

A Method to Align Three - Dimensional Chromatographic Profiles of LC/MS Experiments

Rovshan G. Sadygov*, Fernando Martin Maroto*, Manor Askenazi+, Leo E. Bonilla+, Andreas F.R. Hühmer* • Thermo Electron Corporation

*San Jose, CA, +Boston, MA, US.

Overview

Purpose: Development of an algorithm to align ion chromatograms of LC-MS/MS experiments

Methods: Fast Fourier Transform analysis is done for "crude" alignment in the first step. Dynamic programming on correlation coefficients is used to generate detailed alignment during the second step.

Results: We present results of the application of our algorithm to real sample chromatograms. The program produces accurately aligned chromatograms, where the common species from different samples elute concurrently.

Introduction

Precise alignment of chromatography traces is a necessary procedure in comparative and quantitative proteomics. In particular, in differential analysis for biomarker discovery, where detection of minor changes in analyte composition is required, the necessity of accounting for chromatographic shifts is well recognized. Due to physical processes like aging of separation columns and variations in sample buffers, sample analytes frequently do not elute at the same retention time in consecutive chromatographic separations. Most of the current approaches to chromatographic alignment utilize base peak intensities for peak overlay. These studies align the main features in the base peak chromatograms based on some function that minimizes (or maximizes) differences (or similarities). However, these techniques do not take into account the mass scan information available in LC-MS/MS experiments. In a recent study, J. T. Prince and E.M. Marcotte¹ suggested use of mass scans and dynamic programming to generate time warping functions. The use of mass scan information is potentially a more accurate approach as alignment of every time point is not only based on a single mass-intensity pair, e.g., base peak, but rather on features of entire mass scans.

In this presentation we discuss the chromatographic alignment algorithm, the methods used in the approach and results of its application to real chromatograms.

Methods

Description of algorithm: We use a two-step alignment procedure in our algorithm. At the first step, we align the chromatograms "crudely". This step is fast and is done to align main features in the chromatogram. The criteria used in this step is to maximize the overlap between two base peak chromatograms. This is accomplished by the use of Fast Fourier Transform (FFT). Base peak based correlation values are generated for different time shifts. The array element with maximum overlap is chosen as a starting point for the second, more comprehensive and resource-intensive step.

In the second step, we calculate the correlation coefficients between the "crudely" aligned chromatograms. Since we have already "crudely" aligned the chromatograms, we do not need to generate the correlation values between all scans of the two chromatograms. Rather, we generate correlation values for the appropriate peaks. The process of generating correlation values is resource intensive. Once we have generated correlation values, we determine a path in the two-dimensional correlation matrix that maximizes the sum of the correlation coefficients. In general, this is a quadratic process, which for large number of scans can take a long time. However, the use dynamic programming considerably simplifies the process and the alignment path can be generated in a realistic time frame.

It is noteworthy, that we do not require any user input values or a parameter file in our approach. The algorithm itself determines the necessary alignment parameters (e.g. slack) and applies them. To indicate the quality of the time warping, we print out a probabilistic score that is a characteristic of the alignment. The score varies between zero and one. Zero indicates no reliable match, and one means a complete alignment and can only be achieved in self-alignment.

When there are more than two chromatograms to align, the first chromatogram is assumed to be a reference, and the rest are aligned to it. The approach has been implemented in the program ChromAlign. The program runs under Windows® XP.

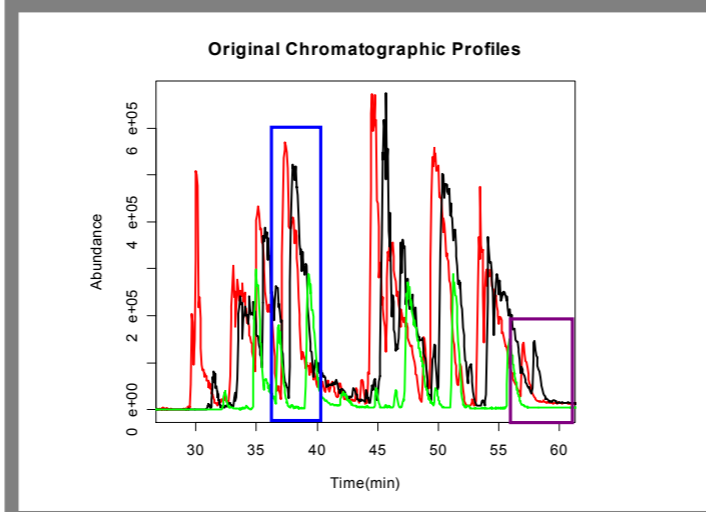
Experimental Data: 9 proteins (Horse myoglobin, bovine serum albumin, chicken egg lysozyme, chicken egg ovalbumin, bovine carbonic

anhydrase, bovine β -casein, horse cytochrome C, α -Lactalbumin, rabbit (all from SIGMA, St. Louis), were digested individually with serine proteases and then combined to obtain a mixture of 8 proteins with individual concentrations between 5fmol to 600fmol for those proteins. An additional protein digest of horse apomyoglobin was added into this mixture at variable concentration levels between 25pmol and 250fmol (as indicated in the figure legends) to achieve peptide concentrations with >3 orders of magnitude dynamic range against a low level peptide mixture background.

