local maximum is considered as a candidate for a monoisotopic peptide peak. Peaks that do not pass either of the conditions are considered noise peaks and are removed from the list of local maxima.

Peptide peak validation

The relative heights of peaks corresponding to a peptide depend only on the distribution of the isotopic variants of the peptide. The prior knowledge about the isotopic distribution allows us to discriminate in a mass spectrum between a series of valid peptide peaks and peaks originating from noise. The expected probabilities of the isotopic distribution are estimated by using the method proposed by Valkenborg *et al*.[13] This method predicts the ratios between the expected probabilities of the successive isotopic variants, based on a set of theoretical peptides constructed from the average amino acid *averagine* proposed by Senko *et al*.[14] To validate the candidate for a monoisotopic peptide peak, and to distinguish it from a series of noise peaks, the Pearson's chi-squared test statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{3} \frac{(R_E(i) - R_O(i))^2}{R_E(i)} \qquad (2)$$

where $R_E(i)$ is the ith expected isotopic ratio and $R_O(i)$ is the ith isotopic ratio obtained from the observed peaks in the mass spectrum. For example, the second observed isotopic ratio $R_O(2)$ is calculated as the height of the third observed peak divided by the height of the second observed peak, etc. If the obtained value of this goodness-of-fit measure χ^2 is smaller than 0.15, the candidate peak is considered as a valid monoisotopic peptide peak. The threshold of 0.15 is obtained via an empirical simulation study and is kept constant for the analysis of our case study data. It should be noted that this threshold is not generally applicable to other LC-MS settings because of the diversity of mass spectrometers and the possible experimental settings. For instance, an increased amount of noise can lead to a different threshold value. For this reason, one should conduct an experiment to determine the optimal threshold for monoisotopic peptide peak detection before applying this method to other LC-MS settings.

To validate a candidate for a monoisotopic peptide peak in an observed mass spectrum, only the first four probabilities of the isotopic distribution are considered for this goodness-of-fit measure (see circles in Fig. 2b). The peaks corresponding to the first four isotopic variants of a peptide are normally easily found in a mass spectrum. However, one can argue that for low-mass peptides, the fourth isotopic peak may not be observed when the abundance is low. In these situations one can consider using three peaks, while setting the height of the fourth (undetected) peak to zero. Note that this is a conservative approach, as the fourth isotopic variant will contribute to the Pearson's χ^2 error in Eqn (2). This can be interpreted as a penalization for not finding the fourth peak. An alternative approach is to use three peaks for the validation of a peptide, but then we need to define another

threshold. However, because the validation of a peptide is based on a smaller amount of information, we need to accommodate for the higher number of noise peaks that will be selected as a valid peptide peak and, in turn, this will result in a stricter threshold.

Peptide quantification and normalization

For each valid monoisotopic peptide peak, the ion count I of the data points composing the isotopic cluster is calculated as a measure of the relative peptide abundance in a fraction. Because the data quality is not uniform between the consecutive mass spectra, we need to correct for sample degradation, laser intensity variations, and fluctuations in ionization efficiency, spotting, and crystallization. Therefore, the ion count is normalized by the total ion count (TIC) of a spectrum. The TIC is calculated as the sum of all intensity measurements composing the spectrum after baseline correction. One can consider alternate normalization schemes, such as, e.g., normalization based on a partial ion count. In this variant, the normalization term ignores the region below 1000 Da to avoid influence of matrix and solvent-related intensity measurement. However, this strategy was found inferior compared to TIC normalization.

TIC normalization works well for individual mass spectra, but may not be as meaningful in LC-MS experiments, where not the same amount of peptide elutes in the different fractions. Therefore, an extra normalization step is required to perform an additional inter-spectrum normalization beside the intra-spectrum normalization. The issues about the normalization of high-dimensional LC-MS data are addressed in the Section on Assembling Peptides.

Mass calibration

Because of the spectrometer's high mass resolution in the range between 500 to 4000 Da, the peaks corresponding to the isotopic variants of a peptide are separately discernable in the mass spectrum. By contrast, in low-resolution mass spectra, a peptide will appear as a single peak shape. Because of the high mass resolution, the monoisotopic mass can be used to define the location of the peptide in the mass dimension, instead of, e.g., the centroid mass. However, the time-of-flight measurement with a MALDI-TOF mass spectrometer is often affected by an error. According to Grass *et al.*,[15] this error can reach up to 1 Da. Normally, the obtained mass spectra are calibrated by the mass spectrometer, but when the machine calibration fails (i.e., ppm $>$ 100), we perform an additional mass calibration via the quadratic transformation in Eqn (3) using the mass information of internal standards (calibrants). Obviously, this method works only when the internal standards are well detectable. The internal standards are selected on the basis of two criteria:

1. the approximate theoretical location ($\pm$5 Da) of the internal standard;
2. the theoretical mass difference between the internal standards.

We argue that, if the internal standards are present within a 100 ppm error interval, not much improvement can be obtained by performing a mass calibration. If the location of the internal standards exceeds the 100 ppm error interval, we perform a mass correction and indicate the fraction as 'ill calibrated'.

The calibration makes use of the quadratic relation between the time-of-flight and m/z, as described by[16]:

$$\text{TOF} = \beta_1 m/z + \beta_2 \sqrt{m/z} + \beta_3 \tag{3}$$

Instead of using the time-of-flight measured in nanoseconds, we use the dimensionless term *tick* (or channel) to indicate the time interval wherein ionized molecules collide with the detector of the mass spectrometer. The tick value indicates the position of the corresponding m/z value in the data vector. If the location (TOF in (3)) of the five internal standards in the data is known, then under model (3), parameters β_1, β_2, and β_3 can be calculated by the least squares method using the five theoretical m/z-values (+1.00783 Da for the MALDI-proton) of the internal standards displayed in Table 1.

Next, we calibrate the mass of a peptide, on the basis of the obtained values of β_1, β_2, and β_3, by computing

$$m/z_{\text{cal.}} = -\frac{\beta_2}{2\beta_1} \pm \sqrt{\left(\frac{\beta_2}{2\beta_1}\right)^2 - \frac{(\beta_3 - \text{TOF})}{\beta_1}} \tag{4}$$

with TOF denoting the location of the validated monoisotopic peak in the datavector (tick-value).

