Electronic Journal of Statistics
Vol. 8 (2014) 1703–1713
ISSN: 1935-7524
DOI: 10.1214/14-EJS900

Proteomics profiles from mass spectrometry*,†

Inge Koch

School of Mathematical Sciences, The University of Adelaide, Australia
e-mail: inge.koch@adelaide.edu.au

Peter Hoffmann

School of Molecular and Biomedical Science, The University of Adelaide, Australia
e-mail: peter.hoffmann@adelaide.edu.au

and

J. S. Marron

Department of Statistics and O.R., University of North Carolina
Chapel Hill, NC, 27599-3260, USA
e-mail: marron@unc.edu

Abstract: Proteomics is a rapidly growing research area within bioinformatics which focuses on quantification of peptide concentrations and on the identification of proteins and peptides. In quantitative proteomics the identification of biomarkers from peptide concentrations is important for diagnostic purposes and treatment of diseases.

The goal of this paper is to facilitate research in this area, by providing a test bed for comparison of 1D curve registration methods. This is done in a novel way, by providing not only curves, but also an answer key as to how the peaks should align. In the following papers a number of approaches to this problem are given, and the answer key provides unusually useful insights into how the methods compare.

For this reason, we consider proteomics mass spectrometry profiles which are part of a larger study into the identification of biomarkers in Acute Myeloid Leukaemia (AML). For these profiles large ion counts result in large peaks, but these peaks may occur at different retention times for different profiles. The first step in the quantification of peptides in proteomics profiles is the alignment of the 1D curves of total ion count (TIC).

The paper includes a description of proteomics mass spectrometry profiling, and considers profiles from five patients with AML. It outlines the preprocessing steps we applied to the multiple TIC samples from each patient, and introduces the reference peptides. The retention times of the reference peptides are known for each profile, and using these times as an answer key makes the 1D TIC curves a particularly informative test bed for curve registration.

Keywords and phrases: Mass spectrometry, one-dimensional curve registration, peak alignment, proteomics, total ion counts.

Received August 2013.

*This is an original research paper.

†Related articles: 10.1214/14-EJS900A, 10.1214/14-EJS900B, 10.1214/14-EJS900C, 10.1214/14-EJS900D and 10.1214/14-EJS900E; rejoinder at 10.1214/14-EJS900REJ.

1. Introduction

Proteomics includes the study of proteins, the smaller peptides and their expression levels. Availability of high-accuracy mass spectrometry (MS) instruments in proteomics research, which allow efficient acquisition of mass data, has led to the development of methods for quantitative proteomics. Of particular interest is the quantification of peptide concentrations which includes comparisons of peak intensities across multiple data sets.

The mass data consist of detected peptide ions which are recorded for a range of retention times and mass-to-charge intervals. In this paper we focus on total peptide ion count – or total ion current (TIC) – as functions of retention times only, and we refer to these data as mass spectrometry (MS) curves or MS profiles, or simply as profiles or curves. For different profiles large peaks may occur at different retention times – partly a consequence of the set-up and external conditions influencing the measurements. An alignment of the profiles is therefore an important first step in an analysis of such data.

We consider liquid chromatography-mass spectrometry (LC-MS) profiles of total ion count which were collected as part of a larger study into the identification of biomarkers in Acute Myeloid Leukaemia (AML) at the University of Adelaide in 2010–11, see Ho [3] and references therein. For these data a number of *reference peptides* with known retention times are available. Our mass profiles include the reference peptides which makes these one-dimensional profiles an unusually informative test bed for curve registration.

Acute Myeloid Leukaemia is an aggressive haematological malignancy in which haematopoietic progenitors are arrested in an early stage of development. Extensive cell proliferation of immature blasts accumulates in bone marrow and peripheral blood and eventually in the liver, spleen and the central nervous system. See Lowenberg *et al.* [5]. The conventional therapy for patients diagnosed with AML is chemotherapy which has a survival rate of less than 50% over five years, and which is even less effective for patients over 60 years old. See Koschmieder *et al.* [4]. Alternative treatments of AML include the therapeutic use of all-trans retinoic acid (ATRA) which has benefitted patients with a distinct subtype of the disease. See Czibere *et al.* [2]. So far, there is not enough evidence to determine whether other patient groups would benefit from ATRA. Clinically useful biomarkers are required to assist therapeutic decision making and prediction of therapeutic responses for each individual patient with AML. In the biomarker discovery study at the University of Adelaide, we try to find proteins through an MS profiling approach which is able to distinguish between patients who respond and those who do not respond to chemotherapy and treatment with ATRA.

