J. Am. Soc. Mass Spectrom. (2012) 23:764–772
DOI: 10.1007/s13361-011-0334-2

RESEARCH ARTICLE

Retention Time Alignment of LC/MS Data by a Divide-and-Conquer Algorithm

Zhongqi Zhang

Process and Product Development, Amgen Inc, One Amgen Center Drive, Thousand Oaks, CA 91320, USA

Abstract

Liquid chromatography-mass spectrometry (LC/MS) has become the method of choice for characterizing complex mixtures. These analyses often involve quantitative comparison of components in multiple samples. To achieve automated sample comparison, the components of interest must be detected and identified, and their retention times aligned and peak areas calculated. This article describes a simple pairwise iterative retention time alignment algorithm, based on the divide-and-conquer approach, for alignment of ion features detected in LC/MS experiments. In this iterative algorithm, ion features in the sample run are first aligned with features in the reference run by applying a single constant shift of retention time. The sample chromatogram is then divided into two shorter chromatograms, which are aligned to the reference chromatogram the same way. Each shorter chromatogram is further divided into even shorter chromatograms. This process continues until each chromatogram is sufficiently narrow so that ion features within it have a similar retention time shift. In six pairwise LC/MS alignment examples containing a total of 6507 confirmed true corresponding feature pairs with retention time shifts up to five peak widths, the algorithm successfully aligned these features with an error rate of 0.2%. The alignment algorithm is demonstrated to be fast, robust, fully automatic, and superior to other algorithms. After alignment and gap-filling of detected ion features, their abundances can be tabulated for direct comparison between samples.

Key words: Iterative, Pairwise, Proteomics, Metabolomics, Multiple samples, Comparison, Quantitation, Data analysis

Introduction

Liquid chromatography-mass spectrometry (LC/MS) has been increasingly used in the analysis of complex mixtures in many fields, including proteomics [1], metabolomics [2], full characterization of therapeutic proteins [3, 4], characterization of complex cell-culture raw materials and media, and investigation of protein conformation and dynamics by hydrogen/deuterium exchange [5, 6], and other labeling techniques [7, 8], etc. Many of these applications involve comparison of multiple samples of complex mixtures. When multiple samples are quantitatively compared, the components of interest not only need to be detected and identified, their retention times also need to be aligned so that the same component in different runs will be grouped together for appropriate comparison.

Many different algorithms have been developed for retention time alignment. These algorithms have been extensively reviewed [9–14]. Most of these alignment algorithms belong to one of two general types [12]. The first type aligns retention time on the chromatography profile without peak detection, and the second type aligns ion features detected from the LC/MS data. Most of these algorithms are complicated and difficult to implement. This article describes a simple iterative algorithm, based on the divide-and-conquer technique, for alignment of ion features detected from LC/MS data. Alignment of ion features is preferred for most of our applications because it facilitates

Correspondence to: Zhongqi Zhang; *e-mail:* zzhang@amgen.com

Received: 9 September 2011
Revised: 20 December 2011
Accepted: 30 December 2011
Published online: 2 February 2012

output of a tabulated list of features for direct comparison between samples, which is usually the final goal of LC/MS analyses. The alignment algorithm is shown to be fast, robust, fully automatic, and superior to available algorithms. Because the algorithm can be written into a simple recursive function, it is easy to implement.

Method

Retention time shifts frequently occur in LC/MS runs. In order to compare corresponding components in different runs quantitatively, corresponding ion features in each sample must be time aligned. The following describes a simple iterative algorithm for aligning corresponding ion features in LC/MS runs. The algorithm is based on the divide-and-conquer approach, which is a commonly used computing technique that recursively subdivides a complex problem into simpler problems.

The Divide-and-Conquer Algorithm for Retention Time Alignment

The alignment algorithm described here applies on ion features (monoisotopic *m/z*, charge, retention time, peak width, and peak height) extracted from each LC/MS raw data file. In the first step, all ion features above a certain user-defined signal-to-noise ratio (S/N) threshold are detected from each LC/MS run. Details of the feature detection process have been described previously [15]. Briefly, nearby full MS scans in the LC/MS run are averaged by applying a moving Gaussian filter function to improve the S/N of each scan. After averaging, MS ion detection is performed on each scan. For low-resolution data, ion detection is performed using an algorithm similar to previously described [16]. For high-resolution data where all isotopic peaks are resolved, ion detection is achieved by examining the isotopic pattern of each ion; a successful determination of the charge state of the ion (based on the isotopic pattern) indicates a positive ion detection. After ion detection is completed for all the scans, the selected-ion chromatogram (SIC) is constructed for each detected ion. A detectable chromatographic peak in the SIC indicates a positive detection of a sample ion feature. For every detected ion feature, its retention time, monoisotopic *m/z*, charge, peak width, and peak height are calculated and recorded. An ion feature is selected as a retention time anchor feature if its monoisotopic peak is visible, it is well resolved from neighboring peaks in its SIC, and its chromatographic peak width is not significantly wider than the width of a typical peak in the run.

After the list of anchor features is obtained for each sample, these features are then aligned using the algorithm described below. The purpose of the retention time alignment routine is to identify corresponding ion features in different LC/MS runs and adjust their retention times to match the retention times in a reference run. A reference run is a representative run from the runs analyzed together.