We can already mention that in the considered case studies (see Sections on Bovine Cytochrome C Mass Spectra and COFRADIC Mass Spectra) the mass calibration performed by the mass spectrometer was satisfactory and did not require any additional adjustments. Therefore, the proposed mass calibration method was validated on a set of ill-calibrated mass spectra, for which the method was proven to work correctly. This can be observed from the heat maps in Fig. 3, where the dots indicate the local maxima found in the spectra. The x-axis indicates the mass of the local maxima and the y-axis indicates the fraction (or mass spectrum), in which the local maxima were found. The intensity of the local maxima is indicated by a grayscale; white is low intense and black is high intense. In panel (a) of Fig. 3, local maxima are shown for the region near the second internal standard (IS2 in Table 1). The dark gray vertical stripe at mass 1046.5 Da is the peptide peak corresponding to the monoisotopic variant of internal standard IS2. It can be observed that for fractions with numbers $\geq$ 193, the internal machine calibration fails to work. This can be seen by the scattering of, e.g., the monoisotopic peptide peak across the mass region. After applying the calibration method described in this section, based on internal standard IS2, IS3, IS4, and IS5 (see Table 1) the local maxima are aligned, as can be seen from panel (b) of Fig. 3. This means that the vertical stripes are now reconstructed. Note that the external mass calibration did not work for the mass spectra near fraction 200. The horizontal middle gray band at this location indicates a decrease in intensity, which may be caused by inappropriate ionization due to laser fluctuations or poor crystallization. Therefore, we could not trace back the internal standards required to recalibrate the mass spectrum for the fractions near 200. Probably, this is also the reason why the internal machine calibration lost track of the internal standards and could not recover mass calibration after the problem was solved.

Assembling peptides

After processing of all the mass spectra, the validated monoisotopic peptide peaks, separated over different fractions, but

Journal of MASS SPECTROMETRY

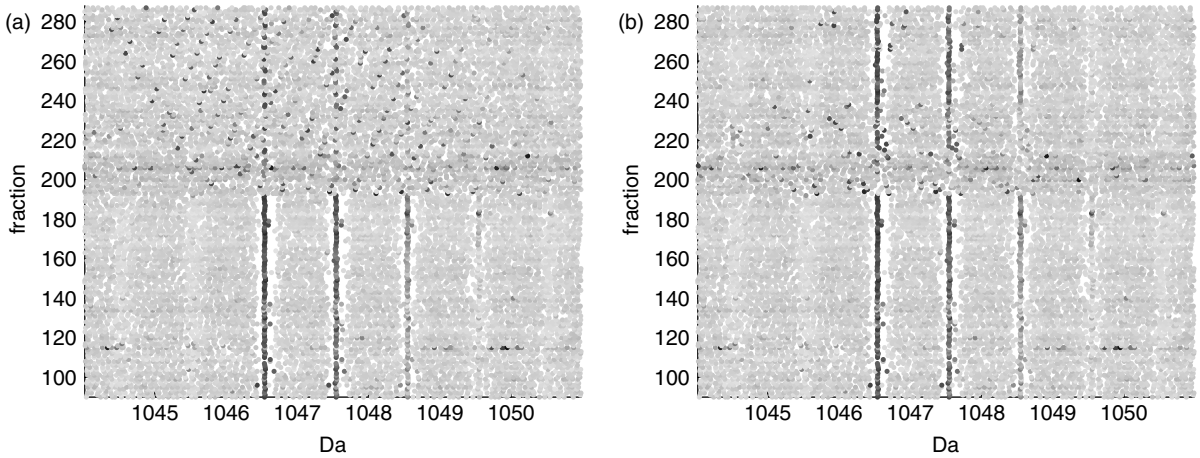

Figure 3. Partial heat map of the local maxima present in the COFRADIC mass spectra from a biological sample. The focus is on the second internal standard for which the stripes are the successive peaks corresponding to the isotopic distribution. Panel (a) presents the heat map before mass calibration, while panel (b) presents the mass-calibrated heat map.

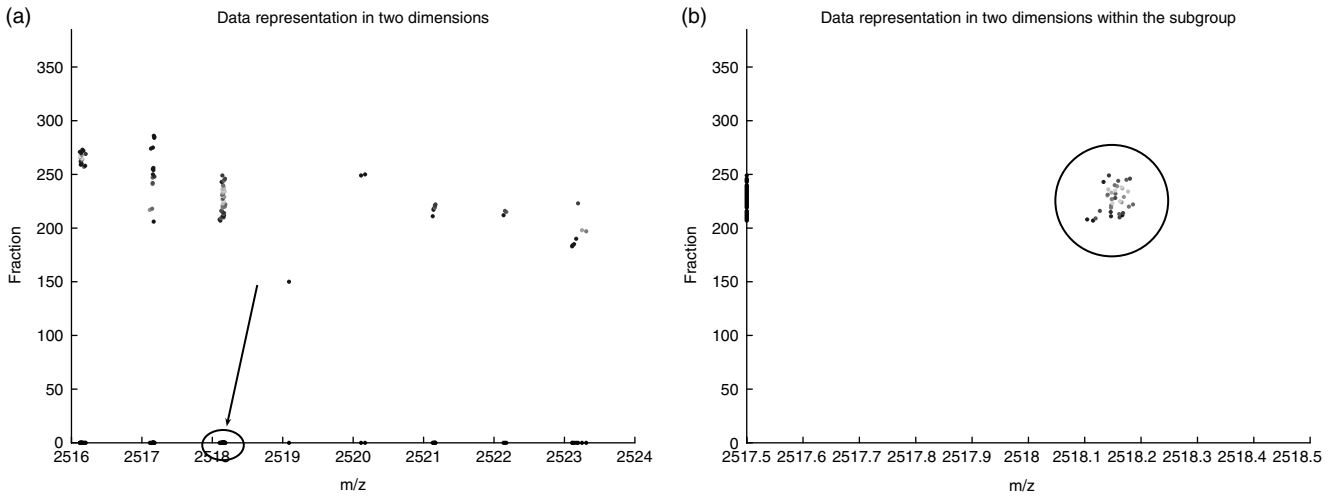

Figure 4. Heat map of the valid monoisotopic peptide peak for the mass range near 2518 m/z. (a) Valid peptide peaks are projected on the m/z axis, where a clear separation between the clusters can be observed. (b) The cluster at 2518 m/z is projected at the retention time axis. No further separation is possible, so the cluster of valid monoisotopic peptide peaks is assembled as a peptide.