This paper is organised as follows. Section 2 gives a brief introduction to proteomics, including the LC-MS approach. Section 3 describes LC-MS for the AML data and discusses techniques and aims relevant to these data. In Section 4 we describe the data in more detail, including the preprocessing steps of the AML data. Section 5 suggests how to assess the performance of alignment methods and proposes a visualisation tool for illustrating and comparing the success of alignment methods.

2. Proteomics

The proteome is the entire complement of proteins including their splice forms and post translational modifications of a cell, a tissue or a body fluid at any given point in time. Proteomics is the large scale qualitative and quantitative analysis of the proteome and the field which aims to describe the proteome of all biological organisms.

The word *proteome* was coined by Marc Wilkins in 1994 out of the words protein and genome, see Wasinger *et al.* [7], and the word *proteomics* was coined in 1997 in analogy with genomics. The proteome is a very complex mixture of proteins and for that reason proteomics uses separation methods with high resolution. 2D gel electrophoresis was the first method to separate proteins on a large scale and to detect them using protein staining. Separated proteins can be identified by mass spectrometry. Proteins are very long chains of amino acids with a large molecular weight. In order to accurately measure and identify proteins by mass spectrometry, proteins are cleaved (digested) into peptides which is routinely done with the enzyme trypsin. The resulting tryptic peptides are of appropriate molecular weight and are ideal for analysis by mass spectrometry.

Peptide identification. Tryptic peptides can be measured very accurately by mass spectrometry, and modern MS instruments are able to isolate specific peptides by their masses and fragment them. The generated fragmentation spectra of each peptide can be used to identify the peptide's amino acid sequence, and if this sequence is long enough it will be specific for exactly one protein in the proteome. This means the protein can be identified using a protein database containing all known proteins. For human and mouse samples all proteins are known because the genome of these organisms are sequenced, and the proteome can be translated from the genome.

LC-MS. Electro spray ionisation (ESI) – liquid chromatography (LC) – mass spectrometry is commonly used for identification and quantification of peptides. ESI-LC-MS refers to the separation of the complex peptide mixture which results from a tryptic digest: separation of the peptides is accomplished by liquid chromatography, and the peptides are ionised by electro spray and their masses measured by a mass analyser in the mass spectrometer. If the peptides are also isolated and fragmented in order to identify their sequences, then the process is called ESI-LC-MS/MS, or simply LC-MS/MS or LC-MS2.

The acquisition of the AML data is accomplished with LC-MS – the first step in liquid chromatography – which results in measurements of ion counts at retention times. For each retention time there is a mass spectrum with peaks at different mass-to-charge values. The sum of the intensities across all mass-to-charge values is the total ion count or total intensity recorded with LC-MS. See also Section 3. LC-MS does not involve a secondary fragmentation or, more generally multiple fragmentations, inherent in LC-MS/MS or LC-MSn respectively, where n-1 is the number of fragmentations following LC-MS. LC-MS has evolved as a powerful method for identification and quantification of peptides

and proteins. Because of its high peak resolution in the two dimensions mass-to-charge and retention times, LC-MS is suitable for comparing peak intensities between MS samples based on retention times and mass-to-charges values. Experimental drifts in mass-to-charge values and retention times typically occur in LC-MS, and may be non-linear across samples. These drifts complicate direct comparisons of multiple LC-MS samples, and their complexity makes a successful alignment of LC-MS data difficult. For an overview and a review of recent approaches in LC-MS, see America and Cordewener [1].

Alignment methods in LC-MS. A number of alignment methods have been developed in the proteomics literature specifically for LC-MS data. Table 1 of America and Cordewener [1] contains a list of algorithms and their public or commercial availability. The algorithms typically focus on aligning peptides or individual peaks and apply to two samples at a time. We briefly describe the main ideas of two typical algorithms: SIEVE and *SuperHirn*. Both algorithms work with peaks observed at retention times and mass-to-charge values as obtained with LC-MS.