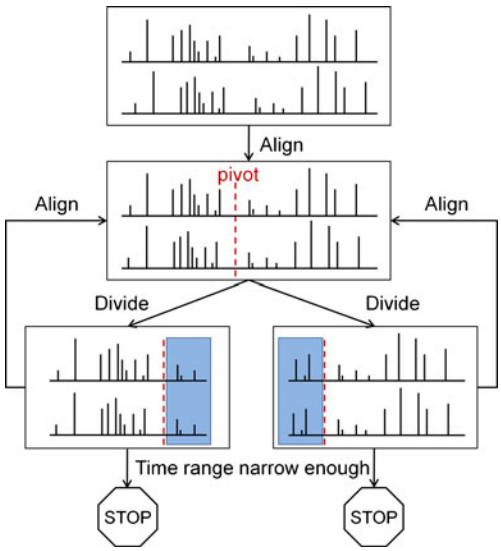

Figure 1. Flowchart describing the iterative algorithm for retention time alignment of detected ion features in two LC/MS runs. The shaded are the extended region for proper alignment of features near the pivot

Figure 1 shows the flowchart of the algorithm. In this iterative algorithm, anchor features in the sample run are first aligned with anchor features in the reference run by applying a single constant shift of retention time, i.e., a constant value is subtracted from the retention times of all features in the sample run. This constant time shift is obtained by optimizing the similarity of the two chromatograms after applying a retention time shift as described below. After this initial crude alignment, the sample chromatogram is then divided near the middle (the pivot) to create two shorter chromatograms with different time ranges. The two shorter chromatograms are then aligned to the corresponding time range of the reference chromatogram the same way as the entire chromatogram. Each shorter chromatogram is further divided into two even shorter chromatograms and aligned with the reference chromatogram. This process continues until there are no more than 10 anchor features left in each time range, or each time range is no wider than 1/4 of a typical peak width, in which case the alignment is complete. Because this routine is an iterative process that can be achieved by a simple recursive function, the algorithm is easy to implement.

The pivot used to divide each chromatogram is usually set at the most feature-free time point near the middle of the chromatogram. Abnormalities may occur near the pivot. For example, when the pivot divides two isomers into separate chromatograms, the two isomers may align to the same isomer in the reference run. To avoid abnormalities near the pivot, each shorter chromatogram is extended on both sides by including ion features within the maximum allowable retention time shift for that iteration. The derived time shift, however, is not applied to the features in the extended time range.

This alignment algorithm effectively divides a chromatogram into hundreds of small time windows, within which all

ion features have a constant retention time shift. This approach avoids the use of any linear or nonlinear functions, which are frequently shown not ideal for many situations [17]. The divide-and-conquer algorithm shown here applies to retention time shift profiles of any shapes.

Implementation

In practice, each detected ion feature in a sample run is assigned a retention time shift by aligning it to the reference run. Let Δt_0 be the retention time shift determined from the previous iteration, and Δt be the retention time shift to be determined for the current iteration, the score function used to evaluate the match between the two chromatograms after retention time adjustment is shown in equation 1. When two chromatograms are compared, the retention time shift (Δt) that has the maximum values of s is determined to be the determined shift for that part of the chromatogram. Because the chromatogram has already been adjusted by Δt_0 in the previous iteration, all ion features within this part of the chromatogram are assigned to have a retention time shift of $\Delta t_0+\Delta t$, and this $\Delta t_0+\Delta t$ value will be used as the Δt_0 value in the next iteration.

For any anchor feature i in run S (sample run) and anchor feature j in run R (reference run) with the same determined monoisotopic m/z and charge, the score function to evaluate the similarity of the two chromatograms s is defined as

$$s = \frac{\sum_{i,j}\left\{\exp\left[-11.1(t_i^S - \Delta t_0 - \Delta t - t_j^R)^2 \Big/ \left(w_i^S w_j^R\right)\right]\left(I_i^S I_j^R\right)^{\frac{1}{2}}\right\}}{\left(\sum_i I_i^S \sum_j I_j^R\right)^{\frac{1}{2}}} \quad (1)$$

where w_i^S and w_j^R stand for the chromatographic peak widths (in SIC) for features i and j, respectively. I represents intensities (peak heights in SIC) for features i and j. Intensities are considered in the score function to ensure a match in the general pattern of the two chromatograms. The exponential term represents a Gaussian function which reaches maximum value of 1 when all anchor features in S are adjusted to have the same retention time as their counterparts in R ($t_i^S - \Delta t_0 - \Delta t - t_j^R = 0$). In this case, s becomes the similarity score between run S and R, similar to the spectral similarity score described previously [18, 19]. The factor of –11.1 inside the exponential term is used so that the Gaussian function has a value of 0.5 when the adjusted retention time differs from the reference retention time by 1/4 of the peak width. The tolerance for determining whether two monoisotopic m/z values are equal is instrument and ion dependent. For high-resolution Orbitrap data, the tolerance is determined to be 1/4 of the theoretical MS peak width at half height, plus 0.02/charge for large peptides (mass>500 u) or plus 2 ppm of the m/z for small molecules (mass≤500 u). For low-resolution data acquired on Thermo Scientific (San Jose, CA) LTQ, the m/z tolerance is set at 0.2 for singly charged ions and 0.5/charge for multiply charged ions.