Results

In Figure 1, we present base peak chromatographic profiles of three raw files obtained from the separation of a nine protein mix spiked with two concentrations of horse apomyoglobin digest. Chromatographic traces in red and black correspond to samples containing 25pmol of spiked horse apomyoglobin digest. The third chromatographic trace, colored in green, was obtained for a protein mixture with a horse myoglobin concentration at 250fmol, i.e., 100 times smaller concentration.

FIGURE 1. The red and black colored chromatograms correspond to 25 pmol protein concentration and the green profile represents the trace for a protein sample at 250 fmol concentration.

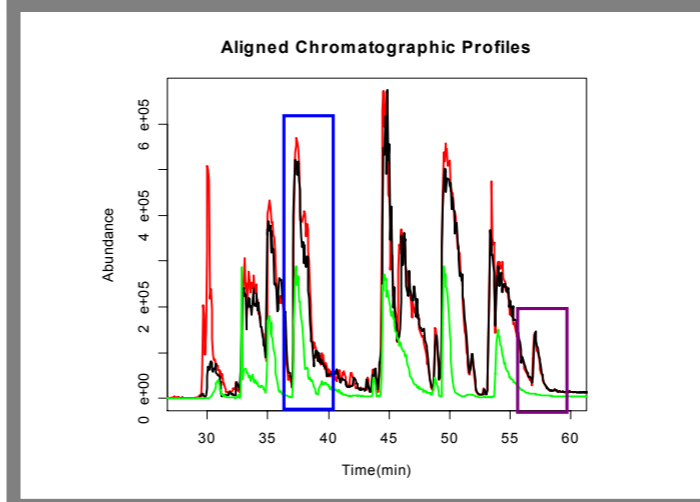


We have chosen these sample concentrations in this example to demonstrate the two major characteristics of the ChromAlign algorithm. The red and black chromatographic traces demonstrate that ChromAlign perfectly overlays nominally identical peak traces, i.e., from protein mixtures of the same concentration, where only time shifts, rather than shifts in intensity are expected (blue box in Figure 2).

More importantly, comparison of the green traces with the other two chromatographic profiles in Figure 2 indicates that ChromAlign also performs well for traces from samples of vastly different concentrations, where large differences in peak intensity in addition to time shifts are expected (purple box). The alignment scores for the red and black profile and for the red and green profiles are 0.74 and 0.62, respectively. The observed high scores are explained by the fact that the protein contents of the samples are identical with large number of common peptides eluting. Alignment between the samples whose protein concentration differs 100 fold (red and green profile) is not as strong. It is expected that some peaks in the sample that is 100 times less concentrated are absent compared to the more concentrated sample leading to a somewhat lower alignment score.

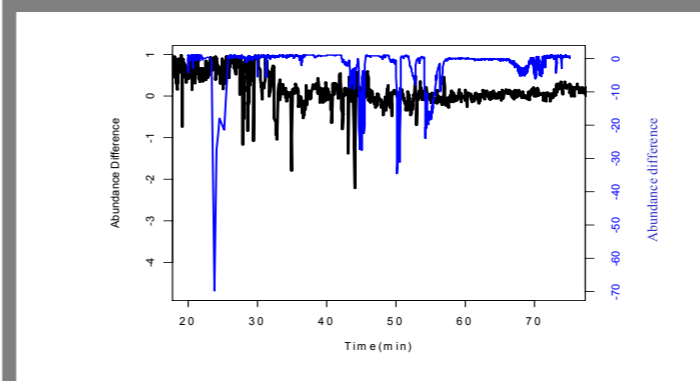
In Figure 3, we show the normalized differences between base peak chromatograms of the samples with the same protein concentrations (red and black lines in Figure 1 and 2). The abundance differences before (blue line in Figure 3) and after the alignment (black line in Figure 3) are about 20 fold. In

FIGURE 2. Base peak LC/MS/MS ion chromatograms after alignment with ChromAlign. The alignment score calculated by ChromAlign is 0.74 for the red and black trace, which is indicative of a high quality alignment.



