國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Department or Graduate Institute of Networking and Multimedia
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

搭配頻寬估測方法在核密度估測液相層析質譜儀資料之滯留時間校正與一維質子核磁共振代謝體圖譜之相位校正

The Retention Time Alignment for Nontargeted LC/MS Analysis Using Kernel Density Estimation with a Novel Bandwidth Estimator and Phase Correction of Metabolomic 1D ^{1}H-NMR Spectra

劉家瑋

Jia-Wei Liu

指導教授：曾宇鳳 博士

Advisor: Y. Jane Tseng, Ph.D.

中華民國 104 年 7 月

July, 2015

誌謝

在此僅感謝指導教授曾宇鳳老師的熱心教導，讓我有系統地學習代謝體學及認識其相關議題，讓我有機會參與代謝體的國際會議，看見更多有趣的代謝體學的新興議題，同時也要感謝藥學系郭錦樺老師的指導以及郭老師實驗室同學在實驗上的幫忙，讓我可以在足夠的資料中開發出合適的演算法。

研究過程中，感謝代謝體組學長們教導以及幫助，使我能專注於研究上以及順利找到問題癥結使研究能順利進行。感謝實驗室三源學長的指導，讓我在每次討論中看見演算法可以再改進的地方，或是找到研究上遇到瓶頸可能的解法，以及在我遇到挫折時給我的鼓勵；感謝國清學長的指導，讓我可以很快地熟悉資料處理流程，以及指出開發演算法的可以改進的地方；感謝天爵學長的鼓勵以及幫忙，讓我在遇到挫折時可以很快地站起來；感謝乃文學姐、東銘學長、宇彥還有 Brendan 的互相幫忙還有勉勵；感謝實驗室學長姐、學弟妹們的幫忙。

最後要感謝最親愛的家人，謝謝你們一直以來對我的支持與信任，如果沒有你們在背後的支持，就不會有現在的我。

中文摘要

本篇論文呈現兩套我們研發的演算法，用來解決偵測小分子訊號時面臨的計算問題，它們是由代謝體的應用發展而來。

在本篇論文的第一個部分，我們發展一套用於液相層析質譜儀之訊號滯留時間校正工具 – *LAKE*，它可以校正層析質譜訊號的滯留時間(retention time)。代謝體學(metabolomics)分析上常用層析法為高效液相層析儀，因其具有系統穩定性佳、分析結果再現性高之優點，但因梯度沖提時易發生滯留時間偏移，倘若樣本內包含多元化合物，層析圖譜之滯留時間偏移將導致化合物之辨識錯誤率提高，另外現有滯留時間校正工具仍無法有效處理多批次資料之滯留時間校正，因此出現滯留時間未校正(misalignment)之情形發生，因此需要發展可處理多批次資料之滯留時間校正工具。 LAKE 將偵測到之波峰資料依照樣本之資料相似性由高到低依序進行滯留時間之校正，在每一輪校正過程中會將波峰依序從質荷比(m/z)到滯留時間進行分組，再對分組完的質荷比-滯留時間群(m/z-RT group)內的滯留時間做頻寬選擇(bandwidth selector)，並使用估計出的頻寬對該組資料之滯留時間進行核密度估計(kernel bandwidth estimator)，作為滯留時間再分組的根據。每一輪結束後都會將各波峰之質荷比以及滯留時間更新為該組的平均質荷比以及平均滯留時間。LAKE 可應用於在外生性化合物混和樣本，外生性化合物添加於體液樣本，以及含有多種複雜的內生性化合物樣本訊號於多批次資料之滯留時間校正。

在本篇論文的第二個部分，我們發展一個自動化的一維質子核磁共振圖譜(1D ^{1}H-NMR)相位校正(phase correction)演算法 – *PHASION*，它能夠自動將多筆一維質子核磁共振圖譜完成圖譜相位校正。

在一維質子核磁共振圖譜的相位誤差來自於機器本身，是一種不可避免的誤差，需要再進行後續處理前解決此誤差消除，將頻譜還原。目前大部分研究人員所使用的圖譜相位校正方法多仰賴有經驗之使用者，依照手動方式調整參數，以求得各圖譜之最佳相位校正結果。由於這些校正方法，很容易會因為人為的因素而產生不同的校正結果，並造成後續處理結果上的差異，我們為求達到自動化的需求並且可客觀的校正一維質子核磁共振代謝體圖譜的相位，在此提出此新的自

動化相位校正方法。 PHASION 藉由選擇穩定訊號的圖譜區段，計算該區段圖譜在相位校正後之基線穩定程度作為分數，並搭配 Nelder-Mead Simplex Optimizer 最佳化搜尋方式，求出該圖譜之最佳校正之相位角度。

關鍵字：液相層析質譜儀、滯留時間校正、核密度估計、質子核磁共振圖譜、相位校正、代謝體學

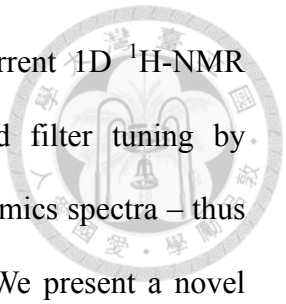
Abstract

This dissertation presents two developed algorithms for solving computational problems of detecting small molecules in the field of metabolomics analysis.

In the first part of this dissertation, we present the tool – LAKE, which is a tool for detected peak alignment to align retention time for chromatographic methods coupled to spectrophotometers such as high performance liquid chromatography for metabolomics works. The existed tools for retention time correction still can't properly aligning retention times of detected peaks from multiple batches and some detected peaks are left misalignment. LAKE resolves peak shifts from high data similarity to low data similarity. In each turn, detected peaks would be clustered in mass-over-charged (m/z) dimension and then retention time (RT) dimension. For each m/z-RT cluster, bandwidth used in RT density estimation with kernel density estimation (KDE) is estimated with bandwidth selector. At the end of each turn of retention time shift resolution, the m/z and RT of detected peaks would be updated with average m/z and average RT of the m/z-RT group before next turn of detected peak alignment. LAKE can be applied to aligning retention time from mixed exogenic compounds samples, multiple exogenic compounds added in biofluid samples and complicate endogenous compounds contained metabolomics samples in multiple batches.

In the second part of this dissertation, we present the tool – PHASION, which is a tool for automatic phase correction on multiple 1D proton nuclear magnetic resonance (^{1}H-NMR) spectra for metabolomics works.

The phase error is an unavoidable error happened when FID signal is recorded, after Fourier transformed into spectrum mixed with phase error. The phase correction is to find zeroth-order and first-order phase error to make misphased spectrum into



phase-corrected spectrum before any further data processing. Current 1D ^{1}H-NMR phase correction methods usually require manual parameter and filter tuning by experienced users to obtain desirable results from complex metabolomics spectra – thus becoming prone to correction variation and biased quantification. We present a novel alternative method, PHASION, for automatically estimating the phase angles of 1D ^{1}H-NMR metabolomics data. PHASION finds optimal phase angles by calculating proposed objective score for relative stable segments of spectrum and calculates the score for baseline of spectrum phased with phase angles (PH0, PH1) and approach to the optimal phase angles for the spectrum with Nelder-Mead Simplex Optimizer.

KEYWORDS: liquid chromatography/mass spectrometry, retention time alignment, kernel density estimation, proton nuclear magnetic resonance spectrum, phase correction, metabolomics.

Contents

誌謝 i

中文摘要 ii

Abstract iv

Contents vi

List of Figures viii

List of Tables xx

Chapter 1 LAKE: a Peak Alignment Tool for Nontargeted LC-MS Based Metabolomics 1

1.1 Introduction 1

1.2 Materials 6

1.2.1 *Chemicals* 6

1.2.2 *Sample Preparation* 7

1.2.3 *Chromatographic and Mass Spectrometric Analysis* 8

1.2.4 *Data Preparation* 10

1.3 Theoretical Basis 10

1.3.1 *Grouping Peaks of Technical Replicates from the Same Sample* 11

1.3.2 *Grouping Peaks of Sample from the Same Batch* 18

1.3.3 *Grouping Peaks from Different Batches* 23

1.3.4 *Performance Evaluation of LAKE on Peak Alignment* 25

1.3.4.1 Performance Evaluation of LAKE on Peak Alignment 25

1.3.4.2 Performance Evaluation of LAKE on noise introduced peak alignment. 25

1.4 Results 31

1.4.1 *LAKE and XCMS Algorithms Using Forensics Drugs* 31

1.4.2 *LAKE and XCMS Algorithms Using Metabolomics Data Set with Introduced Different Types of Noise* 38

1.5 Discussion 58

1.5.1 *Comparison of LAKE and XCMS Algorithms Using Forensics Drugs* 58

1.5.2 *Comparison of LAKE and XCMS Algorithm on Metabolomics Data Set* 61

1.6 Conclusion 66

Chapter 2 PHASION: PHASing Intrinsically On NMR Spectrum 68

2.1 Introduction 68

2.2 Material 75

2.3 Theoretical Basis 77

2.3.1 *Data Pre-processing* 78

2.3.2 *Nelder-Mead Optimizer* 79

2.3.3 *Scoring Function* 80

2.3.4 *Performance Evaluation of PHASION on Phase Correction* 84

2.4 Result and Discussion 85

2.4.1 *Convergence of Nelder-Mead Optimization* 85

2.4.2 *Comparison of Different Pre-processing Methods* 87

2.4.3 *Comparison of Performance on Synthesized Spectra with Gaussian Noise Introduced* 90

2.4.4 *Comparison of Performance on Complex Metabolomic Plasma Samples* 96

2.5 Conclusion 97

Table of Abbreviations 99

Appendix 100

Reference 101

List of Figures

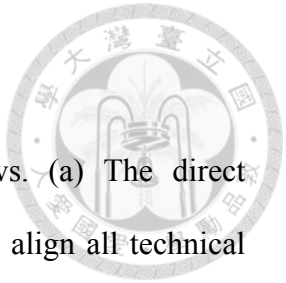


Figure 1.1: Illustration of direct and Lake alignment workflows. (a) The direct alignment, a commonly seen alignment workflow. First, align all technical repeats. Then, generate a peak table. (b) The LAKE alignment. First, align technical repeats from the same sample. Next, align the samples from the same batch. Finally, align all batches and generate a peak table. 11

Figure 1.2: The workflow of grouping peaks of technical replicates from the same sample. (a) The workflow of grouping of detected peaks of similar m/z values and RT values construction with suggested relative mass difference tolerance and RT tolerance, respectively. (b) The workflow of RT regrouping with kernel density estimation. 12

Figure 1.3: Pseudo code for algorithm of RTclusteringOnMzclusteredGroup 14

Figure 1.4: Density distributions of estimated retention time using different kernels: (a) the estimated density distribution with bandwidth h=0.3 using Epanechnikov, Gaussian and triangular kernel; (b) the Gaussian kernel; (c) the triangular kernel; (d) the Epanechnikov kernel. The mean of estimated density distribution using Epanechnikov (red vertical line) is closest to the mean in the real data (green vertical line is the mean of red data points in the rug plot). It is more robust even noise (black data points in the rug plot) exists. 16

Figure 1.5: Pseudo code for algorithm of RTRegroupByCheckingMultimodal 18

Figure 1.6: The workflow of grouping peaks of sample from the same batch. (a) The workflow of grouping detected peaks with similar m/z values construction with suggested m/z tolerance and RT tolerance. (b) The workflow of RT distribution estimation on detected peaks with similar m/z values by kernel

density estimation. (c) The workflow of peak table regroup. 19

Figure 1.7: Pseudo code for algorithm of RTRegroupByUCV 21

Figure 1.8: Pseudo code for algorithm of MergingMisalignedGroupdinPT 23

Figure 1.9: The workflow of grouping peaks from all batches. (a) The workflow of grouping detected peaks with similar m/z values construction with suggested m/z tolerance and RT tolerance. (b) The workflow of peak table regroup. 24

Figure 1.10: Misaligned peaks in 4th batch of the aligned compound (m/z, RT) = (174.0822, 191.07) in positive ionized mode of the metabolomics data set. Different colors represent peaks from different groups. The red peak group contains peaks from 1st batch to 3rd batch and some peaks from 4th batch. The black peak group contains peaks from 4th batch for peaks in 4th batch with larger batch difference in RT dimension which makes peak misalignment when using existed alignment algorithm. 26

Figure 1.11: The batch difference in the aligned compounds of QC samples. Each blue line represents z-scores of an aligned compound over all QC samples. (a) The m/z variation among batches in z-score. X-axis is the injection order of QC samples. (b) The RT variation among batches in z-score. X-axis is the injection order of QC samples. 28

Figure 1.12: The noise introduced data sets. (a) The data set with only local shift in both m/z and RT dimension. (b) The data set with both local shift in both m/z and RT dimension and global shift among batches in both m/z and RT dimension. 29

Figure 1.13: Different types of global shift introduced to the 129th compound. The (m/z, RT) of the 129th compound is (507.8494, 848.50). The first, second, third

and fourth batch are colored in red, orange, green and blue, respectively. ..31

Figure 1.14: The peak alignment by XCMS with different parameters. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 33

Figure 1.15: The peak alignment by LAKE with different parameters. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 34

Figure 1.16: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. 35

Figure 1.17: Peaks of selected compounds before and after XCMS alignment. (Left panel): Data distribution in m/z-RT panel, (Right panel): Misaligned peaks in peak alignment done by XCMS are marked with blue circles. 36

Figure 1.18: Peaks of selected compounds before and after LAKE alignment (Left panel): Data distribution before alignment in m/z-RT plot, (Right panel): Misaligned peaks in peak alignment done by LAKE are marked with blue circles. 37

Figure 1.19: The peak alignment by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the

alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 38

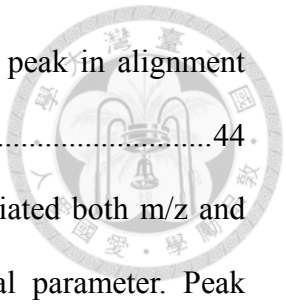
Figure 1.20: The peak alignment by LAKE with estimated parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 39

Figure 1.21: Peaks of selected compounds before and after XCMS alignment (Left panel): Data distribution before alignment in m/z-RT plot, (Right panel): Misaligned peaks in peak alignment done by LAKE are marked with blue circles. 40

Figure 1.22: Peaks of selected compounds before and after LAKE alignment (Left panel): Data distribution before alignment in m/z-RT plot, (Right panel): Misaligned peaks in peak alignment done by LAKE are marked with blue circles. 41

Figure 1.23: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. 42

Figure 1.24: The peak alignment on BR+GNLM1-2 data set (exaggerated both m/z and RT difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned



with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 44

Figure 1.25: The peak alignment on BR+GNLM1-3 data set (alleviated both m/z and RT difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 45

Figure 1.26: The peak alignment on BR+GNLM1-4 data set (exaggerated m/z difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 46

Figure 1.27: The peak alignment on BR+GNLM1-5 data set (alleviated m/z difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 47

Figure 1.28: The peak alignment on BR+GNLM1-6 data set (exaggerated RT difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result

respectively. 48

Figure 1.29: The peak alignment on BR+GNLM1-7 data set (alleviated RT difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 49

Figure 1.30: The peak alignment on BR+GNLM1-2 data set (exaggerated both m/z and RT difference among batches) by LAKE with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively. 50

Figure 1.31: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing. 51

Figure 1.32: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The

black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing. 52

Figure 1.33: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing. 53

Figure 1.34: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing. 54

Figure 1.35: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing. 55

Figure 1.36: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE

alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing. 56

Figure 1.37: Peaks from 129th compound after XCMS alignment. Misaligned peaks in peak alignment done by XCMS are marked with blue circles. The (m/z, RT) of the 129th compound is (507.8494, 848.50). 57

Figure 1.38: Peaks from 129th compound after LAKE alignment. Misaligned peaks in peak alignment done by LAKE are marked with blue circles. The (m/z, RT) of the 129th compound is (507.8494, 848.50). 58

Figure 1.39: The PCA of two peak table done by two different peak alignment algorithms. 66

Figure 2.1: Data processing from a raw FID to a spectrum. A spectrum is the Fourier transformed FID. Before further data processing, the spectrum need to be phase corrected. The phase correction can be done by finding the optimal phase angle (PH0, PH1) which minimizes an objective function. (a) The curve represents the real part of the FID. (b) The curve represents the imaginary part of the FID. (c) The curve represents the real part of the spectrum. (d) The curve represents the imaginary part of the spectrum. (e) The curve represents the phased spectrum. 69

Figure 2.2: The workflow of NMR data processing from the raw FID to the spectrum before applying the statistical analysis. 70

Figure 2.3: The distortion in the spectrum which is caused by the imbalance in quadrature detectors. 71

Figure 2.4: The effect of adjusting the phase angle when phasing a spectrum. (a) The effect of only adjusting PH0 of the spectrum. The numbers indicate the degree of PH0 used in phasing the spectrum. (b)The effect of only adjusting PH1 of the spectrum. The numbers indicate the degree of PH1 used in phasing the spectrum. 72

Figure 2.5: Phase distortion observation via ACD autosimple phase corrected spectrum. The following subfigures are the distributions of (a) the PH0 distortion and (b) the distribution of the PH1 distortion observed from 30 random selected samples. 77

Figure 2.6: The detailed of workflow of PHASION. 78

Figure 2.7: The Nelder-Mead simplex optimizer. The triangle SLM for Nelder-Mead simplex optimizer. The Nelder-Mead simplex optimizer is decided by three scaling factors: alpha, beta and gamma. Alpha is the reflection factor (default 1.0), beta is the contraction factor (0.5) and gamma is the expansion factor (2.0). The red triangle would transform its shape during optimization and moving toward the optimal solution of searching space.....80

Figure 2.8: The illustration of how sdBLdiff is calculated. The black and red lines represent the phased spectrum and unphased spectrum, respectively. The green vertical lines are boundary lines of the highest signal in the spectrum. The red arrow indicates the minimum negative intensity in 1st part of spectrum. The blue arrow indicates the standard deviation of the negative intensities in 1st part of spectrum. 82

Figure 2.9: The illustration of how the area between zero and first deciles of two parts of spectrum is calculated. The black and red lines represent the phased spectrum and unphased spectrum, respectively. The green vertical lines are

boundary lines of the highest signal in the spectrum. The green colored area represents the area of between zero and first deciles of two parts of spectrum. 83

Figure 2.10: The illustration of how the normalized averaged summed negative area is calculated. The black and red lines represent the phased spectrum and unphased spectrum, respectively. The red colored area represents the summed area of negative intensity of the unphased spectrum. 84

Figure 2.11: Different optimization searching approaches for selecting the optimal phase angle. The green vertical lines are boundary lines of the highest signal in the spectrum. (a) The spectrum phased with the optimal phase angles searching with initial phase angles assigned with (ph0*, ph1*), (ph0*, 0) and (0, 0) are colored in red, black and blue, respectively. (b) The score distribution over (PH0, PH1). The optimal phase angles searching with different initial phase angles assigned with (ph0*, ph1*), (ph0*, 0) and (0, 0) are labeled as green star, black circle and blue triangle, respectively. Yellow and red colored block represent the score of given (ph0, ph1) is high and low respectively. 86

Figure 2.12: Effect of proposed data processing steps for phasing: phase optimization using different pivot points. The green vertical lines are boundary lines of the highest signal in the spectrum. (a) The phased spectrum with pivot point set at arbitrary point (ppm=2.49). (b) The phased spectrum with pivot point set at chemical shift=0 (ppm=0). (c) The phased spectrum with pivot point set at the highest signal in the spectrum (ppm=4.8). 87

Figure 2.13: Effect of proposed data processing steps for phasing: phase optimization using partial spectrum. Two green vertical lines represent the location of

water signal. The phased spectrum with complete spectrum and partial spectrum are colored in red and black, respectively. 88

Figure 2.14: Effect of proposed data processing steps for phasing: phase optimization using different types of penalties. Two green vertical lines represent the location of water signal. (a) The phased spectrum with no penalty and normal penalty are colored in red and black, respectively. (b) The phased spectrum with 10% penalty and normal penalty are colored in green and black, respectively. (c) The phased spectrum with 10 times penalty and normal penalty are colored in blue and black, respectively. 89

Figure 2.15: Effect of proposed data processing steps for phasing: phase optimization with the smile elimination. Two green vertical lines represent the location of water signal. The phased spectrums with smile artifact and without smile artifact are colored in red and black, respectively. 90

Figure 2.16: Evaluation of different autophasing algorithms on synthesized data. The synthesized data with the PH0 distorted with Gaussian noise (sd=1). 90

Figure 2.17: Evaluation of different autophasing algorithms on synthesized data. The synthesized data with the PH1 distorted with Gaussian noise (sd=1). 91

Figure 2.18: Evaluation of different autophasing algorithms on synthesized data introduced with the PH0 distorted with Gaussian noise (sd=1). 92

Figure 2.19: Evaluation of different autophasing algorithms on synthesized data introduced with the PH1 distorted with Gaussian noise (sd=1). 92

Figure 2.20: Evaluation of different autophasing algorithms on synthesized data with normalized SSE with average over 35 synthesized data on each chemical shift range. The bar represents the standard deviation of 35 synthesized data. The results of the synthesized data introduced with the PH0 distorted with

Gaussian noise (sd=1) are shown in left panel. The results of the synthesized data introduced with the PH1 distorted with Gaussian noise (sd=1) are shown in right panel. The result of ACD, Chenomx and PHASION phased result is colored in red, black and green, respectively. 93

Figure 2.21: Evaluation of different autophasing algorithms on synthesized data with normalized SSE and Proposed Score. The results of the synthesized data introduced with the PH0 distorted with Gaussian noise (sd=1) are shown in left panel. The results of the synthesized data introduced with the PH1 distorted with Gaussian noise (sd=1) are shown in right panel. 94

Figure 2.22: PHASION phasing spectrum with baseline not equals to zero.....95

Figure 2.23: PHASION phasing with spectrum baseline equals to zero.....95

Figure 2.24: Comparison of different phasing methods on real sample. (a), (b) and (c) are the phased spectrum. (d), (e) and (f) are the baseline removed phased spectrum.....96

Figure 2.25: Evaluation of different autophasing algorithms on synthesized data with proposed score. 97

List of Tables

Table 1 Algorithm comparisons 5

Table 2 Different Global shift introduced to each batches for different data sets.....30

Table 3 Gaussian noise introduced data set (GNLM) 43

Table 4 Global shift among batches noise introduced data set (BR+GNLM1) 58

Chapter 1 LAKE: a Peak Alignment Tool for Nontargeted LC-MS Based Metabolomics

1.1 Introduction

Liquid chromatography-mass spectrometry (LC-MS) has been widely used in metabolomics studies associated with environmental and stress,[1-3] functional genomics,[4-6] biomarker discovery,[7-10] and integrative systems biology.[11-14] The metabolomics studies based on LC-MS can be further divided into targeted or untargeted analysis. In a targeted analysis, only metabolites of interests (targets) would be measured.[15-19] In an untargeted analysis, the goal is to find any metabolites regardless known or unknown functions and molecular structures associated with the questions asked. While known compounds are usually widely studied biochemically, the unknown metabolites are considered the small molecule can be detect with reproducible result but the chemistry identify still not elucidated yet.[20] The untargeted analysis is a powerful tool for understanding biochemistry and metabolism in biological systems with the special ability to identify potential novel biomarkers. For example, untargeted metabolomics approaches were used to identify novel substrates of different enzymes, including N-acyl taurines for fatty acid amide hydrolase[21] and fatty acids for different families of human cytochrome P450 enzymes.[22] In addition, it was applied successfully to identify both known and novel vitamin E metabolites down-regulated upon activation of pregnane X receptor, a member of the mammalian nuclear receptor superfamily.[23] In bacterial systems, an untargeted metabolomics approach was recently used to compare hydrophobic metabolite profiles of P. aeruginosa strains lacking a functional pyochelin gene cluster.[24] These experiments revealed that this cluster regulates many metabolites,

in addition to pyochelin, including a family of novel metabolites that were characterized as 2-alkyl-4,5-dihydrothiazole-4-carboxylates (ATCs). Moreover, it successfully applied to find out key metabolites for protecting bacteria from the high proton concentration and metal-rich environment in biofilms growing in pH ~0.9 acid mine drainage.[25]

The major difference for targeted and untargeted analysis is the number of metabolites that can be detected in each study. Take the research on the role of sarcosine played in prostate cancer progression[26] as an example, only about 800 metabolites of interested were measured, which is less than 2000,[27] an average number of metabolites can be detected in a untargeted analysis. While only focused metabolites of interest can be reduce the number of metabolites to be detected and save time. It loses the general view of metabolites associated with the interesting mechanism. In practical clinical metabolomics studies, comparisons of metabolites between hundreds of LC/MS runs at a time is often seen. However, the peaks from different samples represent the same metabolite may be different due to mass-to-charge ratio (m/z) deviation and retention time (RT) deviation. The reason for m/z deviation is related with mass resolution of different mass spectrometers and the magnitude of m/z deviation is generally small and predictable.[27] The effect of m/z difference is often considered once using a range of mass tolerance used when matching compounds in database searching.[28] On the other hand, the common reasons for RT deviation are the competition between sample and solvent, the instability of condition when generating chromatogram, pressure fluctuation, column temperature variations, and column aging.[29, 30] To matching peaks representing the same analyte from different samples. alignment is required.[31]

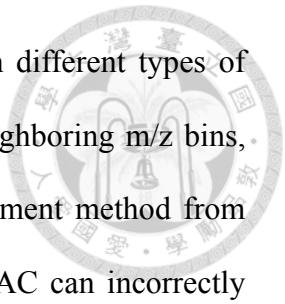
As a result, the alignment step in both targeted analysis and untargeted analysis can impact subsequent analysis greatly.[32] However, the complexity of alignment in targeted and untargeted analysis is different. The alignment in targeted study can be individually

processed on the interested compounds with certain tolerance in m/z and RT range while no targeted compounds can be used in untargeted analysis.

RT deviation is a non-linear shift and tends to have greater magnitude than m/z deviations.[33] Moreover, the RT deviation of two samples from the same experiment tend to become larger with increasing periods of time between the experiments and even larger between measurements obtained from different LC/MS instruments.[33] As a result, same compound from different samples would not always have same m/z value and retention time even in the technical replicate.[34] Therefore, solving the different RT in detected peaks among different samples and aligning detected peaks across different samples properly becomes important for the further statistical analysis.

Current LC-MS peak alignment methods can be further divided into two categories: (1) aligning chromatograms,[30, 35, 36] and (2) aligning peak features.[13, 37-52] Peak alignment using direct chromatogram alignment aligns two chromatograms m/z-RT without identifying peaks.[37] Most common methods are warping or its modified methods such as correlation optimized warping (COW), [53] and dynamic time warping (DTW).[30, 35, 36, 54] Warping finds the best mapping relation between two time axes with the minimum distance by stretching or shrinking segments of chromatograms.[55]

The current mainstream LC-MS RT alignment is curve resolution, or aligning by peak features (m/z ,RT and intensity)[36]. It is much faster than warping methods[45 56]. Peaks with the most similar m/z and RT between samples would be considered as the same peaks.[37, 55] Algorithms such as XCMS,[31] MetaboAnalyst,[43] MZmine,[38] metalign,[48] MZmine RANSAC,[39, 45] apLCMS,[46] openMS,[47] msInspect,[13] SpecArray,[49] XAlign,[50] SuperHirn,[51] GPMS[57] and etc. all belongs to this category, so do commercial softwares, such as MS-resolver[58] (Pattern Recognition Software, Bergen, Norway), and MarkerLynx[59] (Waters, Massachusetts, USA).



