A Novel Two-Stage Alignment Method for Liquid Chromatography Mass Spectrometry-Based Metabolomics*

Xiaoli Wei[1], Xue Shi[1], Seongho Kim[2], Craig McClain[3,4,5,6], and Xiang Zhang[1]

[1] Departments of Chemistry, University of Louisville, Louisville, KY 40292
{jujuxiao,xueshisx}@gmail.com, xiang.zhang@louisville.edu
[2] Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40292
biostatistician.kim@gmail.com
[3] Medicine, [4] Pharmacology & Toxicology, [5] Alcohol Research Center, [6] Robley Rex Louisville VAMC, University of Louisville, Louisville, KY 40292
craig.mcclain@louisville.edu

Abstract. We report a novel two-stage alignment algorithm that contains full alignment and partial alignment, for the analysis of LC-MS based metabolomics data. The purpose of full alignment is to detect landmark peaks that present in all peak lists to be aligned. These peaks were first selected based on m/z value and isotopic peak profile matching. After removing peaks with large Euclidian distance of retention time from the potential landmark peaks, a mixture score was calculated to measure the matching quality of each landmark peak pair between reference peak list and a test peak list. After optimizing the weight factor in the mixture score, the value of minimum mixture score of all landmark peaks was used as the threshold for peak matching in the partial alignment. A local optimization based retention time correction method was used to correct the retention time changes between peak lists during partial alignment. The two-stage alignment method was used to analyze a spiked-in experimental data and further compared with literature reported algorithm RANSAC implemented in MZmine.

Keywords: LC-MS, two-stage peak list alignment, local optimization.

1 Introduction

Metabolomics is the study of low molecular weight molecules (i.e., metabolites) found within cells and biological systems. It aims to measure and interpret the complex time-related concentration, activity and flux of large sets of metabolites in biological samples. Several types of instruments have been utilized to analyze

* This work was supported by NIH grant 1RC2AA019385 through the National Institute on Alcohol Abuse and Alcoholism. This work was also partially supported by National Institute of Health (NIH) grant 1RO1GM087735 through the National Institute of General Medical Sciences (NIGMS).

D.-S. Huang et al. (Eds.): ICIC 2012, LNCS 7389, pp. 152–159, 2012.

metabolites including nuclear magnetic resonance (NMR), gas chromatography-mass spectrometry (GC-MS), and liquid chromatography-mass spectrometry (LC-MS). Each type of instrumental analysis provides limited coverage of the metabolites, and therefore, only generates a partial metabolite profile of each sample. Currently, significant challenges remain in almost every aspect for the application of metabolomics to biomedical research. Among these, the lack of accurate and efficient bioinformatics tools for the processing of metabolomics data has become a critical bottleneck to the progress of metabolomics. Many data analysis steps are involved in deciphering the mass spectrometry data, including data preprocessing, metabolite identification, quantification, network and pathway analysis.

Peak alignment is a key step of data preprocessing in LC-MS based metabolomics. It recognizes peaks generated by the same metabolite occurring in different samples from the millions of peaks detected during the course of an experiment [1, 2]. A large volume of information-rich data can be generated in a LC-MS based metabolomics study. To carry out the alignment procedure, several bioinformatics tools have been developed including *XCMS2* [3], *centWave* [4], *MZmine2* [5], *MZedDB* [6], *OpenMS* [7]. However, the accuracy of peak alignment remains challenge in metabolomics.

The objective of this work was to develop a novel approach for high accuracy peak alignment. For this reason, we developed a two-stage alignment method to align the peak lists generated from high-resolution mass spectrometry for metabolomics study. The developed method has been implemented in *MetSign* [8] and used to analyze a set of spiked-in experimental data acquired on a LC-MS system. The performance of this method was compared with existing software packages *MZmine*.