originating from the same peptide, must be linked in order to quantify the abundance of the peptide in the mixture. For this purpose, a two-step clustering algorithm is proposed to collect the information about a peptide scattered over different fractions. To this aim, we assume that peaks that appear in subsequent fractions and that have approximately the same mass are related to the same peptide eluted over the fractions. The clustering algorithm consists of two consecutive steps:

1. The validated monoisotopic peaks, situated in a three-dimensional space (intensity measure, mass, elution time), are projected on the mass dimension. As can be seen in panel (a) of Fig. 4, the data are well separated over the m/z-axis. Hence, in the mass dimension, clustering is performed using a threshold of 0.3 Da for the size of the gap that separates the data in the case study.
2. For a cluster obtained from the previous step, the elution time dimension (fraction) is now considered (see panel (b) in Fig. 4). Within each subgroup, the data are projected on the time dimension. If peptide peaks are found in subsequent fractions, then these peaks are assumed to be generated by the same peptide and hence clustered in a single group. In this step, a gap of 37.5 s is allowed to occur, equivalent to not observing the peptide in 5 consecutive fractions. The missing gap of 37.5 s is generally accepted by chromatographic experimentalists.

For each group (cluster), a vector with descriptive statistics is registered containing information about the mean m/z, retention time, relative abundance, missing observations, etc. Clusters that manifest themselves over more than 25 min (or 200 fractions) are discarded from the data because they are assumed to come from chemical contaminations, matrix, or internal standards. The two-step clustering algorithm is able to perform the clustering on the complete set of validated peptide peaks, and does not require working on local parts of the LC-MS data.

In the next paragraphs, we provide further detail regarding the calculation and normalization of the peptide abundances. Usually, the intensity of a peak representing a peptide in a mass spectrum is mainly influenced by laser intensity, matrix crystallization, ion suppression, ionization efficiency of a peptide, and the absolute

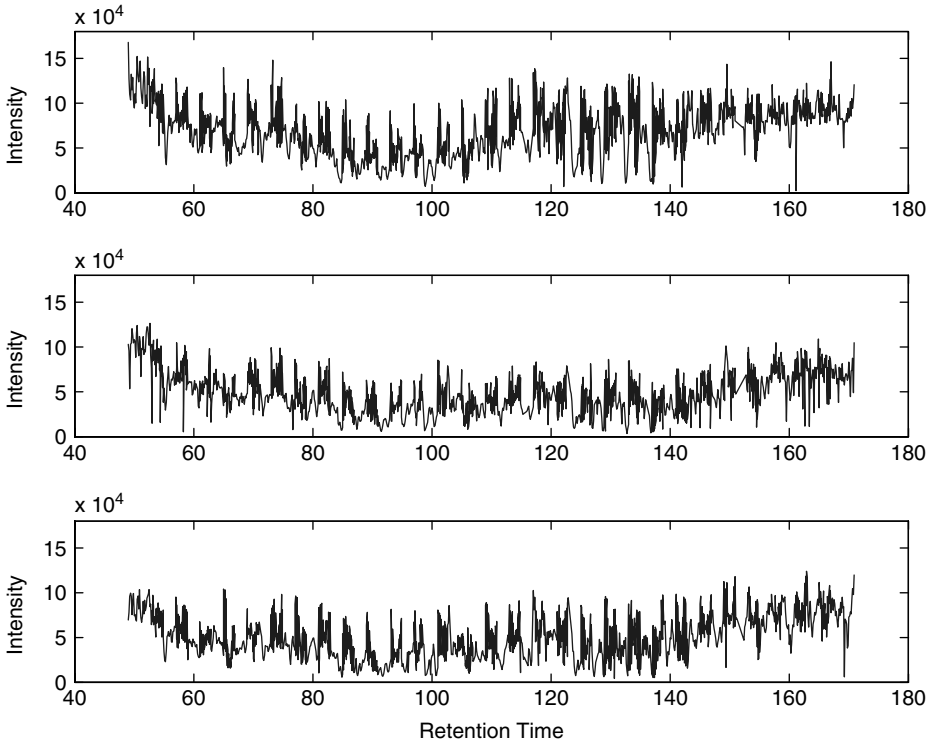

Figure 5. Normalized intensity of the third internal standard (IS3) over the retention time dimension for the three COFRADIC replicates.

abundance of the peptide in a fraction. Because optimization of peptide separation ensures low peptide density for each MALDI spot, we must adjust the laser intensity in order to ensure that the peptides give rise to a clear MS1 signal. In the COFRADIC case example, the laser intensity is tuned once for each LC-MS run to avoid signal saturation. When saturation occurs, the information about the isotopic distribution is lost and our de-isotoping algorithm fails to work. This, of course, affects only the quantification of peptides that correspond to very abundant proteins. The compromise made while tuning gives a priority to accurate quantification of low-abundance proteins. This means that the mass spectral signal will saturate occasionally for a few abundant proteins. Another aspect to reduce nuisance in the data is to ensure reproducible crystallization. To this aim, the spotting of the biological sample onto MALDI plates is optimized and automated. Further, the ionization efficiency depends on the atomic composition of a peptide and the peptide's ability to absorb ions. This factor is not of crucial importance, because we do not want to absolutely quantify the peptide abundance, but rather want to relatively compare the abundance of particular peptides across biological conditions. As already mentioned, beside the TIC-normalization, the normalization of the peptide abundances in high-dimensional LC-MS data needs extra attention. Note that we measure the relative abundance of the peptide in a fraction.[17] Therefore, in order to obtain the relative abundance P_j of peptide j in the whole sample, we need to weight the peptide abundance measurements, or equivalently, ion count I_{ij} in fraction i by the proportional content w_i of fraction i in the total biological sample:

$$P_j = \sum_{i=rt_1}^{rt_2} \frac{w_i}{\sum_k w_k} I_{ij} \qquad (5)$$

where rt_1 and rt_2 are, respectively, the first and the last fraction, in which the peptide eluted, while index k goes through the total number of fractions in the COFRADIC experiment.