Ideally, the pivot (the time point used to divide a chromatogram) is selected so that it is not in a feature-crowded region, and the resulting two shorter chromatograms have similar time ranges and contain similar number of ion features. Assuming point 1 divides the chromatogram into two equal time ranges, and point 2 divides the chromatogram into two sections containing equal number of ion features, the pivot is selected within the time range of point 1 and point 2 to have the largest time gap between its two adjacent features.

Although the extended time ranges beyond the pivot (Figure 1) helps resolving the abnormalities near the pivot, to further reduced errors caused by abnormalities near the ends of the extended regions, a weight function described in equation 2 is applied on the intensity of ion features in the extended regions so that the features near the ends of the extended time ranges will have smaller impact to the score function.

$$I_{weighted} = I \exp\left[\frac{-2\left(t - t_{pivot}\right)^2}{\Delta t_{\max}{}^2}\right] \quad (2)$$

where I and t are the intensity and retention time of each feature in the extended region, t_{pivot} is the time of the pivot (each end of the normal chromatogram), and Δt_{max} is the maximum allowed retention time shift for the current iteration. $I_{weighted}$ calculated from equation 2 will be used as the intensity values for both R and S in equation 1 when calculating the score function.

The maximum retention time shift for each iteration Δt_{max} is determined as either the user-defined overall maximum retention time shift $\Delta t^0_{\max}$, or 1/3 of the current time window to be aligned, whichever is smaller. Besides the typical chromatographic peak width, which can usually be determined automatically from the raw data, this maximum retention time shift $\Delta t^0_{\max}$ is the only other user-defined parameter required by the described alignment algorithm. However, due to the robustness of the algorithm, the default value, which is set at 1/3 of the time range of the entire LC/MS run, works virtually all the time.

The score function shown in equation 1 is evaluated for different time shifts Δt within the range of $-\Delta t_{\max}$ to $\Delta t_{\max}$. Typically, 50 Δt values are evaluated. Because the optimized Δt is more likely to have a value close to zero, especially when the value of $\Delta t_{\max}$ is large, the Δt values to be evaluated are selected in square scale, for example, 0, 0.01, 0.04, 0.09, 0.16 min, etc.

After retention time alignment, each detected ion feature in the sample run will have a determined retention time shift value. After applying retention time adjustment, corresponding ion features among different runs are then well aligned and can be conveniently grouped together. Because some features detected in one run are not detected in others, a second round of feature detection (gap-filling) is performed to detect any features that are missed in the first round. To perform gap-filling of a feature in run A (any sample or reference run), the

expected retention time of the feature in run A is first calculated based on the average corrected retention times of the feature detected from other runs and the retention time shifts of the nearby features in run A. The SIC of the feature is then constructed near the retention time of interest, followed by peak detection. If a peak is detected in the SIC with retention time differing from the expected retention time by less than half the peak width, then the feature is considered detected in run A. If no peak is detected in this time range, a wider time range is searched and the retention time tolerance is increased to the value determined by equation 3.

$$tol_{gapfilling} = \left(\frac{1}{2} + \ln\frac{\max|\Delta t|}{w}\right)w, \ when \ \max|\Delta t| > w \quad (3)$$

where w is the typical peak width, and $\max|\Delta t|$ is the maximum absolute values of determined retention time shifts for all runs analyzed together. When $\max|\Delta t| \leq w$, $tol_{gapfilling}$ takes the value of $1/2w$. If no peak is detected within this increased time range, the ion feature is considered missing in run A and a peak area of zero is assigned for the feature in run A. After all corresponding features are grouped accordingly, their peak areas can be conveniently tabulated for direct comparison between samples.

The above divide-and-conquer algorithm for retention time alignment is implemented in the custom-built program MassAnalyzer [15] for automated analysis of Thermo Scientific XCalibur LC/MS data. An LC/MS run with the most representative retention times and also with a large number of anchor features is selected automatically by MassAnalyzer as the reference run. For a test of MassAnalyzer for noncommercial research purpose, please contact the author directly.

Experimental

For testing the retention time alignment algorithm for small-molecule mixtures, 25 different lots of soy hydrolysate (DMV International) were analyzed on an Agilent (Santa Clara, CA) 1200 SL system connected to a Thermo Scientific LTQ-Orbitrap high-resolution mass spectrometer. A Waters (Milford, MA) BEH 300 C18 reversed-phase column (1.7 µ particle, 150×2.1 mm) was used for the separation, followed by electrospray ionization. Components were eluted with an acetonitrile gradient (0.5% to 20% in 30 min after the initial condition of 0.5% acetonitrile for 10 min, followed by column washing with 20%–90% acetonitrile) at a flow rate of 0.3 mL/min, with 0.04% trifluoroacetic acid (TFA) in the mobile phase. About 20 µg of sample was injected for each analysis. The MS method was set up to collect one full scan (*m/z* 100–1500) in the high-resolution Orbitrap (resolution=60,000 at *m/z* 400) in centroid mode, followed by two data-dependent collision-induced dissociation (CID) MS/MS scans in the linear trap of the top two most abundant ions, with dynamic exclusion duration of 12 s. Among these runs, five runs with large retention time shifts were selected to demonstrate the alignment algorithm.