However, aligning by peak features algorithms do suffer from different types of limitations. For example, peaks in XCMS might be allocated to neighboring m/z bins, and cause incorrect statistical results.[55] MetaboAnalyst[43] uses alignment method from XCMS, so suffer from the same problems that XCMS has. RANSAC can incorrectly assign matches when high numbers of features are extremely grouped.[60] Metalign could not process with more than 14 samples in its iterative mode and required considerable computing time in the rough mode.[60] MZmine and Xalign require considerable longer computing time when processing metabolomics data.[61] msInspect and SpecArray were reported poor performance on aligning metabolomics data when comparing with other alignment methods.[61] OpenMS were reported with low recall rate if inappropriate reference were selected.[62] SuperHirn requires tandem mass data for alignment.[51] Other than those mentioned above, all alignments are affected by the parameters used greatly, which would be different from one experiment to another and usually required manual selecting parameters such as m/z tolerance, RT tolerance or even penalty values used in the algorithm. Best results from different alignment algorithms usually would require experience or exhaustive evaluation of parameters for each sample.[63] We summarize the mentioned peak alignment algorithms in Table 1.1.

Table 1 Algorithm comparisons

Algorithm name	# of parameter	parameters	Parameter estimation	Issues
Mzmine	4	mz tolerance,weight for m/z , retention time tolerance, weight for RT	No	Long computing time
metalign	2	retention time region, maximal intensity	No	Can't process more than 14 samples
Mzmine RANSAC	6	m/z tolerance, RT tolerance after correction, RT tolerance, RANSAC iterations, minimum number of points, threshold value	No	Incorrectly assign matches when high numbers of features are extremely grouped
apLCMS	2	mz.tol , chr.tol	Yes	RT tolerance overestimated
openMS	3	m/z bucket, precision m/z, precision RT	No	Low recall rate due to inappropriate reference selection
msInspect	2	massWindow,scanWindow	Yes	Poor performance on metabolomics data
SpecArray	0	hard-coded	No	Poor performance on metabolomics data
Xalign	2	m/z variation, retention time variation	No	Long computing time
SuperHirn	2	mass / charge window, retention time window	No	Trapped in local maximum rather than global maximum
GPMS	2	m/z tolerance, RT tolerance	No	The parameter in distance function of nearest peak did not consider scale difference between m/z and RT
MS-resolver	N/A	N/A	N/A	Not free
MarkerLynx	N/A	N/A	N/A	Not free
XCMS	2	bw,mzwid	No	grouping with bin approach cause peak misalignment

One way to avoid manual parameter selection is to design methods that can automatic estimate parameters. The automated estimated parameters are usually generated by two approaches. One is performed by defining a stopping criteria for iterative searching optimal parameters.[64] The other is searching parameters that can give a minimum integrated squared error for distribution models of features in a chromatogram (e.g. m/z and retention time) for grouping similar m/z ions and retention time together for an alignment. [35, 38, 46] Current freely available software such as *MZmine*,[45] *apLCMS*[46] and *SuperHirn*[51] all provide such function. However, methods like apLCMS suffers from inherited expectation-maximization algorithm (EM) [65] problem that the optimization often are trapped at local optima therefore misalign ion

pairs,[65] RANSAC *in MZmine* would assign matches incorrectly when high numbers of features are present,[60] and Superhirn would require larger sample number for correct alignments.[66]

To provide a faster and easy-to-use LC-MS alignment that are not bound to manual selecting parameters and large sample size, we hereby propose a novel approach using Layer Alignment with Kernel density Estimation, *LAKE*. LAKE clusters peaks with similar m/z and RT values first to build pre-processed m/z-RT groups. LAKE estimates RT density for each m/z-RT groups using kernel density estimation (KDE) with unbiased cross-validation bandwidth selector (UCV), a function to estimate the allowed RT shift. LAKE can also solve multiple compounds with similar m/z and RT values. Unlike common alignments using KDE that requires a user input bandwidth to grouping similar RT values in an alignment, LAKE can estimate suitable bandwidth adaptively for similar RT values from chromatograms and therefore improve the alignment results. The performance of *LAKE* was evaluated with two data sets, 12 urine samples spiked with 50 forensic drugs with two concentrations and 251 plasma samples analyzed in 4 batched with 23 QC samples.

1.2 Materials

1.2.1 *Chemicals*

For the NTU MetaCore metabolomics chemical standards library, standards were purchased from Sigma-Aldrich (St. Louis, MO, USA). For the analysis of the NTU MetaCore metabolomics chemical standards library and the analysis of metabolomics study, MS-grade water was purchased from Scharlau (Sentmenat, Spain), and MS-grade acetonitrile was obtained from J.T. Baker (Phillipsburg, NJ). Acetic acid was purchased from Merck (Darmstadt, Germany). Reagents were analytical grade and they were

obtained from Sigma-Aldrich (Steinheim, Germany). Methanol (MeOH) were purchased from Scharlau (Mas d'en Cisa, Barcelona, Spain).

Amphetamine, alprazolam, 7-aminoflunitrazepam, aminorex, bromazepam, butorphanol, 4-bromo-2,5-dimethoxyphenethylamine (2C-B), clonazepam, chlordiazepoxide, clobazam, dihydrocodeine, diazepam, estazolam, fentanyl, flurazepam, flunitrazepam (FM2), heroin, ketamine, lorazepam, lormetazepam, LSD, methamphetamine, 4-methoxyamphetamine (PMA), 3,4-methylenedioxyamphetamine (MDA), 3,4-methylenedioxymethamphetamine (MDMA), 3,4-methylenedioxy-N-ethylamphetamine (MDEA), para-methoxymethamphetamine (PMMA), meperidine, methadone, midazolam, meprobamate, methylephedrine, methylphenidate, norketamine, norephedrine, nitrazepam, nordiazepam, nalorphine, oxazepam, pentazocine, phentermine, phencyclidine (PCP), prazepam, pseudoephedrine, temazepam, tramadol and zolpidem were purchased from Cerilliant (Round Rock, TX, USA). Cocaine hydrochloride, codeine, morphine were purchased from Sigma-Aldrich (St. Louis, MO, USA).

1.2.2 *Sample Preparation*

For the analysis of the NTU MetaCore metabolomics chemical standards library, standards were prepared in organic solvents at a concentration of 500 – 2000 ng ml^{-1}.

A total of 228 plasma samples consisting of three repetitions, each at two time points, from 38 patients were obtained from the National Taiwan University Hospital. Additionally, 23 QC samples were obtained by pooling aliquots from each plasma sample. A volume of 100 μL of human plasma sample was extracted with 400 μL of methanol containing 250 ng mL^{-1} of [D8]-phenylalanine (IS1) and [15N2]-theophylline (IS2) as ISs and the extraction was performed by shaking for 2 min using a

Geno/Grinder 2010 (OPS Diagnostics, LLC, NJ). The extract was then centrifuged at 15 000g for 5 min at 4 °C (Eppendorf Centrifuge 5415R).

The supernatant (375 μL) was collected in a new Eppendorf tube, and the pellet was re-extracted using the same protocol. The plasma extracts were pooled and dried using a centrifugal vaporizer for 4 h (Thermo SpeedVacSPD111 V, Waltham, MA). The residue was reconstituted with 200 μL of 50% methanol and centrifuged at 15 000g for 5 min. The supernatant was filtered with a 0.2-μm filter (Minisart RC 4, Sartorius, Goettingen, Germany) and subjected to LC/TOFMS analysis.

For forensic drug analysis, a sufficient amount of the standard solutions was added to drug-free urine to give spiked urine samples. The drug-free urine samples were donated by healthy volunteers and were verified to not contain drugs before use. Spiked urine samples (100 μL) were diluted with 400 μL of deionized water and centrifuged at 15,000 g for 5 min. The supernatant (200 μL) was then subjected 4 to LC/TOF-MS analysis.

1.2.3 *Chromatographic and Mass Spectrometric Analysis*

All standard mixtures, plasma samples, and forensic drugs were analyzed with an Agilent 1290 U-HPLC system with a 6540-QTOF (Agilent Technologies, Santa Clara, CA, USA) equipped with an electrospray ion source. Mass spectrometry calibration was performed daily before analysis by the infusion of a low concentration of a tuning mix (Agilent Technologies, USA). The mass resolution and the mass accuracies used in this study was 20 000 and 10 ppm, respectively.

For the NTU MetaCore metabolomics human metabolite standard library analysis, two microliters of the standard mixture was injected into an ACQUITY UPLC HSS T3 column (2.1 mm x 100 mm, 1.8 μm) (Waters, Milford, MA, USA). The mobile phase

was composed of 0.1% formic acid in water (solvent A) and ACN (solvent B). The gradient profile was as follows: 0–1.5 min: 2% B; 1.5–9 min: linear gradient from 2 to 50% B; 9–14 min: linear gradient from 50 to 95% B; 14–15 min: 95% B. The flow rate was maintained at 0.3 mL/min. For sample ionization, a Jet Stream electrospray ionization source operated with a capillary voltage of 4,000 V in positive mode and 3,500 V in negative mode. The mass scan range was m/z = 50-1700 m/z. The micro-channel plate detector voltage was set at 720 V. During the analysis, reference masses of 121.0509 and 922.009798 as well as 112.9856 and 966.0007 m/z were used for positive and negative mode mass accuracy correction, respectively.

For plasma analysis, sample was injected into an Acquity HSS T3 column (2.1 mm x 100 mm, 1.8 μm; Waters, Milford, MA) and the column was maintained at 40 °C. The mobile phase was composed of solvent A, water/0.1% formic acid and solvent B, acetonitrile/ 0.1% formic acid. The gradient elution program was as the following: 0−1.5 min, 2% B; 1.5−9 min, linear gradient from 2 to 50% B; 9−14 min, linear gradient from 50 to 95% B; and hold at 95% B for 3 min. The flow rate was 300 μL min^{-1}. For sample ionization, a Jet Stream electrospray ionization source was used with a capillary voltage of 4.0 kV in positive mode. The MS parameters were set as follows: gas temperature, 325 ˚C; gas flow, 5 L/min; nebulizer, 40 psi; sheath gas temperature, 325 °C; sheath gas flow, 10 L/min. A scan range of 50−1700 m/z was set.

For forensic drug analysis, five microliters of sample was injected into an Agilent Poroshell ECC18 column (2.1 mm x 100 mm, 2.7 μm). The mobile phase was composed of 0.1% acetic acid in water (solvent A) and MeOH (solvent B). The gradient profile used for positive ionization detection was as follows: 0–1 min: 2% B; 1–10 min: linear gradient from 2 to 50% B; 10–15 min: linear gradient from 50 to 90% B; 15–17 min: 90% B and then re-equilibration of the column for 3 min. The flow rate was

maintained at 0.4 mL/min, and the injection volume was 5 μL. The parameters of the mass spectrometer for positive and negative ionization mode detection were as follows: sheath gas temperature, 325℃; 5 sheath gas flow, 11 L/min; nebulizer, 45 psi; capillary voltage, 3,000 V; gas temperature, 325℃; drying gas flow, 6 L/min; nozzle voltage, 1,000 V and TOF-MS scan range, 50-1000 m/z. The micro-channel plate detector voltage was set at 720 V. For sample ionization, a Jet Stream electrospray ionization source was operated with a capillary voltage of 3,000 V in positive mode. The flow rate was 0.4 mL/min. The mass scan range was m/z = 50-950.

1.2.4 *Data Preparation*

MS data were converted to mzXML format in centroid mode using Trapper (ISB)[67] The input peak lists for LAKE are processed by XCMS. The centWave[27] method from XCMS was used to detect the peaks with the following parameters, ppm=25,peakwidth=(5,20), and snthresh=5.[68] All further processes and further data analysis were performed using R statistical environment (version 3.1.2).[69]

1.3 Theoretical Basis

Overall Procedures in Main Algorithm

Overall, LAKE comprises of three steps to generate aligned peak table for peak lists includes the following: (a) grouping peaks of technical replicates from the same sample; (b) grouping peaks of samples from the same batch; (c) grouping peaks of batches from the same experiment. The overall procedure of *LAKE* is illustrated in Figure 1.1.

Most alignment treat each sample as equivalent and this is based on the assumption of there are no global shift exist among different batches (Figure 1.1a). LAKE, on the

other hand, group sample based on data similarity, from grouping peaks from the same sample, grouping peaks from the same batch to grouping peaks from the same experiment (Figure 1.1b). LAKE is designed for better processing when data with batch difference or shift in either m/z or RT among batches.

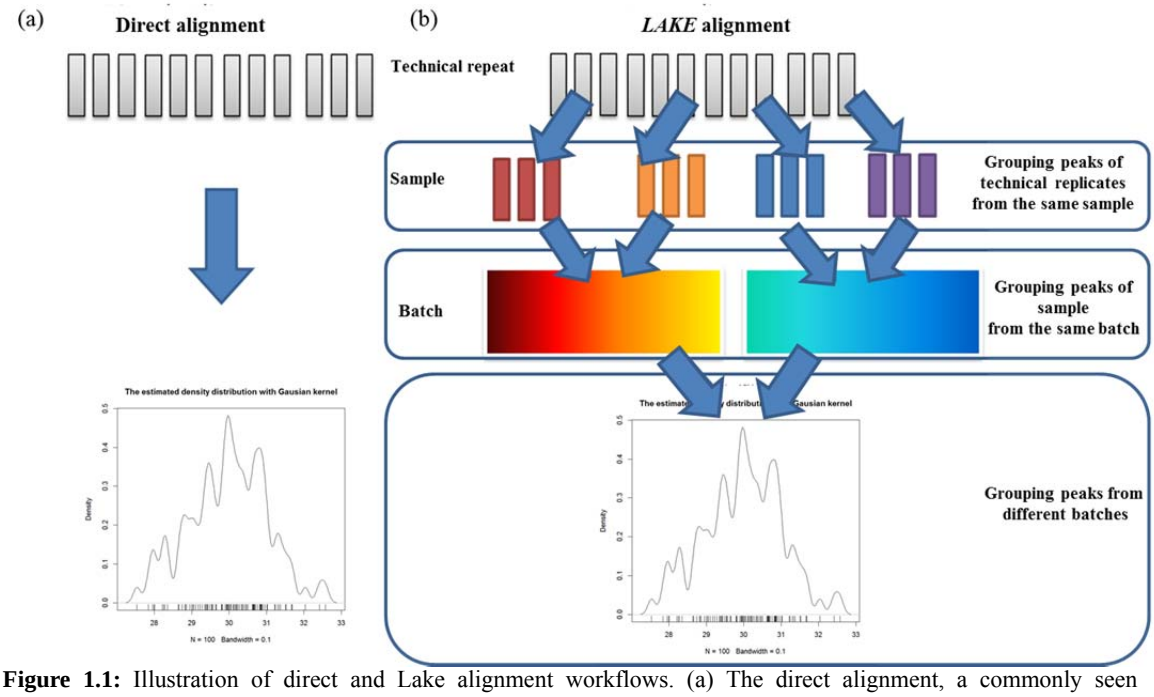


Figure 1.1: Illustration of direct and Lake alignment workflows. (a) The direct alignment, a commonly seen alignment workflow. First, align all technical repeats. Then, generate a peak table. (b) The LAKE alignment. First, align technical repeats from the same sample. Next, align the samples from the same batch. Finally, align all batches and generate a peak table.

1.3.1 *Grouping Peaks of Technical Replicates from the Same Sample*

The first step in LAKE is to group peaks from the same sample with similar m/z and RT values. After the process is done, the m/z value would be updated with average m/z value of the m/z-RT cluster to reduce data complexity on later processing. The following processing is to cluster peaks into m/z-RT groups and update m/z value of peaks with average m/z value of m/z-RT group containing this peak. The process includes: (a) group of detected peaks with similar m/z values construction with suggested relative mass difference tolerance and RT tolerance (Figure 1.2a); (b) RT

regrouping with KDE (Figure 1.2b).

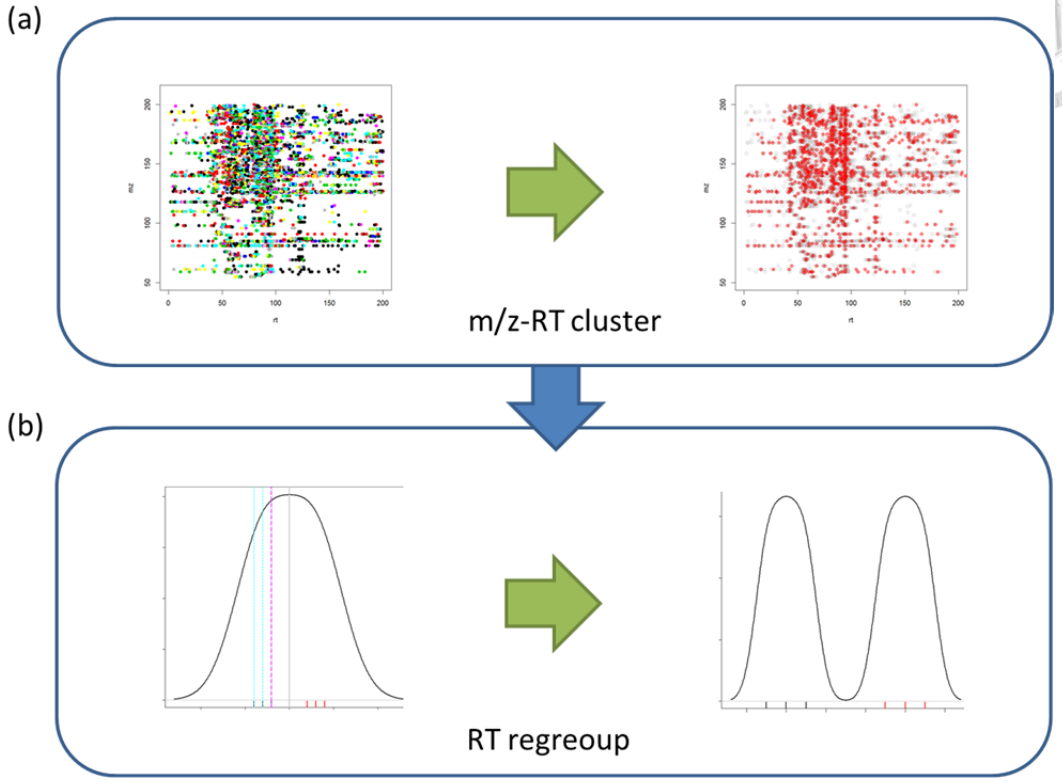


Figure 1.2: The workflow of grouping peaks of technical replicates from the same sample. (a) The workflow of grouping of detected peaks of similar m/z values and RT values construction with suggested relative mass difference tolerance and RT tolerance, respectively. (b) The workflow of RT regrouping with kernel density estimation.

A. Detected Peaks of Similar m/z Values and RT Values Grouping with Suggested M/Z Tolerance and RT Tolerance

In this step, we group peaks with similar m/z and RT values from the same sample. The precise average m/z and RT is an important issue here because the inaccurate average m/z value of peak group caused by inappropriate peak grouping used in the following peak clustering would lead to severe peak misalignment. The average m/z of peak group processed by group.mzclust [70] using error window in parts per million rather than binning using error window in fixed size. However, group.mzclust is still not precise enough for containing peaks from other compounds with similar m/z values but

RT values. So the idea is to apply group.mzclust algorithm on the RT dimension of the processed m/z group before output from group.mzclust to cluster peaks with different RT distribution into different m/z-RT groups for precise average m/z of m/z-RT peak group. The process can be illustrated in upper panel of Figure 1.2. In this step data complexity is reduced by updating m/z value of detected peak (dots in Figure 1.2a left) with average m/z and average RT of m/z-RT group containing this peak (red dots in Figure 1.2a right) after m/z-RT clustering.

The input of the group.mzClust is detected peak list. The detected peak list generated from any peak detection algorithms need to convert into a matrix with at least 3 columns in comma-separated values (CSV) format. The m/z, the RT and the intensity should be stored in first, second and third column of the matrix respectively. The CSV peak list would be imported to R statistical programming language.[69] The processed m/z group is the input for algorithm of RTclusteringOnMzclusteredGroup and the outputs from the algorithm are peak groups with precise average m/z value. The process can be briefly described as follow: (1) Initialization: read in RT values of the peaks clustered in m/z clustered bin to be aligned in an array, RA. Then, form initial cluster using RT peaks from RA using a fixed error window (pseudo code line 2). (2) Form next cluster: form another cluster RT peak from RA using a fixed error window (pseudo code line 4). (3) Two clusters processing: two clusters are evaluated to see if any overlapped between the first cluster and the second cluster, or if any outlier existed in first cluster (pseudo code line 5 to 15). If these situations existed, these can be done by applying Hierarchical Clustering on these cluster to solve either overlapped or outlier (pseudo code line 7 and 17). (4) Output the processed m/z-RT clusters and leave last unprocessed group as first cluster and form second cluster RT peak from RA using a fixed error window (pseudo code line 14, 24 and 25). (5) Repeat previous step 2-4 until no more data left in the

samples. The pseudo code is shown in Figure 1.3.

```
Algorithm RTclusteringOnMzclusteredGroup(R,e)
Input:    RT values of the peaks from the same m/z cluster are sorted in ascending
          order and stored in array R.
          e is the RT tolerance.
Output:   list BR contains m/z-RT groups with proper RT deviation.
1   cur ←1;
2   B_RA←FormRTBin(R,e,cur)
3   while cur < | R |
4       B_RB←FormRTBin(R,e,cur);
5       if CheckBinOverlap(B_RA,B_RB) =TRUE then
6           B_RC←MergeBin(B_RA,B_RB);
7           Result←HierarchicalClustering(B_RC);
8           if |Result| ≥ 1 then
9               B_RA ←Result[|Result|];
10          else
11              B_RA ← NULL;
12          if |Result| ≥ 2 then
13              for i=1 to |Result| -1 do
14                  BR←Add(BR, Result[i] );
15      else if CheckOutlier(B_RA)=TRUE then
16          B_RC←B_RA;
17          Result=HierarchicalClustering(B_RC);
18          if |Result| ≥ 1 then
19              B_RA ←Result[|Result|];
20          else
21              B_RA ← NULL;
22          if |Result| ≥ 2 then
23              for i=1 to |Result| -1 do
24                  BR←Add(BR, Result[i] );
25      BR←Add(BR, BR_A);
26      B_RA ←B_RB;
27  Output BR;
```

Figure 1.3: Pseudo code for algorithm of RTclusteringOnMzclusteredGroup

The original algorithm can be referenced in description in algorithm.[70]

B. RT regrouping with KDE

After m/z-RT peak groups are generated, the m/z-RT peak groups still need to be refined to peak group with less m/z deviation for later updating m/z values with average m/z value of m/z-RT peak group. Because the m/z-RT groups may contain more than

one m/z-RT distribution for having similar m/z values but RT values from different distribution which can't be separated with the RT tolerance used in 1.3.2.a. To make m/z-RT peak groups with smaller deviation in the m/z domain, we calculate the multimodality to get proper bandwidth for KDE on the RT distribution for each m/z-RT peak group to separate one multimodal m/z-RT distribution to multiple m/z-RT distributions.

KDE is a better density estimation when comparing to histogram because histogram has two problems which are estimated density is not smooth and estimated density depends on end points of bin.[71] There are two main factors in affecting the estimated density: kernel type and bandwidth.

In previous studies,[72] different types of kernel functions are introduced with different abilities to describe different data distributions. Figure 1.4 is density distributions of estimated RT using different kernels when data containing noise.

Figure1.4a is the estimated RT distributions with bandwidth h = 0.3 using the Epanechnikov, the Gaussian and the triangular kernel. The other two except for Epanechnikov are both inferred by the noise and the mean of estimated RT distributions become farther away from the original mean of the RT density. It suggests the boundless concept of Gaussian kernel and the sharp response of triangular kernel both are not robust enough. The Epanechnikov kernel is selected for having both features from smooth shape of Gaussian kernel and triangular kernel with limit.

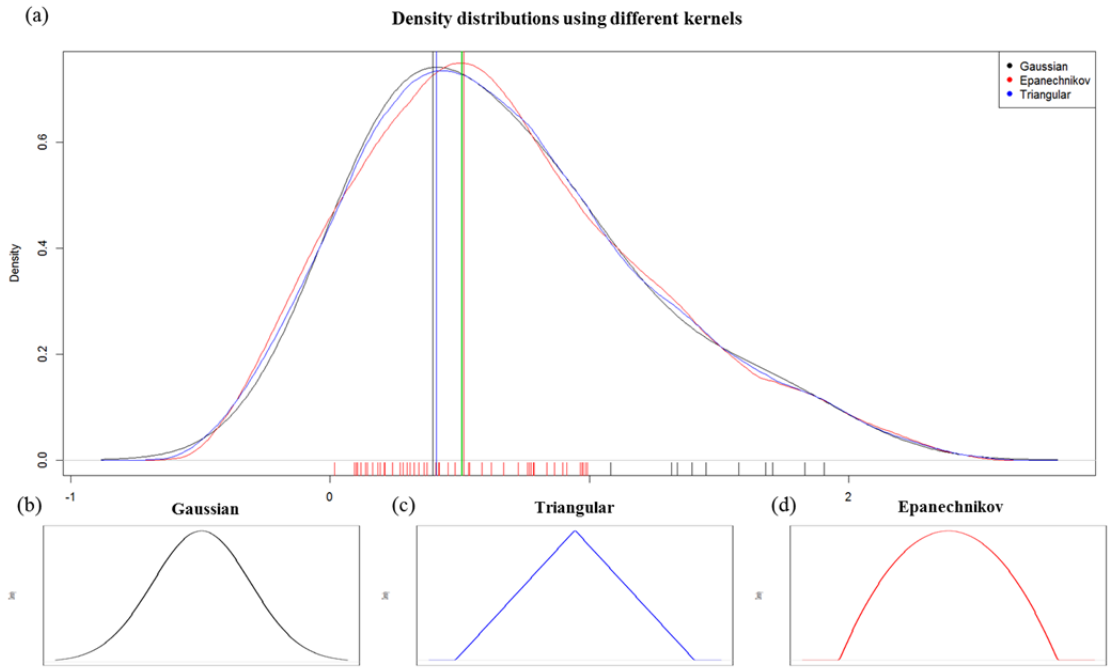
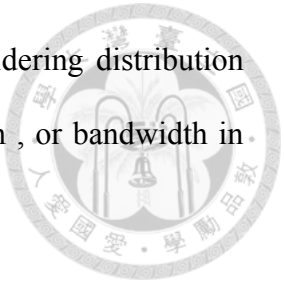


Figure 1.4: Density distributions of estimated retention time using different kernels: (a) the estimated density distribution with bandwidth h=0.3 using Epanechnikov, Gaussian and triangular kernel; (b) the Gaussian kernel; (c) the triangular kernel; (d) the Epanechnikov kernel. The mean of estimated density distribution using Epanechnikov (red vertical line) is closest to the mean in the real data (green vertical line is the mean of red data points in the rug plot). It is more robust even noise (black data points in the rug plot) exists.