Theoretically, an estimate of the content w_i of a fraction could be retrieved from the LC column by means of an optical density (OD). In practice, we use $w_i = 1$, which is equivalent to summing the peptide abundances in the fractions. One could also consider using the LC profile of internal standards to retrieve an estimate of the OD. This, of course, is possible only when the absolute quantity of internal standard in each fraction is kept constant. For example, Fig. 5 presents the measured abundance of the third internal standard (IS3) over several fractions after TIC normalization. The amount of internal standard is the same across the fractions, because it is spiked into the sample after the COFRADIC procedure. Therefore, we expect the same intensity measure in each fraction. However, because of ion suppression, the intensity measure of the internal standard does not appear to be constant. This indicates that the sample density differs across the fractions. However, how this fluctuating intensity measure of the internal standards can be used to estimate the OD of a fraction directly from the data is a topic for further research. It should be noted that the periodic fluctuations in Fig.5 are the result of the secondary chromatographic separation of the 30 primary fractions.

High-level processing

The main goal of quantitative proteomics is to find peptides that are differentially expressed across different biological conditions. Therefore, peptides with inaccuracies in mass and retention time alignment from different LC-MS runs should be matched over the different experimental conditions. To take into account these inaccuracies, a clustering algorithm, similar to the one described previously, is used. Therefore, we assume that clusters

Journal of MASS SPECTROMETRY

of monoisotopic peptide peaks with the same mass coordinates and the same retention time represent the same peptide present in different samples. We allow for a chromatographic time shift of 1 min, which corresponds to a misalignment of eight fractions between the chromatographic separations.

Next, the peptide's relative abundances are analyzed across various biological conditions. To avoid redundant comparisons, the chemical modifications and isoforms of a peptide should be removed. For this purpose, we could search for possible modifications on the basis of the mass differences caused by the modifications reported at the unimod -database.[18] However, searching for modifications based on mass differences is not very accurate. Therefore, we argue that, during a statistical analysis, the effect of a redundant comparison of modifications and isoforms is negligible. Peptides that are found differentially expressed across different biological conditions are marked as candidate biomarkers. However, caution should be taken during this step. When statistically comparing multiple peptide expressions, one should correct for multiplicity. A similar issue appears, e.g., in the analysis of multiple genes in microarrays. In this context, different methods addressing the multiple testing issue have been developed (e.g., SAM[4]), which can also be applied to the analysis of MS data after proper preprocessing.

Next, via tandem MS, the amino acid sequence of the *N*-terminal peptide that is found to be differentially expressed for various groups of samples can be identified and should be linked to the differently expressed parent protein. To facilitate the peptide characterization, we can use information from the observed elution profile of the specific peptide and choose the fraction where most of the material, i.e., peak with the highest intensity measure, is eluted from the column.

Results

The algorithm described in the Section on Methods has been implemented in MATLAB 7.1 release 14SP3 and applied to two case examples.

In the Section on Bovine Cytochrome C Sample, we discuss the performance of the de-isotoping algorithm on the 384 bovine cytochrome C mass spectra. In the Section on Human Blood Sample, we assess the clustering and assembly algorithm using a human blood sample of a healthy volunteer, processed three times (i.e., with technical replicates) by the COFRADIC methodology.

Bovine cytochrome C sample

After processing the 384 spectra with the algorithm described in the Section on Methods, 12 out of 17 peptides present in the bovine cytochrome C tryptic digest and internal standard mixture (see Table 1) were consistently classified as valid monoisotopic peptide peaks. The algorithm was unable to find the peptides at 573.3, 634.4, 678.4, and 806.5 Da, and had difficulties with consistently finding the fifth internal standard, i.e., the peptide at 3494.7 Da. A possible reason for missing the low-mass peptides are the interfering peaks produced by the matrix (range 500–650 Da) and the solvent (range 750–850 Da). For the peptide in the higher mass range, poor ionization efficiency and decreasing resolution can be a reasonable explanation. However, note that the fifth internal standard is consistently found in the mass spectra from the COFRADIC replicates, as will be discussed later.

From Table 1, which lists the peptides in the purchased peptide mixture, we should expect to find 17 peptides in the mass spectra, but we find more. The monoisotopic masses of the 52 peptides found in more than 20% of the spectra are listed in Table 2. However, we particularly focus on the validated peptide peaks that were consistently found in 90% of the 384 spectra. Besides the 12 valid monoisotopic peptide peaks, which could be linked to peptides in the purchased sample, the algorithm nominated 23 extra valid monoisotopic peptide peaks found in 90% of the spectra. These extra 23 findings might be due to possible modifications. For example, chemical modification such as oxidation (+15.99 Da), sodium adduct (+21.98 Da), tryptophan oxidation to formylkynurenin (+31.99 Da), and tri-oxidation (+47.97 Da) could be a reasonable explanation for the peptide with sequence 'GITWGEETLMEYLENPKK'. These possible modifications for the latter peptide are indicated in Table 2 by bold numbers and are based on the mass differences as reported by the unimod database.[18]

Table 2. Peptides found in more than 20% of the 384 bovine cytochrome C tryptic digest mass spectra