For testing the alignment algorithm for protein digests, seven different lots of an IgG2 antibody (Amgen, Thousand Oaks, CA, USA) were digested with trypsin at 37 °C for 2 h, after reduction and alkylation with iodoacetamide, using a procedure similar to the method described previously [20]. To create differences in the samples, three lots were treated with Peptide-*N*-Glycosidase F (QA-Bio, Palm Desert, CA, USA) to remove the glycans in the IgG2 antibody. The tryptic digests were analyzed on an Agilent 1200 SL system connected to a Thermo Scientific LTQ-Orbitrap mass spectrometer. An Agilent 1.8 µm particle rapid-resolution reversed-phase column (SB C18, 2.1×150 mm) was used for the analysis. Peptides were eluted with a gradient of 1% to 20% acetonitrile in 38 min, followed by 20%–40% acetonitrile in 60 min, with 0.02% TFA in each mobile phase, at a flow rate of 0.2 mL/min. The mass spectrometer was set up to acquire one high-resolution full scan in centroid mode at 60,000 resolution (at *m/z* 400), followed by three concurrent data-dependent CID (normalized collision energy 35%) MS/MS scans of the top three most abundant ions, with dynamic exclusion duration set at 10 s. About 30 µg of each digest was injected into the LC/MS/MS system for analysis. Among these runs, two runs with largest retention time shifts with and without deglycosylation were selected to test the alignment algorithm.

To test the performance of the algorithm on low-resolution data, a non-reduced Lys-C digestion was performed on another IgG2 antibody, followed by a post-digestion disulfide reduction. The digestion and reduction procedure was performed automatically on an Agilent 1100 autosampler, with details described elsewhere [21]. LC/MS/MS data were acquired on the Agilent 1100 HPLC system connected to a Thermo-Scientific LTQ-XL mass spectrometer equipped with an electrospray ionization source. A Waters BEH 300 C4 column (2.1×150 mm, 1.7 µ particle) was used for the analysis. Proteolytic peptides were eluted at 0.2 mL/min with a gradient of 0.5%–22% acetonitrile in 40 min, followed by 22%–42% acetonitrile in 80 min, then followed by column washing and equilibration, with 0.1% TFA in each mobile phase. MS was set up to acquire one full scan in centroid mode, followed by a data-dependent ultra-zoom scan and a CID MS/MS of the most abundant ion, with dynamic exclusion duration of 5 s. Each of the non-reduced and reduced samples was injected five times, and the two runs with most severe retention time shifts were selected for the test of the alignment algorithm.

Data Analysis

Data analysis was performed on MassAnalyzer, a custom-built program for automated detection, identification, retention time alignment, and peak area calculation of ion features detected in LC/MS experiments [15]. User-adjustable parameters for ion feature detection and alignment included the S/N threshold, typical chromatographic peak width, and maximum retention time shifts. Typical chromatographic peak width was determined automatically by MassAnalyzer from the raw data, and the default value of 1/3 of the

chromatographic time range was used as the maximum retention time shifts for all analyses. As a result, the user only needed to specify the minimum S/N of ion features to be detected. An S/N threshold of 10 was used for all LC/MS analysis described in this work. After alignment of all detected features, the detected features in all LC/MS runs were combined into a list, and a second round of feature detection (gap-filling) was performed to detect all features in the list that were not detected in the first round. Finally, MassAnalyzer output a list of ion features, including their retention time, masses, retention time shifts, and peak areas etc., with corresponding features tabulated in the same row.

To evaluate the accuracy of alignment algorithms, a set of "ground truth" of corresponding ion features was established for each set of data. To establish the ground truth, MassAnalyzer first performed ion feature detection with S/N threshold of 10. The MS/MS of ions with the same m/z and charge states were compared. If the similarity score [18] between the two MS/MS was greater than 0.9 for soy hydrolysate samples or 0.7 for protein digests, a corresponding ion feature pair was considered detected. When ambiguities were present, corresponding ion features were manually selected by examining the elution profile of nearby features. The MS/MS information, however, was not used in any of the tested alignment algorithms.

For comparing the described algorithm to other retention time alignment algorithms, the data sets described here were also processed by the Join aligner [22, 23] and the RANSAC aligner in MZmine 2 [24]. The two alignment algorithms were considered among the best available alignment algorithms [10, 25]. Ion features were first detected by MZmine 2 with a minimal signal of 10^4 for the Orbitrap data and 10^3 for the LTQ data, followed by retention time alignment. After alignment with either the Join aligner or the RANSAC aligner, gap-filling was performed using the "peak finder" routine with retention time correction and a retention time tolerance of 0.4 min.

