Robust Peak Detection and Alignment of nanoLC-FT Mass Spectrometry Data

Marius C. Codrea[1], Connie R. Jiménez[2], Sander Piersma[2], Jaap Heringa[1], and Elena Marchiori[1]

[1] Centre for Integrative Bioinformatics VU (IBIVU)
Department of Computer Science
mcodrea@few.vu.nl, heringa@few.vu.nl, elena@cs.vu.nl
[2] Cancer Center Amsterdam
c.jimenez@vumc.nl, S.Piersma@vumc.nl
Vrije Universiteit Amsterdam, The Netherlands

Abstract. In liquid chromatography-mass spectrometry (LC-MS) based expression proteomics, samples from different groups are analyzed comparatively in order to detect differences that can possibly be caused by the disease under study (potential biomarker detection). To this end, advanced computational techniques are needed. Peak alignment and detection are two key steps in the analysis process of LC-MS datasets. In this paper we propose an algorithm for LC-MS peak detection and alignment. The goal of the algorithm is to group together peaks generated by the same peptide but detected in different samples. It employs clustering with a new weighted similarity measure and automatic selection of the number of clusters. Moreover, it supports parallelization by acting on blocks. Finally, it allows incorporation of available domain knowledge for constraining and refining the search for aligned peaks. Application of the algorithm to a LC-MS dataset generated by a spike-in experiment substantiates the effectiveness of the proposed technique.

1 Introduction

Computational analysis of proteomic datasets is becoming of crucial relevance for discovery of reliable and robust candidate biomarkers. In particular, quantitation of changes in protein abundance and/or state of modification is the most promising, yet most challenging aspect of proteomics. In recent years label-free LC-MS methods that quantify absolute ion abundances of peptides and proteins have emerged as promising approaches for peptide quantitation and profiling of large numbers of clinical samples [1].

Briefly, peptides are subjected to (multi-dimensional) liquid chromatography for separation. Each peptide fraction is then analyzed on an LC-MS system. Each LC-MS run of a sample generates a pattern of very high input dimension consisting of one intensity (relative abundance) measurement for each pair of molecular mass-to-charge ratio (m/z) and retention time (RT) values. Ideally, the same molecules detected in the same LC-MS instrument should have the

same retention time, molecular weight, and signal intensity. However, in practice this does not happen due to experimental variations. As a consequence, patterns generated by LC-MS runs need to undergo a number of processing steps before they can be comparatively analyzed. Such processing steps include normalization [5, 17], background subtraction [8], alignment [4, 13, 16, 17], and peak detection (e.g., [9]). Several tools and algorithms for processing and for difference analysis of LC-MS datasets have been introduced (e.g., [2, 3, 7, 9, 11, 15, 20, 17]).

In this paper we focus on peak detection and alignment. As well explained in the overview paper by Listgarten et al [10], alignment algorithms involve either (i) the maximization of an objective function over a parametric set of (generally linear) transformations, or (ii) non-parametric alignment based on dynamic programming, or (iii) combination of these methods like piecewise transformations. They act either on the full pattern or on features (peaks) selected beforehand; they may or may not use the signal intensity and they may or may not incorporate scaling. Most of alignment algorithms require a reference template, to which all time series are aligned. Peak detection is usually performed in an ad-hoc manner [10], involving either a comparison of intensities with neighbours along the m/z axis [19] or detection of coinciding local maxima [18].

Whether one should perform peak detection before [14, 17] or after [12] alignment has not been clearly established. In this paper we circumvent this issue by performing both tasks at the same time by means of a novel clustering algorithm. The motivations for a new algorithm rely also on the desire to overcome drawbacks of alignment algorithms, such as the need for a reference template, or the assumption of a given (local or global, usually linear) transformation in RT dimension. The versatility of clustering for simultaneous alignment and peak detection has been already recognized by Tibshirani et al. in [16]. They align MALDI (a technique that generates two dimensional patterns of m/z and intensity values from each sample) data along the m/z axis by applying one-dimensional hierarchical clustering with complete linkage for constructing the dendrogram and a specific cutoff for extracting clusters representing peaks.