Mass	%	Mass	%	Mass	%
568.1	100	1322.7	90.1	1633.6	100
650.1	23.7	1340.7	42.4	1649.6	100
779.4	100	1367.7	99.59	1656.6	82.6
817.3	89.8	1377.8	89.3	1672.9	100
861.1	24.0	1419.8	32.3	1820.7	91.7
964.5	100	1434.8	100	2010.0	100
986.5	22.9	1438.8	72.9	2026.0	100
1046.5	100	1456.7	100	2032.0	103.4
1099.5	41.4	1478.7	93.0	2042.0	98.4
1100.5	21.4	1488.7	20.1	2058.0	97.1
1124.6	99.7	1504.5	62.8	**2138**.1	100
1152.6	98.4	1505.5	26.0	**2154**.1	100
1168.6	100	1562.9	100	**2160**.1	97.4
1184.6	100	1567.8	77.3	**2170**.1	98.4
1196.6	96.1	1584.8	100	**2186**.1	95.3
1212.6	99.7	1588.8	41.9	2465.2	101.3
1296.7	100	1589.7	46.9		
1306.7	100	1606.8	99.7		

Another possible explanation is that some of the found peptides are artifacts of the proteolytic background.[19] This means that the tryptic digest is not 100% correct, such that trypsine fails to cleave arginine or lysine. This can result in more peptides in the mixture than reported by the supplier. To support this statement, we used the MS-digest tool from Protein Prospector to perform an *in silico* tryptic digest of bovine cytochrome C allowing for 10 miscleavages. In a mass range between 500 and 4000, this resulted in 99 peptides containing up to eight miscleavages. On the basis of the monoisotopic mass location of these peptides, 5 out of 99 could be linked to peptides that are observed in the mass spectra (indicated by underscore in Table 2).

In order to rigorously investigate the possibility of post-translational modification or proteolytic background, we should perform a tandem MS step on each of the additional peptides found by the proposed algorithm. However, we could validate our finding using an alternate method. We calculated the exact isotopic distribution for the peptides on the basis of the atomic composition reported by Protein Prospector or the unimod-database, and the method proposed by Rockwood.[20] Next, we calculated the Pearson χ^2 errors, as in Eqn (2), with the observed peak ratios from

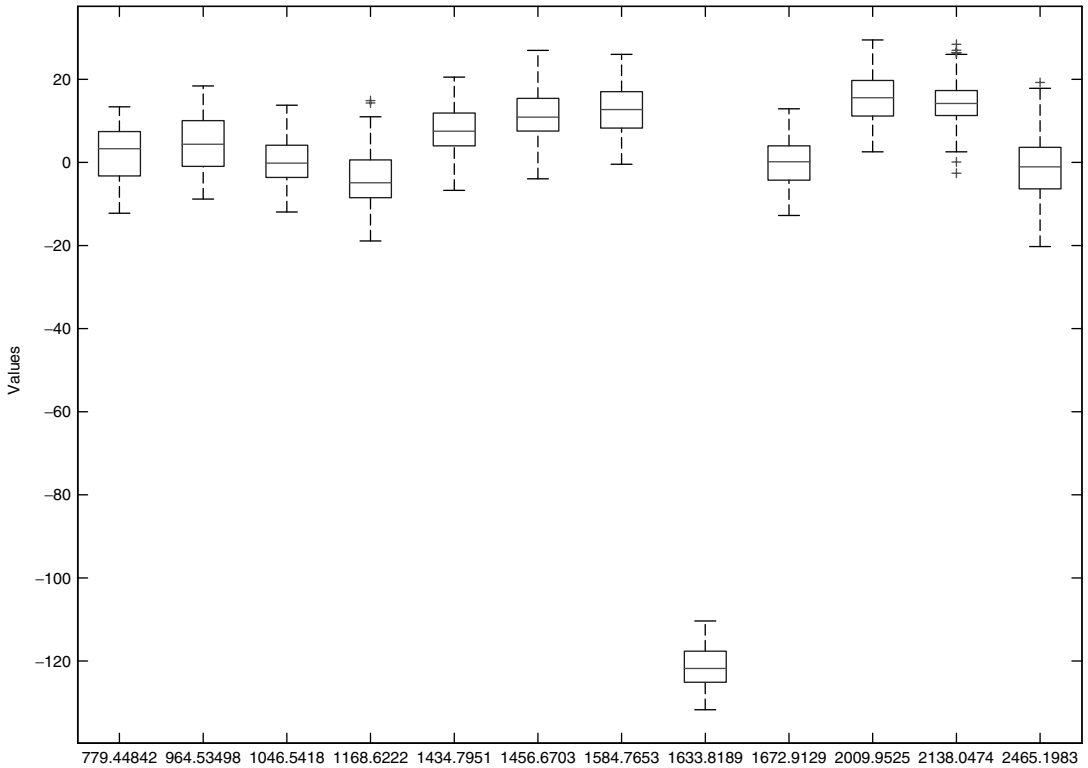

Figure 6. Box plots of the ppm for the 12 found bovine cytochrome C tryptic peptides consistently found in the 384 MALDI-TOF mass spectra.

the 384 spectra and the exact isotopic ratios. These errors were compared with the errors obtained by using the average isotopic ratio, as explained in the Section on Peptide Peak Validation . As expected, the Pearson χ^2 errors using the exact isotopic ratios were smaller than the errors obtained by using the average isotopic ratios. This confirms that the peptides observed in the spectrum have the same atomic composition as the peptides reported by Protein Prospector or the unimod database. For example, Fig. S1 in Supporting Information displays a histogram of the errors for the peptide with a mass of 1296.7 Da (see Table 2). It can be observed that the errors obtained via the isotopic distribution based on the atom composition (panel (b)) are much closer to zero than the errors obtained via the predicted average isotopic ratios (panel (a)). This also supports our argument in the Section on Methods that the information about the isotopic distribution is retained when reducing the data, i.e., the information about peak heights is enough for a satisfactory detection of peptide-related peaks in a mass spectrum.

To assess the appropriateness of the TIC normalization discussed in the Section on Peptide quantification and Normalization, we calculated the coefficient of variation (CV), CV $= \sigma/\mu$, with σ being the sample standard deviation and μ the sample mean of the peptide abundances measured in the 384 spectra for the 12 detected monoisotopic peptide peaks, reported by the supplier of the bovine cytochrome C mixture. Table 1 presents the values of the CV computed before (left) and after (right) the TIC normalization. It can be observed that the global TIC normalization decreases the variability of peptide intensities, which is a desired effect. Note that the CV is approximately constant across the mass range. This indicates that the mass of the peptide does not influence the relative magnitude of the variability of the abundance measure.