The bandwidth selector only works with larger data size, not the case with data size about the number of technical replicate of a sample has. So we have proposed another bandwidth selecting algorithm, RTRegroupByCheckingMultimodal, to solve this problem. The whole process is containing the following process can be described as follows: (1) Mode number calculation: The mode number of *i*th m/z-RT group (m_i) is calculated with equation 1 to see if this m/z-RT group needs RT regroup,

$$m_i = \frac{n_i}{r_i} \tag{1}$$

where n_i is the peak number in the *i*th m/z-RT group and r_i is the replicate number in this sample in the *i*th m/z-RT group. If m_i is higher than 1.5, further processing is needed (grey lines in Figure 1.2b). (2) Bandwidth calculation: the new bandwidth for the RT density of the *i*th m/z-RT group (h_i) can be calculated with m_i with equation 2



and solve multimodal in RT density of the m/z-RT group by considering distribution interval can be approximated by about 6 times of standard deviation , or bandwidth in our case.

$$h_i = \frac{(\max(RT_i) - \min(RT_i))}{ceiling(m_i)/2/3} \tag{2}$$

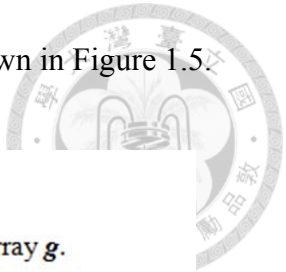
where RT_i is the set of RT values of the peaks in ith m/z-RT group (h_i : the interval between two cyan dash lines). (3) Density estimation: RT density is estimated by KDE with Epanechnikov kernel and bandwidth h_i,. The start point and end point of estimated RT density are extended with 3 times of standard deviation of RT values of the peaks in the m/z-RT group to avoid RT values fall on boundaries and treated as outliers in later processing. To avoid overfitting, the estimated density is applied with moving average with window size (w_i) on the estimated RT density.

$$w_i = \max(round(h_i / k_i), 2) \tag{3}$$

$$k_i = (\max(RT_i) - \min(RT_i))/512 \tag{4}$$

where k_i is the resolution of estimated RT density. (4) Mode detection: The locations of multiple modes in estimated RT density are determined by local maximal detection. The process is iterative and only process the largest mode (mode with highest density) of current estimated RT density. The stopping criterion of local maximal detection is set as 65% of the original estimated RT density. (5) Peak grouping: The interval is decided by two points from the highest point of the detected mode slide down from each side until the density goes up again. The RT values fall within this interval are collected into the newly created m/z-RT group. The density within the interval is then set to zero. Assigning peaks into detected mode with highest density of the current estimated RT density is repeated until the highest density of the current estimated RT density is lower than the threshold. The new m/z values for peaks in each m/z-RT group are updated

with average m/z value of the m/z-RT group. The pseudo code is shown in Figure 1.5.



```
Algorithm RTRegroupByCheckingMultimodal(g,n,r)
Input:    RT values of peak from the same m/z-RT group stored in array g.
          n is the number of sample with peak in the m/z-RT group g.
          r is the number of technical replicate from the same sample.
Output:   Generate processed m/z-RT peak groups stored in list G with smaller
          peak group deviation in m/z dimension.
1   nModes ← n / r ;
2   if nModes >1.5 then
3      h ← (max(g)-min(g))/ceiling(nModes)/2/3;
4      from←min(g)-3*sd(g)
5      to←max(g)+3*sd(g)
6      K ← KDE(h , g , Epanechnikov, from,to) ;
7      res ← (max(g)-min(g))/512 ;
8      w ← max(h/res, 2) ;
9      sK ← MovingAverage(K,w) ;
10     G ← PeakDetection(sK) ;
11  else
12     G ← g ;
13  Output G ;
```

Figure 1.5: Pseudo code for algorithm of RTRegroupByCheckingMultimodal

1.3.2 *Grouping Peaks of Sample from the Same Batch*

The second step of LAKE is to group peaks from the same batch with similar m/z and RT values. In this step, we can get possible compound with enough data size to get m/z and RT distribution of aligned compound and peak table can be generated at this stage. The process includes: (a) group of detected peaks with similar m/z values construction with suggested relative mass difference tolerance and RT tolerance (Figure 1.6a); (b) RT distribution estimation on detected peaks with similar m/z values by kernel density estimation (Figure 1.6b); (c) peak table regroup (Figure 1.6c).

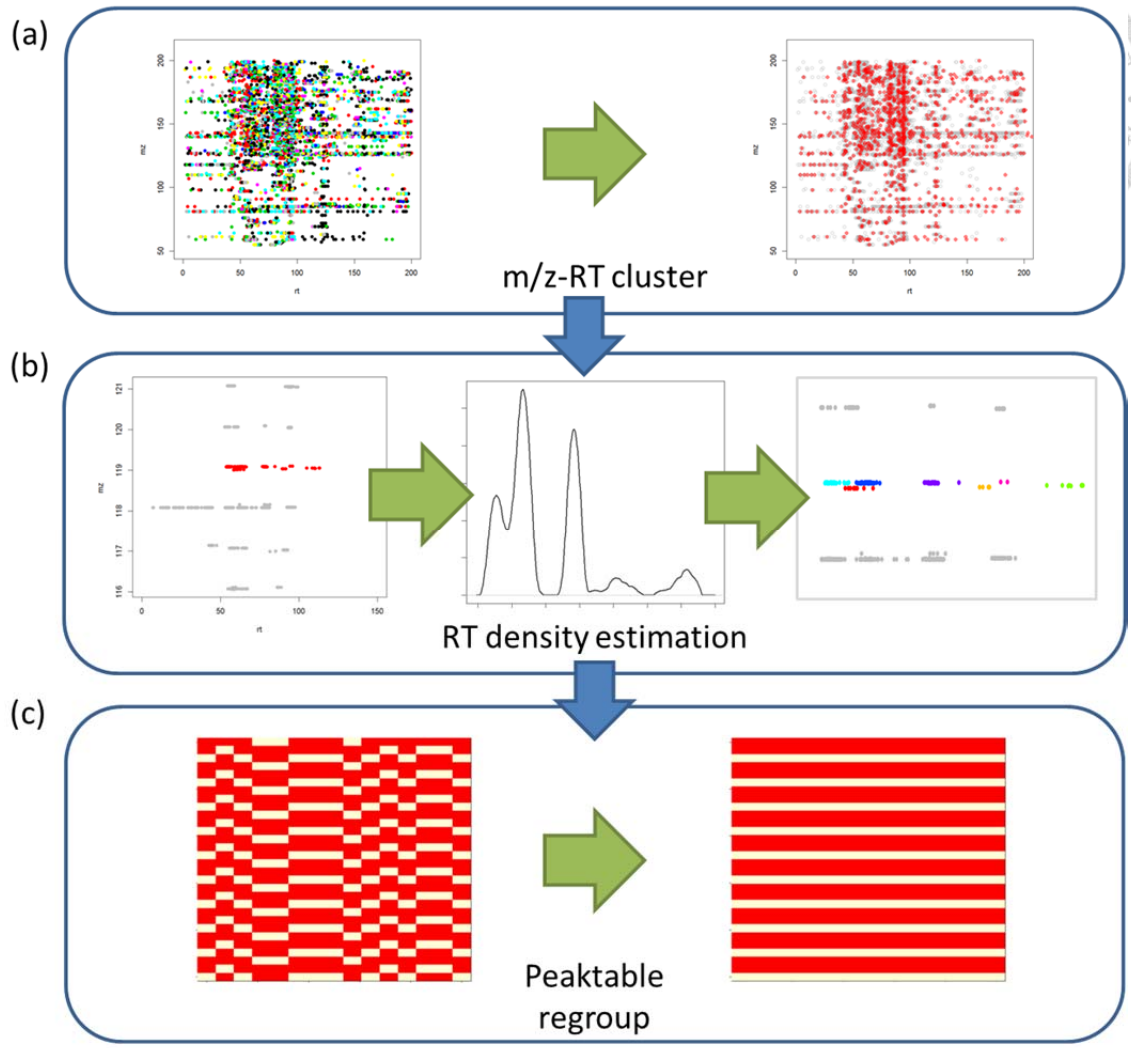


Figure 1.6: The workflow of grouping peaks of sample from the same batch. (a) The workflow of grouping detected peaks with similar m/z values construction with suggested m/z tolerance and RT tolerance. (b) The workflow of RT distribution estimation on detected peaks with similar m/z values by kernel density estimation. (c) The workflow of peak table regroup.

A. Detected Peaks of Similar m/z Values and RT Values Grouping with Suggested M/Z Tolerance and RT Tolerance

We take the result from previous stage, grouping updated m/z and RT values of peaks from the same batch. The processing is the same as we mentioned in 1.3.1.a. The m/z-RT groups containing peaks from the same batch are then generated.

B. RT Distribution Estimation on Clustered Peaks with Similar M/z Values by KDE

To reveal the potential compounds in m/z-RT groups (Figure 1.6b rightmost subfigure), we apply local maximal detection on the KDE-estimated RT density for each m/z-RT peak groups (Figure 1.6b second subfigure). The process is similar to 1.3.1.b except for the bandwidth used in KDE is decided by bandwidth selector, unbiased cross-validation bandwidth selector.

The bandwidth selection is especially important for selecting smaller or larger bandwidth would cause estimated density undersmoothed or oversmoothed, respectively. Both can cause problem when identify the modes in estimated density of the data. The existed bandwidth selectors are try to find the bandwidth which minimizes the optimality criterion asymptotic mean integrated squared error (AMISE). Three categories bandwidth selector are NRD (Silverman's 'rule of thumb'), cross-validation, and plug-in method. The NRD and plug-in methods are often oversmooth estimated density while cross-validation method is tend to undersmooth estimated density.[73] The cross-validation method is used because it's easier to merge peak groups rather than divide the group with unknown number of modes existed in the data.

The algorithm, RTRegroupByUCV, is almost the same as 1.3.1.b but for unbiased cross-validation (UCV) bandwidth selector is used in this process (pseudo code line 1). To solve undersmoothed or oversmoothed when inappropriate bandwidth is calculated by UCV, we set boundaries on UCV calculated bandwidth. The bandwidth (h_{ucv}) is generated by using UCV as bandwidth selector with bandwidth boundaries (h_{min}, h_{max}), the acceptable range of bandwidths (pseudo code line 2).

$$h_{ucv} = \begin{cases} h_{min} & , \text{when } h_{ucv} < h_{min} \\ h_{ucv} & , \text{when } h_{min} \le h_{ucv} \le h_{max} \\ h_{max} & , \text{when } h_{ucv} > h_{max} \end{cases} \quad (5)$$

where h_{min}=2 , h_{max}=5 are decided from experience. The pseudo code is shown in Figure 1.7.

Algorithm RTRegroupByUCV(g)
Input: RT values of peak from the same m/z-RT group stored in array g.
Output: Generate processed m/z-RT peak groups stored in list G with smaller deviation in RT dimension.

```
1   bw ← UCV(g) ;
2   bw ←AcceptableBwCheck(bw) ;
3   from←min(g)-3*sd(g)
4   to←max(g)+3*sd(g)
5   K ← KDE(bw,g,Epanechnikov,from,to) ;
6   Res ← (max(g)-min(g))/512 ;
7   w ← max(bw/Res,2) ;
8   sK ← MovingAverage(K,w) ;
9   G ← PeakDetection(sK) ;
10  Output G ;
```

Figure 1.7: Pseudo code for algorithm of RTRegroupByUCV

C. Peak Table Regroup

The misaligned peak group is a common problem when grouping with error window in either relative mass difference or in fixed absolute mass difference. The misaligned peak groups can be found by checking if any peak group with similar average m/z and average RT in the m/z-RT groups (Figure 1.6c left). Therefore, the misaligned peak groups can be regrouped based on the feature and processed in following algorithm, MergingMisalignedGroupdinPT : (1) Calculating coverage rate for each m/z-RT group (CR_i) with equation 6.

$$CR_i = \frac{n_i}{N} \tag{6}$$

where n_i is the number of sample with peak grouped in ith m/z-RT group, N is the total number of sample in the batch (pseudo code line 5). (2) Finding the m/z-RT group with

coverage rate lower than 80% and sorting m/z-RT group by their coverage rate in descending order and stored in list L1. The peak table, P0, only contains m/z-RT groups with coverage rate higher than or equal to 80% (pseudo code line 6-9). (3) For the m/z-RT group with highest coverage rate (PG_A) in L1, find m/z-RT groups with similar average m/z and average RT with m/z tolerance +/-15 ppm and RT tolerance +/- 6 sec from experience. Store these groups into list L2 and sorted by their coverage rate in descending order (pseudo code line 12-15). Then stored PG_A into list L3. (4) If there is more than one group in L2, start from the group with highest coverage rate in L2, if not, proceed to step 5. For m/z-RT group with the highest coverage rate in L2 (PG_B), calculate how many peaks would be overlap ratio if PG_A and PG_B are merged ($OR_{A,B}$) with equation 7 (pseudo code line 19).

$$OR_{A,B} = \frac{\sum_i n_{O(A_i,B_i)}}{N}$$
$$n_{O(A_i,B_i)} = \begin{cases} 1 & \text{if } A_i > 0 \ \& \ B_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $n_{O(A,B)}$ is the number of overlapped peak number when merging PG_A and PG_B (pseudo code line 18). If the $OR_{A,B}$ is under 50%, these two groups are merged update the coverage rate, and average m/z and average RT of PG_A and store PG_B in L3. For tie-breaking in merging two peak groups, the peak with highest intensity is used. Then, remove PG_B from L2. Otherwise no merged is applied, remove PG_B from L2. This step is repeated until there is no group in L2. Stored the merged peak group into L0 (pseudo code line 16-24). (5) Remove peak groups were merged into the peak group from L1 (pseudo code line 25). (6) Repeat step 3-5 until there is no peak group in L1. (7) Arrange these merged peak groups in L0 into peak table and attached with rest of peak table (pseudo code line 26-27). The pseudo code is shown in Figure 1.8.

Algorithm MergingMisalignedGroupsinPT(P,N, sp)
Input: peak table P with N aligned compounds with sp samples.
Output: Generate peak table $P0$ with misaligned compounds properly merged.

```
count←1
for i = 1 to N do
    Gi←GetGroup(P,i);
    ni ← SampleNumberInPeakGroup(Gi) ;
    CR[i]← ni / sp ;
    if CR[i]< 0.8 then
        L1←Add(L1,Gi);
    else
        L0←Add(L0,Gi);
L1←Sort(L1,CR,descending);
while |L1| > 0
    PGA←GetHighestCR (L1) ;
    L2←FindPossibleMisalignedPeakgroup(PGA) ;
    L2←Sort(L2,CR,descending);
    L3←Add(L3,PGA);
    while |L2| > 0
        PGB←GetHighestCR (L2) ;
        no(A,B) ←OverlappedPeakNumber(PGA,PGB);
        ORA,B← no(A,B)/ N ;
        if ORA,B< 0.5 then
            PGB←Merge(PGA ,PGB );
            L3←Add(L3,PGB);
        L2←Remove(L2,PGB);
L0←Add(L0,PGA);
L1←Remove(L1,L3);
P0←GeneratePeakTable(L0) ;
Output P0;
```

Figure 1.8: Pseudo code for algorithm of MergingMisalignedGroupdinPT.

1.3.3 *Grouping Peaks from Different Batches*

The third step of LAKE is to group peaks from the same sample with similar m/z and RT values. After the process is done, peak table is generated. The process includes: (a) group of detected peaks with similar m/z values construction with suggested relative mass difference tolerance and RT tolerance (Figure 1.9a); (b) Peak table regroup (Figure

1.9b).

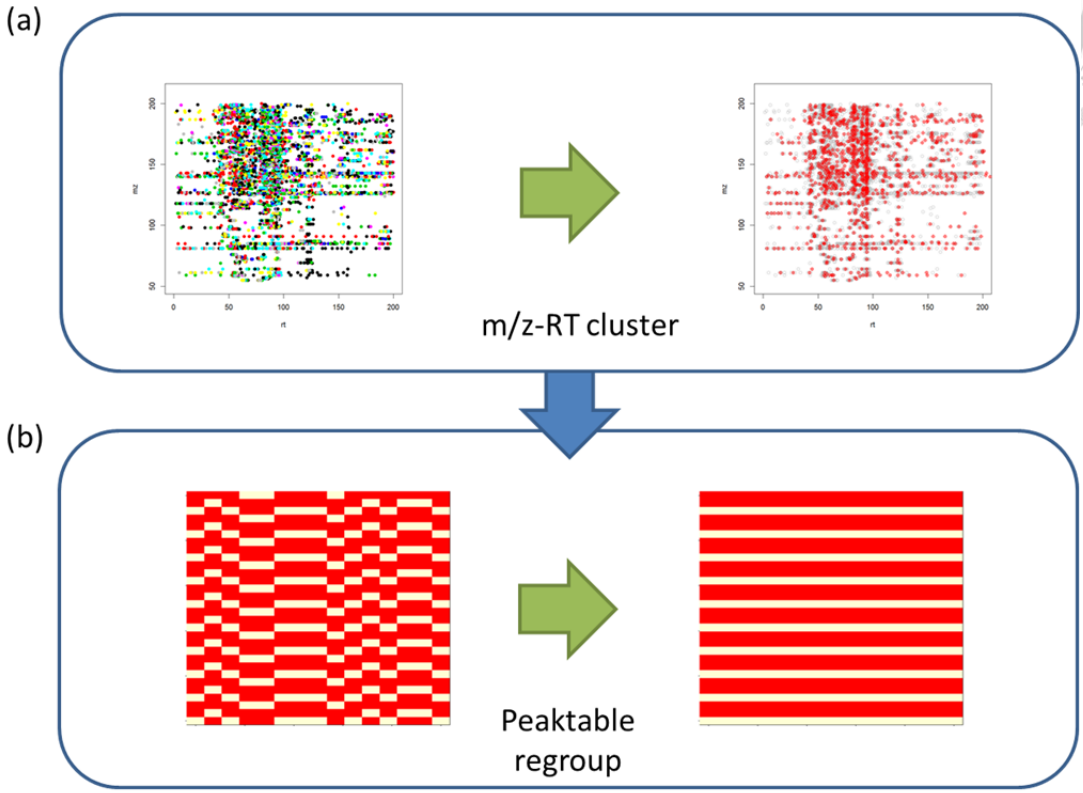


Figure 1.9: The workflow of grouping peaks from all batches. (a) The workflow of grouping detected peaks with similar m/z values construction with suggested m/z tolerance and RT tolerance. (b) The workflow of peak table regroup.

A. Group of Detected Peaks with Similar m/z Values Construction with Suggested M/Z Tolerance and RT Tolerance

In this step, average m/z and average RT of m/z-RT group from each batch are treated as m/z and RT of detected peaks in this new peak list. The intensity of the new peak list is replaced by the id number of the m/z-RT group in the batch. The peak list from each batch is now grouped by the m/z-RT clustering algorithm we mentioned in 1.3.1.a. After the alignment result is generated, the m/z-RT groups from different batch are grouped according to the alignment result.

B. Peak Table Regroup

After alignment of the average peak group list from each batch is done, peak table

regroup we mentioned in 1.3.2.c is applied to the result we just processed. Then the peak table for the experiment is generated.

1.3.4 *Performance Evaluation of LAKE on Peak Alignment*

In the already proposed alignment algorithms, they are often divided into two types, peak list alignment and chromatogram alignment. Our proposed algorithm is in the peak list alignment, so we select the most commonly used algorithm in this category, XCMS with default alignment algorithm.[31] The algorithm is implemented in a function name group.density in XCMS package.

1.3.4.1 Performance Evaluation of LAKE on Peak Alignment

To evaluate the performance of LAKE for untargeted peak alignment, we spiked 50 forensic drugs with two concentrations (high, 10 times of cut off values listed in Table A-1; and low, 1 time of cut off values of each forensic drug) into 6 urine samples. Among them 50 spiked forensic drugs, 47 well-behaved, with good peak shape and can be detected in most of the samples, spiked detected compounds are used as the evaluation data set.

1.3.4.2 Performance Evaluation of LAKE on noise introduced peak alignment.

To evaluate the performance of LAKE for noise existed data set, a metabolomics data set is selected for analyzed in four batches and large global shift in RT exist between 4^{th} batch and other batches which make some aligned compounds with most misaligned peaks from 4^{th} batch after using existed alignment algorithm (Figure 1.10). In metabolomics data set, three technical replicates for each sample, average peak number in each replicate is 3,837, total sample number is 76 and 20 samples in each

batch. The batch number for this experiment is four.

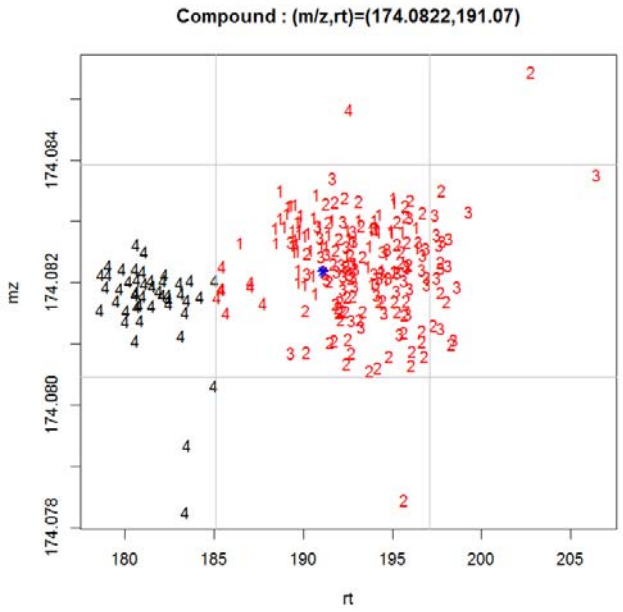


Figure 1.10: Misaligned peaks in 4th batch of the aligned compound (m/z, RT) = (174.0822, 191.07) in positive ionized mode of the metabolomics data set. Different colors represent peaks from different groups. The red peak group contains peaks from 1st batch to 3rd batch and some peaks from 4th batch. The black peak group contains peaks from 4th batch for peaks in 4th batch with larger batch difference in RT dimension which makes peak misalignment when using existed alignment algorithm.

The metabolomics data set is used to generate a ground truth for evaluation when the data set is introduced with two different kinds of noises, local shift and global shift among batches. The local shift is that the deviation curve of RT and the deviation curve of m/z added to the peaks of each technical replicate. The global shift among batches is that time offset and m/z offset added to the peaks from the same batch.

To generate ground truth, we find peaks alignment result can be seen in both XCMS and LAKE algorithm. We then compare these results and pick out aligned compounds with following criteria: (1) The aligned compound consisted of exactly the same peaks in both aligned results; (2) The aligned compound consisted of peaks from all samples. We then got 130 aligned compounds are satisfied with the criteria

mentioned above. The 130 aligned compounds were used to evaluate peak alignment when noise exists.

To generate local shifts, the m/z and RT values from the 130 selected aligned compounds are used for calculated linear model between mean and standard deviation for m/z and RT, respectively. Take m/z for example, we take the slope and intercept, two coefficients from the linear model of m/z we created in last step, with following formula we can calculate ideal standard deviation for specific m/z value *v*. The standard deviation of RT is calculated with the same method by replacing the formula with two coefficients from the linear model of RT we created in last step.

$$sd_v = a_{lm} \times v + b_{lm} \tag{8}$$

where a_{lm} is the slope of linear model, b_{lm} is the intercept of linear model, and v is the value you want to calculate for its standard deviation. The calculated standard deviation is used for noise generating. For example, the noise of specified value v is generated by picking one value from the normal distribution centered at v with standard deviation sd_v. We compute standard deviation for each m/z and RT in peak list. The noise of v is then added back to the v to get the local shift noise introduced data set. The difference between the original data and the data with local shift noise introduced can be seen in Figure 1.12a. The Gaussian based local shift noise is derived from the aligned compounds, so the deviation is acceptable.

To generate global shift among batches, we observed data and get the possible range of global shift among batches. The global shift among batches is based on the 130 selected compounds used as ground truth. The batch difference in m/z and RT dimension is added according to the observation of aligned compounds in QC samples shown in Figure 1.11. The global shift among batches introduced data set in m/z and RT still are within acceptable deviation range. The local shift noise in m/z and RT

introduced in six data sets are the same, the only difference among these data sets is the batch difference in m/z and RT. There are six different global shift noise introduced in the data set: (1) alleviate batch difference in both m/z ad RT, (2) exaggerate batch difference in both m/z and RT, (3) only alleviate batch difference in m/z,

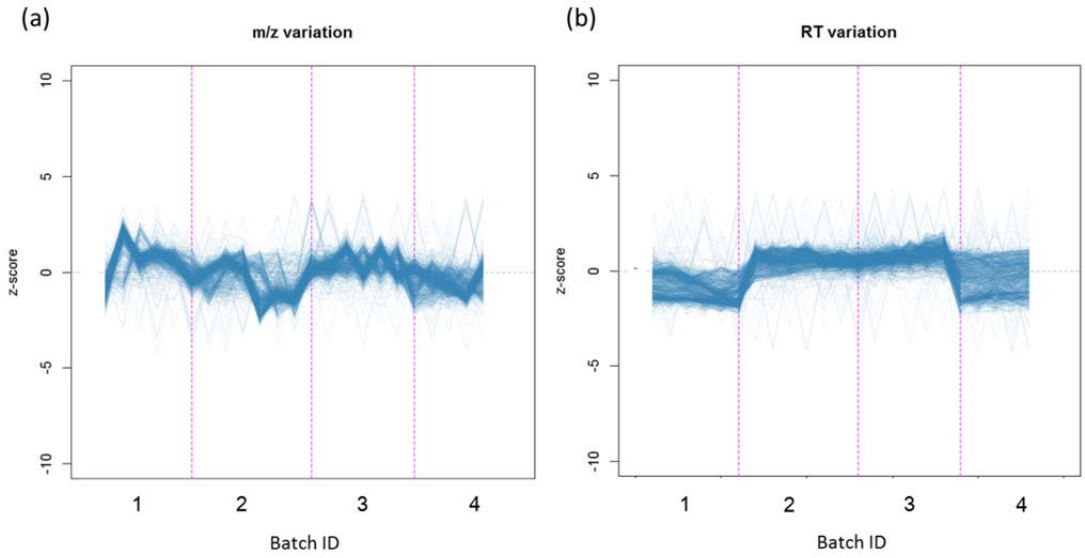


Figure 1.11: The batch difference in the aligned compounds of QC samples. Each blue line represents z-scores of an aligned compound over all QC samples. (a) The m/z variation among batches in z-score. X-axis is the injection order of QC samples. (b) The RT variation among batches in z-score. X-axis is the injection order of QC samples.

(4) only exaggerate batch difference in m/z, (5) only alleviate batch difference in RT and (6) only exaggerate batch difference in RT. The global shift among batches in m/z can be ranged from -10 to 10 ppm while the global shift among batches in RT can be ranged from -4 to 4 second. For example, the global shift among batches applied in one of data set in m/z is (4,-4,0,-4) which means for the m/z values in 1^{st}, 2^{nd} ,3^{rd} and 4^{th} batch are added with 4,-4,0 and -4 ppm, respectively. The global shift among batches in one of data set in RT is (-2,0,0,-2) which means for the RT values in 1^{st}, 2^{nd} ,3^{rd} and 4^{th} batch are added with -2, 0, 0 and -2 second, respectively. The difference between original data and data with both local shift noise and global shift among batches introduced can be seen in Figure 1.12b.

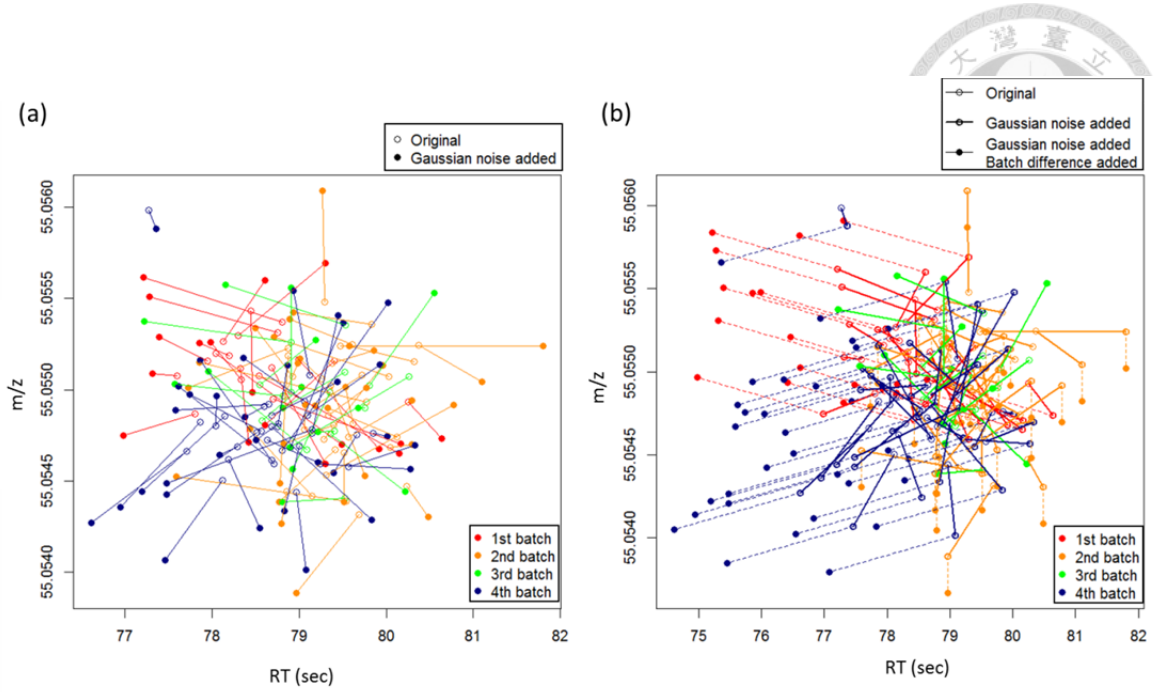


Figure 1.12: The noise introduced data sets. (a) The data set with only local shift in both m/z and RT dimension. (b) The data set with both local shift in both m/z and RT dimension and global shift among batches in both m/z and RT dimension.