The algorithm proposed here, called Peak Detection and Alignment (`PDA`), acts on blocks of m/z values. Blocks are obtained by splitting the runs along the m/z axis (e.g., in blocks of equal size). Each block is processed individually. The input of `PDA` is a set of (blocks of) runs described by a list of triplets (m/z, RT, run id) consisting of 'm/z' value, 'RT' value and run identifier. The output of `PDA` is a set of clusters of (m/z, RT) pairs representing peaks, together with information about their signal intensity.

Novel features of `PDA` include:

1. a similarity measure for comparing features, where a feature is a triplet (m/z, RT, id). Weights are associated to each of the three attributes to specify their relevance;
2. a cluster merging strategy;
3. a cluster refinement procedure for handling peaks occurring near the boarder of blocks.

Application of PDA to a dataset generated by a spike-in experiment substantiates its effectiveness. The results indicate that PDA provides a flexible, efficient and robust approach for simultaneous peak detection and alignment of high-throughput label-free LC-MS data.

2 Materials and methods

2.1 nanoLC-FT Mass Spectrometry Proteomic Data

Efficiently identifying and quantifying disease- or treatment-related changes in the abundance of proteins in easy accessible body fluids such as serum is an important area of research for biomarker discovery. Currently, cancer diagnosis and management are hampered by lack of discriminatory and easy obtainable biomarkers. In order to improve disease management more sensitive and specific biomarkers need to be identified. In this light, the simultaneous detection and identification of multiple biomarkers (*molecular signatures*) may be more accurate than single marker detection. Therefore, great promise holds in combining global profiling methods, such as proteomics, with powerful bioinformatics tools that allow for marker identification. The additional dimension of separation provided by coupling nanoliquid chromatography to high-performance Fourier transform mass spectrometry (nanoLC-FTMS) allows for profiling large numbers of peptides and proteins in complex biological samples at great resolution, sensitivity and dynamic range.

In LC-MS, the *chromatographic column* separates peptide mixtures based on one or more physicochemical properties prior to MS. An *ionization source* converts eluting peptides into ions which are separated by the *mass analyzer* on the basis of m/z ratios. The *detector* then registers the relative abundance of ions at discrete m/z values. Therefore, LC-MS yields values indicating that, at a particular time, an ion with a particular m/z value was detected with a particular intensity. In the proteomics community, the unit of measure of the m/z axis is one Dalton (Da), defined as 1/12 of the mass of one atom of carbon-12.

The dataset was generated as follows. Angiotensin 1 (1296.68 Da) was spiked at 1 fmol/μl into a background of a 50 fmol/μl tryptic digest consisting of bovine cytochrome *c*, hen egg lysozyme, bovine serum albumin, bovine apo-transferrin and Escherichia coli β-galactosidase. 1 μl of the peptide mixture was loaded on column (Pepmap C18 100 Å, 3 mm, 75 μm x 150 mm) and was separated at 250 nl/min in a 50 minute linear gradient of 10-40% buffer B (80% acetonitrile/0.05% formic acid). Eluting peptides were detected using an LTQ-FT hybrid mass spectrometer; mass spectra were acquired every second at resolution 100000. In total 12 samples were generated, consisting of six replicates for each class (spiked, unspiked).

Figure 1 (a) illustrates a sample spectrum comprising the set of intensities (relative abundance of ions) measured at different m/z values across a particular range. Such spectra are measured during the chromatographic period at discrete time points, producing what is often called a 'run'. The chromatographic period

is also referred to as elution time or retention time (RT). 2D possible visualizations of runs are given in panels (b and c) of Fig. 1, where the intensity at each RT-m/z location is represented by a gray-level. The spectrum in (a) is the one indicated by the vertical line at 1420 seconds in (b). The panels (d and e) are close-up views within the range where the spiked-in peptide is located.