We also evaluated other normalization schemes, such as normalization with the TIC of the region above 1000 Da to avoid the area with matrix peaks, and the normalization with the ion count of the validated peptides. Both methods appeared to perform worse in reducing the abundance variability.

To study the internal machine mass calibration, we calculated the ppm for the 12 found peptides (Table 1) for the 384 replicates, presented as box plots in Fig. 6. No additional mass calibration was needed, because the internal standard peaks were within an interval of 100 ppm around their expected masses. Performing a mass calibration on these data would not yield an increased mass precision. It can be observed from Fig. 6 that the peptide coming from the bovine cytochrome C digest at 1633.8189 m/z is biased with approximately 120 ppm, which corresponds to a mass shift of 0.2 Da. The reason for the mass shift is unclear.

To conclude, the constructed de-isotoping algorithm was capable of distinguishing between valid peptides and noise peaks in the case study.

Human blood sample

In this section, we describe the application of the proposed algorithm to the human blood sample case study. In a complex biological sample, there is an increased probability that two peptides with the same mass would elute from the RP-HPLC column at the same time. However, because of the peptide separation properties of the COFRADIC methodology, the sample complexity is reduced. Hence, mass spectra become less crowded and the probability of overlapping peptides is small. The peptide separation also increases the ability to detect low-abundant serum proteins, because of an increased sensitivity of the mass spectrometer.

Journal of MASS SPECTROMETRY

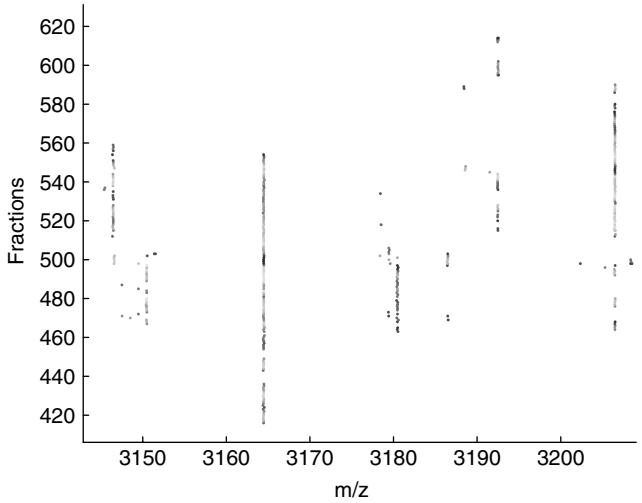

Figure 7. Closeup of an arbitrary mass region in the heat map for the first COFRADIC replicate.

Table 3. Summary statistics for the three technical COFRADIC replicates

	Replicate 1	Replicate 2	Replicate 3
Valid monoisotopic peaks	66209	62040	58516
Total ion count ($\times 10^9$)	8.441	6.103	5.903

An arbitrary 'closeup' of the heat map is depicted in Fig. 7. In this figure, the *y*-axis (ordinate) indicates the time, at which the material, visualized by the mass spectrum, eluted from the column. Again, the gray value is an indication of the abundance, with black for high abundance measurements. From this plot it is clearly seen that a peptide can elute over multiple fractions. Close observation of the color profile for the peptide at mass 3165 Da indicates intensity fluctuation in the LC dimension. The subtle distinction as to whether these fluctuations are intrinsically present in the data, or are just an artifact of the ion suppression, or correspond to a set of overlapping peptides in LC dimension, would require more advanced methods and is a topic for further research.

We were able to detect approximately 60 000 valid monoisotopic peptide peaks in each COFRADIC replicate. The precise numbers are given in Table 3. It can be observed that the number of valid monoisotopic peptide peaks per run decreases with the observed TIC. This might be caused by sample degeneration or by laser fluctuations between the processing of the COFRADIC experiments.

Note that the, approximately, 60 000 found monoisotopic peptide peaks refer to the number of peptide peaks that are found in one COFRADIC experiment. However, peptides are eluted over multiple fractions and need to be assembled into one single group representing the abundance of a peptide in the sample. Further, in order to compare peptides across the different conditions, they should be first matched between multiple samples. Recall that for this purpose the clustering algorithm, described in the Section on Assembling Peptides, was developed. First, the clustering algorithm combines the valid monoisotopic peptide peaks from the same peptide. Second, it performs a clustering to match identical peptides on the basis of mass location and retention time. Because of the complexity of the data, it is likely

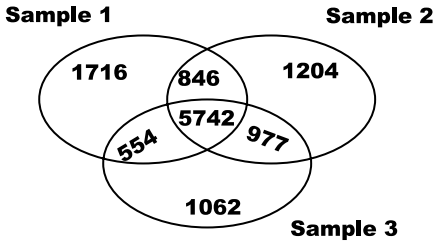

Figure 8. Venn diagram with the number of peptides found in the three technical COFRADIC replicates; 5742 peptides were found in all three COFRADIC runs.

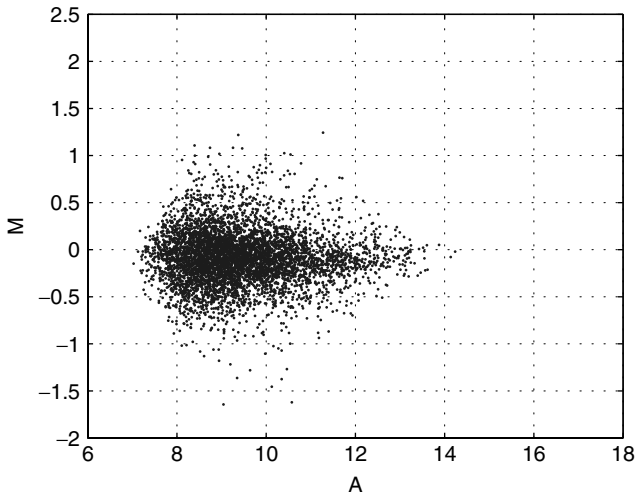

Figure 9. MA plot for the 5742 peptides found in for replicate 1 *vs* replicate 2.