The name of different type of noise introduced metabolomics data set is named after how we process the data set and described as follow: (1) local shift noise introduced data set: GNLM-1, GNLM-2, GNLM-3 and GNLM-4. (2) global shift among batches introduced data set: BR+GNLM1-1, BR+GNLM1-2, BR+GNLM1-3, BR+GNLM1-4, BR+GNLM1-5 and BR+GNLM1-6. The global shift introduced in m/z and RT for each batch is list in Table 1.2. The effect of different global shift among batches affect data is shown in Figure 1.13. The data set BR+GNLM1-2 and BR+GNLM1-3 are to show how manipulate m/z and RT difference among batches would affect alignment result. The data set BR+GNLM1-4 and BR+GNLM1-5 are to show how manipulate m/z difference among batches would affect alignment result. The data set BR+GNLM1-6 and BR+GNLM1-7 are to show how manipulate RT difference among batches would affect alignment result.

Table 2 Different Global shift introduced to each batches for different data sets

	global shift introduced to each batch in m/z (unit : ppm)				global shift introduced to each batch in RT (unit : sec)			
Data set	Batch 1	Batch 2	Batch 3	Batch 4	Batch 1	Batch 2	Batch 3	Batch 4
BR+GNLM1-1	0	0	0	0	0	0	0	0
BR+GNLM1-2	4	-4	0	-4	-2	0	0	-2
BR+GNLM1-3	-4	4	0	4	2	0	0	2
BR+GNLM1-4	4	-4	0	-4	0	0	0	0
BR+GNLM1-5	-4	4	0	4	0	0	0	0
BR+GNLM1-6	0	0	0	0	-2	0	0	-2
BR+GNLM1-7	0	0	0	0	2	0	0	2

To evaluate the performance of the two algorithms, we used recall, precision as measurements. Recall is one of the most commonly used measurements of peak aligning algorithms [30]. The recall of the peak aligning algorithm is defined as $\frac{TP}{TP+FN}$, where *TP* is the abbreviation of true positive, which is the number of true aligned peaks reported by the algorithm, and *FN* is the abbreviation of false negative, which is the number of peaks that should be aligned together but not aligned together by the algorithm. The true peaks are defined by analysts. Another measurement can evaluate type I error (false positive) is the precision value. The precision of the peak aligning algorithm is defined as $\frac{TP}{TP+FP}$, where *FP* is the abbreviation of false positive, which is the number of peaks aligned together by the algorithm but not supposed to be aligned together. An ideal peak aligning algorithm should generate a peak table with both high recall and precision values.

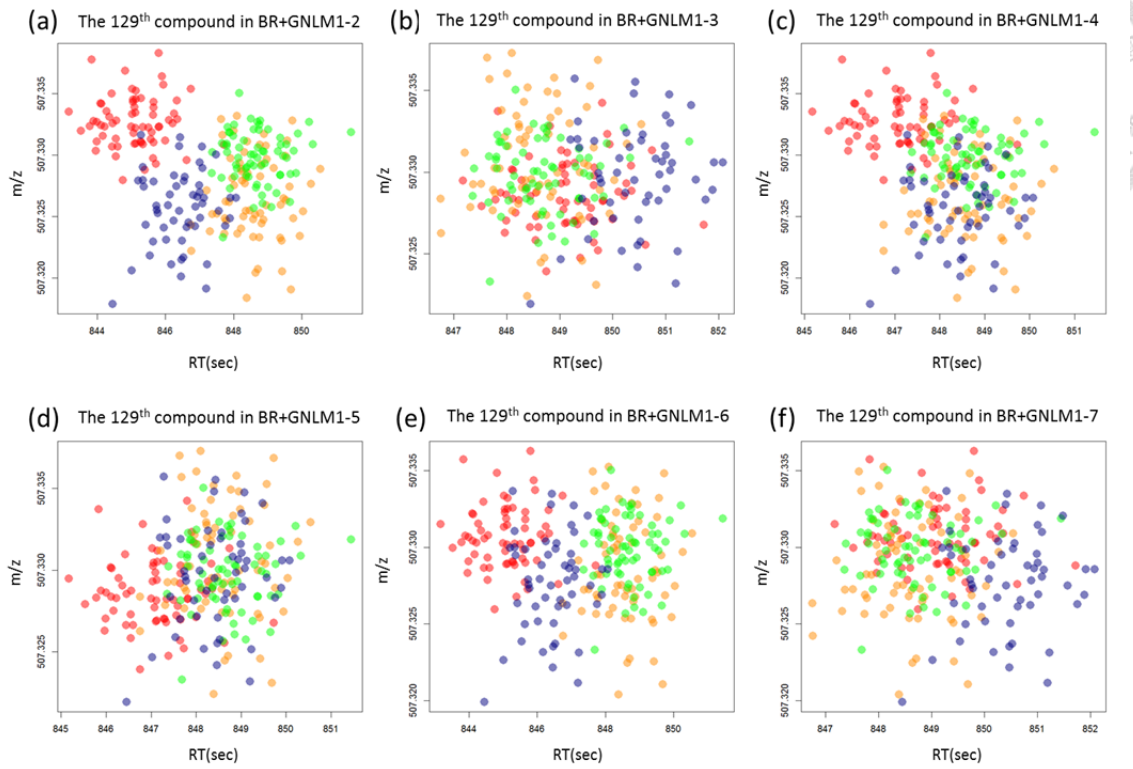


Figure 1.13: Different types of global shift introduced to the 129th compound. The (m/z, RT) of the 129th compound is (507.8494, 848.50). The first, second, third and fourth batch are colored in red, orange, green and blue, respectively.

1.4 Results

1.4.1 *LAKE and XCMS Algorithms Using Forensics Drugs*

LAKE can be used for peak alignment for general metabolomics studies. To compare and evaluate performance of *LAKE* and *group.density* for aligning detected peaks in a practical case, we took forensic drugs spiked in urine samples as evaluation data set. The relative mass difference tolerance and RT difference tolerance were estimated by *LAKE*. Since *group.density* has no default parameter estimation method, we have to find optimal relative mass difference tolerance and RT tolerance for peak alignment in XCMS rather than using recommended parameters from the journal of Nature Protocols for metabolomics data[68] directly. The following *group.density* parameters were used for performance evaluation: mzwid of 0.01, 0.015 and 0.02; bw of

1, 2 and 3. The mzwid value smaller than 0.01 is not considered for causing peaks from the same compound aligned into different peak groups. All combination of these two selected parameters generates nine parameter pairs (mzwid, bw) and generate nine different results. After performance comparison, the optimal parameter pair with best result is selected. The best result is the alignment result with least misaligned peaks and least multiple peaks from the same sample aligned together among the alignment results using different parameter pairs. No missing peaks or least missing peaks would be focused because multiple peaks from the same sample can be properly handled by eliminating the unwanted peaks while missing peaks would cost relative more effort to recover. The nine parameter pairs used in experiments are also tested in LAKE as a comparison.

To visualize the alignment result of spiked forensic drugs, the Figure 1.14 is plotted. The better alignment result is the plot with least blue blocks (misaligned peaks). In XCMS alignment results with different parameter pairs (Figure 1.14), the misaligned peaks can be seen in these nine parameter pairs. The number of misaligned peak change when adjusting m/z tolerance while adjusting RT tolerance did not affect the number of misaligned peak. However, the number of misaligned peak did neither increase nor decrease when m/z tolerance increase. The number of multiple peaks from same sample aligned together did not increase or decrease when adjusting m/z tolerance or RT tolerance. The optimized parameter pair (mzwid, bw) is (0.015, 2).

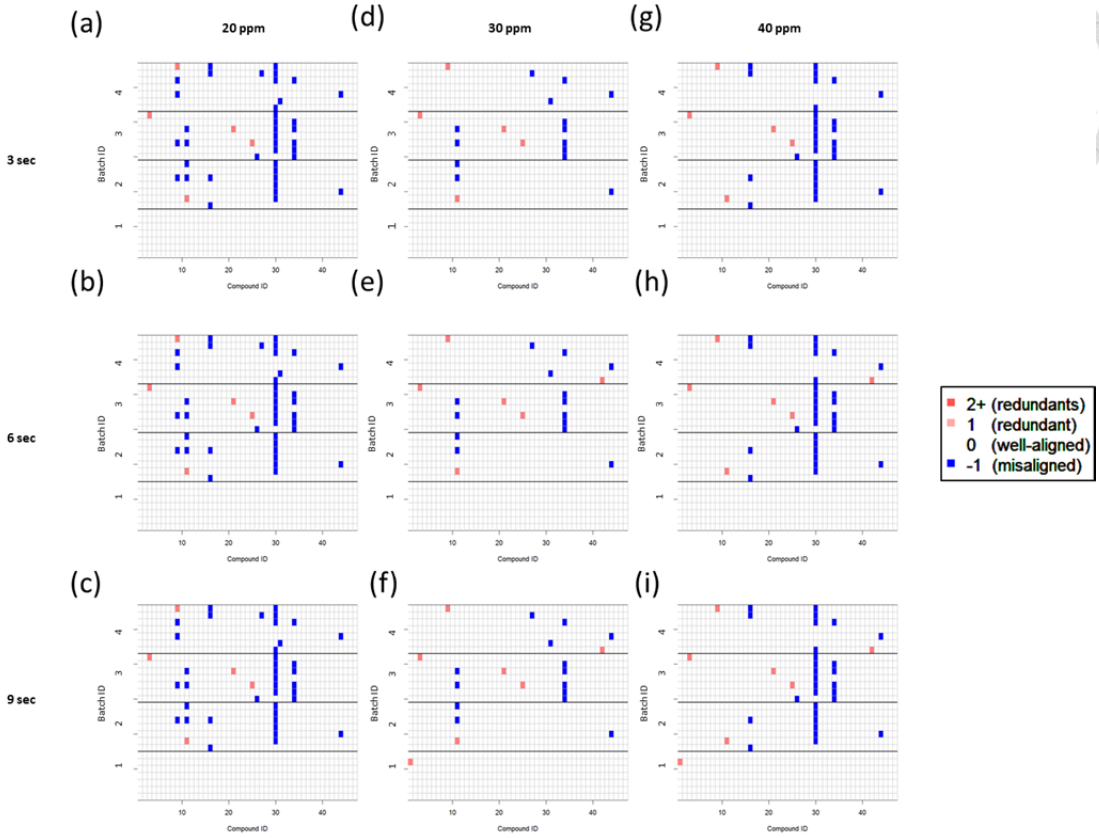


Figure 1.14: The peak alignment by XCMS with different parameters. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

In LAKE alignment results with different parameter pairs (Figure 1.15), the misaligned peaks still can be seen in the results using nine different parameter pairs, but the number of misaligned peaks in LAKE alignment result is less than that in XCMS alignment results. The number of misaligned peaks decreases when increasing m/z tolerance. The number of multiple peaks from same sample aligned together increase when increasing RT tolerance.

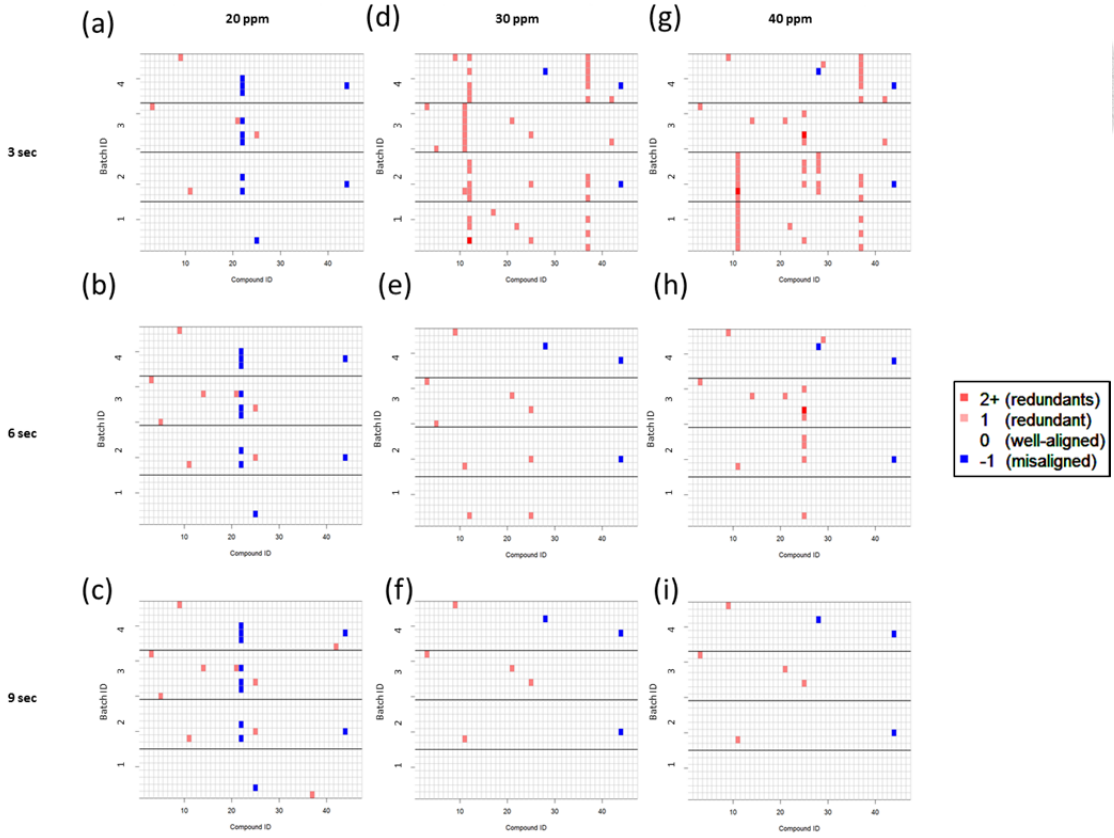


Figure 1.15: The peak alignment by LAKE with different parameters. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

To visualize the misaligned peaks in the alignment results of LAKE and XCMS with optimal parameters, the misaligned peaks from spiked compounds are plotted in m/z versus z-score plots and RT versus z-score plots and placed in Figure 1.16 upper panel and lower panel, respectively. The better alignment result should be with least X marks in the following plots (least misaligned peaks).

From m/z vs z-score plot (Figure 1.16 upper panel), the m/z values in each batch are equally scattered within the same range, and difference between each batch is relative small.

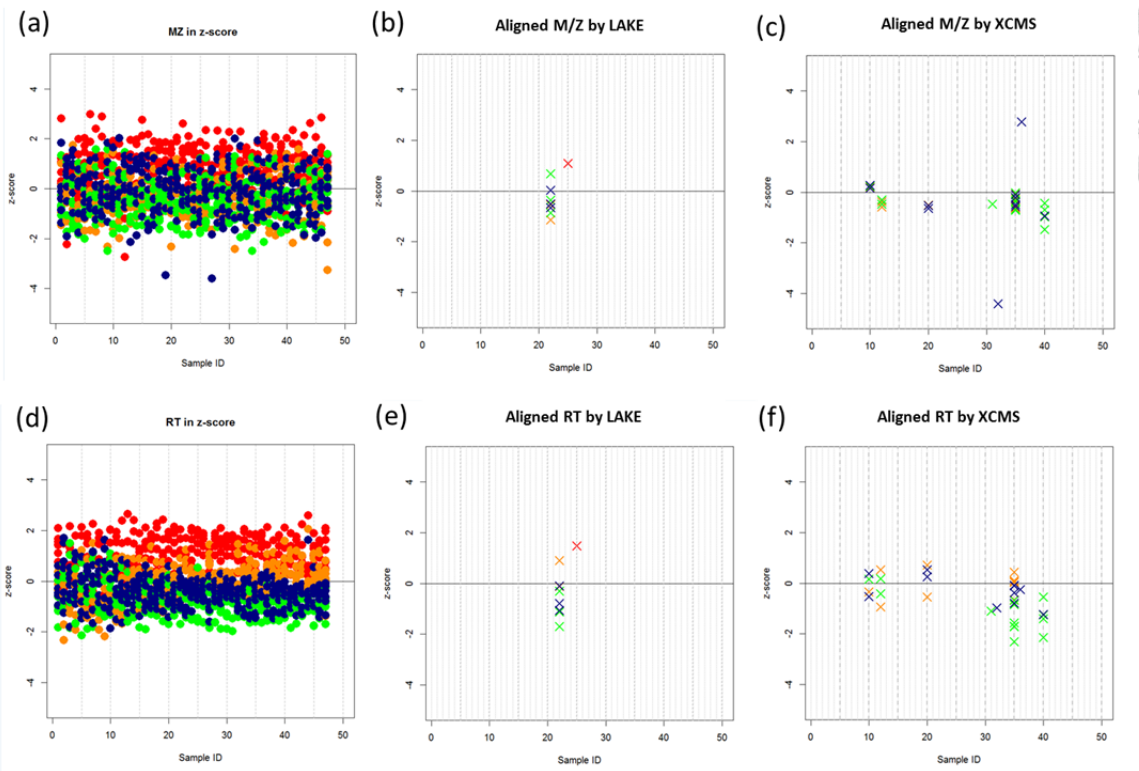


Figure 1.16: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch.

From RT vs z-score plot (Figure 1.16 lower panel), we can see RT values in 1st batch are different from the rest of batches. The batch difference becomes clearer from compound ID 20 to compound ID 50. The z-score of RT values in each batch are not equally scattered within the range but clustered with the RT values from the same batch. The sorted z-score of RT value for each batch is 1st, 2nd, 4th and 3rd batch in descending order.

In XCMS, peaks from the 2nd, 3rd and 4th batch are found misaligned on seven compounds. In LAKE, misaligned peaks can be found in all batches but only on two compounds.

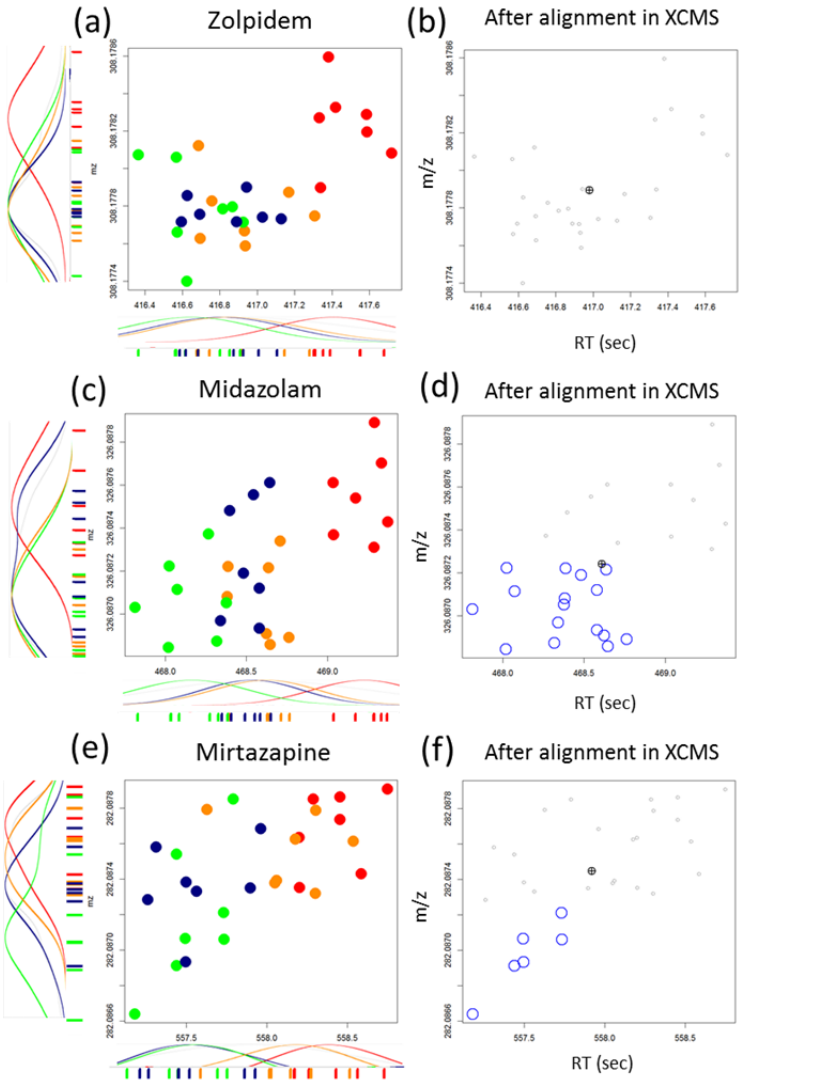


Figure 1.17: Peaks of selected compounds before and after XCMS alignment. (Left panel): Data distribution in m/z-RT panel, (Right panel): Misaligned peaks in peak alignment done by XCMS are marked with blue circles.

To find out the reason of misaligned peaks in alignment result by different algorithm, the compounds with misaligned peaks are plotted in m/z versus RT plot and label the misaligned peaks with blue circles. From previous alignment results, we pick out three aligned compounds which are shown different when comparing two results done by different algorithms with their optimal parameter pairs. We found that misalignment peaks in XCMS alignment result is shown peak group split into two peak groups with invisible straight hard cutline (Figure 1.17).

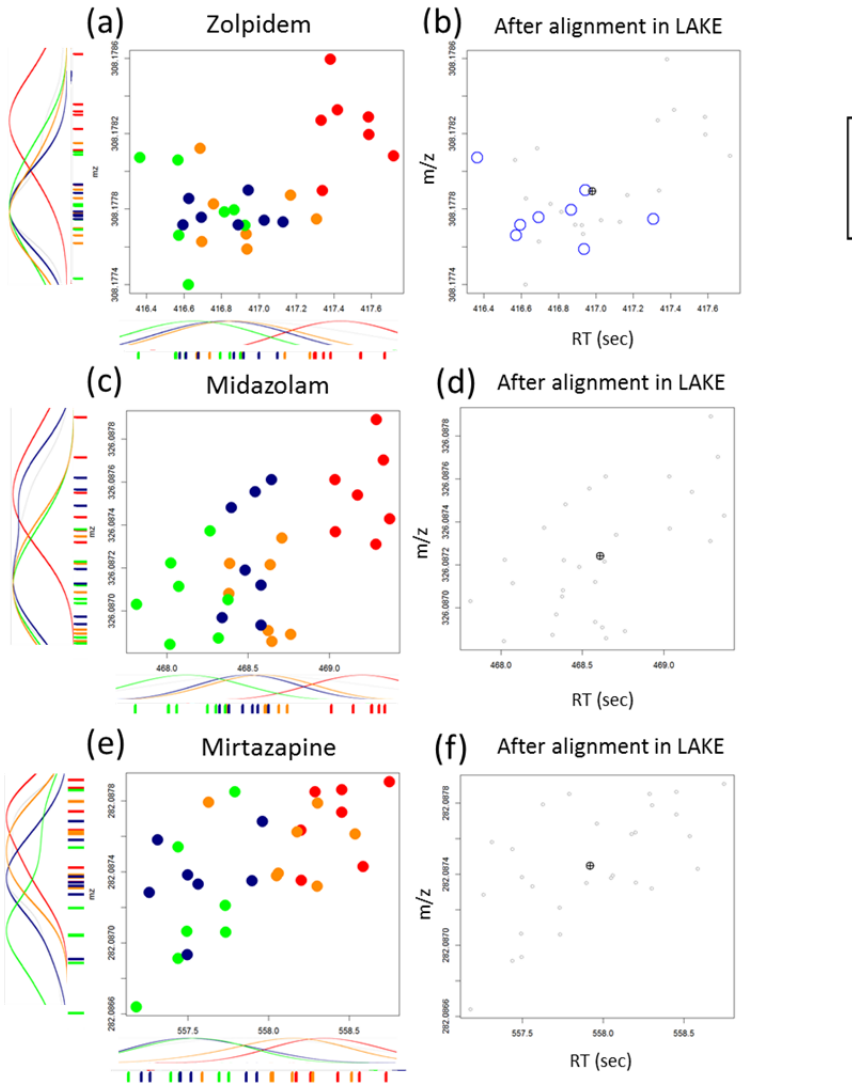


Figure 1.18: Peaks of selected compounds before and after LAKE alignment (Left panel): Data distribution before alignment in m/z-RT plot, (Right panel): Misaligned peaks in peak alignment done by LAKE are marked with blue circles.

Most of misaligned peaks in XCMS alignment are from 2nd and 3rd batch. Misalignment peaks in LAKE alignment result is shown peak group with some peaks misaligned rather than split up situation in XCMS alignment result. The misaligned peaks in LAKE alignment are from 2nd, 3rd and 4th batch (Figure 1.18).

1.4.2 *LAKE and XCMS Algorithms Using Metabolomics Data Set with Introduced Different Types of Noise*

In the data introduced with local shift noise in both m/z and RT dimension, four data sets introduced noise with four different random seeds show similar alignment results. The GNLM-3 data set is selected to show the alignment results done by two different algorithms for this case with average performance among the four data sets.

In XCMS alignment results with optimal parameter pairs (Figure 1.19), the misaligned peaks can be seen in four batches. However, the number of misaligned peak did neither increase nor decrease when m/z tolerance increase. The misaligned peaks frequently found in one of its technical replicate from the last sample in 4th batch.

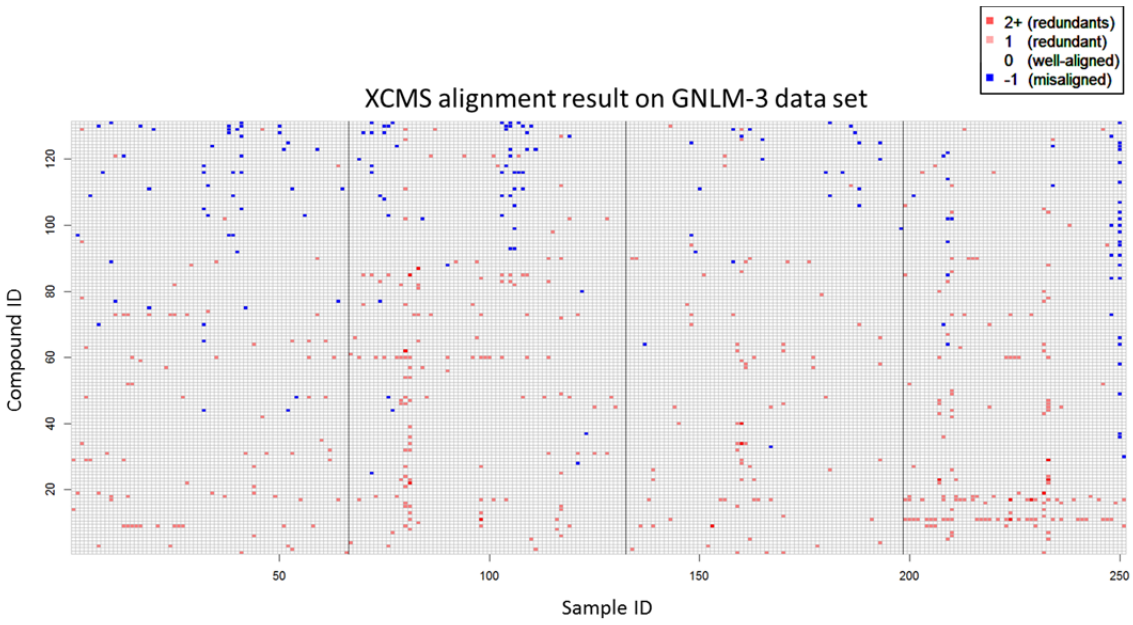


Figure 1.19: The peak alignment by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

In LAKE alignment results with estimated parameter (Figure 1.20), the number of misaligned peaks in LAKE alignment result is less than that in XCMS alignment results. With relative more multiple peaks aligned in the same sample at 18th detected

compound in 1st and 2nd batch.