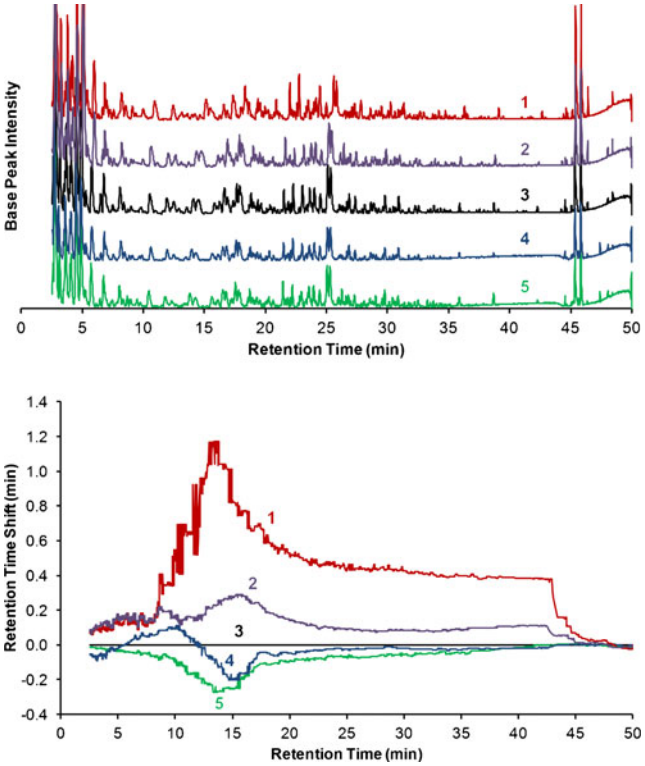

Figure 2. Determined retention time shifts for five LC/MS runs of different lots of soy hydrolysate. The top panel shows the base-peak chromatograms of the five runs before alignment; the bottom panel shows the determined retention time shifts as compared to the reference run-3. Default parameters were used for determining the retention time shifts (typical peak width 0.27 min, maximum retention time shift 15.8 min)

Results

Soy Hydrolysate on High Resolution MS

The retention time alignment algorithm is demonstrated by applying the algorithm to reversed-phase LC/MS analyses of five lots of soy hydrolysate. Soy hydrolysate contains a large number of components, including amino acids, short peptides, carbohydrates, vitamins, and other nutrients, and is widely used as a source of nutrients for cell culture media. The composition of soy hydrolysate is similar to the samples often encountered in metabolomics experiments. Table 1 shows the number of detected ion features and alignment ground truths in these LC/MS runs. Run-3 is selected by MassAnalyzer as the reference run. Figure 2 shows the determined retention time shifts of detected ion features, as related to their retention times, in the five runs. The retention time shift profiles shown in Figure 2 are clearly nonlinear. Because the divide-and-conquer algorithm divides the chromatogram into hundreds of narrow time windows, each having different retention time shifts, it is able to align retention time shift of any profile. For example, when aligning run-1 against run-3, after recursion depths of 7 to 10, the divide-and-conquer process divided the chromatogram into 220 narrow windows, each having a time range from 0.01 to 1.5 min and containing 11 to 196 ion features. A total of 425 alignment steps were performed, with determined retention time adjustment for each step between –0.37 to 0.40 min.

Table 1. Number of Detected Ion Features and Alignment Ground Truths in LC/MS Runs of Soy Hydrolysate

	Run-1	Run-2	Run-3 (reference)	Run-4	Run-5
Features with S/N>10 (detected in the first round)	4210	3295	3350	2861	3617
Number of anchor features	3899	3108	3029	2568	3269
Total features detected (after gap-filling)	5093	5093	5093	5093	5093
Ground truth for alignment	1354	1391	N/A	1335	1328

Therefore, the maximum retention time shift (user-defined parameter) would work for any value in the range of 0.40 min to 15.8 min, although 15.8 min was used in this case, demonstrating the robustness of the algorithm.

To evaluate the accuracy of the determined retention time shifts, the determined retention time shift profile of all detected ion features in soy hydrolysate run-1 (with the most severe retention time shift among the four runs) is compared with the 1354 true retention time shifts determined from the ground truth (Figure 3). It is seen that the determined retention time shift profile accurately reflects the true profile of retention time shift in the run.

Protein Digest on High Resolution MS

To evaluate the algorithm for aligning larger ions such as seen in a protein digest, a tryptic digest of an IgG2 antibody, with and without deglycosylation, were analyzed by LC/MS on an Orbitrap high-resolution mass spectrometer. The data were analyzed by MassAnalyzer for ion feature detection and alignment. The digest without deglycosylation was automatically selected as the reference run. Figure 4 shows the determined retention time shifts of 2125 detected ion features in the deglycosylated run compared with the true retention time shift determined from the 1008 ground truth ion features. It can be seen again that the determined time shift profile matches the true retention shift profile for the protein digest.