that a peptide in one condition will match within the threshold criterion with a nonidentical peptide. After applying the clustering strategy to the 60 000 monoisotopic peptide peaks for the three COFRADIC replicates, we found 18 736 matched peptide clusters representing a peptide in the three replicates. There were 4995 clusters containing only a single valid monoisotopic peptide peak. Those 4995 isolated peaks were regarded as noise and were removed from the data. To further eliminate possible noise peaks we considered only clusters with a peptide present in at least two consecutive fractions. This assumption is acceptable given the thorough separation of peptides over multiple fractions. Components that eluted over more than 25 min were considered contaminants (solvent, matrix, or internal standards) and were removed from the data. After this filtering step, we obtained 12 101 matched peptide clusters. The Venn diagram in Fig. 8 summarizes the peptides found in the three replicates. Note that only 5742 peptides were jointly found in all three COFRADIC runs. The problem with inconsistent finding of the other peptides can be related to the low intensity measurements of the peptides.

To study the variability between the three COFRADIC runs, we focus on the 5742 peptides found in the three experiments. The agreement between any two replicates is investigated by using the MA plots (see Bland and Altman[21]). In Fig. 9, the MA plot for the peptide abundances are shown. The ordinate *M* and abscissa

A for peptide *j* are calculated as

$$M = log_{10}(P_j) - log_{10}(P'_j) \quad (6)$$

$$A = \frac{\log_{10}(P_j) + \log_{10}(P'_j)}{2} \quad (7)$$

where P_j is the sum of (TIC-normalized) intensities for peptide j obtained from Eqn (5) for a particular COFRADIC run. The prime indicates the abundance of the same peptide P'_j, observed in another COFRADIC run. The plots show the difference for log intensity *versus* the mean log intensity for each peptide in the paired spectra. Ideally, the plots would show a symmetric scatter of points around the horizontal line at zero. This would suggest a simple additive measurement error with a constant variability, and without a systematic bias. The MA plot for two arbitrary COFRADIC runs is presented in Fig. 9. The plot is representative for the other COFRADIC runs (data not shown). Note that the scattered points are centered at the zero line, indicating that there is no bias in the different technical replicates. From Fig. 9 one can observe that the paired abundance measures exhibit a clear 'drop' shape, which indicates that the variability changes with the overall intensity level. One consequence of this fact is that it might be easier to detect differences for very abundant peptides (i.e., high-intensity peptides), as the variability for these peptides is smaller. Unfortunately, it would be difficult to assess whether low-abundant peptides are differentially expressed between different conditions, because measures of low intensity are much more affected by noise.

Furthermore, usually the mean, and sometimes the variance, is reported when quantifying results for the CV, calculated from the peptide abundances. However, these are descriptive statistics valid mainly for symmetric distributions. For skewed distributions, it is advisable to present a histogram plot or, at least, present the median and quartiles, in order to reflect the true distribution of the CVs. The mean value alone would not provide an adequate description of the distribution. From the three technical replicates, a CV for the 5742 peptide abundance P_j is computed with [0.025, 0.25, 0.50, 0.75, 0.975] quantiles equal to [0.0548, 0.1948, 0.3102, 0.4719, 0.8938] and mean 0.354, which indicates that the distribution of the CV is skewed. Also note that 50% of the peptides have a CV larger that 0.3102. This means that many replicates may be required to detect differentially expressed proteins.

To quantify the sensitivity of the COFRADIC methodology, i.e., the ability to detect low-abundant as well as high-abundant peptide peaks, we used the *dynamic range* expressed in decibels (dB), which was calculated as:

$$10 \log_{10} \frac{P_{max}}{P_{min}} \quad (8)$$

with P_{max} and P_{min} denoting the maximum and minimum peptide abundance measured in an experiment, obtained via Eqn (5). A dynamic range of 37 dB was found for all COFRADIC replicates. This means that the order of magnitude of difference in detected concentrations is approximately equal to 5000.

The internal mass calibration was checked by calculating the ppm for the internal standards IS2, IS3, IS4, and IS5. All ppm values were below 100, indicating that the calibration performed by the acquisition software was accurate.

Conclusions

This paper mainly focuses on the preprocessing of raw ASCII MS1 data files generated from a high-dimensional LC-MS setting, in this case, *N*-terminal COFRADIC. To this aim, prior knowledge about the peptide's isotopic distribution is used to turn the LC-MS data into a protein/peptide list, such that the abundances across different LC-MS runs can be easily compared by using classical statistical methods. A prerequisite for the proposed method is that the data are produced by a highly accurate, high-resolution LC-MS system. The advantage of working with high-resolution mass spectra is that a peptide will appear as a series of peaks with peak heights proportional to the probability of occurrence of the isotopic variants of the peptide. This enables us to discern peaks generated by error from those due to a peptide. When this is not the case, careful adjustments should be made to the proposed strategy. For example, Valkenborg *et al*.[22] describe a method that is able to extract features from low-resolution MS data for clinical diagnosis.

The algorithm proposed in this manuscript has been implemented in MATLAB, and can be flexibly adapted to other specifications. Processing one mass spectrum takes approximately 2.6 s on a DELL Latitude D505. Currently, adjustments to the algorithm are implemented in order to increase the speed, and to ultimately obtain a tool for real-time mass spectra preprocessing, such that the algorithm can be efficiently included in the pipeline of a laboratory information management system (LIMS). The clustering algorithm assembles the peptides and matches them across the three technical COFRADIC replicates in 3.96 s.

Two experiments were specially designed purely for the evaluation of the proposed analysis strategy. After applying the algorithm to the bovine cytochrome C and COFRADIC data sets, we could conclude that the removal of the normal error term ε in (1) by undecimated continuous wavelet transform[11] has only a minor effect on the intensity of the peptide peaks. The baseline correction is fast and yields results comparable to, e.g., LOWESS. A disadvantage of this baseline removal is the need to truncate the resulting negative values at zero, which hampers the noise estimation. In some regions an infinite signal-to-noise ratio is detected because of zero noise. An infinite signal-to-noise ratio is not an informative measure; therefore, one might consider a more conservative baseline removal, such as a moving minimum filter. The proposed mass calibration works well, but does not improve the accuracy when the machine calibration is within a 100 ppm error interval. On the other hand, the mass calibration was proven to work correctly on a separate ill-calibrated data set.