Each analyte (organic specimen under study) gives rise to more than one peak in the measured spectra. Firstly, an analyte A with mass m can appear with different charges (e.g., A^{+1}, A^{+2}, A^{+3}) and because the actual quantities being measured are m/z ratios, it gives peaks at the m/z values: $m/1$, $m/2$, $m/3$. Secondly, because of the natural isotopic occurrence of the chemical elements, different peaks are generated by the same analyte A at the same charge state. For example, about 1.1% of the carbon atoms are carbon-13 isotopes instead of carbon-12. These have an extra neutron in the nucleus and therefore a mass higher by 1 Da. The spacing between the isotopic peaks depends on the charge of the ions. A group of isotopic peaks is shown in Fig. 1(d), where the distance between peaks is 1/3 Da, due to the +3 charge. The number of isotopic peaks and their relative intensity depend on the chemical composition of the analyte and therefore the distribution (pattern) of the isotopic peaks is mass dependent. Note that our algorithm detects and align isotopic peaks. In order to retrieve peptides from isotopic peaks procedures for assembling isotopes into peptides can be used, such as the one implemented in msInspect [3].

2.2 Methods

`PDA` takes a set of runs as input and outputs a set of (isotopic) peaks. The search strategy for detecting and aligning peaks acts on a three dimensional search space with dimensions given by m/z value, RT and run identifier.

`PDA` performs the following sequence of steps:

1. *Feature extraction*: Extract features from individual runs (in our experiments we use MZmine [7]).
2. *Block construction*: Split runs into blocks along the m/z axis.
3. Apply the following two steps to each block:
   (a) *Features clustering*: Generate a common set of clusters using weighted hierarchical clustering.
   (b) *Peak construction*: Generate peaks from clusters.
4. *Peak refinement*: Refine peaks located nearby splitting points of blocks.

1. Feature extraction. We use a routine from MZmine [7] called "Centroid peak detection" for extracting features from runs. A feature in a run is defined here as the (m/z, RT) pair of coordinates of a local signal intensity maximum. These coordinates are computed by first detecting local maxima in each scan (spectrum). Certain features can then be discarded if their intensity is smaller than a chromatographic threshold of the extracted chromatogram (XIC) that contains the local maxima. An XIC is constructed by summing all intensities across RT present within a m/z range (or bin).

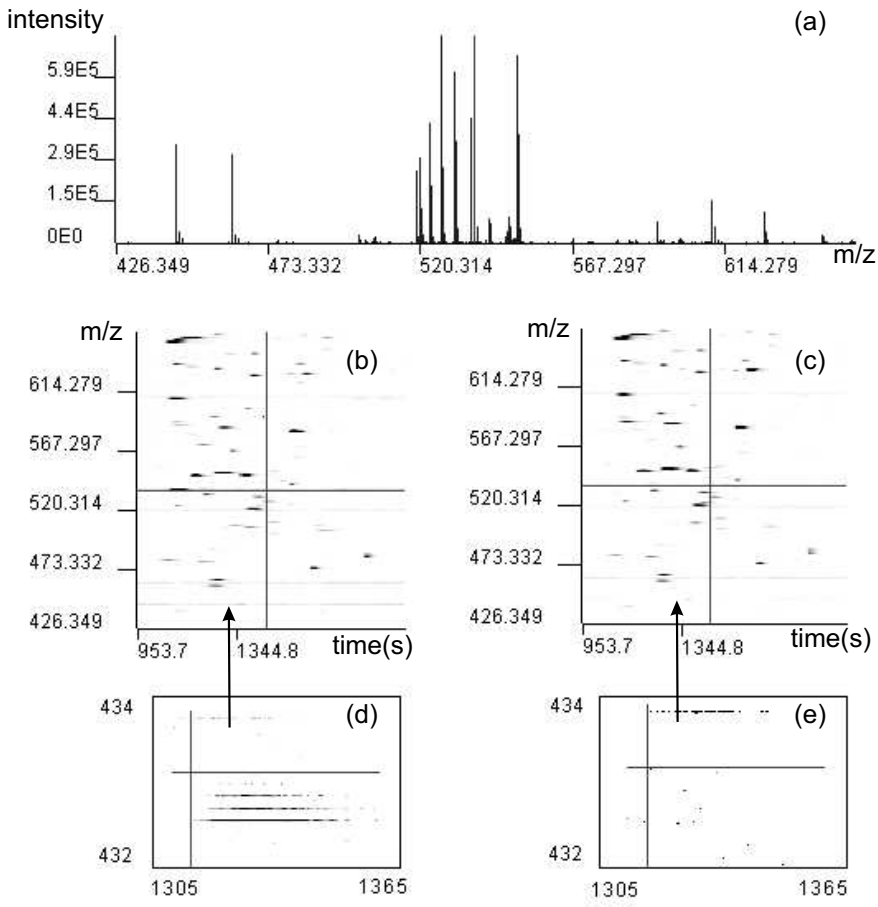

Fig. 1. (a) a sample spectrum (relative abundance of ions) measured at RT=1420s indicated by the vertical line in (b). (b and c) 2D visualizations of two runs from control and spiked-in samples. (d and e) are close-up views of the two runs within the range 1305-1365 s and 432-434 Da where the spiked-in peptide is located.