2 Experimental Section

2.1 Spiked-in Samples

A mixture of 30 compound standards was prepared at a concentration of 100 µg/mL for each compound. The standards included 11 fatty acid (benhenic acid, tricosanoic acid, stearic acid, myristic acid, nonadecanoic acid, heptadecanoic acid, adipic acid, heneicosanoic acid, nonanoic acid, butyric acid, linoleic acid), 5 triglycerides (trilauroyl-glycerol, trimyristin, tripalmitin, tricaprylin, tricaprin), 9 phospholipids PC(16:0/16:0), PC(16:0/14:0), PC(12:0/12:0), PC(6:0/6:0), LysoPC(16:0/0:0), LysoPC(10:0), PC(20:4(5Z,8Z,11Z,14Z)/16:0), PC(18:2(9Z,12Z)/18:2(9Z,12Z)), PC(24:1(15Z)/24:1(15Z)), and 5 other small molecules (caffeine, L-tryptophan, lidocaine, creatine, trans-4hydroxyl-L-proline). 10 µL of the standard mixture was added to a 100 µL sample of metabolite extract of mouse liver, and dichloromethane: methanol (v/v = 2:1) was then added to the sample vials to make the total volume up to 200 µL.

2.2 LC-MS Analysis

A LECO Citius LC-HRT high resolution mass spectrometer equipped with an Agilent 1290 Infinity UHPLC with a Waters Acquity UPLC BEH hydrophilic interaction chromatography (HILIC) 1.7 µm 150 × 2.1 mm column was used in this work. The sample was loaded in H_2O + 5 mM NH_4OAc + 0.2% acetic acid (buffer A) and separated using a binary gradient consisting of buffer A and buffer B (90/10 acetonitrile/H_2O + 5 mM NH_4OAc + 0.2% acetic acid). Flow rate was set at 250 μL/min on the column with 100% B for 4 min, 45% B at 12 min holding to 20 min, 100% B at 21 min and holding to 60 min for the gradient. The Citius was operated with electrospray ionization in positive ion mode. The system was optimized in high resolution mode (R = 50,000 (FWHM)) and was mass calibrated externally using Agilent Tune Mixture (ATM). The mass spectrometry was operated in both full mass analysis and tandem MS/MS mode to acquire molecular *m/z* value and the corresponding MS/MS spectrum. The spiked-in sample was analyzed 6 times on the LC-MS system.

3 Theoretical Basis

The peak alignment method was developed as a two-stage algorithm: full alignment and partial alignment. The goal of full alignment is to recognize landmark peaks, which are defined as a set of metabolite peaks present in every sample. In the partial alignment stage, the peaks in a test sample that are not recognized as the landmark peaks are aligned.

Let $S = \{S_1, S_2, ..., S_i, ..., S_{n+1}\}$ be the sample set, and $n+1$ is the total number of samples to be aligned. After selecting the reference peak list (RPL) in a random manner, the rest of peak lists are considered as test samples, which can be written as $S = \{RPL, t_1, t_2, \ldots, t_i, \ldots, t_n\}$. Each of the test peak lists is aligned to the RPL, respectively.

Considering two peak lists $\{RPL, t_i\}$, all *m/z* value matched peak pairs between these two peak lists can be selected using a user defined *m/z* variation. If a peak can be matched to multiple peaks in the other peak list, the peak pair with the minimum retention time difference is selected as the most probable match and the other matches are discarded. Therefore, the *m/z* matched peak pairs can be recorded as $\{(r_1, s_1), (r_2, s_2), \ldots, (r_p, s_p)\}$, where r is a peak from RPL, s is a peak from t_i, and p is the total number of the *m/z* matched peak pairs. The *m/z* matched peak pairs are further filtered based on the Euclidean distance of retention time between r and s, i.e., $d_j = |r_j - s_j|$ with a confidence interval of 95%. The peak pairs filtered by retention time are represented as $\{(r_1, s_1), (r_2, s_2), \ldots, (r_m, s_m)\}$ and $m \le p$. This process is iteratively operated on all the test samples, respectively.

A mixture similarity score S_m was developed to measure the matching quality between two peaks as follows:

$$S_m(d_i, \Delta_i | w) = w * \exp\left(-1.6 * \frac{d_i - d_{min}}{d_{med} - d_{min}}\right) + (1 - w) * \frac{1}{1 + \Delta_i} \tag{1}$$

where d_i is the Euclidean distance of retention time between the i th matched peak pair, d_{min} and d_{med} are the minimum and median retention time distance among all matched peaks in the two peak lists, respectively, Δ is the absolute value of *m/z* difference between the i th matched peak pair, and w is a weight factor and $0 \leq w \leq 1$.