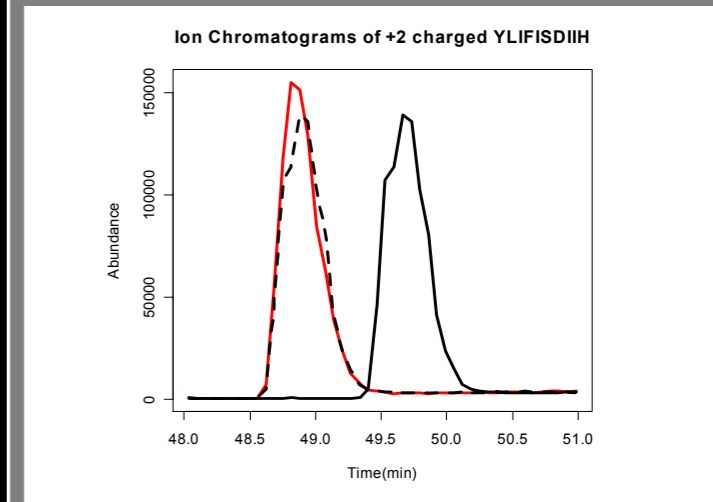
the time interval between 30 and 60 minutes, in which many of the peptides elute, the abundance difference after the alignment varies around zero, while for unaligned chromatograms this difference is around one minute, demonstrating that the algorithm pairs the correct ions.

FIGURE 3. Abundance difference between base peak chromatograms before (blue line and blue y axes) and after alignment (black line and black y axes).



To illustrate usefulness of the alignment procedure in comparative studies, we chose a reconstructed ion chromatogram of a doubly charged peptide species with $m/z = 661.09$. It is present in both samples at approximately equal concentrations and was chosen for identification by tandem mass spectra in each run by Data Dependent™ analysis. Figure 4 shows the reconstructed ion chromatograms of this peptide before (black and red solid lines) and after (black broken and red solid lines) the alignment. It is apparent in Figure 4 that the ion chromatograms of the same sample at equivalent concentration in both samples before the alignment are completely time separated. After the alignment with the ChromAlign algorithm, the reconstructed ion chromatograms fully overlap. The complete overlap of both ion chromatograms indicates that they belong to the same analyte species, which is also confirmed in this case by their MS/MS spectra. Database search of the MS/MS spectra identifies those peaks as the myoglobin peptide "YLIFISDIIH" with SEQUEST® cross-correlation scores of 3.94 and 4.05 in those samples. This example also shows

FIGURE 4. Ion chromatograms of a +2 charged myoglobin peptide. The extracted ion chromatograms of this ion are shown as non-aligned (solid black and solid red lines) and aligned (solid red and broken red lines).



the necessity of time alignment in experiments where sample difference are sought based on differences in chromatographic profiles only.

ChromAlign took about 100 seconds to align these three chromatograms on a computer with Pentium® 4 R processor (3.40 GHz) running under Windows XP. The application utilizes about 200 MB of a RAM for manipulation of these chromatograms with 3359 and 3520 mass scans with mass range up to 2000 m/z units. If the first step of the alignment (FFT based correlation) is skipped, the alignment takes about ten percent longer time to run. The relatively small computational advantage of the "crude" alignment in this case is explained by the fact that overall these chromatograms are misaligned by about 1-1.5 minutes. For situations when the misalignment between two traces is even larger, the advantage from the "crude" alignment is more significant.

Conclusions

- We present a new alignment algorithm, ChromAlign and the results of its application to differential analysis in comparative and quantitative proteomics.
- ChromAlign precisely overlaps chromatograms that are shifted in the time and/or peak intensity dimension.
- The use of full mass scan information instead of a single mass-intensity pair for chromatographic alignment is a more accurate approach to alignment of three-dimensional chromatographic profiles of LC/MS experiments.
- The application of FFT calculations in combination with dynamic programming in a two step alignment approach provides an efficient implementation of the chromatographic alignment algorithm.

References

1. J. T. Prince and E. M. Marcotte. Chromatographic Alignment of Multi-Dimensional Mass Spectra using Interpolated Dynamic Time Warping 53rd ASMS Conference, San Antonio Texas, 2005.
2. J. K. Eng, A. L. McCormack, J. R. Yates III. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database

Acknowledgements

Microsoft and Windows are registered trademarks of Microsoft Corporation. SEQUEST is a registered trademark of the University of Washington. Intel and Pentium are registered trademarks of Intel Corporation. All other trademarks are the property of Thermo Electron Corporation and its subsidiaries.