2. Block construction. To speed up the processing, runs are split into blocks along the m/z-axis. Clustering and peak extraction are then performed on each block independently, thus (possibly) processed in parallel. The peak refinement step is applied to peaks detected nearby block boarders, in order to deal with the initial splitting of true peaks.

3(a). Features clustering. This step constitutes the core of `PDA`. The particularities of the LC-MS measurements lead us to introduce a weighted distance within the clustering algorithm. The following properties of LC-MS measure-

ments were used for setting the values of weights. m/z measurements are rather reproducible across different runs. The RT of the same analyte may, however, greatly differ among distinct runs, even in the order of tens of seconds. It is therefore sensible to assign a larger weight to the m/z coordinate than to the RT, thus enforcing small m/z variations while allowing for discrepancies in time. Moreover, since our aim is also to align peaks over different runs, it is also desired to group features from different runs. To this aim we use also the run identifier as attribute for comparing features.

Our clustering operates on the set of all features, augmented with a 'run id' component. Searching for peaks in this three-dimensional space is performed by hierarchical clustering with average linkage and a novel similarity measure defined below. Since the desired number of peaks is not known a-priori, we have to estimate it. Estimation of the number of clusters can be performed by a model selection criterion, like BIC or AIC (see e.g., [6] for an experimental comparison of the effectiveness of these criteria). Here we adopt a more efficient knowledge-based approach. Starting from a high number of clusters we apply a heuristic for joining clusters which relies on domain knowledge about the minimal distance in m/z and RT dimension between two valid peaks.

Clusters are generated as follows. Firstly, in order to overcome problems related to different m/z and RT scale, the standardized scores of the values (also called z-scores) are used. Informally, for a value x (here m/z or RT value) from a set X, the corresponding z-score measures how far the observation is from the mean in units of standard deviation. The similarity measure used in `PDA` is defined as follows. For two features $F_1 = (mz1, RT1, r1)$, $F_2 = (mz2, RT2, r2)$

$$sim(F_1, F_2) = sqrt(w1 * (mz1 - mz2)^2 + w2 * (RT1 - RT2)^2) + w3 * I(r1, r2)$$

where $I(r1, r2)$ is 1 if $r1 = r2$ and zero otherwise, and $w1, w2, w3$ are weights associated to the three attributes. We elaborate on the choice of the weights values in the next section.

The dendrogram generated by hierarchical clustering is used for clustering features. A (loose) user-given upper bound M on the desired number of clusters is employed for computing the cutting threshold of the dendrogram. The output of the feature clustering step consists of M 3-dimensional clusters.

3(b). Peak construction. In this step nearby clusters are merged to form peaks by means of the following algorithm. The distance between the mean m/z values and mean RT values of each pair of centroid clusters is computed and used in an iterative procedure that generate peaks. The procedure starts from a given centroid (selected at random) as initial part of a peak. Then, centroids are incrementally added to that peak if their distance from at least one element of the actual peak is smaller than a threshold in m/z and RT. The construction of a peak terminates when no centroid can be added to it. Then the construction of a new peak can begin using the remaining centroids. The process terminates when the set of centroids becomes empty.

4. Peak refinement. The peak refinement step processes those peaks whose features result distributed in clusters belonging to neighbour blocks. Because the runs are split into independent, non-overlapping blocks, features that belong to the same true peak may lie on either side of a certain m/z cutting point. These features will be assigned to different clusters even if they are close to each other (corresponding clusters fulfil the peak construction constraints). Therefore, we introduce a peak refinement step that joins such clusters. This is accomplished by repeating the peak construction routine on the set of cluster centroids, or more efficiently, on centroids close to block boarders.