The commercially available SIEVE algorithm is part of the instrumentation used in Ho [3]. The first part of the SIEVE algorithm deals with aligning spectra; the other parts deal with feature matching and protein expression ratio calculations which are not relevant for this paper. SIEVE requires the selection of a reference sample, say R, and compares each sample, S, with the reference sample. For each pair of time points $(t_{R,j}, t_{S,k})$ in the full range of retention times of R and S respectively, the correlation coefficient of marginal mass-to-charge spectra at $t_{R,j}$ and $t_{S,k}$ respectively is calculated. In a second step, SIEVE exploits additional information in the form of internal reference points with known retention times for specific peak intensities in order to find optimal pairs of retention times which are consistent with the alignment of the reference peaks. The time points of the sample are then aligned with those time points of the reference sample that yield highest correlation. This process is repeated for each sample.

The publicly available algorithm *SuperHirn* of Mueller et al. [6] first finds peaks and sorts them in decreasing order. The peak list is compared with a template which is typically obtained from an LC-MS/MS fragmentation, and the order of the peaks is adjusted in accordance with the size of the peaks in the template. This template-matching is necessary to the success of the algorithm, as it provides valuable information about the size of the peaks which might be obscured by noise in the LC-MS records. In a second step, peaks with the same mass-to-charge value are combined as '*features*' across retention times. The peak-finding, peak-sorting and peak-combining steps are carried out separately for each sample, and result in a ranked list of features. For two samples at a time, common features are identified, and a similarity score is calculated which uses a robustified Spearman score. Once all pairwise similarity scores have been calculated, the alignment is accomplished in a hierarchical way: the pair with the highest score is regarded as most similar, and the two samples are merged. The merged sample replaces the two previous samples, similarity scores are updated

and the next best pair is merged. This process continues until all samples have been merged, and are therefore aligned.

These algorithms typically work well, but they are limited in a number of ways:

1. Calculations are carried out for pairs of samples, and cannot be easily adapted to dealing with more than two samples simultaneously.
2. Additional information relating to known peptides and their retention times or fragmentation patterns is integrated into the alignment process.

In this paper we are interested in alignment based merely on the 1D TIC curves. However, the availability of reference peptides allows us to assess and compare the performance of curve registration methods.

3. The AML data

LC-MS for AML biomarker discovery. For the AML data, Ho [3] used label-free LC-MS which refers to an approach that does not require isotopic labels as references for quantification, but still enjoys the high peak resolution common to LC-MS.

In the AML biomarker discovery study six patients were tested: three responders and three non-responders. One of the responders was excluded from the final analysis because of problems with the liquid chromatography system during the analysis. For the discovery of biomarkers by MS profiling with label-free LC-MS, samples from each patient were analysed multiple times. Isolated blast cells were lysed protein extracted and an equal amount of protein was digested with trypsin separately for each patient sample, thereby circumventing a need for a subsequent normalisation of the data. Digested tryptic peptides were analysed, and the complex mixture of tryptic peptides was separated via a reversed-phase column using a long 240 minute gradient. At each time point peptides were eluted from the column and sprayed for detection in the MS instrument. The recorded profiles contain peaks which represent peptides eluted at a particular time point, the retention time. Peaks which are identified as potential peptides are isolated and fragmented for identification. The retention times of the peptides in the 240 minute gradient had unusually large shifts. Whilst these large shifts make it difficult to align the TIC curves, they give rise to data that are ideal for testing and for comparing multiple methods of alignment.

Quantification of AML profiles. In order to quantify changes in AML profiles, the area under the peaks of every potential peptide is compared for all runs which result from each sample of each patient. A quantitative comparison for each potential peptide requires the time alignment of the traces or profiles over all runs, the TIC curves, since the peptides should have the same mass-to-charge values and should elute from the column at the same retention time. After alignment the aim is to find peptides which have different abundances

in the responding and non-responding patient samples. Such peptides and their related proteins could act as biomarkers for distinguishing responders from non-responders.