The peaks that are present in every test peak list and are matched to the same peak in the RPL are used to optimize the value of weight factor w for the alignment of a test peak list and the RPL by maximizing the value of $\sum_{i=1}^{k} S_m(d_i, \Delta_i \mid w)$, k is the number of matched peaks between the test peak list and the RPL, and w is set as 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95, respectively. After optimizing the weight factor w, the value of S_m is calculated for each matched peak pair between the test peak list and the RPL, followed by an outlier detection in S_m^j, $j = 1, \ldots, k$. By iteratively considering pair set $\{RPL, t_i | i = 1, \ldots, n\}$, the landmark peaks $\{(r_1, t_{11}, \ldots, t_{n1}), \ldots, (r_m, t_{1m}, \ldots, t_{nm})\}$ are obtained. The minimum mixture score S_m^{min} among all the test peak lists is then used as a threshold value in the partial alignment.

To perform the partial alignment, the retention time value of each landmark peak in the test peak list is assigned to the retention time value of the corresponding landmark peak in the RPL. A local polynomial fitting method is employed to correct the retention time of peaks present between two adjacent landmark peaks. Because multiple landmark peaks can be detected in a set of experimental data, adjusting retention time shifts using two adjacent landmark peaks can correct nonlinear retention time shifts. To correct the retention time of peaks not present between two landmark peaks, an iteratively optimization method is applied to the group of peaks eluted earlier than the first-eluted landmark peak and the group of peaks eluted later than the last-eluted landmark peaks, respectively. In each optimization process, 30% of landmark peaks are randomly selected from $\{(r_1, t_{11}), \ldots, (r_m, t_{1m})\}$ and a polynomial model fitting error is computed as follows

$$\varepsilon = \sum_{i=1}^{N} \mid t_{R,i}^{o} - t_{R,i}^{f} \mid \tag{2}$$

where $t_{R,i}^{o}$ is the original retention time of the i th peak, $t_{R,i}^{f}$ is the fitted retention time of the i th peak, N is the number of peaks in the test peak list at the region of interest. This process is repeated 1000 times and the model with minimum error is selected and used for retention time correction.

After the retention time correction, the partial alignment is applied to all the non-landmark peaks present in each of the test peak lists and aligns them to the peaks present in the RPL, where a mixture score S_m is calculated using equation (1) for each peak pair. A peak pair is considered to be a match if its mixture score is larger than S_m^{min}. It is possible that one peak in the test sample can be matched to multiple peaks in the RPL and *vice versa*. In these cases, the peak pair with the maximum mixture score is kept while the remaining matches are discarded. If there is a peak in the test peak list that cannot be matched to any peaks in the RPL, this peak is considered as a new peak to the RPL and is added to the RPL. The updated RPL is then used to align the peaks in the next test peak list, and this process is repeated until all the test peak lists are aligned.

4 Results and Discussion

The raw instrument data were first converted into *mz*ML format and further reduced to peak lists using *MetSign* software. [8] There are about 2300 peaks detected by *MetSign* software in each sample and about 1100 peaks assigned to a database compound. Of the 30 spiked-in compound standards, most of the compounds were detected based on the match of *m/z* values with a variation window of ≤ 5 ppm and the similarity of isotopic peak profile measured by Pearson's correlation coefficient ≤ 0.75. Table 1 shows the number of compound standards detected in each replicate injection. The variation of the number of detected compound standards was generated by the experimental variation. After *MetSign* processing, a total of six peak lists were generated, and these peak lists were subjected for peak alignment.