Protein Digest on Low Resolution MS

To demonstrate the capability of the algorithm to process low-resolution data as well as samples with larger differences, the non-reduced and reduced Lys-C digest of an IgG2 antibody was analyzed by LC/MS on a low-resolution Thermo Scientific LTQ mass spectrometer. The data were processed with MassAnalyzer for ion feature detection and alignment. The digest with disulfide reduced was automatically selected as the reference run. Figure 5 shows the determined retention time shifts of 1016 detected ion features in the non-reduced run

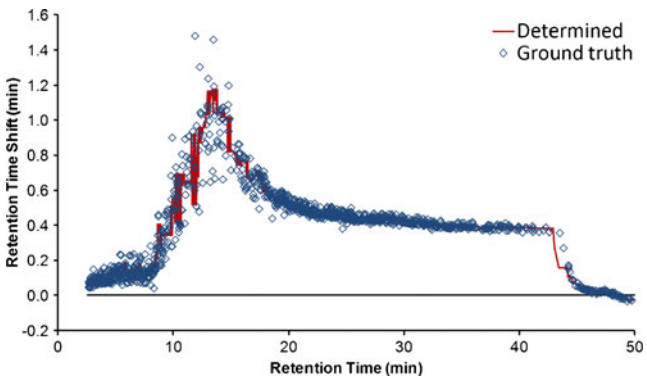

Figure 3. Determined retention time shifts (red line) of soy hydrolysate run-1 compared with the true retention time shift (blue diamond) determined from the ground truth

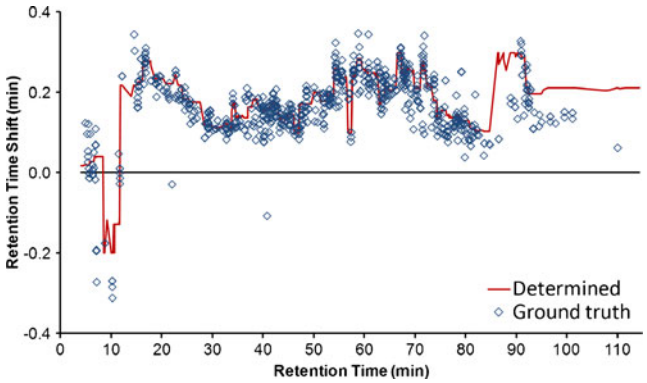

Figure 4. Determined retention time shifts (red line) of an IgG2 tryptic digest compared with the true retention time shift (blue diamond). Maximum retention time shift for alignment is set to the default value of 37.0 min

compared with the true retention time shift determined from the 101 ground truth ion features. It can be seen that the determined time shift profile matches the true retention shift profile for the protein digest.

Quantitative Assessment of the Alignment Accuracy

The accuracy of alignment algorithms can be quantitatively evaluated from their precision and recall values [10], calculated based on the comparison between the alignment results and the ground truth. Precision and recall values are calculated based on the following definitions.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

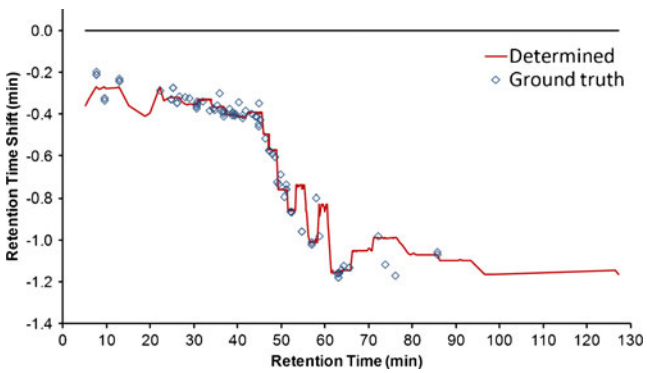

Figure 5. Determined retention time shifts (red line) of a non-reduced IgG2 Lys-C digest against a reduced digest compared with the true retention time shift (blue diamond). Maximum retention time shift for alignment is set to the default value of 48.6 min

Table 2. Performance of Alignment Algorithms as Determined from Comparison to the Ground Truth (GT)

Data set	Performance measure	MZmine 2 (Join)	MZmine 2 (RANSAC)	MassAnalyzer
Soy hydrolysate (Orbitrap)	Total GT detected	592	626	1351
Sample 1 vs. 3	Precision	0.921	0.901	0.996
Total GT=1354	Recall	1.000	0.998	0.996
Soy hydrolysate (Orbitrap)	Total GT detected	527	534	1389
Sample 2 vs. 3	Precision	0.947	0.964	1.000
Total GT=1391	Recall	1.000	1.000	1.000
Soy hydrolysate (Orbitrap)	Total GT detected	498	508	1334
Sample 4 vs. 3	Precision	0.942	0.963	1.000
Total GT=1335	Recall	1.000	1.000	1.000
Soy hydrolysate (Orbitrap)	Total GT detected	539	550	1326
Sample 5 vs. 3	Precision	0.941	0.962	1.000
Total GT=1328	Recall	1.000	1.000	1.000
Antibody digests (Orbitrap)	Total GT detected	871	932	1007
trypsin	Precision	0.745	0.704	0.999
Total GT=1008	Recall	1.000	0.991	0.999
Antibody Lys-C digest (LTQ)	Total GT detected	27	27	100
Disulfide intact vs. reduced	Precision	0.926	0.926	1.000
Total GT=101	Recall	1.000	1.000	1.000
Overall	Total GT detected	3054	3177	6507
Total GT=6517	Precision	0.882	0.875	0.999
	Recall	1.000	0.997	0.999

where *TP*, *FP* and *FN* represent true positive (correct alignment), false positive (aligned to wrong features) and false negative (no corresponding ion feature detected), respectively.