AML patient samples. AML data of five patients were collected for a range of mass-to-charge intervals and retention times. To minimise the effect of experimental drifts in mass-to-charge values and retention times – see the paragraph on LC-MS in the previous section – a first sample was collected for all five patients in a first run or experiment. The experiment was repeated twice with randomly selected patient order, and resulted in a total of three replicates for each patient. The three sets of samples corresponding to the non-responders are referred to as A_i, B_i and C_i, and the two sets of samples of the responders are referred to as X_i and Y_i, where $i = 1, 2$ and 3 indexes the three replicates for each patient.

The randomisation of the experimental order addresses primarily the drift in retention times. A common approach for dealing with the drift in mass-to-charge values is to aggregate peak intensities across all mass-to-charge values – separately for each retention time. This process is typically part of the instrument-internal acquisition and processing, and yields one-dimensional profiles as functions of the retention times. These profiles contain the total ion counts (TIC), and as they are the curves in our registration test bed, we refer to them as the raw data.

4. The TIC alignment problem

In this paper we are interested in aligning all one-dimensional TIC curves simultaneously. The TIC profiles are difficult to align because LC-MS profiles typically exhibit non-linear drifts in retention times, and for these data there appear to be unusually large shifts in retention times of peptides. On the other hand, these data are very suitable for testing different 1D alignment methods, because all samples were spiked with known peptides which are independent of the patients. The identity and retention times of these peptides are known and can be checked before and after alignment. We refer to these peptides as *reference peptides*.

The raw TIC data are noisy, and noise is particularly prominent for the very early and very late retention times. This is a feature of the equipment which requires a 'warming-up period' and typically also has no useful results for times larger than 220 minutes. It was therefore necessary to reduce the range of retention times. In addition we applied a number of other preprocessing steps to the TIC data. The individual steps are:

1. **Define the data at regularly spaced time points.** The intensity measurements were available at the midpoints of contiguous time intervals. The intervals changed in size along the complete time range with an average close to 0.1 minutes. For the time points in the range 5 to 240 minutes, we therefore chose regular steps of 0.1 minutes and interpolated the

data onto this grid, which resulted in a total of 2349 linearly interpolated TIC values.

2. **Log transform.** As common in many analyses of proteomics data, we applied a $\log_{10}$ transform to the interpolated data. This is sensible, because the intensities range over several orders of magnitude.
3. **Median smoothing.** The original and log-transformed data are quite noisy. To reduce the noise, we applied a running median filter to the interpolated data. We examined a number of filter lengths, and settled on a smooth over 9 consecutive points as this amount of smoothing still preserves the peaks well.
4. **Truncation of the time range.** Because of the time drifts that occur during the data acquisition, the three different experiments (runs 1–3) showed a clear shift which was easily recognisable by comparing the times of the large peaks. Some experimentation was necessary to make sure that all important peaks were included in the truncated time range, yet the noisier observations at the beginning and end were excluded. We found the range 20–220 minutes suitable for these data, which reduces the original interpolated data to 2001 time points for each profile.

Figure 1 shows all log-transformed and smoothed profiles. For easier visualisation, we displayed the profiles of the three runs in separate panels. The x-axis shows time, and the y-axis shows the log intensity or the log TIC. For each patient the same colour is used in each of the three panels. Dark blue, lighter blue and cyan refer to the three non-responders, and red and orange refer to the two responders. In the top panel the highest TIC peak is at about 115 minutes for all five patients. In the second run (second panel) the highest peak has drifted to smaller times, but by different amounts for the five patients. We notice a similar behaviour for the third run.

To examine the drift between runs and between patients further, we consider the three profiles of the non-responder A, and the second profile X_2 of the responder X in Figure 2. As in Figure 1, time is shown on the x-axis, and the log TIC on the y-axis.

The two panels in the top row and the left panel in the bottom row of Figure 2 refer to patient A. The three runs for patient A clearly show very similar overall peak patterns, together with a strong time drift as observed in the three panels of Figure 1. In the top left panel, which shows the profile of A_1, the largest peak appears at about 115 minutes, for A_2 in the top right panel we notice it at about 105 minutes, and in bottom left panel, which corresponds to A_3, this peak is observed at about 98 minutes. The bottom right panel shows the profile of X_2. These data were collected in the same run as the A_2 data. We note that the largest peak occurs at about 105 minutes, but this peak is much smaller than the largest peaks of the three A profiles. A major challenge to registration algorithms is that the peak patterns, in terms of relative peak heights, are quite different for this sample. In particular a group of three peaks around the largest peak in the A spectra appears to correspond to a group of three peaks in the X_2 sample for which the two outer peaks are high, but the center peak is much smaller.