Information about the height of the peptide peaks in a mass spectrum is sufficient, while information about the shape of the isotopic peaks can be safely discarded. A disadvantage of working with a peak height only is that the detection of overlapping peaks is difficult. Despite the peptide separation obtained by COFRADIC, it can happen that two peptides with approximately the same mass elute from the column. Therefore, the detection of overlapping peaks may still require attention. The issue of how a series of overlapping peptide peaks can be automatically interpreted and how this can be implemented in an efficient manner to the proposed strategy is a topic of further research.

The two-step clustering algorithm operates on the set of validated monoisotopic peaks and is, in this case, a fast alternative for, e.g., K-means clustering. An additional advantage is that we do not need to specify the number of expected clusters in advance, because it is automatically inferred from the data.

Journal of MASS SPECTROMETRY

It should be noted that the proposed strategy was tailored for the specific case of COFRADIC samples in combination with a high-resolution MS. Therefore, caution should be applied when using the proposed strategy to other experimental settings.

Supporting information

Supporting information may be found in the online version of this article.

Acknowledgements

Financial support from the IAP research network nr P6/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged by the first and last author.

The first author acknowledges support from Bijzonder Onderzoeksfonds Universiteit Hasselt (grant BOF04G01).

We are grateful to the reviewers for their insightful comments, which resulted in an improved manuscript.

References

[1] K. Gevaert, M. Goethals, L. Martens, J. Van Damme, A. Staes, G. R. Thomas, J. Vandekerckhove. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted *N*-terminal peptides. *Nature Biotechnology* **2003**, *21*, 566.

[2] K. Sandra, K. Verleysen, C. Labeur, L. Vanneste, F. D'Hondt, G. Thomas, K. Kas, K. Gevaert, J. Vandekerckhove, P. Sandra. Combination of cofradic and high temperature-extended column length conventional liquid chromatography: a very efficient way to tackle complex protein samples, such as serum. *Journal of Separation Science* **2007**, *30*, 658.

[3] X. Li, E. C. Yi, C. J. Kemp, H. Zhang, R. Aebersold. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Molecular and Cellular Proteomics* **2005**, *4*, 1328.

[4] V. Tusher, R. Tibshirani, G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **2001**, *98*, 5116.

[5] J. J. Jaffe, D. R. Mani, K. C. Leptos, G. M. Church, M. A. Gilette, S. A. Carr. Pepper, a platform for experimental proteomic pattern recognition. *Molecular and Cellular Proteomics* **2006**, *5*, 1927.

[6] L. N. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M. Brusniak, O. Vitek, R. Aebersold, M. Müller. Superhirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **2007**, *7*, 3470.

[7] A. Schmidt, R. Aebersold. High-accuracy proteome maps of human body fluids. *Genome Biology* **2006**, *7*, 242.

[8] M. Hilario, A. Kalousis, C. Pellegrini, M. Müller. Processing and classification of protein mass spectra. *Mass Spectrometry Reviews* **2006**, *25*, 409.

[9] O. Schulz-Trieglaff, R. Hussong, C. Gröpl, A. Hildebrandt, K. Reinert. A fast and accurate algorithm for the quantification of peptides from mass spectrometry data. *Lecture Notes in Bioinformatics* **2007**, *4453*, 473.

[10] D. Valkenborg, P. Assam, L. Krols, G. Thomas, K. Kas, T. Burzykowski. Using a poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. *Rapid Communications in Mass Spectrometry* **2007**, *21*, 3387.

[11] K. Coombes, S. Tsavachidis, J. Morris, K. Baggerly, M. Hung, H. Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from seldi by denoising spectra with the undecimated discrete wavelet transform. *Proteomics* **2005**, *5*(16), 4107.

[12] J. Listgarden, A. Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular and Cellular Proteomics* **2005**, *4*, 419.

[13] D. Valkenborg, I. Jansen, T. Burzykowski. A model-based method for the prediction of the isotopic distribution of peptides. *Journal of the American Society for Mass Spectrometry* **2008**, *19*(5), 703.

[14] M. Senko, S. Beu, F. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. *Journal of the American Society for Mass Spectrometry* **1995**, *6*, 229.

[15] R. Grass, M. Müller, E. Gasteiger, S. Gay, P. Binz, W. Bienvenut, C. Hoogland, J. Sanchez, A. Biaroch, D. Hochstrasser, R. Appel. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* **1999**, *20*, 3535.

[16] M. Guilhaus. Principles and instrumentation in time-of-flight mass spectrometry. *Journal of Mass Spectrometry* **1995**, *30*, 1519.

[17] J. Silva, R. Denny, C. Dorschel, M. Gorenstein, I. Kass, G. Li, T. McKenna, M. Nold, K. Richardson, P. Young, S. Geromanos. Quantitative proteomic analysis by accurate mass retention time pairs. *Analytical Chemistry* **2005**, *77*, 2187.

[18] http://www.unimod.org/ [accessed 11 April 2008].

[19] P. Picotti, R. Aebersold, B. Domon. The implications of proteolytic background for shotgun proteomics. *Molecular and Cellular Proteomics* **2007**, *6*(9), 1589.

[20] A. Rockwood. Relationship of Fourier transforms to isotope distribution calculations. *Rapid Communications in Mass Spectrometry* **1995**, *9*, 103.

[21] J. Bland, D. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307.

[22] D. Valkenborg, S. Van Sanden, D. Lin, A. Kasim, Q. Zhu, P. Haldermans, I. Jansen, Z. Shkedy, T. Burzykowski. A cross-validation study to select a classification procedure for clinical diagnosis based on proteomic mass spectrometry *Statistical Applications in Genetics and Molecular Biology* **2008**, *7*(2), Article 12.