A C++ routine was written to compare the alignment result against the ground truth result to calculate the precision and recall values for each data set, and the results are presented in Table 2. Also shown in Table 2 are precision and recall results when ion feature detection and alignment was performed on MZmine 2 with the Join aligner and RANSAC aligner separately. The Join aligner in MZmine had been previously evaluated to be among the top two best algorithms for metabolomics data [10], and the RANSAC aligner in MZmine 2 [24] had recently been evaluated to be among the top two best overall alignment algorithms [25]. It is clear from Table 2 that MassAnalyzer not only detected twice as many ion features in the ground truth, the precision values are significantly better than both algorithms. For the datasets presented here, the overall error rate (FP and FN) for MassAnalyzer is about 0.2%, compared with ~12% for both alignment algorithms in MZmine 2. The precision and recall values shown in Table 2 for the alignment algorithms employed in MZmine 2 are also similar to the values demonstrated previously for other metabolomics and proteomics data [24]. It is worthwhile to point out that the alignment accuracy represented by precision and recall values is not only affected by the alignment algorithm, but a combined effect of ion feature detection, retention time alignment and gap-filling.

Discussion

The robustness of the iterative alignment algorithm based on divide-and-conquer technique is demonstrated by its remarkable accuracy with very limited user interactions. Table 3 shows some key parameters used by MZmine 2 and MassAnalyzer; these parameters are typically required by most ion feature detection and alignment algorithms. MassAnalyzer, however, determines most of these parameters directly from the raw data. Except for the minimum S/N of ion features to be detected, the only other parameter not determined from the data is the maximum retention time shift, of which a very large default value of 1/3 of the

Table 3. Key Parameters Used for Data Processing by MZmine 2 and MassAnalyzer

Parameters	MZmine 2	MassAnalyzer
For all procedures		
m/z Tolerance	0.01 for Orbitrap, 0.4 for LTQ	Automatically determined based on instrument type and resolution
Feature detection		
Peak width	Minimum 0.1 min	Typical peak width automatically determined from data
Filtering	5-Point Savitzky-Golay	Gaussian with width=1/3 of typical peak width
Signal intensity	$>1\times10^4$ for Orbitrap, $>1\times10^3$ for LTQ	S/N>10 ($>8\times10^4$ for Orbitrap, $>5\times10^3$ for LTQ)
Retention time alignment		
Maximum retention time shift	2 min (both Join and RANSAC)	Default value=1/3 of the chromatography range
Gap-filling		
Retention time tolerance	0.4 min	Automatically set based on maximum determined retention time shift (equation 3)

chromatogram time window is usually applicable and is used throughout the work.

A key reason for the remarkable accuracy and robustness of the algorithm is the lack of error in the early stage of alignment. Because the algorithm is an iterative process, it is reasonable to believe that errors introduced early in the process may propagate to later iterations and cause significant errors in the final results. However, because a large number of ion features are involved when evaluating the score function (equation 1) during earlier iterations, the contribution to the score function of a few misaligned features is minimal. As a result, the chance of incorrect alignment during early iterations is virtually zero.

Incorrect alignment does occur in later stages, when the alignment window becomes very narrow and only a small number of ion features are involved. When components with the same *m/z* and charge (mostly isomers) are presents in this small window, there is a chance that some isomers will be aligned to incorrect isomers. Because a large number of alignment steps have already been performed on this time window in previous iterations, the retention time adjustment during these later steps should be very close to zero. Therefore, incorrect alignment in this stage is controlled by limiting the maximum allowable retention time shift to less than 1/3 of the alignment window, which reduced the chance of misalignment in the later alignment iterations.

Presence of isomers indeed presents a challenge to the alignment algorithm. When isomers are present in a small time window, their SIC will present multiple peaks, which can be misaligned to one another. MassAnalyzer deals with this problem by applying a similarity measure in intensity profiles as described in equation 1. For example, if a component has three isomers in both the control and the case sample, according to our experience, the three isomers in the case sample are most likely the same three isomers in the control sample. A misalignment of the three isomers will generate a total of four or more isomers, which will be penalized by the score function shown in Equation 1. If intensity is not considered in the score function, an isomer can sometimes be misaligned to the wrong isomer with retention time closest to it. In rare cases when there are really four or more isomers, other ion features within this window can usually provide sufficient contribution to the score function to avoid the misalignment.

The algorithm is relatively fast. A typical high-resolution LC/MS run contains 1000 to 10,000 anchor features. On a desktop computer, it usually takes less than 10 s (not counting ion feature detection and gap-filling) to align two runs. The algorithm is also easy to implement due to its recursive nature. For example, in the C++ implementation in MassAnalyzer, the alignment part (not counting ion feature detection and gap-filling) takes ~100 lines of codes.

Limitations of the algorithm include the requirement of sufficient similarity between the samples to be aligned. Secondly, like most alignment algorithms, the accuracy of alignment decreases with more severe retention time shift. The algorithm assumes that components with the same retention time will have the same retention time shift, which may not be always true, especially when the samples have very different matrix. When the true retention time shift of a component differs from the determined retention time shift by more than the amount described in equation 3, a false-positive or false-negative will be generated, depending on whether an isomer is present nearby. Lastly, as described earlier, closely eluting isomers present a challenge to the algorithm.