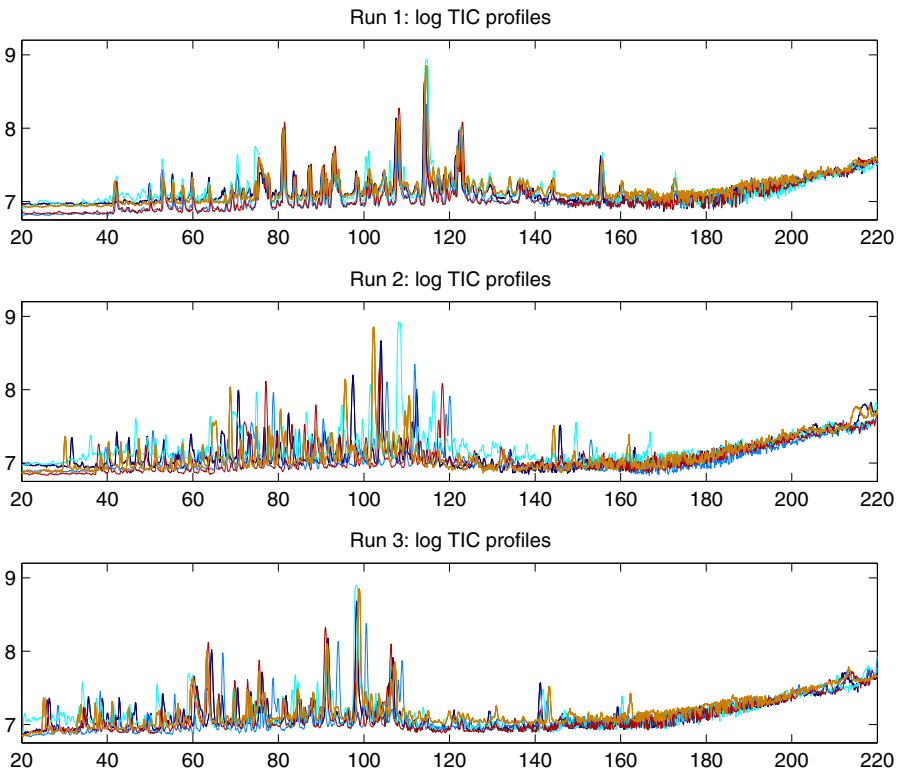

FIG 1. *The three panels show the log transformed TIC profiles for a particular run and all patients versus time. Blue colours refer to non-responders, and red/orange colours refer to the two responders. The samples show generally similar patterns, but with clear time drifts, especially for runs 2 and 3.*

Next we study the region around the largest peaks of these four profiles more closely and zoom in to the range $[90, 130]$ minutes, chosen to include the three peaks in each case. Figure 3 displays these zoom-ins of the four profiles of Figure 2, shown in the same order and using the same colours as in Figure 2. The three peaks, indicated by blue vertical line segments at their retention times and indexed 9, 10, and 11 are of special interest; they correspond to peptides with known masses which allows very accurate measurement of their retention times by mass spectrometry. These peptides are a subset of the reference peptides which hold the answer key for the curve registrations. In the bottom right panel, the retention times of the three reference peptides are traced across all four samples – separately for each of the three peptides – and the actual times are marked by dots along each trace: the three blue dots with the lowest y-value show the times of the A_1 sample, the three blue dots with the second lowest y-value correspond to the times of the A_2 sample, and those with the third lowest y-value refer to the times of the A_3 sample. Finally, the retention times