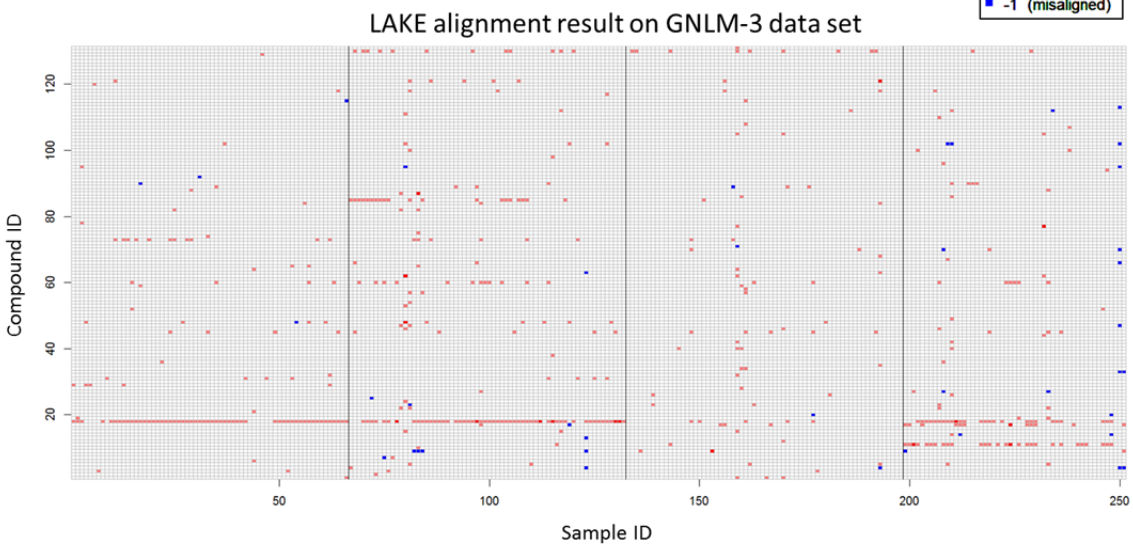


Figure 1.20: The peak alignment by LAKE with estimated parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

Three compounds from the 131 selected compounds we defined as ground truth for further investigation on difference of alignment result between two algorithms. In the result of XCMS alignment (Figure 1.21), most misaligned peaks found in 1st batch for having different distribution when compared with the other batches in m/z dimension.

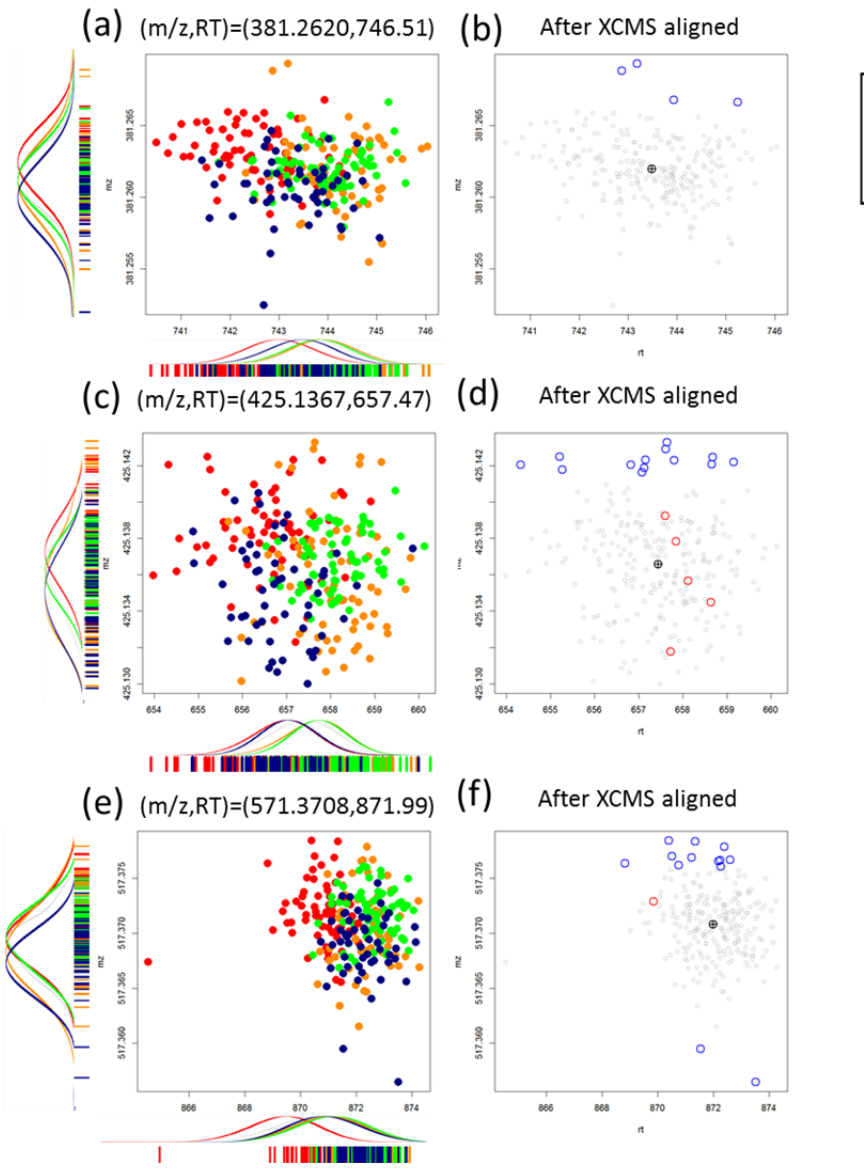


Figure 1.21: Peaks of selected compounds before and after XCMS alignment (Left panel): Data distribution before alignment in m/z-RT plot, (Right panel): Misaligned peaks in peak alignment done by LAKE are marked with blue circles.

In the result of LAKE alignment (Figure 1.22), misaligned peaks are not from any specific batch.

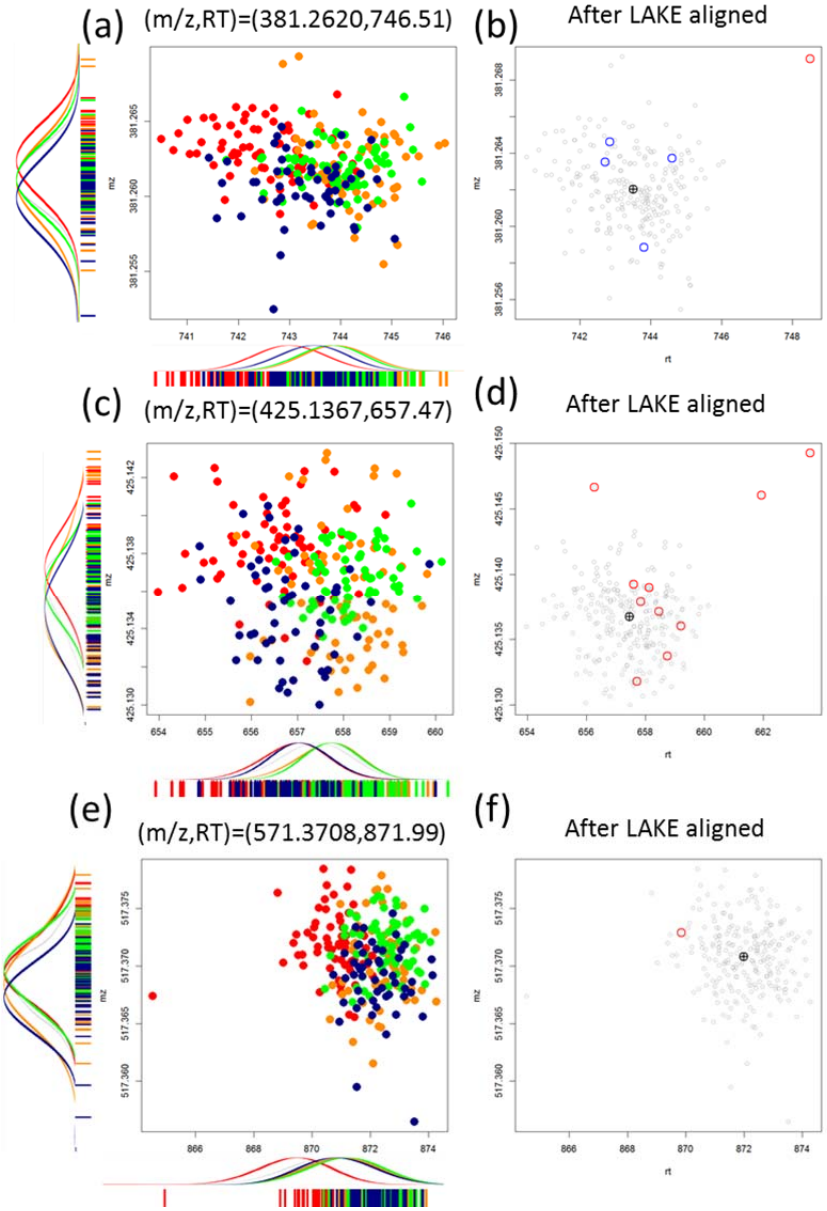


Figure 1.22: Peaks of selected compounds before and after LAKE alignment (Left panel): Data distribution before alignment in m/z-RT plot, (Right panel): Misaligned peaks in peak alignment done by LAKE are marked with blue circles.

To compare the alignment results of LAKE and XCMS with optimal parameters, the misaligned peaks from the 131 selected compounds defined as ground truth are plotted in m/z vs z-score plots and RT vs z-score plots and placed in Figure 1.23 upper panel and lower panel, respectively.

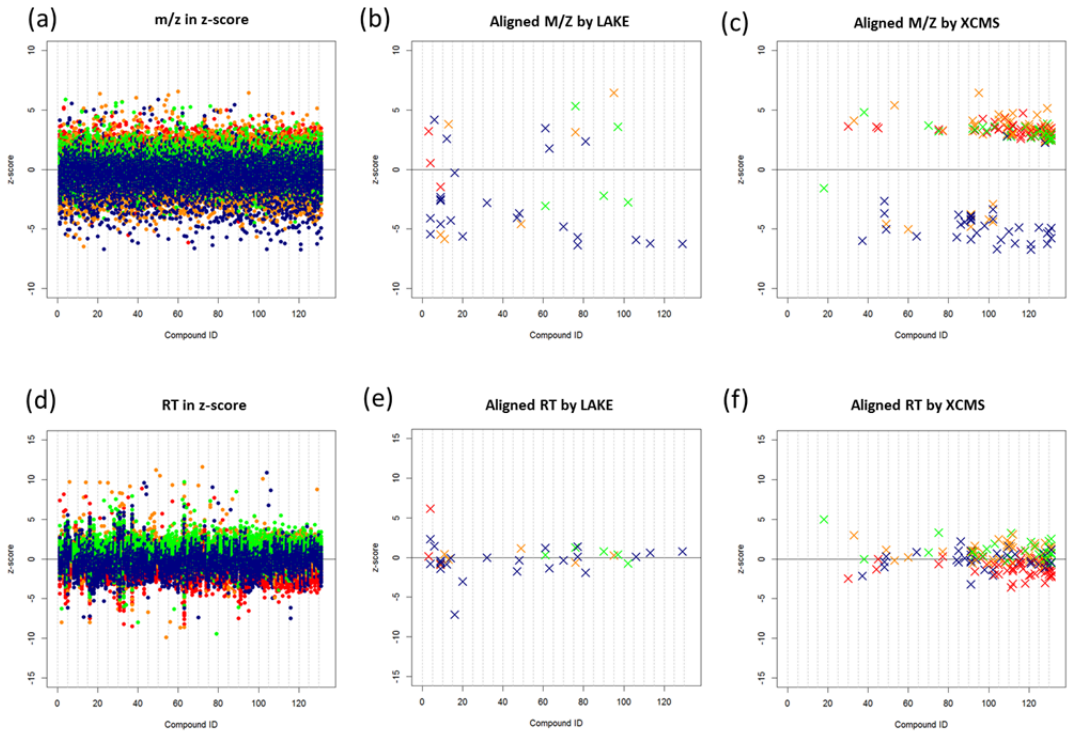


Figure 1.23: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch.

In the m/z versus z-score plots (Figure 1.23 upper panel), misaligned peaks in LAKE is scattered in all batches but majorly from 4th batch while misaligned peak number in XCMS increase with m/z value of aligned compound (The compound ID is a label after the sorted list with increasing m/z value where the compound with compound ID 1 is the compound with smallest m/z value among all 131 compounds while the compound with compound ID 131 is the compound with largest m/z value among all 131 compounds). There is no pattern in the z-score of misaligned peaks in LAKE alignment result. The z-score of misaligned peaks in XCMS are generally either higher than 2.5 or lower than -2.5.

In the RT versus z-score plots (Figure 1.23 lower panel), misaligned peaks in LAKE is scattered in all batches while misaligned peak number in XCMS increase with

m/z value of aligned compound. The z-score of misaligned peaks in LAKE with no rule to be found here but the z-scores of misaligned peaks in LAKE are generally fall between 2.5 and -2.5. The z-score of misaligned peaks in XCMS is generally fall between 2.5 and -2.5.

Table 3 Gaussian noise introduced data set (GNLM)

	LAKE			*group.density*		
Data set	Precision	Recall	F-score	Precision	Recall	F-score
GNLM-1	98.46	99.89	99.17	98.71	99.39	99.05
GNLM-2	98.55	99.67	99.11	98.58	99.35	98.96
GNLM-3	98.43	99.88	99.15	98.66	99.43	99.04
GNLM-4	98.78	99.86	99.32	98.57	99.35	98.96
Average	98.56	99.83	99.19	98.63	99.38	99.03

The performance of alignment on different data sets with local shift noise introduced is shown in Table 1.3. The average precision of LAKE on the four data sets is lower than that of XCMS (98.56<98.63). The average recall of LAKE on the four data sets is higher than that of XCMS (99.83>99.38). The average F-score of LAKE on the four data sets is higher than that of XCMS (99.19>99.03).

To compare the performance of alignment algorithm when global shift among batch is introduced, BR+GNLM1-X data sets are selected as validation data set which include different situations either alleviate or exaggerate the difference among batches in m/z or RT dimension. In the following visualized alignment results, the better alignment result is the plot with least blue blocks (misaligned peaks).

In XCMS alignment result on BR+GNLM1-2 data set (exaggerated both m/z and RT difference among batches) (Figure 1.24), misaligned peaks can be seen in all batches, however the misaligned peaks are majorly from 1st batch and the compound with large m/z value. And multiple peaks from one sample aligned together can be frequently seen in 19th compound in all batches.

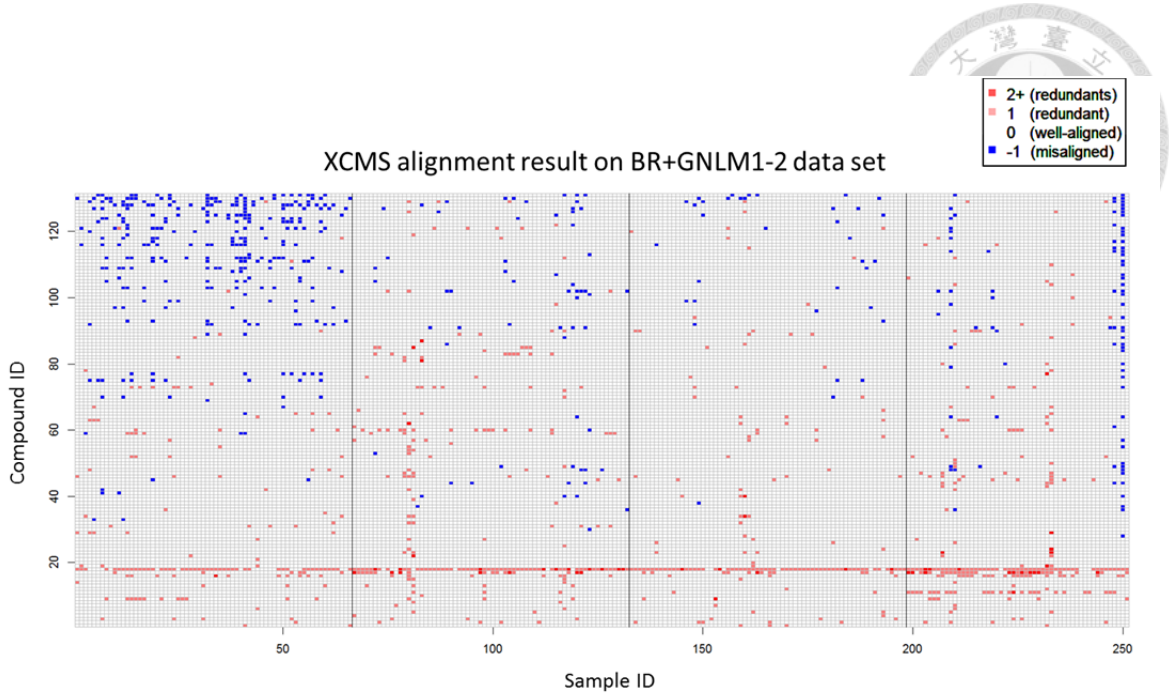


Figure 1.24: The peak alignment on BR+GNLM1-2 data set (exaggerated both m/z and RT difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

In XCMS alignment result on BR+GNLM1-3 data set (alleviated both m/z and RT difference among batches), misaligned peaks can be seen in all batches. The misaligned peaks are majorly from 2nd batch and the compound with large m/z value.

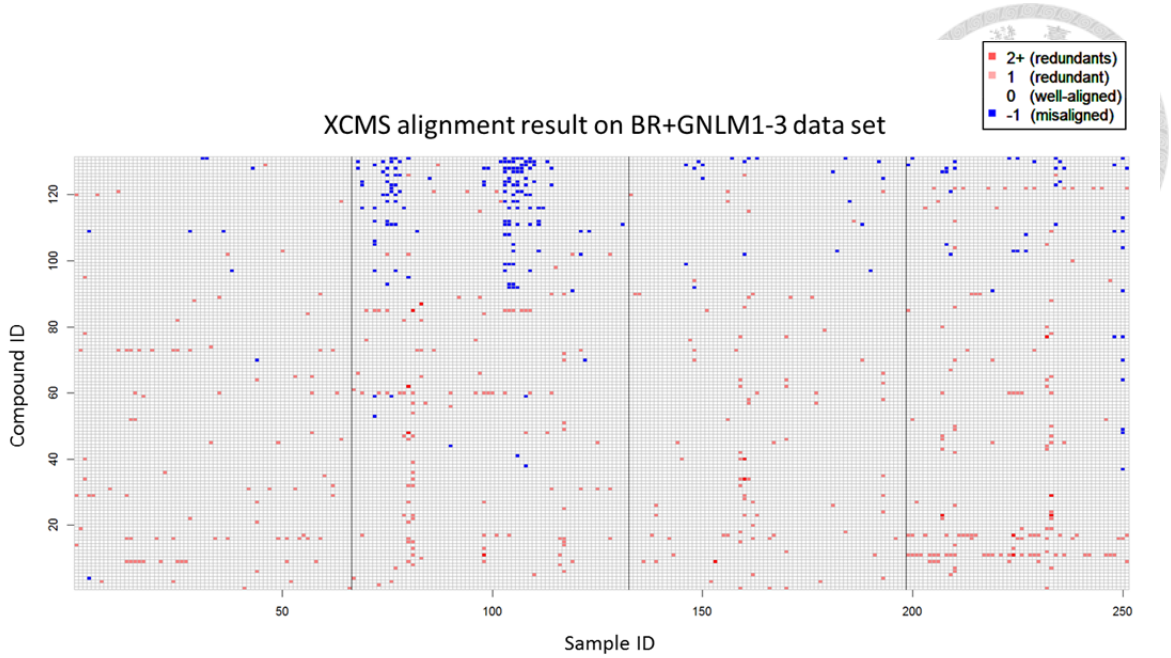


Figure 1.25: The peak alignment on BR+GNLM1-3 data set (alleviated both m/z and RT difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

In XCMS alignment result on BR+GNLM1-4 data set (exaggerated m/z difference among batches), misaligned peaks can be seen in all batches, however the misaligned peaks majorly from 1st batch and the compound with large m/z value.

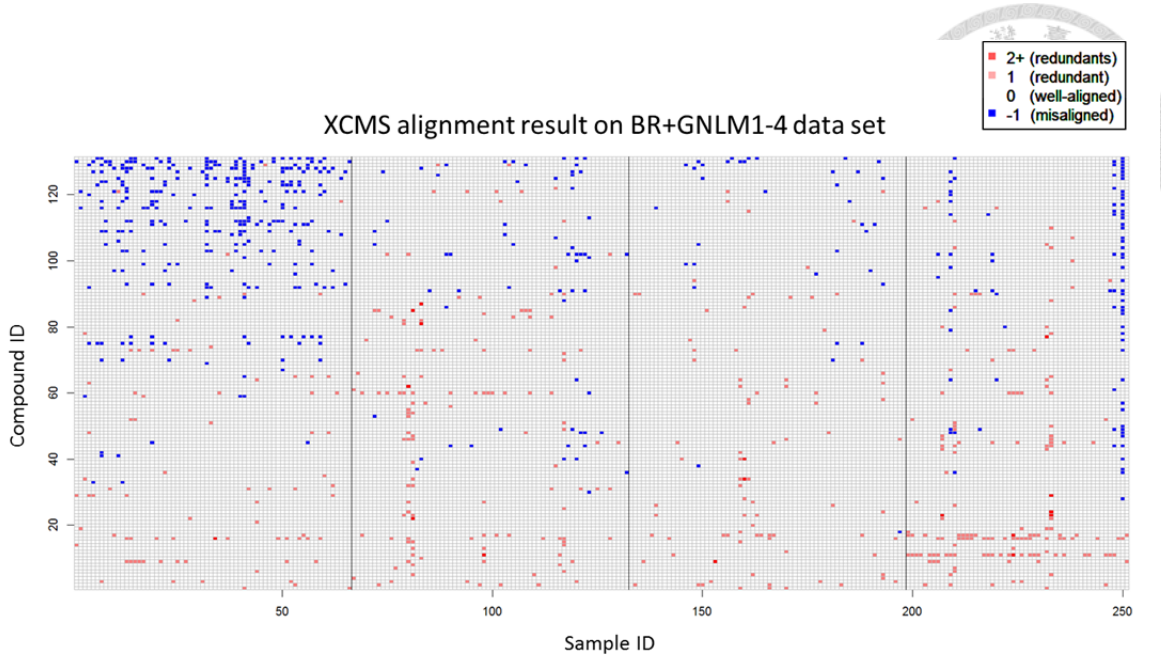


Figure 1.26: The peak alignment on BR+GNLM1-4 data set (exaggerated m/z difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

In XCMS alignment result on BR+GNLM1-5 data set (alleviated m/z difference among batches), misaligned peaks can be seen in all batches, however the misaligned peaks majorly from 2nd batch and the compound with large m/z value.

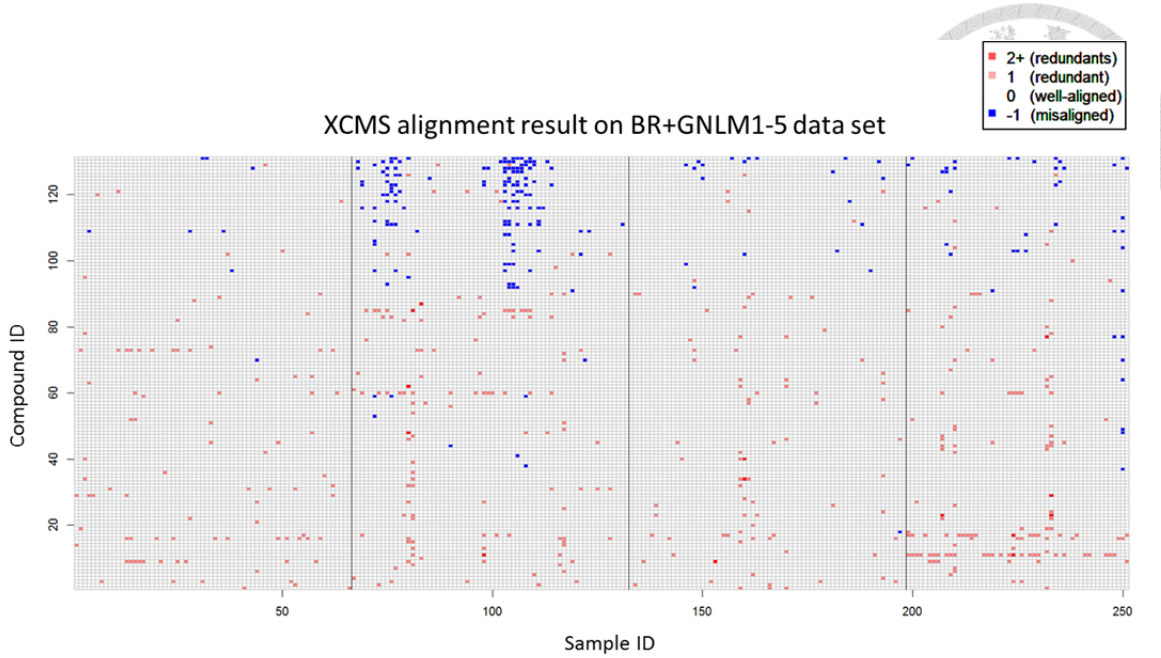


Figure 1.27: The peak alignment on BR+GNLM1-5 data set (alleviated m/z difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

In XCMS alignment result on BR+GNLM1-6 data set (exaggerated RT difference among batches), misaligned peaks can be seen in all batches. And multiple peaks from one sample aligned together can be frequently seen in 19th compound in all batches.

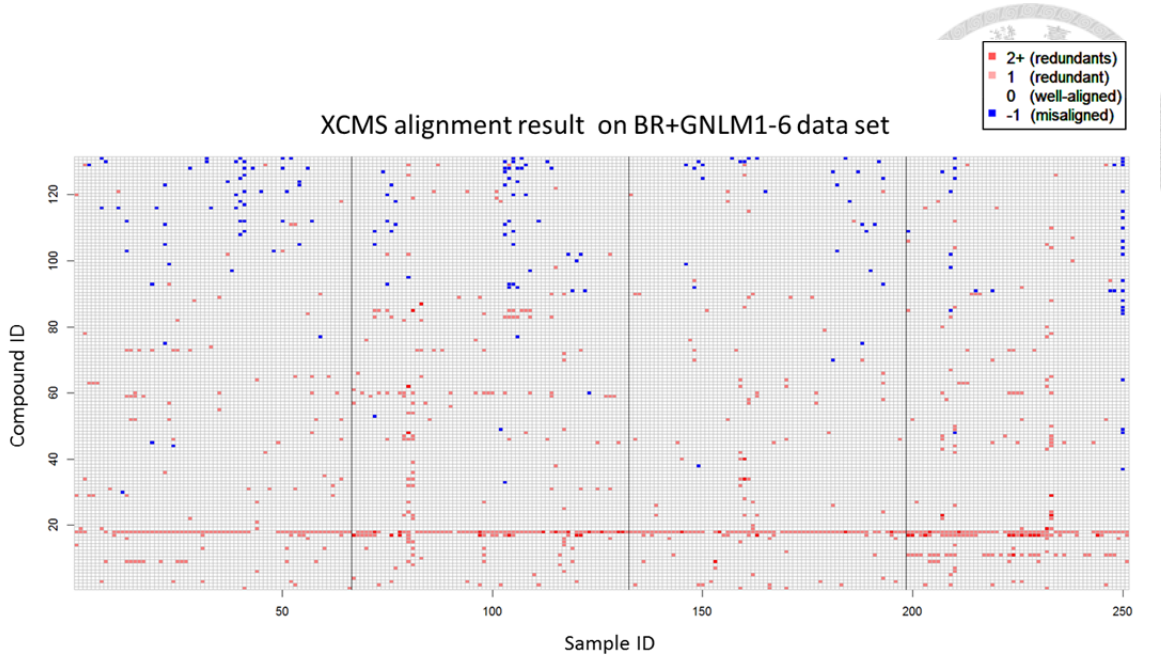


Figure 1.28: The peak alignment on BR+GNLM1-6 data set (exaggerated RT difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

In XCMS alignment result on BR+GNLM1-7 data set (alleviated RT difference among batches), misaligned peaks can be seen in all batches.