Table 1. The number of compound standards detected in six replicate injections

Index of replicate injection	1	2	3	4	5	6
No. of compound standards identified in each injection	25	23	23	24	24	24

During the full alignment, a total of 283 landmark peaks were detected with retention time ranges from 91.52 s to 643.14 s. Of the 283 landmark peaks, 18 were the peaks generated by the spiked-in compound standards while the remaining landmark peaks were generated by the metabolites extracted from mouse liver. Figure 1 shows the effectiveness of retention time correction to the non-landmark peaks during the partial alignment. Even though the experiments of the six replicate injections were performed under the identical experimental conditions, the retention time of each compound still drifted between injections and such a retention time drift is not linear. Therefore, the local polynomial fitting method is able to correct the retention time of peaks present between two adjacent landmark peaks.

To compare the alignment accuracy of the two-stage alignment method, the experiment data were also processed using publically available software *MZmine*. *MZmine* software has two alignment methods, Join aligner and RANSAC aligner. Join aligner is a simple alignment method, which aligns detected peaks in different samples

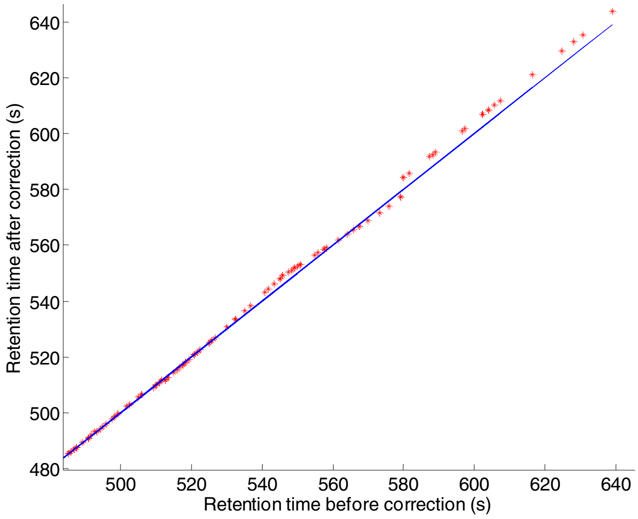

Fig. 1. Local optimization based retention time correction for compounds eluted between two landmark peaks. The value of each red star in x-axis is the retention time of a compound before correction and the value in y-axis is after retention time correction. The solid blue line is a guideline depicting a situation of no retention time correction.

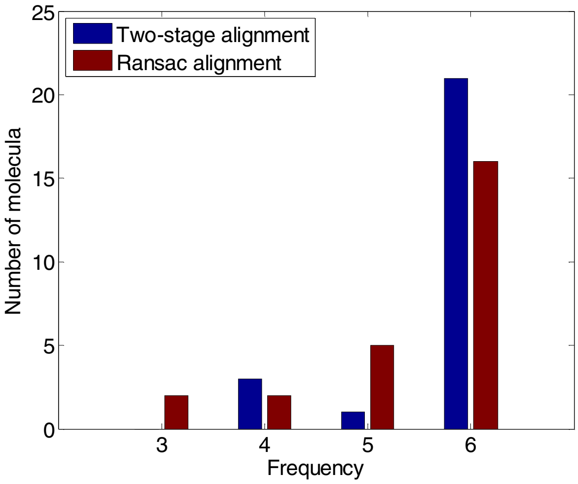

Fig. 2. Alignment results of spiked-in compound standards by two-stage method (blue) and RANSAC method of MZmine2.5 (red)

through a match score. RANSAC aligner is an extension of the Join aligner. It includes a method of retention time correction to adjust the retention time shift in all peak lists. Therefore, we chose the RANSAC alignment in *MZmine2.5* for comparison. Figure 2 depicts the alignment results of the two-stage alignment method and the RANSAC method. Based on the experimental design, all of the spiked-in compound standards should be correctly aligned. In the two-stage alignment, a total 21 peaks of the spiked-in standards are aligned in all six injections, while RANSAC only fully aligned 16 compound standards. Furthermore, all the spiked-in compound standards were aligned in at least 4 peak lists of the 6 replication injections by our method, while RANSAC still had 2 compound standards aligned in only three injections.