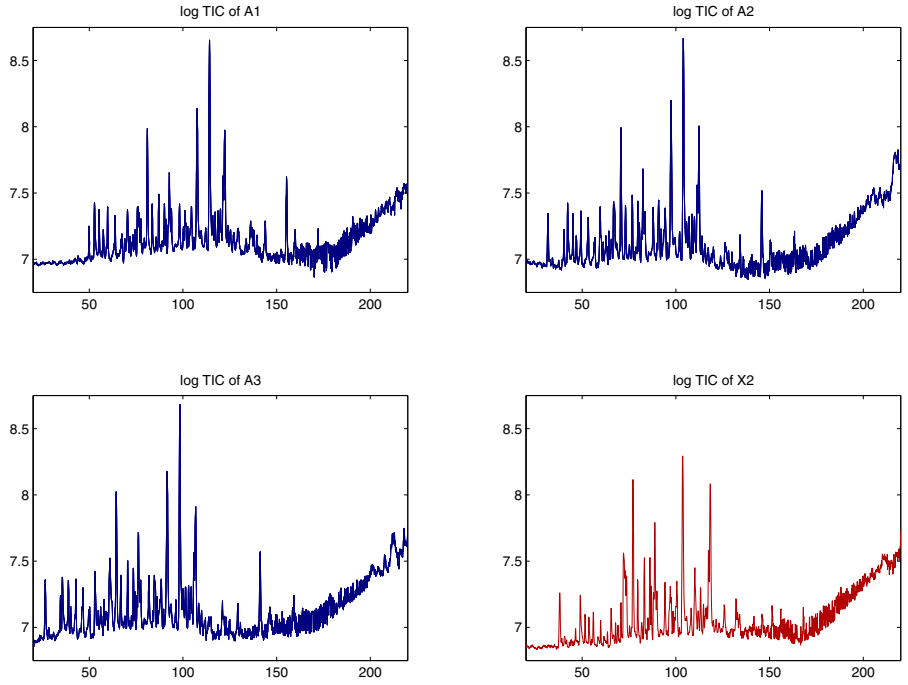

FIG 2. *Log transformed TIC profiles of the samples A_1, A_2, A_3 and X_2 with the log intensity on the y-axis versus time on the x-axis. The panels show a time drift for the A samples, and a noticeably different peak pattern for the X_2 sample.*

of the corresponding peptides of the X_2 sample are shown as three red dots, with numbers 9, 10 and 11 next to these dots. We can see a clear time drift in the A profiles of the numbers 9–11. In each of these three profiles, the three peaks are about the same size and the same relative distance from each other. It is surprising to see that the largest peak of the X_2 profile is peak number 9, while peak number 10 is comparatively small, indicating that peptide 10 is far less prevalent in X_2. If this lower abundance of peptide 10 is also seen in the profiles of the other responders, this peptide could potentially lead to the identification of a biomarker for distinguishing responders from non-responders.

The different shape of the three peaks in profile X_2 poses a challenge for alignment methods, as all methods that are discussed in the following contributions are applied blindly to the data, that is, without knowing that the small peak at about 110 minutes arises from the same peptide as the largest peaks in the A profiles.