The computational complexity of `PDA` is dominated by the *Features clustering* step, that is quadratic in the number of features to be clustered. This number corresponds to the number of intensities occurring in a block. Then on the average, when blocks are processed in parallel and assuming features are uniformly distributed, `PDA` (worse case) computational complexity is $O((N/nb)^2 + n^2_{centroids})$ where N is the number o features, nb the number of blocks, and $n_{centroids}$ the number of centroids obtained after application of step 3(b).

3 Experimental Analysis

In the feature extraction step with "Centroid peak detection" (MZmine [7]), we used a fine bin size (0.02 Da) and a loose chromatographic threshold of 40%, which caused a data reduction from the original measured intensities $\sim 1.8 \times 10^7$ per run to about 4.3×10^5. We split the runs at each 5 Da and produced blocks containing a relatively small number of features, in the order of thousands. The upper bound on the number of clusters M is set to one tenth of the total number of intensities present in each block. Finally, weights of the three attributes m/z, RT, and 'run id' were set to 2, 0.01 and 0.1, respectively. These values are suitable for the type of high resolution LC-MS measurements used in this study. Slight changes of the values did not markedly affect the clustering results. As long as the the m/z weight was significantly larger than the RT and the 'run id' weights, diversity within a cluster was kept moderate and the results were consistent.

PDA results using different m/z and RT weights are illustrated in Fig. 2, while Fig. 3 shows resulting clusterings based on different values of 'the run' id weight. For the purpose of illustration, we only show results on small yet representative m/z, RT regions. Unit weights for the m/z and RT axes (regular Euclidean distance) lead to undesired split of the peak on the right in Fig. 2(a). Using the m/z axis alone ($w = (1\ 0\ 0)$) produces clusters that contain peaks of arbitrary different RTs (b). The RT axis alone is also not sufficient for accurate peak detection, since peaks are split in favor of small RT deviations (c). The output of PDA using the weights selected in our experiments is shown in (d).

The 'run id' attribute is used to discourage shifted values (in either m/z or RT) from joining a cluster if other features from the same run already match well with features from other runs. However, care should be taken about the effects on the true peaks. This is illustrated in Fig. 3 (a and b), where different

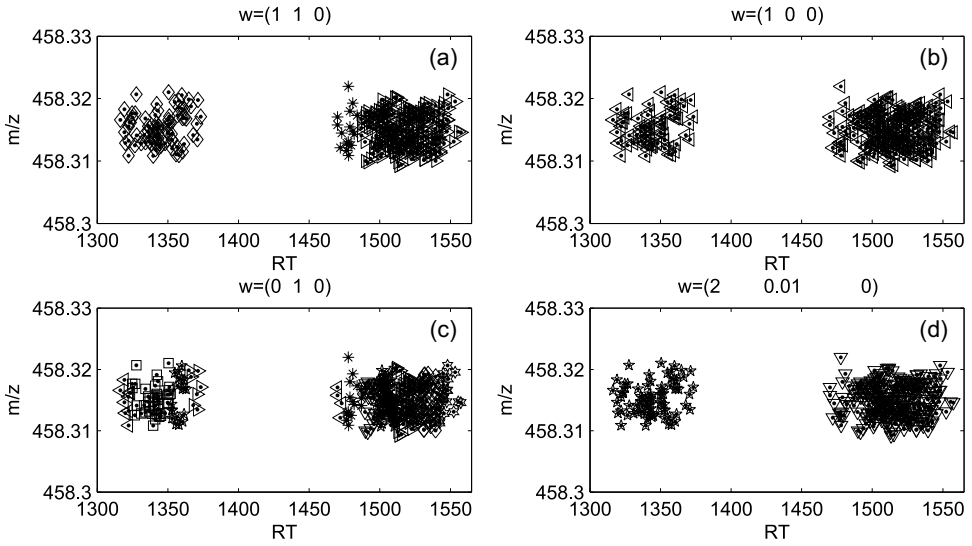

Fig. 2. The influence of the m/z and RT weights on the clustering results (see text for details). Different markers stand for different cluster membership of the features.

markers indicate that the features originate from different runs and large markers indicate cluster centers. A relatively high weight value for 'run ids' can result in clusters that contain more than one true peak. For example, in panel (b) the top two and the bottom three peaks are grouped in two clusters, while the five correct clusters should appear as in (a). Panels (c) and (d) show a zoomed view on a different region of the run. Note that the features in (c) belong to a cluster that lies beyond the shown RT limit and therefore its center is at about 1282 seconds. This cluster even comprises the peak on the left in Fig. 2 which has similar m/z values but obviously different RTs. A small 'run id' weight value helps in isolating sparse features, as desired (see (d)).