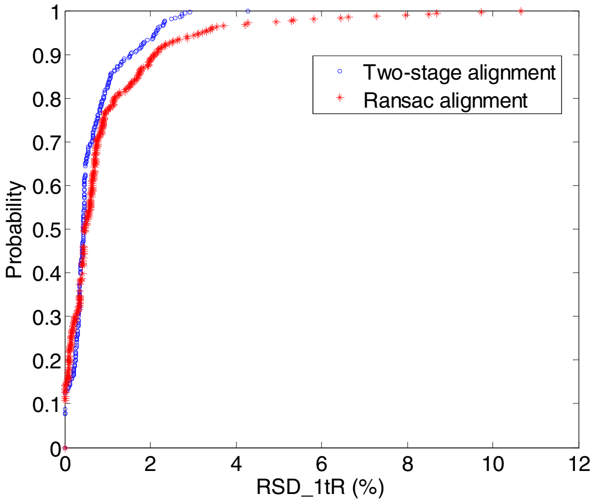

Fig. 3. The comparison of RSD value between two-stage alignment we proposed and RANSAC alignment in MZmine2.5

Figure 3 shows the distribution of relative standard deviation (RSD) of all aligned peaks by the two testing alignment algorithms. The maximum RSD of aligned peaks in the two-stage alignment method is 4.2%. Manual validation shows that this alignment is correct. There are 12 compounds were aligned by RANSAC with a retention time RSD larger than 4%. The maximum retention time RSD reached 10.7%, which is much larger than the retention time variation caused by the experiments. Such a large retention time variation was caused by the inaccuracy of peak alignment. From the comparative analysis, we conclude that the two-stage alignment method outperforms the RANSAC alignment by providing high accuracy of peak alignment for the analysis of LC-MS based metabolomics data.

5 Conclusions

A novel two-stage alignment algorithm, containing full alignment and partial alignment, was developed for high accuracy peak list alignment for LC-MS based metabolomics. The full alignment detects landmark peaks that are present in all the peak lists. During this process, the potential landmark peaks were first selected based on the *m/z* and isotopic peak profile matching. After removing outliers based on the Euclidian distance of retention time from the potential landmark peaks, a mixture score method was employed to evaluate the match quality of each landmark peak pair between the reference and the test sample peaks. The value of minimum mixture score of all landmark peaks was used as the threshold of peak matching during partial alignment, in which local optimization based retention time correction was employed to correct the retention time changes between peak lists. The performance of the two-stage alignment method was tested by analyzing a spiked-in experimental data and further compared with literature reported algorithm RANSAC implemented in *MZmine2.5*. The comparison demonstrates that our two-stage alignment method outperforms the RANSAC algorithm for high accuracy of peak alignment.

References

1. Zhang, X., et al.: Data pre-processing in Liquid Chromatography-mass Spectrometry-based Proteomics. Bioinformatics 21(21), 4054–4059 (2005)
2. Wang, B., et al.: DISCO: Distance and Spectrum Correlation Optimization Alignment for two-dimensional Gas Chromatography time-of-flight Mass Spectrometry-based Metabolomics. Anal. Chem. 82(12), 5069–5081 (2010)
3. Benton, H.P., et al.: XCMS2: Processing Tandem Mass Spectrometry Data for Metabolite Identification and Structural Characterization. Anal. Chem. 80(16), 6382–6389 (2008)
4. Tautenhahn, R., Bottcher, C., Neumann, S.: Highly Sensitive Feature Detection for High Resolution LC/MS. BMC Bioinformatics 9, 504 (2008)
5. Pluskal, T., et al.: MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-based Molecular Profile Data. BMC Bioinformatics 11, 395 (2010)
6. Draper, J., et al.: Metabolite Signal Identification in Accurate Mass Metabolomics Data with MZedDB, an Interactive m/z Annotation Tool Utilising Predicted Ionisation Behaviour 'rules'. BMC Bioinformatics 10, 227 (2009)
7. Sturm, M., et al.: OpenMS-An Open-source Software Framework for Mass Spectrometry. BMC Bioinformatics 9 (2008)
8. Wei, X.L., et al.: MetSign: A Computational Platform for High-Resolution Mass Spectrometry-Based Metabolomics. Analytical Chemistry 83(20), 7668–7675 (2011)