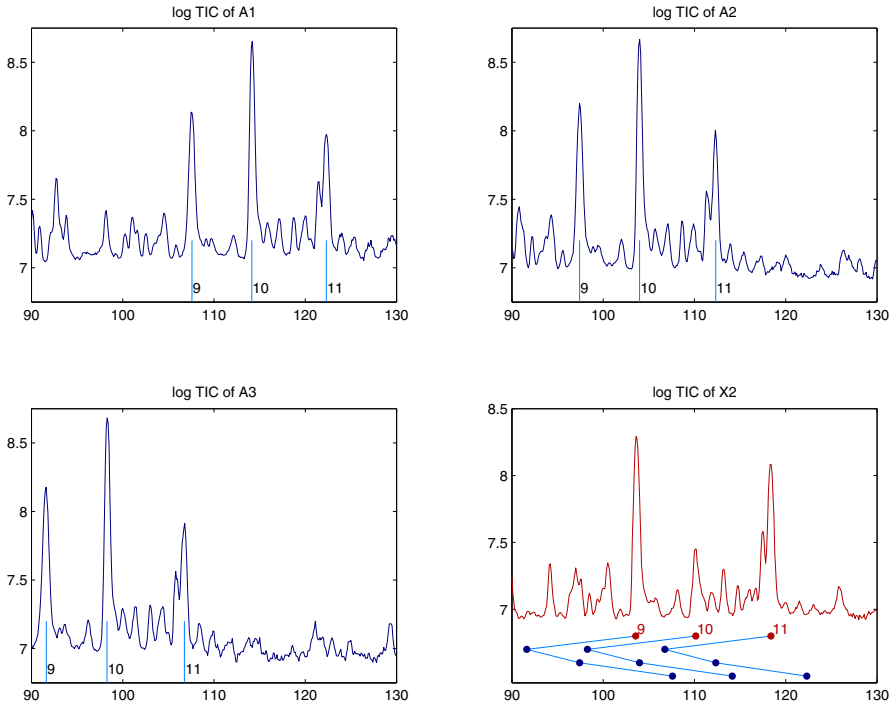

FIG 3. *Zoomed-in log transformed TIC profiles of the samples A_1, A_2, A_3 and X_2 with the log intensity on the y-axis versus the time interval $[90, 130]$ minutes on the x-axis. The panels show a close correspondence of both tall and short peaks, as well as how the central large peak in the A samples is much smaller in X_2. The x-values of reference peptides 9–11 are shown separately for each of the A samples, and are shown as blue and red dots in the bottom right panel.*

5. Assessment of alignment performance

The three figures of the previous section show a time drift and variability in the intensity of the large peaks across runs and patients. For these data, relative heights of these peptide peaks are the desired focus of the analysis and the various time drifts should be regarded as a nuisance to be removed. Thus this is a classical registration problem of time alignment, but in the challenging case of differing peak heights.

As noted above, a very useful feature of this data set for the comparison of registration methods is that there are 14 reference peptides whose time locations are known for each sample. These reference peptides are not necessarily seen as large peaks in each profile, however their mass-to-charge ratios and retention times provide excellent criteria for assessing the quality of curve registration.

A range of statistics could be applied to the vectors of warped times of the reference peptides over all 15 profiles. These statistics include the median devi-

ation for each reference peptide, the mean variance across all reference peptides or other measures derived from these statistics. Such indicators quantify the quality and performance and provide an insight in the overall performance of particular registration methods.

We also suggest use of a visual tool for performance assessment. We assign numbers 1 to 14 to the reference peptides of each profile. Before alignment, the 14 numbers appear at different time points across the different runs and patients, as can be seen in the four profiles in Figure 3 which concerns reference peptides 9–11. In addition, the bottom right panel traces the times of the reference peptides 9–11 of the four spectra before alignment and shows the variability in the drift between samples for these reference peptides. After alignment we recommend plotting the numbers of the 14 warped reference peptides again, and then comparing the location of these numbers across all profiles. The goal is to align the profiles in such a way that all 14 reference peptides appear at the same time points across the 15 profiles.

Acknowledgement

The authors are grateful to the Mathematical Biosciences Institute for financial support which made the meeting possible which generated interesting analyses of this data set. The authors thank the reviewer for helpful comments.

References

[1] America, A. H. P. and Cordewener, J. H. G. (2008). Comparative LC-MS: A landscape of peaks and valleys. *Proteomics* 8, 731–749.

[2] Czibere, A., Grall, F. and Aivado, M. (2006). Perspectives of proteomics in acute myeloid leukemia. *Expert Rev. Anticancer Ther.* 6, 1663–1675.

[3] Ho, Y. Y. (2011). *Protein profiling of acute myeloid leukemia-specific membrane proteins using label-free liquid chromatography-mass spectrometry.* Honours thesis in Chemistry, The University of Adelaide (available from peter.hoffmann@adelaide.edu.au).

[4] Koschmieder, S., Rosenbauer, F., Steidl, U., Owens, B. M. and Tenen, D. G. (2005). Role of transcription factors C/EBPα and PU.1 in normal hematopoiesis and leukemia. *Int. J. Hematol.* 81, 368–377.

[5] Lowenberg, B., Downing, J. R., Burnett, A. N. and Engl, J. (1999). *Medicine* 341, 1051.

[6] Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M. Y., Vitek, O., Aebersold, R. and Mueller, M. (2007). SuperHirn – A novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 7, 3470–3480.

[7] Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L. and Humphery-Smith, I. (1995). *Electrophoresis* 16, 1090.