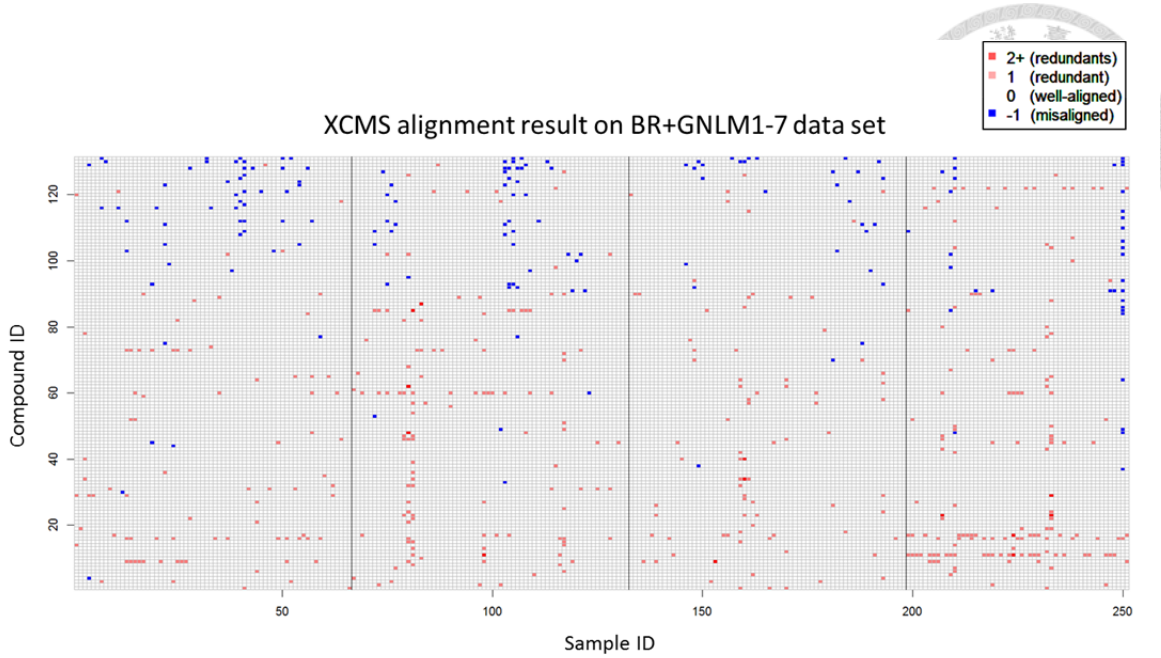


Figure 1.29: The peak alignment on BR+GNLM1-7 data set (alleviated RT difference among batches) by XCMS with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

In LAKE alignment result on BR+GNLM1-2, misaligned peaks can be seen in all batches but the number of misaligned peaks in LAKE is relative low when compare to that in XCMS. And multiple peaks from one sample aligned together can be frequently seen in 19th compound in 1st, 2nd and 4th batches. The rest of data sets (BR+GNLM1-3, BR+GNLM1-4, BR+GNLM1-5, BR+GNLM1-6 and BR+GNLM1-7) share the exactly same result, so BR+GNLM1-2 is selected and show to represent LAKE alignment result on all other data sets.

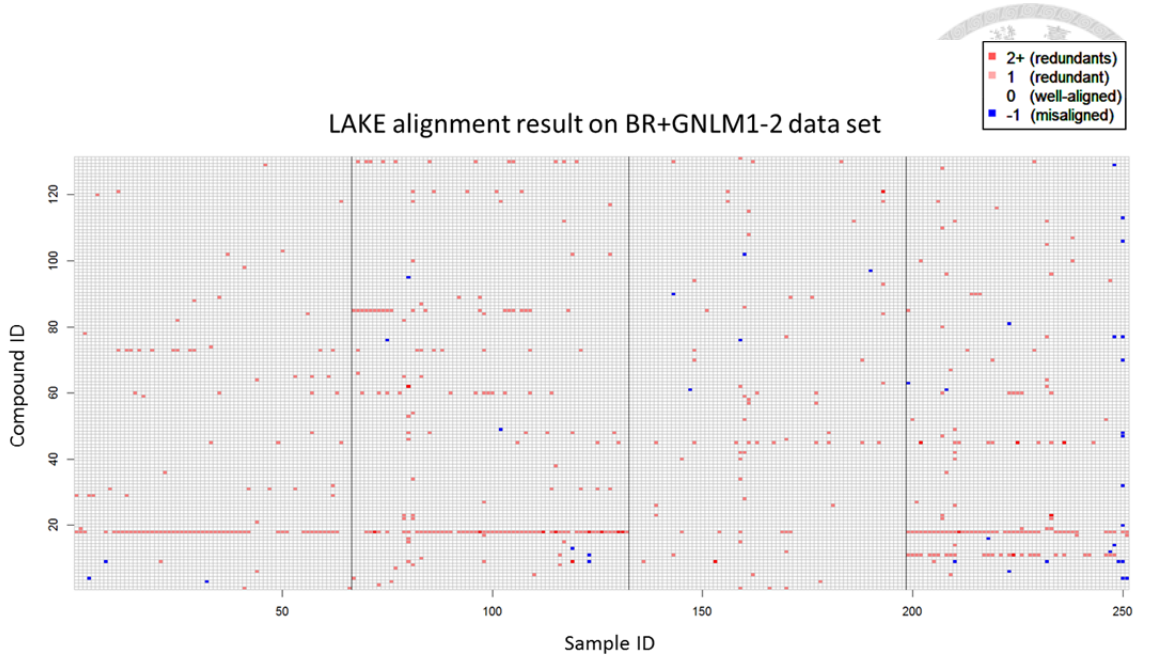


Figure 1.30: The peak alignment on BR+GNLM1-2 data set (exaggerated both m/z and RT difference among batches) by LAKE with optimal parameter. Peak alignment plot of compound versus sample number. The colors in the plot represent the alignment result, where red, white and blue indicates aligned with redundant peak, well-aligned peak and misaligned peak in alignment result respectively.

The distributions of 131 compounds which selected as ground truth with different global shift among batches introduced in m/z versus z-score plot and RT versus z-score plot are to demonstrate how different data distribution can affect alignment result done by different algorithms. The alignment results with optimal parameter from each algorithm are selected and compared. In Figure 1.31, the alignment results of BR+GNLM1-2, the data set with exaggerated both m/z and RT difference among batches, are shown in m/z versus z-score plots (Figure 1.31 upper panel) and RT vs z-score plots (Figure 1.31 lower panel).

In the m/z versus z-score plots (Figure 1.31 upper panel), misaligned peaks in LAKE are scattered in all batches but majorly from 4th batch while misaligned peak number in XCMS increase with m/z value of aligned compound. There is no pattern on the z-score of misaligned peaks in LAKE alignment result. The z-score of misaligned peaks in XCMS are generally either higher than 2.5 or lower than -2.5.

In the RT versus z-score plots (Figure 1.31 lower panel), the z-scores of misaligned peaks in LAKE are generally fall between 3 and -5. The z-score of misaligned peaks in XCMS is generally fall between 3 and -5.

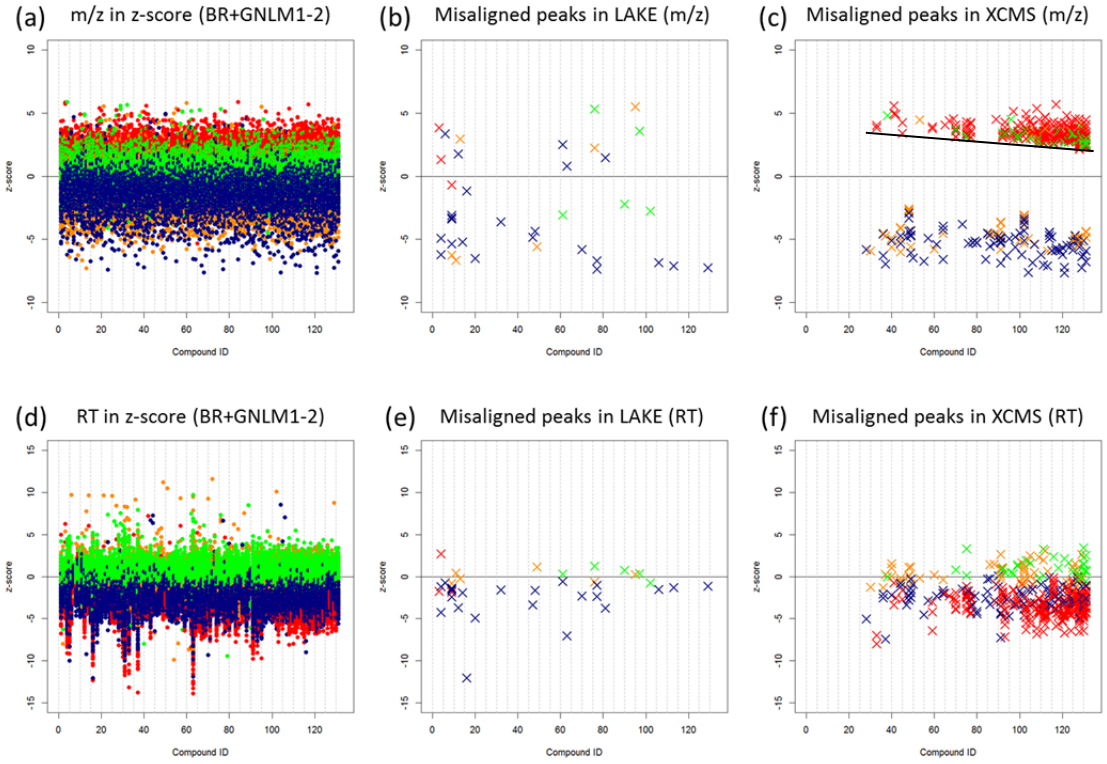


Figure 1.31: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing.

The alignment results of BR+GNLM1-3, the data set with alleviated both m/z and RT difference among batches, are shown in Figure 1.32. In the m/z versus z-score plots (Figure 1.32 upper panel), misaligned peaks in LAKE is scattered in all batches but majorly from 4th batch while misaligned peaks in XCMS majorly from 2nd batch. The misaligned peak number in both alignment results are less than that in BR+GNLM1-2. In the RT versus z-score plots (Figure 1.32 lower panel), the z-score of misaligned

peaks in LAKE are generally fall between 5 and -2. The z-score of misaligned peaks in XCMS is generally fall between 5 and -2.

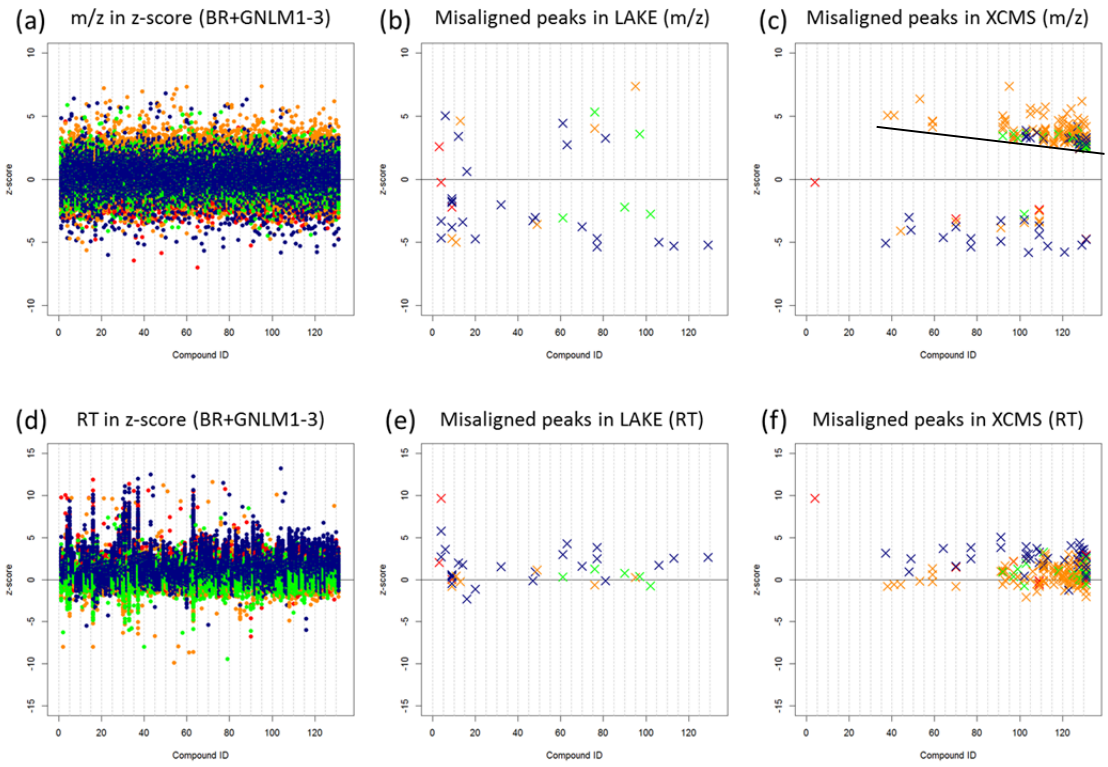


Figure 1.32: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing.

The alignment results of BR+GNLM1-4, the data set with exaggerated m/z difference among batches, are shown in Figure 1.33. In the m/z versus z-score plots (Figure 1.33 upper panel), misaligned peaks in LAKE is scattered in all batches but majorly from 4th batch while misaligned peaks in XCMS majorly from 1st batch. The misaligned peak number in both alignment results are the same as that in BR+GNLM1-2. In the RT versus z-score plots (Figure 1.33 lower panel), the z-score of misaligned peaks in LAKE are generally fall between 3 and -5. The z-score of

misaligned peaks in XCMS is generally fall between 3 and -5.

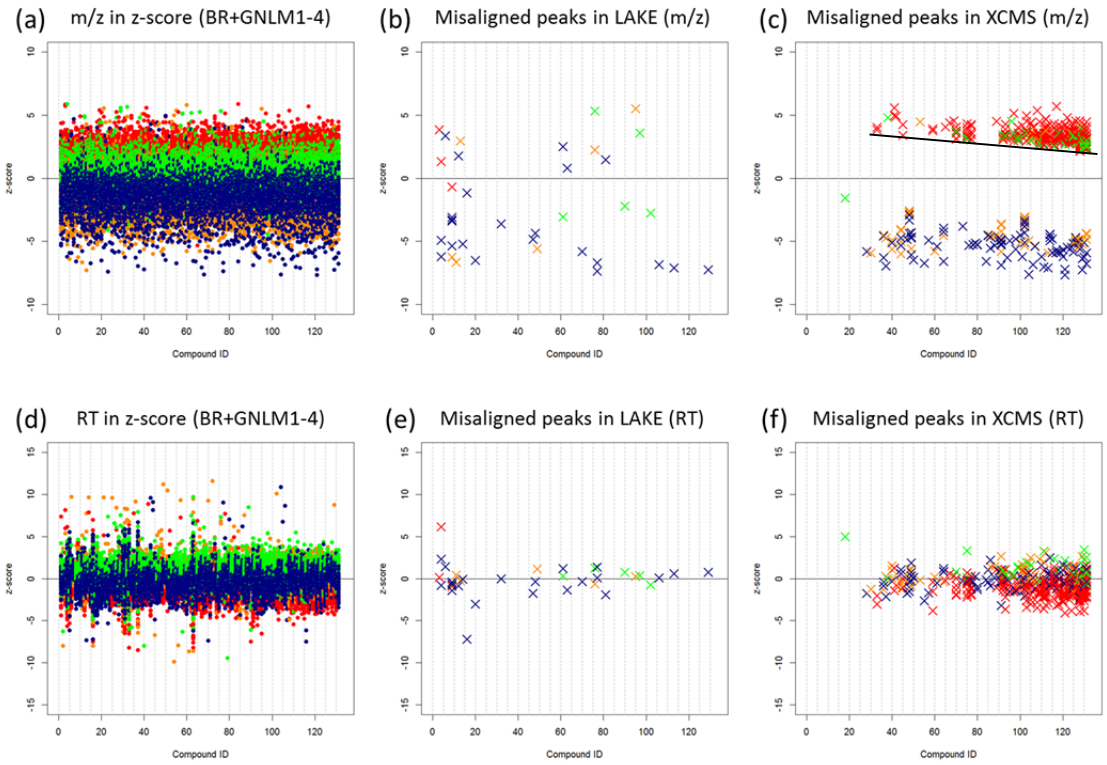


Figure 1.33: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing.

The alignment results of BR+GNLM1-5, the data set with alleviated m/z difference among batches, are shown in Figure 1.34. In the m/z versus z-score plots (Figure 1.34 upper panel), misaligned peaks in LAKE is scattered in all batches but majorly from 4th batch while misaligned peaks in XCMS majorly from 2nd batch. The misaligned peak number in both alignment results are less than that in BR+GNLM1-2. In the RT versus z-score plots (Figure 1.34 lower panel), the z-score of misaligned peaks in LAKE are generally fall between 3 and -3. The z-score of misaligned peaks in XCMS is generally fall between 3 and -3.

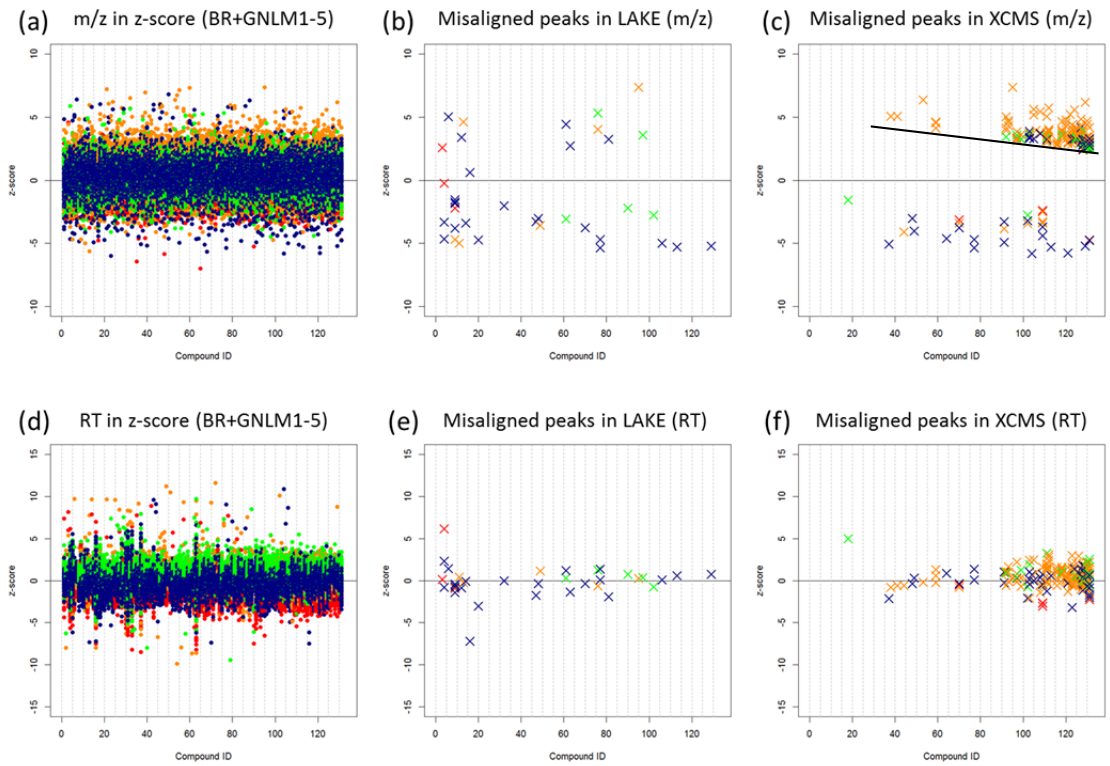


Figure 1.34: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing.

The alignment results of BR+GNLM1-6, the data set with exaggerated RT difference among batches, are shown in Figure 1.35. In the m/z versus z-score plots (Figure 1.35 upper panel), misaligned peaks in LAKE is scattered in all batches but majorly from 4th batch while misaligned peaks in XCMS majorly from 4th batch. The misaligned peak number in both alignment results are less than that in BR+GNLM1-2. In the RT versus z-score plots (Figure 1.35 lower panel), the z-score of misaligned peaks in LAKE are generally fall between 2 and -4. The z-score of misaligned peaks in XCMS is generally fall between 2 and -4.

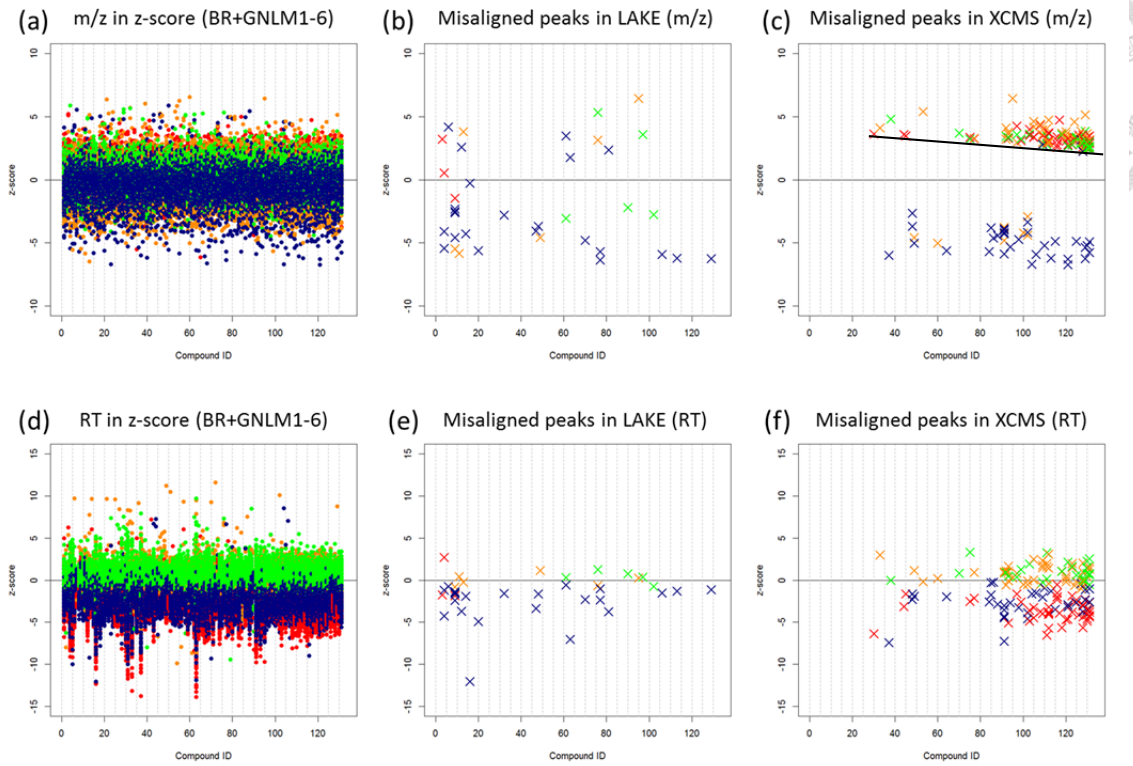


Figure 1.35: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing.

The alignment results of BR+GNLM1-7, the data set with alleviated RT difference among batches, are shown in Figure 1.36. In the m/z versus z-score plots (Figure 1.36 upper panel), misaligned peaks in LAKE is scattered in all batches but majorly from 4th batch while misaligned peaks in XCMS majorly from 4th batch. The misaligned peak number in both alignment results are less than that in BR+GNLM1-2. In the RT versus z-score plots (Figure 1.36 lower panel), the z-score of misaligned peaks in LAKE are generally fall between 4 and -1. The z-score of misaligned peaks in XCMS is generally fall between 4 and -1.

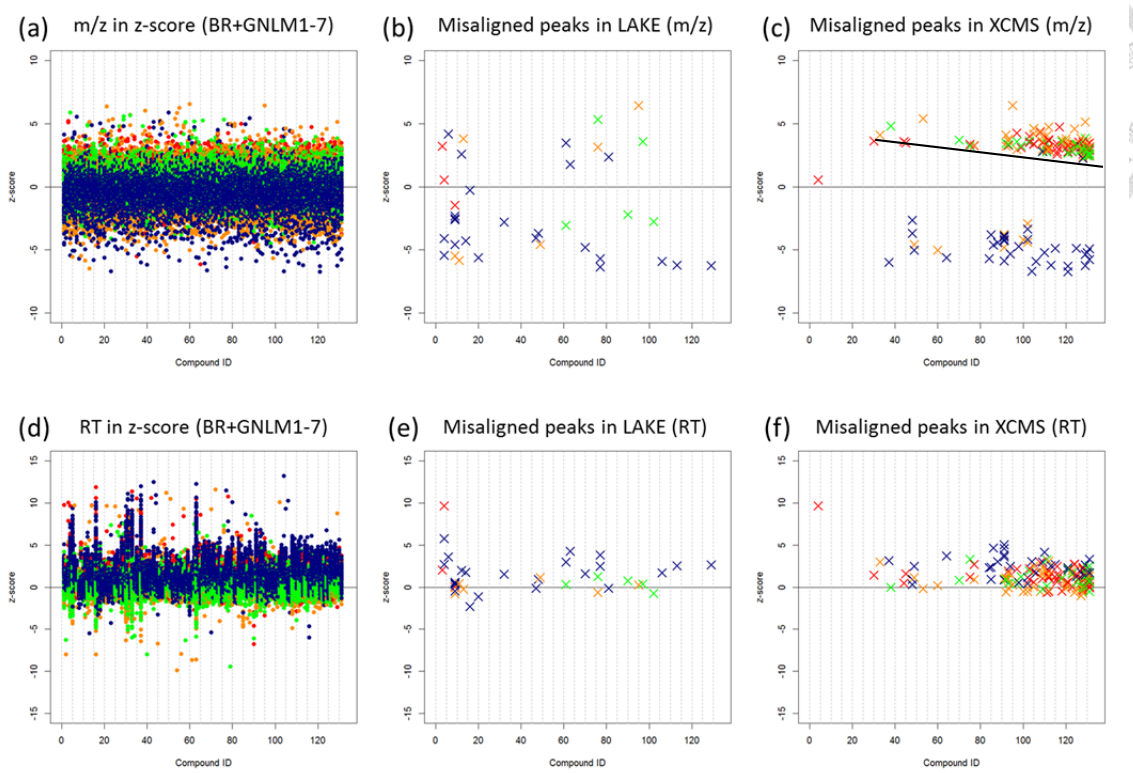


Figure 1.36: The comparison between two different algorithms. The original m/z and RT for each compound is shown in left panel. In the middle panel, LAKE alignment result is shown. In the right panel, XCMS alignment result is shown. X mark represents misaligned peak. Red: peak in 1st batch, Orange: peak in 2nd batch, Green: peak in 3rd batch, and Blue: peak in 4th batch. The black line in (c) indicates the decreasing z-score boundary of misaligned peaks when m/z of aligned peaks increasing.