Since the feature clustering step uses an upper bound M on the number of clusters, features that belong to the same true peak can be assigned to different clusters. To fix this problem, the peak construction step 3(b) was used with thresholds for m/z and RT of 0.01 Da and 10 seconds, respectively. The m/z threshold relates to the minimum distance between two different peaks along the m/z axis, and can be calculated from the instrument set-up and the type of measurements. Two distinct isotopic peaks can be 0.01 Da apart if they originate from an analyte with charge +10, which is not practical. The RT threshold is chosen such that within the time window it defines, two different compounds with the same m/z value do not elute simultaneously. Our investigation revealed that for this dataset double peaks are observed within 30 seconds only after increasing the m/z window to 1 Da. It is therefore obvious that at 0.01 Da, all peaks are separated within a time window of 20 seconds.

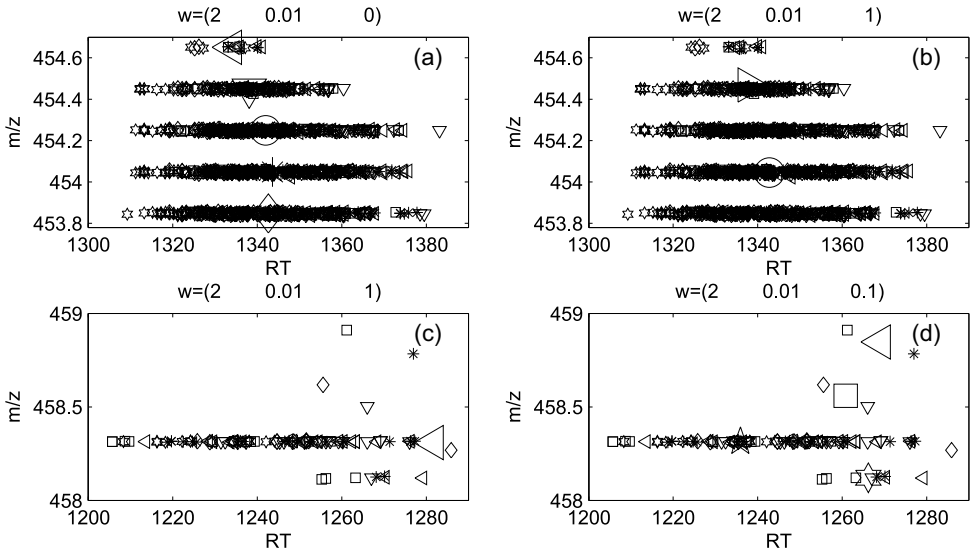

Fig. 3. The influence of the run identifier weight on the clustering results (see text for details). Different markers indicate that the features originate from different runs, large markers indicate cluster centers.

Results

The spiked-in peptide, Angiotensin 1, has theoretical m/z value 432.90032 Da and RT value of 21.75-22.76 minutes, the latter estimated from the experiment by our domain expert. The peptide gives rise to four isotopes (named here P_i, $i = 1 \ldots 4$) with decreasing intensity and duration, located at 432.90032 Da (highest intensity isotope), 433.29, 433.57, 433.91. While $P1, P_2, P_3$ were accurately detected by the `PDA`, P_4 was missed already by `PDA`'s first step (feature extraction). This means that although the local maxima detection method is simple, care should be taken in the choice of its parameters.

The results of `PDA` zoomed in a region containing the spiked-in peptide, are shown in Figure 4. The hierarchical clustering produced a large number of clusters belonging to the same isotope (a and c). Because the peak construction step operates on the cluster centroids, their distance may exceed the estimated RT threshold resulting in few peaks belonging to one isotope. More specifically, P_1 and P_2 both resulted represented by two peak clusters, with 31, 159, and 19, 117 features, respectively. P_3 was detected as single peak with 33 features. All clusters contain features from all the six spiked-in runs.