Conclusions

A simple retention time alignment algorithm is developed. The algorithm is easy to implement, robust, fully automated, and demonstrated superior to currently available algorithms for the type of data frequently encountered in biopharmaceutical industry. The level of accuracy of the described algorithm is important for characterization of biopharmaceuticals, when errors of alignment are not tolerated. The algorithm has been used routinely in our laboratories for various LC/MS applications, including identification and quantification of amino acid substitutions and modifications in therapeutic proteins, characterization of complex cell-culture raw materials and media, as well as metabolomics and some proteomics experiments.

Acknowledgment

The author thanks Jason Richardson and Bhavana Shah for collecting some of the data to test the algorithm, and Pavel Bondarenko for helpful discussions during the development of the algorithm. This work was funded by Amgen Inc.

References

1. Yates, J.R., Ruse, C.I., Nakorchevsky, A.: Proteomics by mass spectrometry: Approaches, advances, and applications. *Annu. Rev. Biomed. Eng.* **11**, 49–79 (2009)
2. Bedair, M., Sumner, L.W.: Current and emerging mass-spectrometry technologies for metabolomics. *TrAC Trends Anal. Chem.* **27**, 238–250 (2008)
3. Srebalus Barnes, C.A., Lim, A.: Applications of mass spectrometry for the structural characterization of recombinant protein pharmaceuticals. *Mass Spectrom. Rev.* **26**, 370–388 (2007)
4. Zhang, Z., Pan, H., Chen, X.: Mass spectrometry for structural characterization of therapeutic antibodies. *Mass Spectrom. Rev.* **28**, 147–176 (2009)
5. Zhang, Z., Smith, D.L.: Determination of amide hydrogen exchange by mass spectrometry: A new tool for protein structure elucidation. *Prot. Sci.* **2**, 522–531 (1993)
6. Engen, J.R.: Analysis of protein conformation and dynamics by hydrogen/deuterium exchange MS. *Anal. Chem.* **81**, 7870–7875 (2009)
7. Xu, G., Chance, M.R.: Hydroxyl radical-mediated modification of proteins as probes for structural proteomics. *Chem. Rev.* **107**, 3514–3543 (2007)
8. Novak, P., Giannakopulos, A.E.: Chemical cross-linking and mass spectrometry as structure determination tools. *Eur. J. Mass Spectrom.* **13**, 105–113 (2007)
9. Katajamaa, M., Orešič, M.: Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A* **1158**, 318–328 (2007)
10. Lange, E., Tautenhahn, R., Neumann, S., Gröpl, C.: Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinforma.* **9**, 375 (2008)

11. America, A.H.P., Cordewener, J.H.G.: Comparative LC-MS: A landscape of peaks and valleys. *Proteomics* **8**, 731–749 (2008)
12. Vandenbogaert, M., Li Thiao Té, S., Kaltenbach, H.M., Zhang, R., Aittokallio, T., Schwikowski, B.: Alignment of LC MS images, with applications to biomarker discovery and protein identification. *Proteomics* **8**, 650–672 (2008)
13. Åberg, K.M., Alm, E., Torgrip, R.J.O.: The correspondence problem for metabonomics datasets. *Anal. Bioanal. Chem.* **394**, 151–162 (2009)
14. Boccard, J., Veuthey, J.-L., Rudaz, S.: Knowledge discovery in metabolomics: An overview of ms data handling. *J. Sep. Sci.* **33**, 290–304 (2010)
15. Zhang, Z.: Large-scale identification and quantification of covalent modifications in therapeutic proteins. *Anal. Chem.* **81**, 8354–8364 (2009)
16. Zhang, Z., Marshall, A.G.: A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* **9**, 225–233 (1998)
17. Podwojski, K., Fritsch, A., Chamrad, D.C., Paul, W., Sitek, B., Stuhler, K., Mutzel, P., Stephan, C., Meyer, H.E., Urfer, W., Ickstadt, K., Rahnenfuehrer, J.: Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* **25**, 758–764 (2009)
18. Zhang, Z.: Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76**, 3908–3922 (2004)
19. Zhang, Z.: Prediction of electron-transfer/capture dissociation spectra of peptides. *Anal. Chem.* **82**, 1990–2005 (2010)
20. Ren, D., Pipes, G.D., Liu, D., Shih, L.-Y., Nichols, A.C., Treuheit, M. J., Brems, D.N., Bondarenko, P.V.: An improved trypsin digestion method minimizes digestion-induced modifications on proteins. *Anal. Biochem.* **392**, 12–21 (2009)
21. Richardson, J., Shah, B., Xiao, G., Bondarenko, P.V., Zhang, Z.: Automated in-solution protein digestion using a commonly available high-performance liquid chromatography autosampler. *Anal. Biochem.* **411**, 284–291 (2011)
22. Katajamaa, M., Orešič, M.: Processing methods for differential analysis of LC/MS profile data. *BMC Bioinforma.* **6**, 179 (2005)
23. Katajamaa, M., Miettinen, J., Orešič, M.: Mzmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **22**, 634–636 (2006)
24. Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M.: Mzmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinforma.* **11**, 395 (2010)
25. Voss, B., Hanselmann, M., Renard, B.Y., Lindner, M.S., Kothe, U., Kirchner, M., Hamprecht, F.A.: SIMA: Simultaneous multiple alignment of LC/MS peak lists. *Bioinformatics* **27**, 987 (2011)