One compound from the ground truth we defined earlier is selected from all six different global shifts among batches introduced data sets (from BR+GNLM1-2 to BR+GNLM1-7) for further investigation on the difference of alignment result between two algorithms. In the result of XCMS alignment on data set BR+GNLM1-2 and BR+GNLM1-4 (Figure 1.37a, c), most misaligned peaks found in 1st batch for having different distribution when compared with the other batches in m/z dimension. In the result of XCMS alignment on data set BR+GNLM1-3 and BR+GNLM1-5 (Figure 1.37b, d), most misaligned peaks found in 2nd batch for having different distribution when compared with the other batches in m/z dimension. In the result of XCMS alignment on data set BR+GNLM1-6 and BR+GNLM1-7 (Figure 1.37e, f), most misaligned peaks

found in 1st and 2nd batch for having different distribution when compared with the other batches in m/z dimension.

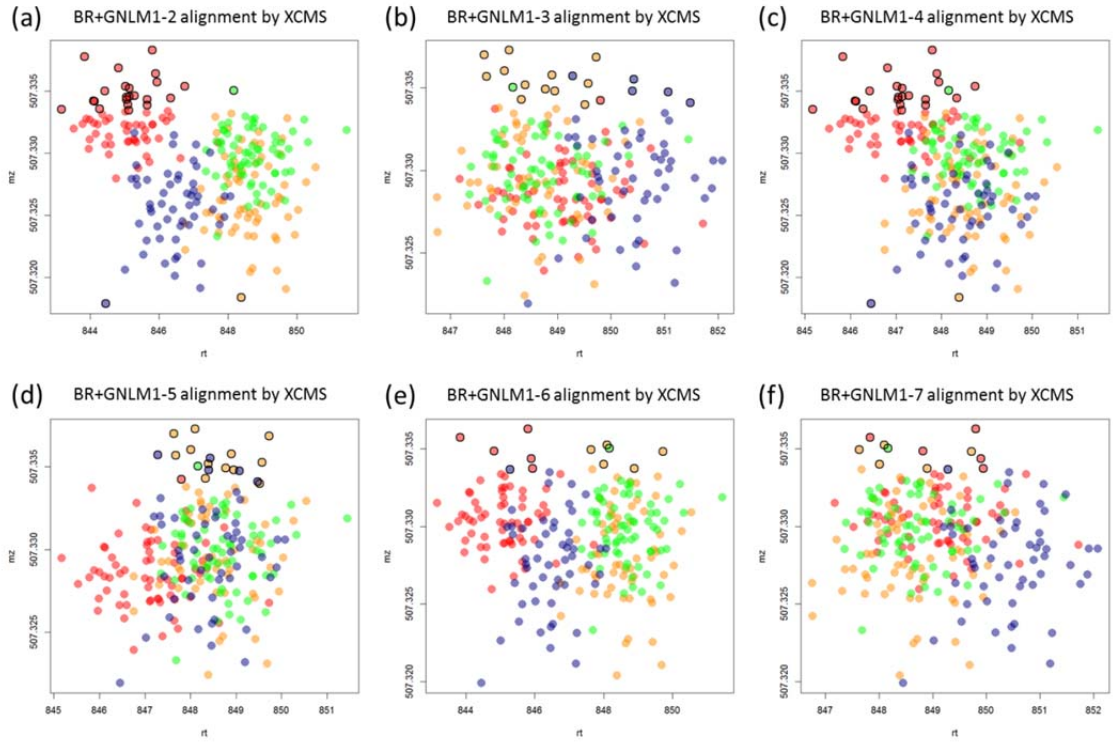


Figure 1.37: Peaks from 129th compound after XCMS alignment. Misaligned peaks in peak alignment done by XCMS are marked with blue circles. The (m/z, RT) of the 129th compound is (507.8494, 848.50).

In the result of LAKE alignment (Figure 1.38), misaligned peaks are not detected.

The performances of alignment on different data sets are shown in Table 1.4. The average precision of LAKE on the seven data sets is higher than that of XCMS (98.46>98.34). The average recall of LAKE on the four data sets is higher than that of XCMS (99.89>99.08). The average F-score of LAKE on the four data sets is higher than that of XCMS (99.17>98.71).

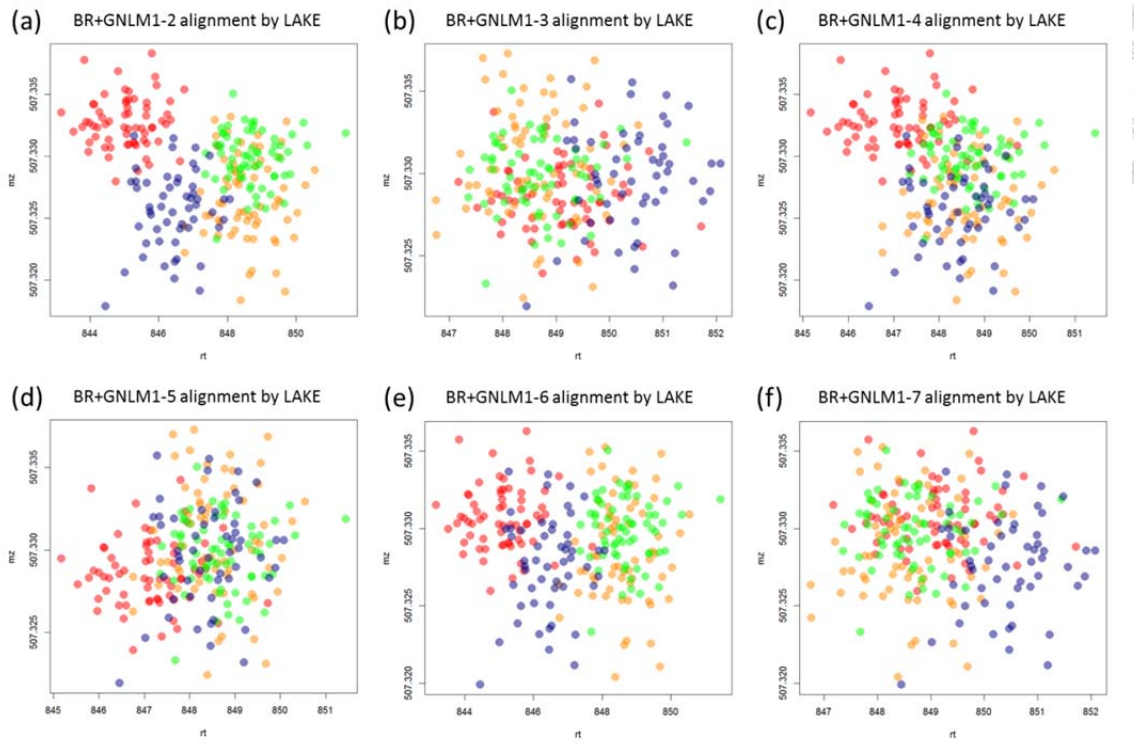


Figure 1.38: Peaks from 129th compound after LAKE alignment. Misaligned peaks in peak alignment done by LAKE are marked with blue circles. The (m/z, RT) of the 129th compound is (507.8494, 848.50).

Table 4 Global shift among batches noise introduced data set (BR+GNLM1)

	LAKE			*group.density*		
Data set	Precision	Recall	F-score	Precision	Recall	F-score
BR+GNLM1-1	98.46	99.89	99.17	98.71	99.39	99.05
BR+GNLM1-2	98.46	99.89	99.17	97.49	98.46	97.97
BR+GNLM1-3	98.46	99.89	99.17	98.68	99.25	98.96
BR+GNLM1-4	98.46	99.89	99.17	98.54	98.46	98.5
BR+GNLM1-5	98.46	99.89	99.17	98.67	99.25	98.96
BR+GNLM1-6	98.46	99.89	99.17	97.66	99.39	98.52
BR+GNLM1-7	98.46	99.89	99.17	98.65	99.39	99.02
Average	98.46	99.89	99.17	98.34	99.08	98.71

1.5 Discussion

1.5.1 *Comparison of LAKE and XCMS Algorithms Using Forensics Drugs*

Inappropriate peak alignment can be found in most alignment result done by existed peak alignment algorithm which might lead to misinterpreted statistical

analyzed result. Common errors are misaligned peaks and redundant peak alignment.

Among these errors results from inappropriate peak alignment, misaligned peak is more important because these lead to the aligned peak table with missing data, which can be avoided by using optimized parameter for the same peak alignment algorithm or choosing a better peak alignment algorithm. Though missing data can be applied with statistical procedures before analyzing data with missing values, the preprocessed data can still result in misinterpreted result.

Therefore choosing optimal parameters on the proper peak alignment algorithm becomes important in aligning peaks from multiple samples from different batches. Figure 1.14 shows XCMS alignment results of the 50 forensic drugs spiked in urine samples using nine parameter pairs. To get the XCMS alignment result with least misaligned peaks is to adjust mzwid (m/z tolerance) rather than adjusting bw (RT tolerance) because adjusting bw did not greatly increase or decrease the peak number of the misaligned peak from Figure1.14 a, b and c. However, the misaligned peak number changes when adjusting mzwid is still unknown from Figure 1.14a, d and g, the misaligned peak number did keep neither decrease nor increase when increasing m/z tolerance.

To have deeper look on misaligned peaks in XCMS, Figure 1.17 XCMS alignment on three selected compounds for having different results when comparing with LAKE alignment result. In Figure 1.17b, d and f, all three selected compounds show the misaligned peaks seems to be separated from the clustered m/z-RT group with an invisible line, which implies the peak misalignment in XCMS might be caused by m/z binning with fixed m/z interval.

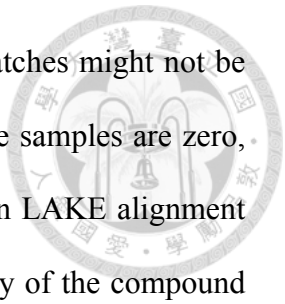
Using fixed interval in m/z binning is inappropriate when grouping compounds with different m/z value for different m/z with different m/z deviation. One fixed m/z

bin might be proper for aligning peaks one compound but might be too big for another compound and aligning multiple peaks from the same sample. The same fixed m/z bin might be too small for the compound and lead to peak misalignment. Even though m/z bin size is big enough but the mean m/z of the compound might not always locate at center of m/z bin decided by detected min and max m/z value and user defined bw value. The reason why increasing mzwid in XCMS did not always decrease misaligned peak number in XCMS alignment result is now clear. Increasing mzwid value might be a temporary solution for some compounds but not a global solution for all compounds, as we can see how adjusting mzwid affect the alignment in Figure1.14 we just showed earlier.

In LAKE alignment result, Figure 1.15 are shown on how adjusting parameters affect LAKE alignment result. The adjusting RT tolerance did not explicit affect misalignment peak number from Figure 1.15 a, b and c. From figure 1.15a, d and g, the misaligned peak number did decrease when increase m/z tolerance in LAKE alignment algorithm. The misaligned peak numbers in LAKE alignment result can be decreased by increasing m/z tolerance which is consistent with our expectation.

The number of misaligned peak in LAKE is relative less than that in XCMS in general and can be seen in both Figure 1.16 and Figure 1.18. Because using relative mass difference when m/z clustering in LAKE alignment, the cut lines separating misaligned peaks from the clustered peak group in XCMS alignment (Figure 1.17) not happened the LAKE alignment result. The misaligned peaks in LAKE alignment algorithm are not aligned with the clustered group because the distance between peaks and the center of estimated peak group are larger than the estimated tolerance and treated as outliers.

The misaligned peaks in XCMS and that in LAKE affect statistical processing in



various degrees. In XCMS alignment result, peaks from different batches might not be aligned together, in the end, the intensities of the compound in some samples are zero, which can significant affects when applying statistical processing. In LAKE alignment result, peaks from different batches are aligned together, the intensity of the compound in some samples are represented by fewer replicates but not all zero ,which has slight affect when applying statistical processing.

1.5.2 *Comparison of LAKE and XCMS Algorithm on Metabolomics Data Set*

To test average performance on two alignment algorithms, the random noise introduced data set are selected as evaluation data set. Figure 1.19 shows most misaligned peaks in XCMS alignment result are from the compounds with large m/z value (the compounds with large compound ID, m/z >300) and aligned multiple peaks from the same sample together when aligning compounds with small m/z value (the compounds with small compound ID).

Figure 1.20 shows LAKE alignment on the same data set with much less misalignment peaks but equal or even more aligning multiple peaks from the same sample together when comparing to XCMS alignment result.

In the selected three aligned compounds, Figure 1.21 shows the misaligned peaks in XCMS and the peak group are separated by invisible cut lines and in this case we found there are m/z differences among batches: peaks in the 1st batch are distant away from the rest of batches, and most misaligned peaks in the three selected compounds are from 1st batch. However the difference among batches in m/z value did little or none on the LAKE alignment result (Figure 1.22).

Figure 1.23 shows the misaligned peaks in the data set are majorly from these

compounds with large m/z value and there seems exist a threshold in z-score. If the z-score of the peak is over the threshold, the peak is treated as outlier and not aligned with the other peaks from the same compounds. The threshold of z-score in m/z of XCMS alignment result decrease when m/z value of the compound increased. This might imply it is inappropriate using fixed absolute mass difference when m/z grouping.

To show average performance of two algorithms on data with global shift among batches, In Figure 1.24 and Figure 1.25 shows how global shift among batches existed in both m/z and RT dimension would affect the XCMS alignment result. When the global shift among batches existed in both m/z and RT dimension, the misaligned peak number and the peak number of multiple peaks from the same samples aligned together would be affected. The misaligned peaks in BR+GNLM1-2 can be seen in Figure 1.24, and the reason why most misaligned peaks are from 1st batch can be observed from Figure 1.13a, the peaks from 1st batch are distant away from the rest of batches in m/z dimension. In BR+GNLM1-3, most of misaligned peaks are from 2nd batch because from Figure 1.13b, peaks in 2nd batch has larger m/z deviation than other batches do ,in other word, peaks in 2nd batch has higher chance of being treated as outliers when cluster peaks of the same compound from other batches when using XCMS alignment algorithm. Redundant peaks alignment found in BR+GNLM1-2 for all four batches are with little overlapped with other batch and is very likely to be aligned with other peaks from compounds with similar m/z and similar RT values from all batches (8 second RT range). Redundant peaks alignment found in BR+GNLM1-3 is less than that in BR+GNLM1-2 for peaks in all batches share similar RT distributions (5 second RT range). However, the real cause of redundant peak alignment is not known by adjusting m/z only or adjusting RT only. This can only be answered after the following discussion on adjusting each dimension, m/z or RT, at a time.

In Figure 1.26 and Figure 1.27 show how global shift among batches existed in only m/z dimension would affect the XCMS alignment result. When the global shift among batches existed in only m/z dimension is large (BR+GNLM1-4), the misaligned peak number would increase. Even though the global shift among batches existed in only m/z dimension is small (BR+GNLM1-5), the total misaligned peak number for all batches did decrease but not the misaligned peak number for each batch. The misaligned peak number can be seen from 2nd batch in BR+GNLM1-5 (Figure 1.27) which was not seen in BR+GNLM1-4 (Figure 1.26).

In Figure 1.28 and Figure 1.29 show how global shift among batches existed in only RT dimension would affect the XCMS alignment result. Comparing these results, the misaligned peaks in two data sets are almost identical (blue blocks position in Figure 1.28 and 1.29). When the global shift among batches existed in only RT dimension is large (BR+GNLM1-6), the peak number of multiple peaks from the same sample aligned together would increase. The global shift among batches existed in only RT dimension is small (BR+GNLM1-7), the total misaligned peak number for all batches would decrease.

From the previous results, the global shift among batches in m/z and RT dimension arise different alignment problems in XCMS alignment result. The global shift among batches in m/z dimension affects the misalignment peak number (FN in performance evaluation) in XCMS alignment result. The global shift among batches in RT dimension affects the peak number of multiple peaks from the same sample aligned together (FP in performance evaluation) in XCMS alignment result.

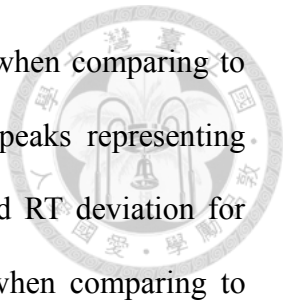
The misalignment in XCMS alignment results of one selected compound introduced with different global shift among batches is shown in Figure 1.37. The invisible cut lines separate misaligned peaks and clustered group mentioned in previous

discussion can still be seen in the result of the data set. Figures 1.31-1.36 again show that absolute mass difference used in m/z binning of XCMS alignment is not appropriate. The misaligned peaks in m/z value versus z-score shows minimum absolute z-score of misaligned peak decrease when m/z value of the compound increases (Downward black lines for outlining the cut lines in m/z z-score versus compound ID plot in Figures 1.31-1.36).

In LAKE alignment result on data set introduced with different global shift among batches, LAKE alignment results of seven different data sets are exactly the same. Therefore one of result is selected and shown in Figure 1.30.

The reason of LAKE alignment on seven different data with the exactly same alignment results is that despite the difference among batches, the m/z-RT data within the same batch is exactly the same for each batch. Therefore, if the average m/z and average RT values, calculated in aligning peaks from the same batch, from all batches are not far from others, the alignment results should be the same as the result on the data without global shift among batches. Therefore, no more misaligned peaks found in data with different global shift among batches in either m/z or RT dimension as shown in Figure 1.31-1.36,1.38.

In the metabolomics data set adding random shift with global shift among batches in both m/z and RT for each batch (BR+GNLM1-X, X=1…7), the noise of global shift among batches has stronger effect on XCMS alignment result than that on LAKE. The misaligned peaks in XCMS alignment result is because the algorithm aligns all peaks at a time without considering the difference among batches. LAKE algorithm aligns these peaks according to the similarity of profiling condition. The similarity of profiling condition is the similarity of condition when peaks from different samples are detected. The peaks representing same compound from the same sample should have smaller m/z



and RT deviation for peaks detected under the same matrix effect when comparing to peaks representing same compound from different samples. The peaks representing same compound from the same batch should have smaller m/z and RT deviation for peaks detected under the same condition within the same batch when comparing to peaks representing same compound from different batches. In general, the deviation of m/z and RT among different samples is smaller than that among different batches. Therefore, LAKE peak alignment algorithm starts from aligning peaks from the same samples, peaks from the same batch and finally peaks from all batches. The peaks represent the same compound from different samples with different global m/z shift and RT shift from each batch can be properly clustered only when not acceptable large m/z and RT tolerance used in XCMS peak alignment. LAKE can properly cluster these peaks without increasing m/z and RT tolerance by only aligning average m/z and average RT values of clustered peak groups from each batch. The global shift among batches in the test data has least effect on LAKE alignment result.

The compounds with batch misalignment can be seen in few compounds when peak alignment on complex sample. Take the plasma samples as a study case; the compound with batch-misaligned peaks over all compounds for each batch is 2% in average. LAKE can find the misaligned peaks and recover 6% peaks for batch-misaligned compound in each batch in average. The difference between the peak table generated by LAKE peak alignment and the peak table generated by XCMS peak alignment can be shown in PCA in Figure 1.39 with 2% difference on the explained variance of first PCA component.

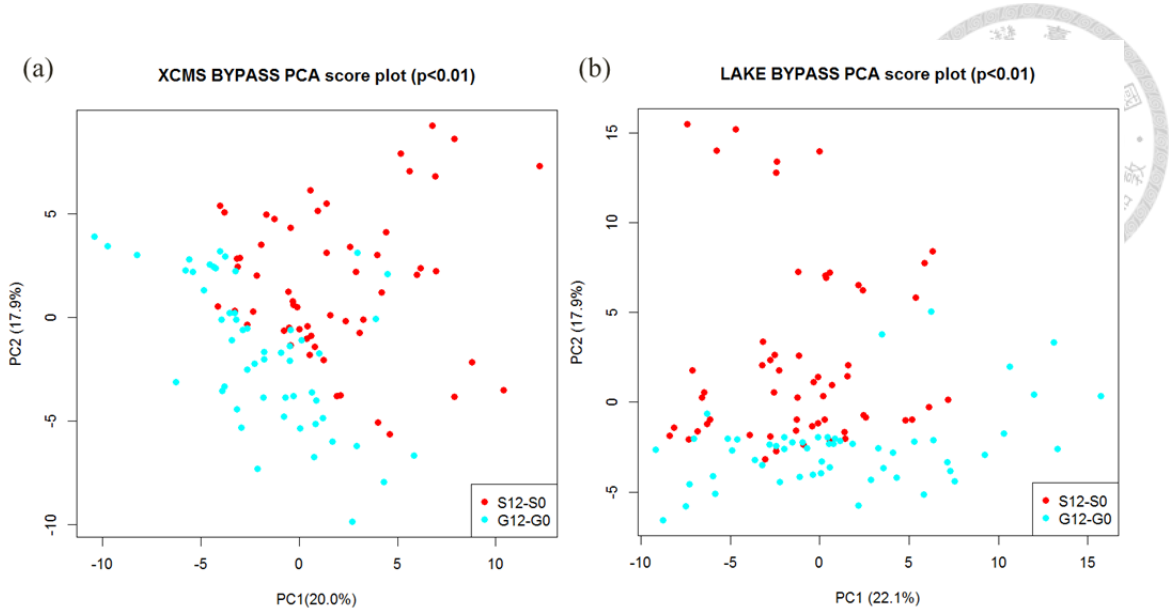


Figure 1.39: The PCA of two peak table done by two different peak alignment algorithms.

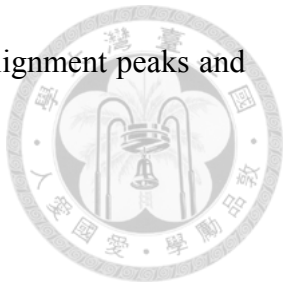
1.6 Conclusion

Peak alignment is a critical step data processing because wrong alignments can lead to wrong statistical analysis. LAKE, a new peak alignment algorithm for LC/TOF-MS-based data without user specified parameter and save the time on looking for optimal parameters for peak alignment.

LAKE performs peak alignment based on data similarity and performs peak alignment from clustering the technical replicates, clustering the samples from the same batch, and merging into an aligned peak table. LAKE, a density-based approach with Epanchnikov kernel function to directly divide group of peaks properly, estimates different bandwidths for different compound mixture with similar m/z value, and generates aligned peak table with less misaligned peaks or aligned peaks from different compounds.

LAKE successfully performed peak alignment on the data with different types of noise existed in experiment. Hence, analyzing LC/TOF-MS-based metabolomics data

with LAKE can get peak table with less zero value because of misalignment peaks and lead to more reliable statistic analyzed conclusion.



Chapter 2 PHASION: PHASing Intrinsically On NMR Spectrum

2.1 Introduction

Proton nuclear magnetic resonance (^{1}H NMR) has been a dominated tool in biomedical research[74-85] and the analytical workhorse for metabolomics research.[78] Scientists from clinical and toxicological research used ^{1}H NMR as the tool for structural characterization/identification of metabolites.[78] Biofluid analyzed in ^{1}H NMR become a common method because nucleus is ubiquitous in organic molecules and the specific functional groups of compounds. We can know the information about the chemical structure, the chemical environment, the dynamic molecular motions and molecular interactions of the molecules to which they are attached from these signals.[78, 86-90] There is growing interest in analyzing samples with ^{1}H NMR in metabolomics for non-destructive, non-selective, cost effective, and only take few minutes per sample, requiring little or no sample pre-treatment or reagents[81, 86, 91] and generating highly reproducible spectra with high signal-to noise ratio.[78] However, there is still strong peak overlap in certain chemical shift ranges even though high field, or high resolution, ^{1}H NMR spectrometer is used.[78] Therefore we need to process the highly reproducible data of biofluid. Actually, the critical step in "Classical" NMR based metabolomics approach is post-experimental data handling including NMR spectra processing, data pre-processing and data analysis.[92]

In most current metabolomics studies, a better data preprocessed result relies heavily on experienced users and large correction differences exist from different people or laboratories.[83] In the following paragraphs, we will introduce what would be done in

data pre-processing.

The raw data generated from ^{1}H NMR is called free Induction decay (FID) spectrum and the Fourier transformed FID is the spectrum we are familiar with. However, the raw FID is generated with multiple sources of interference, such as baseline drifting, the incoherent phase between generated signal and receivers, etc. The data preprocessing is needed before converting a FID to a spectrum. In Figure 2.1, we can briefly know how a raw FID converts into a spectrum.

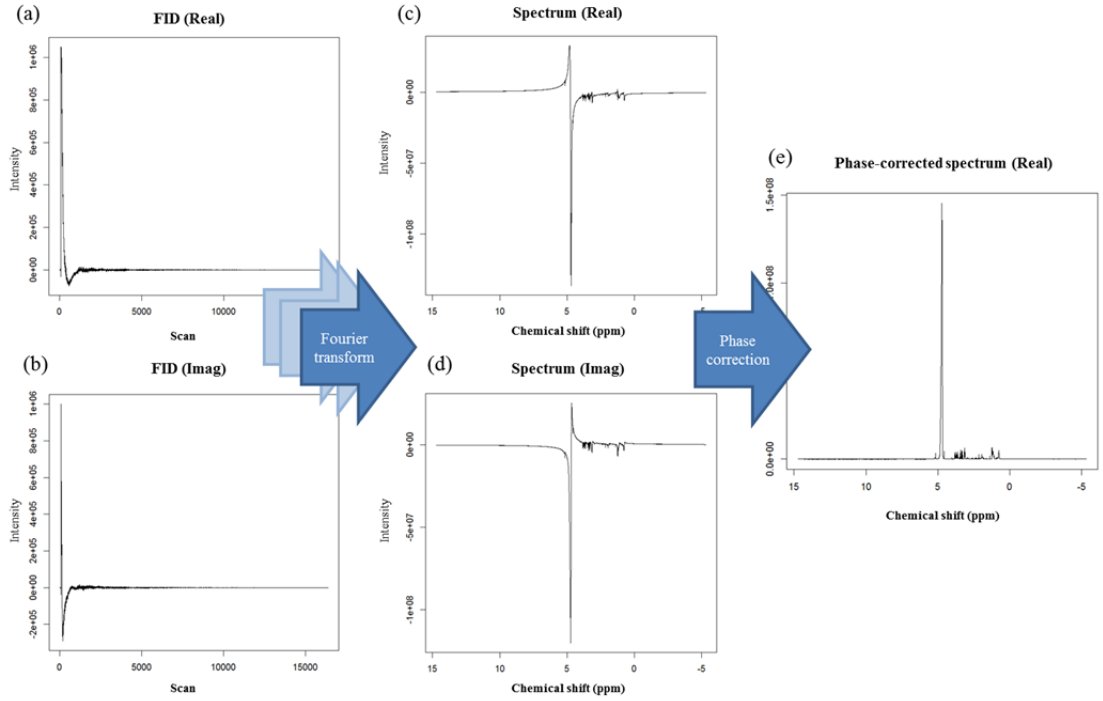


Figure 2.1: Data processing from a raw FID to a spectrum. A spectrum is the Fourier transformed FID. Before further data processing, the spectrum need to be phase corrected. The phase correction can be done by finding the optimal phase angle (PH0, PH1) which minimizes an objective function. (a) The curve represents the real part of the FID. (b) The curve represents the imaginary part of the FID. (c) The curve represents the real part of the spectrum. (d) The curve represents the imaginary part of the spectrum. (e) The curve represents the phased spectrum.