To get an estimate of the actual RT shifts across the runs we computed, for each obtained peak cluster, a value here called 'average time shift'. This value is defined as the mean of the minimum RT (pairwise) distances between features within a cluster belonging to different runs. This measure describes the average shift needed for alignment of features in a given cluster. The minimum pair-wise

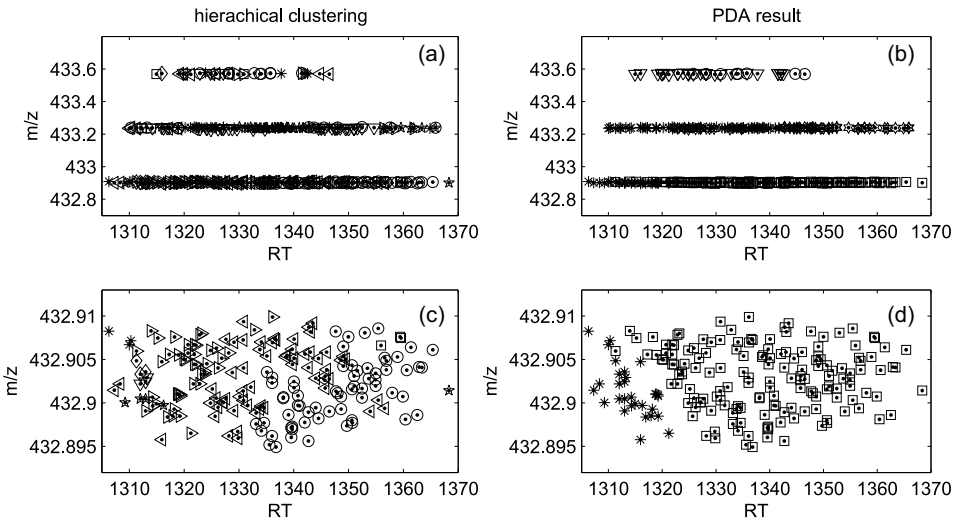

Fig. 4. Results of hierarchical clustering (a, c) and `PDA` (b, d). Occasionally, isotopic peaks can result in more than one cluster as in (d).

distance reflects the amount of time required to consider two sets of features (from two different runs) aligned. The behaviour of average time shift for some values of the RT threshold used in the peak construction step is illustrated in Figure 5. The horizontal line within each box represents the median value. The vertical limits of each box are at the lower quartile and upper quartile values. The rest of the observations extend between the lower and upper whiskers. The large amount of outliers (marked with '+') illustrates the difficulty of the alignment problem and substantiates the robustness of `PDA`. Even large drifts in RT do not obscure the `PDA`. Different RT threshold settings (between 5 and 30 seconds) do not cause significant changes in 'average time shift', which indicates the consistency of our initial assessment for a suitable threshold value. Moreover the figure indicates the good quality of the experimental data. The observation that 75% of the shift values are within a range of less than 10 seconds suggests that in most cases, local small shifts are sufficient for alignment. Thus local alignment schemes seem to be suited for this type of mass spectrometry data.

Finally, to get a flavour about the performance of other algorithms for LC-MS data alignment and peak detection, we consider two popular LC-MS data processing tools: MZmine [7] and MetAlign [2]. They both perform first peak detection followed by alignment. The nine parameters of MZmine's peak detection routine, as well as the twenty parameters of MetAlign were set by our local MS domain expert. We did not succeed in detecting the spike-in peptide using MetAlign. MZmine detected isotopes only in few of the spiked runs, represented by multiple peaks. In particular, the peak detection routine of MZmine detected isotopes in four of the six spiked-in runs, with detection of only one or two of the

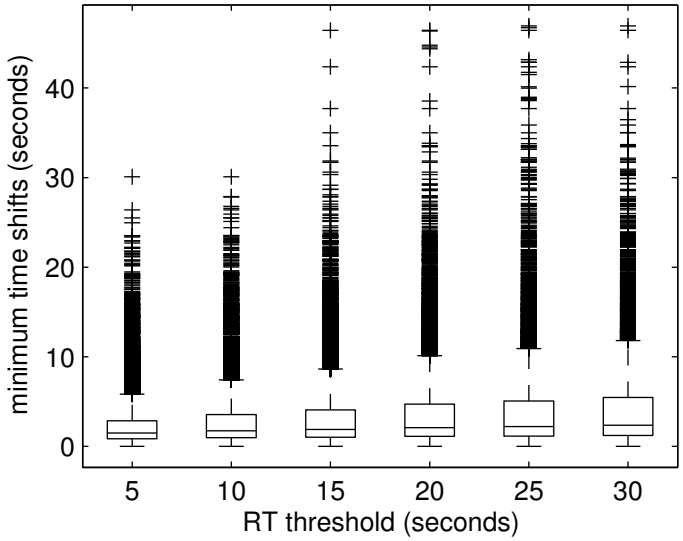

Fig. 5. The behaviour of the minimum shifts among features within each cluster. 75% of the values (within boxes) are smaller than 10 s indicating that local shifts are sufficient for alignment. Different RT thresholds (5 to 30) did not cause significant changes, as the general trend is only slightly increasing. The outliers (marked with '+') shows robustness of `PDA`.

four isotopes. This lead to alignment results where P_1 was detected four times, three times represented by one peak from one run and the other time represented by two peaks from different runs. P_2 was identified by two peaks from different runs and P_1 by a single peak from a single run.

4 Conclusion

In this paper, we proposed a method for simultaneous peak detection and alignment with nanoLCFT-MS data. Results on a spiked-in dataset indicate the effectiveness of this approach. We are currently working on further refinement steps such as: filtering out small cardinality clusters, irrespective of their locations within runs, removing clusters whose diameter in RT is too big (the elution time range of a peptide is limited), and improving the peak construction step. Moreover, we are planning to incorporate this technique in advanced biomarker detection algorithms based on feature selection. Such a comparative profiling implies the identification of clusters that contain relevant features for discrimination between control and case classes. A comparative analysis of results of the extended method on other datasets will provide more through assessment of its power as a computational tool for disease biomarker detection.

References

1. R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, March 2003.
2. A.H. America, et al. Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional lc-ms. *Proteomics*, 6(2):641–53, 2006.
3. M. Bellew, et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution lc-ms. *Bioinformatics*, 2006.
4. D. Bylund, et al. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for parafac modelling of liquid chromatography-mass spectrometry data. *J. Chromatography*, 961:237–244, 2002.
5. S.J. Callister, et al. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.*, 5:277–86, 2006.
6. X. Hu and L. Xu. Investigation on several model selection criteria for determining the number of cluster. *Neural Inform. Proces. - Lett. and Reviews*, 4(1):1–10, 2004.
7. M. Katajamaa, et al. Mzmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 2006.
8. B.M. Kohli, et al. An alternative sampling algorithm for use in liquid chromatography/tandem mass spectrometry experiments. *Rapid Commun. Mass Spectrometry*, 19(5):589–596, 2005.
9. E. Lange, et al. High accuracy peak-picking of proteomics data using wavelet techniques. In *Proc. Pacific Symposium on Biocomputing (PSB-06)*, 243–254, 2006.
10. J. Listgarten and A. Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics*, 4:419–434, 2005.
11. J. Listgarten, et al. Difference detection in lc-ms data for protein biomarker discovery. *Bioinformatics*, 2006. in print.
12. J. Listgarten, et al. Multiple alignment of continuous time series. In *Advances in Neural Information Processing Systems 2005, (NIPS 2004)*, 2005.
13. N.-P. Vest Nielsen, et al. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatography*, 805(1-2):17–35, 1998.
14. D. Radulovic, et al. Informatics platform for global proteomic profiling and biomarker discovery using liquid-chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics*, 3(10):984–97, 2004.
15. C.A. Smith, et al. Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, 78:779 – 787, 2006.
16. R. Tibshirani, et al. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, 20(17):3034–44, 2004.
17. P. Wang, et al. A statistical method for chromatographic alignment of lc-ms data. *Biostatistics*, doi:10.1093/biostatistics/kxl015, 2006.
18. W. Wang, et al. Quantification of proteins and metabolites by mass spectrometry without isotope labeling or spiked standards. *Anal. Chem.*, 75:4818–26, 2003.
19. Y. Yasui, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4(3):449–63, 2003.
20. X. Zhang, et al. Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics*, 21(21):4054–9, 2005.