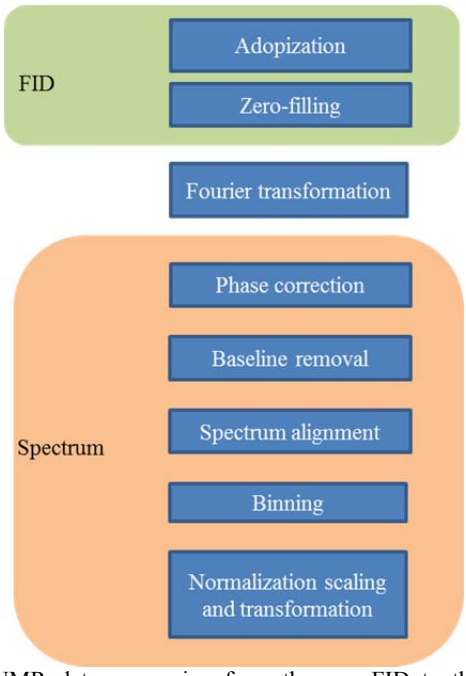


Figure 2.2: The workflow of NMR data processing from the raw FID to the spectrum before applying the statistical analysis.

The basic NMR preprocessing is consisted of two major parts: FID processing and spectrum processing, as shown in Figure 2.2. Before applying Fourier transform to the raw FID, apodization and zero-filling[93] are needed to be done. After the spectrum is generated, phasing, baseline removing, spectrum aligning, binning, and normalization scaling transformation[94] are needed to be done before applying any statistical analysis.

Phasing, or phase correction, is the most important and difficult part to handle among the spectrum processing steps[95] because ill-phased spectrum would lead to inappropriate and poor analyzed results. There is still no acceptable spectrum returned from existing automatic phasing algorithms. In short, the phasing is to flip negative intensity to make whole intensity like the intensity shown in the absorptive mode spectrum, and satisfy with following three criteria[95] : (1) No negative peaks, (2) The baseline is flat and (3) Peaks are symmetrical and narrow.

Phase distortion is phase incoherent recorders when generating raw FID and it gives NMR spectrum negative intensity peaks which are unacceptable in pure

absorption mode. In Figure 2.3 we can see the unphased spectrum with a big negative peak at 4.8 ppm and other countless negative peaks. This problem can be solved by introducing two parameters, zero-order phase error (PH0) and first order phase error (PH1),[95] and find the optimized parameter set (PH0,PH1) then phasing spectrum with the parameter set.

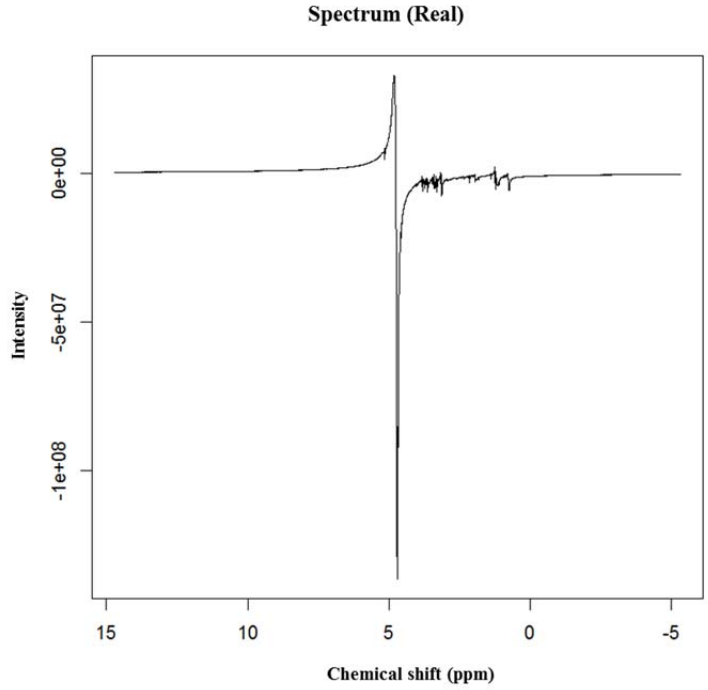


Figure 2.3: The distortion in the spectrum which is caused by the imbalance in quadrature detectors.

The zero-order phase error is induced by reference phase and receiver detector phase,[96] while first order phase error is time delay between excitation and detection, flip-angle variation across the spectrum, and phase shifts from the filter employed to reduce noise outside the spectral bandwidth.[96-99] The zero-order phase error (PH0) can be corrected by adding a constant phase angle to both real part and imaginary part for all time scan, so the error is also called as frequency independent phase error. The first-order phase error (PH1) can be corrected by adding a fixed slope rate of phase change over time, so the error is also called as frequency dependent phase error. In

Figure 2.4, we can see how PH0 and PH1 affect spectrum. The dominated error is PH0 for obvious change of spectrum with change of PH0. Some autophasing algorithms[92] featuring correcting PH0 then correcting PH1 based on effectiveness of altering PH0 and PH1 on spectrum variation.

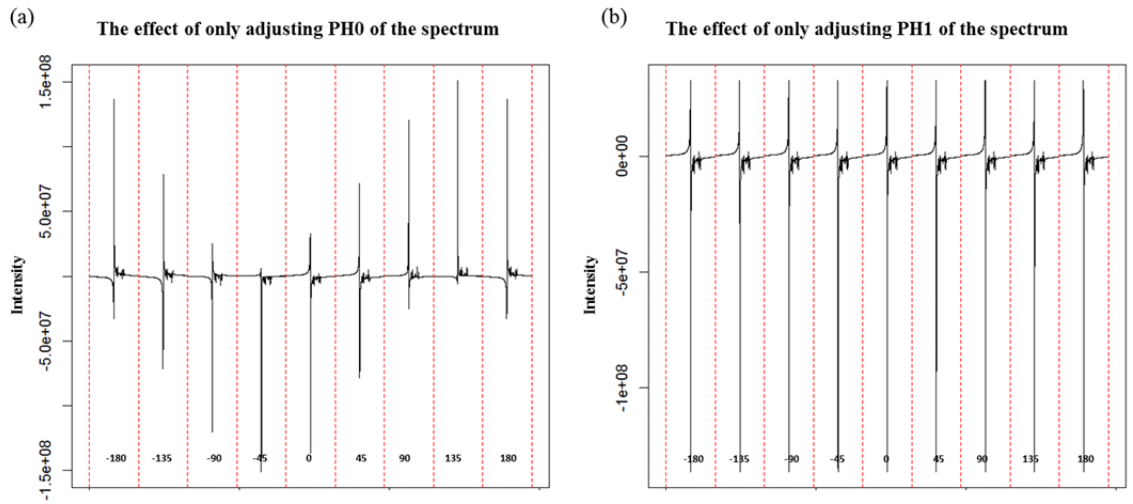


Figure 2.4: The effect of adjusting the phase angle when phasing a spectrum. (a) The effect of only adjusting PH0 of the spectrum. The numbers indicate the degree of PH0 used in phasing the spectrum. (b)The effect of only adjusting PH1 of the spectrum. The numbers indicate the degree of PH1 used in phasing the spectrum.

Phasing is important because ill-phased spectrum will lead to wrong baseline extraction and wrong baseline-removed spectrum which will result in misinterpreted statistical analyzed results. Several phase correction algorithms have been proposed,[92] but some required experienced operators manually phasing one spectrum after another which is time-consuming and hard to correct spectrum with same criteria all the time. Autophasing algorithms were proposed[95] to solve the problems mentioned above and the algorithms with different criteria of autophasing have been proposed.

Peak shape method[100] using peak shape as the scoring function when adjusting PH0 and PH1. However, it was failed for phasing in polymer spectra and wrong assumption of the characteristics of absorption mode.[95] Baseline method is based on

assumption of phased spectrum should be with the one with most baseline points when compared with the other spectrum using different (PH0,PH1) for phasing, so they adjust PH0 and PH1 based on the number of baseline points with automated baseline selection procedures.[101] Modified baseline method using advance baseline selection procedures is also proposed.[102] In the ideal spectrum method is based on assumption of peak are distant away. Sterna *et al*.[103] proposed a method that the spectrum can be phased by finding the best PH0 then adjusting PH1 according to linear least square fitting result of two baseline. Džakula[104] assumed the phased spectrum should be with minimal summed peak area. Therefore, the spectrum can be phased by finding the (PH0, PH1) which with minimal summed peak area. However, these methods failed since not all spectrum with well-defined peaks.[95] Minimal entropy method is proposed based on the assumption of the phased spectrum should be smooth, therefore with minimal entropy of the first derivative of real spectrum. In some implementation, minimal entropy method can combined with negative area penalization.[96] However, it is reported unbalance weight between the entropy and the negative area penalty for the polymer spectra mentioned in de Brouwer's study.[95] The minimal entropy only scoring function is less robust and less accurate.

Other approaches like Hybrid approach, the objective score function consists of MinArea, MaxBaselinepoint and MinNegAreaSquare.[95] The phased spectrum with the least area and the least negative area but it needed relative more time than other existing algorithms did. A much simpler approach is Automics.[92] Phase angles are evaluated from low frequency part and high frequency part of spectrum which are signal-free and calculate the PH0 and PH1 based on the assumption of linear variation of PH1. The method need least time but relative unstable when baseline is severely distorted.[92]

Optimizer selection is an important part in autophasing. The most common

optimizer is Nelder-Mead optimizer.[105] The first one introduce NM optimizer into analytical chemistry are Morgan *et al*..[106] Different optimizers had been tried, such as Powell's steepest descent,[107] and quasi-Newton method[108] but the optimized result was still not satisfied.[101] Someone even doubts the best optimizer for phase correction should not be NM method.[109] However this is still a popular choice of optimizer featuring faster convergence and less parameter tuning and in most cases.

However the searching for optimized parameter set (PH0, PH1) over the big searching space is time consuming and sometimes converge to local optimal depends on initial points, which is risky for not always returning global optimal result. Several systematic searching methods have been proposed, and most of them search based on simplex approach,[105, 110] the main difference between them is different criteria, or objective function, for acceptable corrected spectrum. The corrected spectrum should be in pure absorption mode. The definition of absorption mode as followed:[97, 111]

Customary mode of displaying NMR spectra, in which peaks have nominally Lorentzian shapes.

which featuring without negative intensity, flat baseline, and narrow peak shape.[95]

Recently mixing criteria from the mentioned criteria above is proposed.[95] Since there are still some problems unsolved and sometimes return unsatisfied phase spectrum, which means we need a better phase correction method. We here try to propose a new method to solve the issue of not all spectra can be correctly well.

We are inspired by the objective function of mixing criteria.[95] The "*PHASing Intrinsically On NMR spectrum*," *PHASION*, algorithm is consisted of two parts: (1) the optimizer and (2) the proposed objective score. PHASION is proposed for better phased spectrum, better convergent result, and reducing both computing time and searching time for optimal solution.

2.2 Material

For the analysis of metabolomics study, two groups of subjects including normal controls and bronchiectasis who were collected from National Taiwan University Hospital were enrolled in this study.

A total of 400 μL of 0.9% NaCl in 10% D_2O was added to 200 μL of plasma sample and then mixed well. The resulting samples were centrifuged at 12,000 rcf for 5 min then transferred 550 μL to NMR tube. Plasma ^{1}H-NMR spectra were acquired using both a Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence and a diffusion-edited sequence. The diffusion-edited sequence of the ^{1}H-NMR spectrum characterized the lipid molecules of lipoproteins. ^{1}H-NMR with the CPMG pulse sequence was used to detect the low molecular weight molecules by suppressing most of the broad resonances. Thirty-five samples were selected for demonstrate proposed phase correction algorithm on complex samples.

For the analysis of the synthesized spiking NMR spectra, the 25 pure compounds consist of acetate, alanine, betaine, creatine, creatinine, cytosine, ethanol, formate, fumarate, glutamate, glutamine, glycerol, glycine, hippurate, histidine, leucine, myoinositol, phenol, pyruvate, succinate, taurine, threonine, tyrosine, uracil, and methanol were introduced to generate a set of simulated spectra. The signals of spiking compounds distribute over a spectral range of 0.9–8.5 ppm. These pure compound spectra were produced from our laboratory with the same procedures as the cell extract spectrum or downloaded from the Web site of the Human Metabolome Database (HMDB). Each spectrum was manually aligned, phased, baseline-corrected using the software of Chenomx 7.6, and then output into R v. 3.1.2 for further processing. The signal-free segments were set to zero with an intensity threshold. All of the spectra were

normalized to an equal maximal intensity. The cell extract spectrum was baseline-corrected with BaselineCorrector, but the signal/noise points were assigned manually with the expert-defined signal/noise points. One set of 25 uniformly distributed random coefficients given with U (0.02, 0.2) were generated. One spiked simulated spectra were produced by summation of spectral intensities of the cell extract spectrum and the pure compound spectra multiplied by random coefficients. The spectra were well phased and baseline removed. The imaginary part of these spectra are absent, so we used ACD to calculated corresponding imaginary part of these spectra by phasing with phase angles (PH0, PH1) = (0, 0). We then introduced PH0 and PH1 error here. The introduced phase error is observed from 35 plasma samples and the optimal phase correction by ACD shown in Figure2.5. The phase 0 error is sequence start from 110 to 140 with step size = 5 and phase 1 error is a sequence from -10 to 10 with step size = 5, so the test phased error combination is for 7*5=35 combinations. Two data sets are created: noise existed in phase 0 error (FIXED_PH1) and noise existed in phase 1 error (FIXED_PH0). The introduced noise is Gaussian noise with standard = 1. In FIXED_PH1 data set, the phase 1 error introduced stays the same for all data points in the spectrum but the phase 0 error introduced would be centered at specified ph0 with standard deviation = 1.

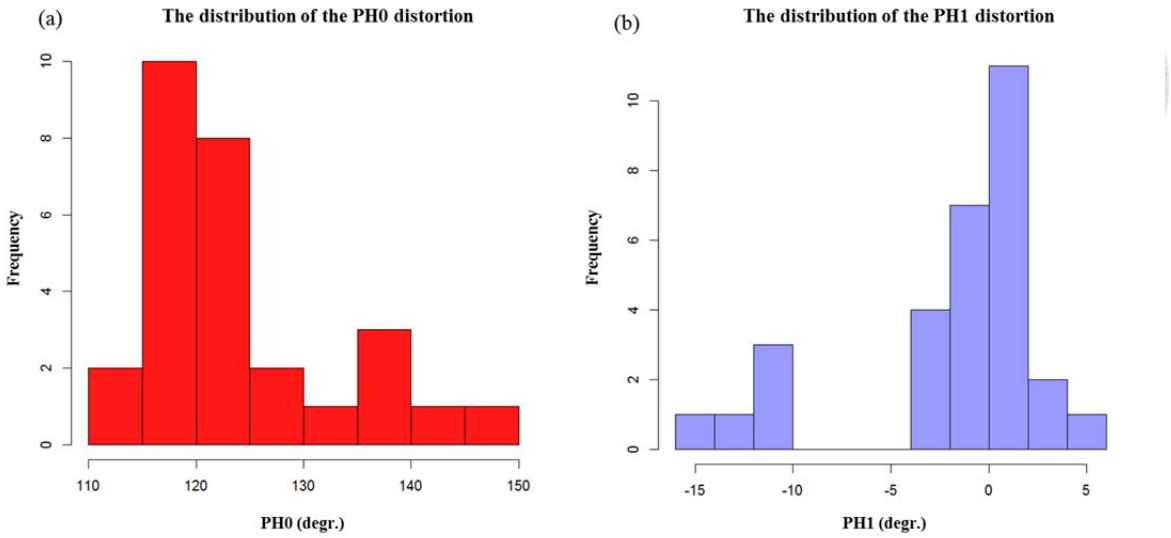


Figure 2.5: Phase distortion observation via ACD autosimple phase corrected spectrum. The following subfigures are the distributions of (a) the PH0 distortion and (b) the distribution of the PH1 distortion observed from 30 random selected samples.

2.3 Theoretical Basis

Each spectrum was generated using ACD/NMR processor version 12.01, and then output into R v.3.1.2, Chenomx 7.6 and ACD for auto phase correction.

The flowchart of PHASION is shown in Figure 2.6. The autophasing is generally consisted of two parts: (1) the optimizer and (2) the proposed objective score. In the optimizer part, the simplex method is the most commonly used method for easy implementation and the converging process with physical meaning, and can return acceptable optimized result. In the proposed objective score part, there are more options and combinations of different objective functions, we will briefly introduce in the following paragraphs.

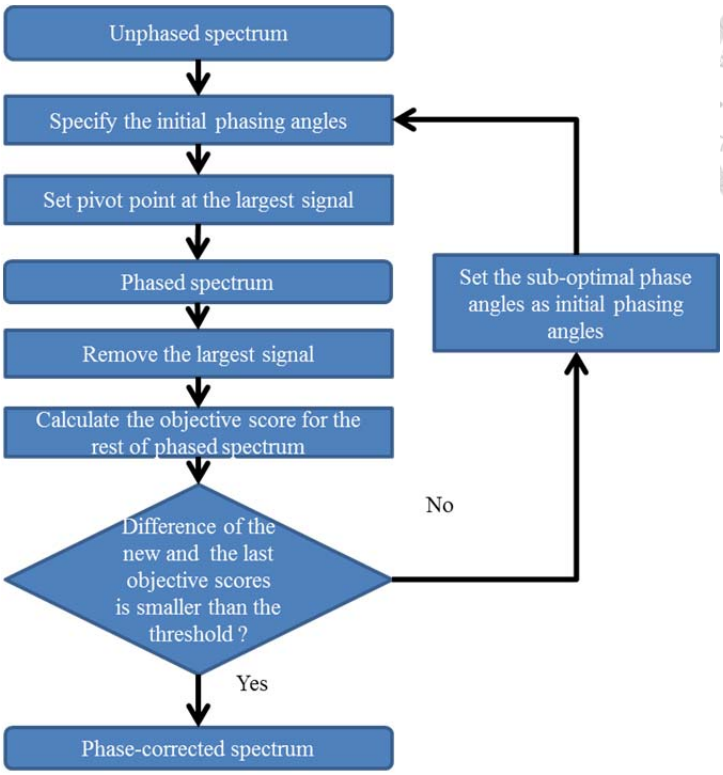


Figure 2.6: The detailed of workflow of PHASION.

2.3.1 *Data Pre-processing*

In data preprocessing stage, we will process Fourier transformed spectrum before searching for optimal PH0, PH1 for autophasing. In data pre-processing stage: (1) Pivot selection, (2) Smile artifact elimination and (3) Partial spectrum selection for autophasing.

Pivot selection is to define the zero degree of PH1. The spectra we processed are spectra without removing water signal. In general, the pivot is set at chemical shift of the highest signal in the spectrum. In most case, the pivot point of ^{1}H-NMR spectrum is set at about 4.8 ppm, the chemical shift of water.

After the pivot point is selected, we remove first and last 800 data points from the spectrum. In most of spectrum, the smile artifacts [112, 113] can be seen at each end of spectrum pointing up for inappropriate data preprocessing on converting from FID to spectrum when the FID is with group-delay. And the smile artifact would affect our

searching for optimal phase angles.

Next step is to keep the spectrum except for the part of spectrum contains largest intensity in the spectrum with window size 0.6 ppm centered at ppm of the largest signal in the spectrum. The part of spectrum contains either water signal or the largest signal in the spectrum is sensitive to the phase change so we select the rest of spectrum, the spectrum without the largest signal in the original spectrum, to find the optimal phase angle for the unphased spectrum. With lesser data points and relative stable spectrum, optimal phase angles searching required less calculation time and the optimal phase angles searching can converge in shorter time.

2.3.2 *Nelder-Mead Optimizer*

The Nelder-Mead optimizer is the most commonly used optimizer in optimal phase angle searching with given scoring function for spectrum autophasing. It's easy to understand and the approach is meaningful when working with appropriate scoring function. Different optimizers have been proposed and someone even doubt the optimizer really converge to the global optimal, or the optimal phase angle for the spectrum. The optimal searching would terminate when iteration criteria is met. The iteration criteria for our proposed algorithm is the rounded off to 2^{nd} decimal place difference of the current phase angles and last phase angles is smaller than 0.1. The parameters for Nelder-Mead are: Alpha is the reflection factor (default 1.0), beta is the contraction factor (0.5) and gamma is the expansion factor (2.0) and is illustrated in Figure 2.7. The initial phase angles selection is an important step for Nelder-Mead simplex optimizer. The Nelder-Mead simplex optimizer can return proper result by starting with proper selected initial phase angles. The initial phase angles are selected by searching the minimum of the proposed objective score over (PH0,0) space. PH0 space

is expanded by a sequence which starts from -180 to 180 with step size =5.

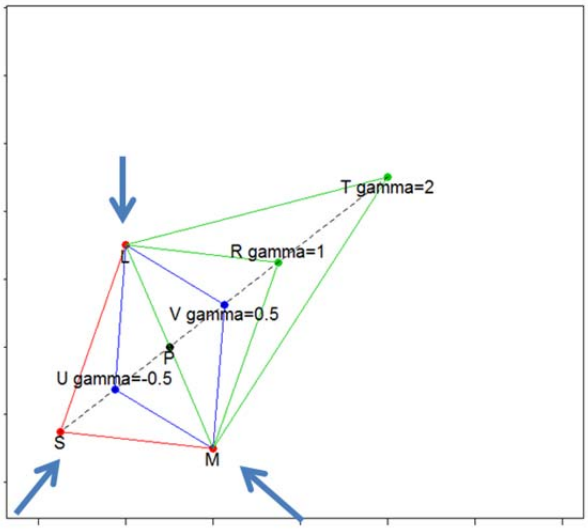


Figure 2.7: The Nelder-Mead simplex optimizer. The triangle SLM for Nelder-Mead simplex optimizer. The Nelder-Mead simplex optimizer is decided by three scaling factors: alpha, beta and gamma. Alpha is the reflection factor (default 1.0), beta is the contraction factor (0.5) and gamma is the expansion factor (2.0). The red triangle would transform its shape during optimization and moving toward the optimal solution of searching space.

2.3.3 *Scoring Function*

The proposed scoring function is composed of two parts: (1) The difference of baseline shoulder and (2) The normalized averaged summed negative area penalty. The penalty in scoring function can be seen in de Brouwer's approach.[95] The penalty would penalize when sharp narrow downward peaks are present in real part of phased spectrum to facilitate optimizer returning least negative peaks in phased spectrum with optimal phase angle pair (PH0, PH1).

The score to describe the baseline differences between two parts spectrum, sdBLdiff, is proposed to find the optimal phase angle which phased spectrum with flat baseline. The sdBLdiff is composed of the distance between the baseline and zero line and the baseline difference between two parts of spectrum. The 1st part of spectrum is

the spectrum from first point of the spectrum to maxintH. The 2nd part of spectrum, is the spectrum from maxintT to last point of the spectrum. maxintH and maxintT are the first and the last point of the largest signal in the spectrum, respectively. A phase-corrected spectrum should be the spectrum with baseline closed to zero and most data points with positive intensities. The sdBLdiff is calculated with the following equation:

$$\text{sdBLdiff} = \frac{\left|\sqrt{\text{sd}(\text{Re}_\text{H}) \times \text{sd}(\text{Re}_\text{T})} - \text{mean}(\text{Re}_\text{H}, \text{Re}_\text{T})\right| \times \left|\text{min}(\text{Re}_\text{H}, \text{Re}_\text{T})\right| + \text{meanBL}}{\text{max(Re)}}$$
$$\text{Re}_\text{H} = \left\{ \text{Re}_\text{i} \in \text{Re} \,\middle|\, 0 < i < \text{maxintH AND Re}_\text{i} < 0 \right\} \quad (1)$$
$$\text{Re}_\text{T} = \left\{ \text{Re}_\text{i} \in \text{Re} \,\middle|\, \text{maxintT} < i < \left| \text{Re} \right| \text{AND Re}_\text{i} < 0 \right\}$$

where Re is the real part of spectrum, Re_H is the set of data points from the 1st part of spectrum with negative intensity, Re_T is the set of data points from the 2nd part of the spectrum with negative intensity.

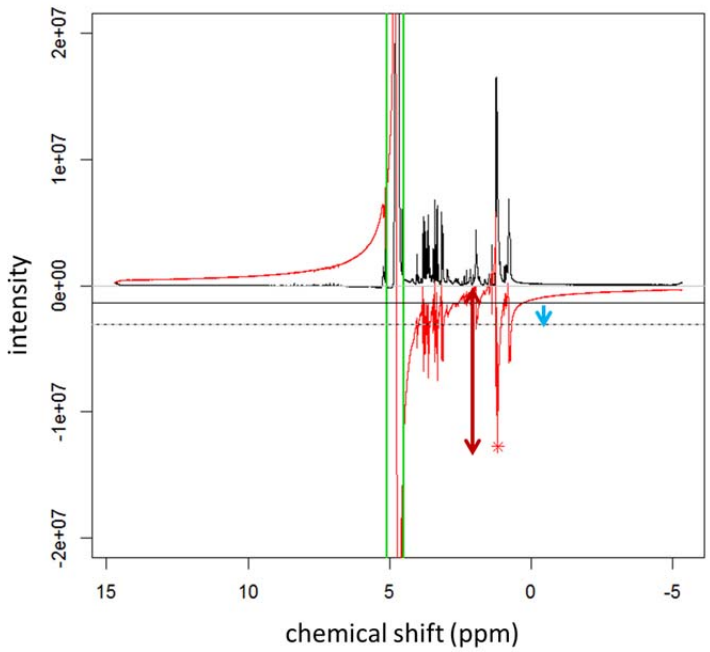


Figure 2.8: The illustration of how sdBLdiff is calculated. The black and red lines represent the phased spectrum and unphased spectrum, respectively. The green vertical lines are boundary lines of the highest signal in the spectrum. The red arrow indicates the minimum negative intensity in 1st part of spectrum. The blue arrow indicates the standard deviation of the negative intensities in 1st part of spectrum.

The meanBL is the score to describe the area between first deciles of 1st part of the spectrum and 2nd part of the spectrum of spectrum and zero line. The baseline of the spectrum can be approximated by computing first deciles of the spectrum, the intensity which is larger than 10% of intensities in the spectrum. A phase-corrected spectrum should be the spectrum with baseline closed to zero, with least summation of green area in Figure 2.9. The meanBL is calculated with the following equation:

$$\text{meanBL} = \left|\text{quantile}(\text{Re}_{1st}, 0.1) \times \text{length}(\text{Re}_{1st})\right| + \left|\text{quantile}(\text{Re}_{2nd}, 0.1) \times \text{length}(\text{Re}_{2nd})\right| \quad (2)$$

where Re_{1st} is 1st part of the spectrum, Re_{2nd} is 2nd part of spectrum.

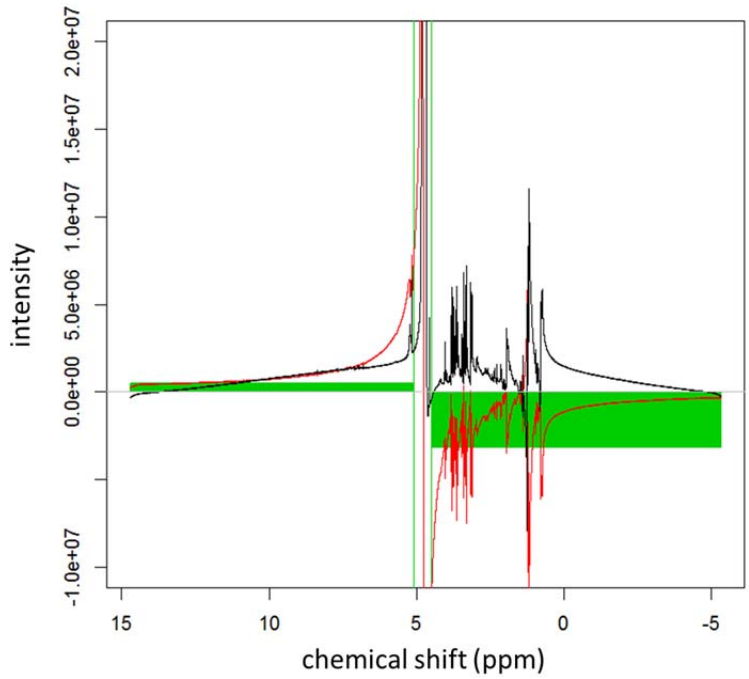


Figure 2.9: The illustration of how the area between zero and first deciles of two parts of spectrum is calculated. The black and red lines represent the phased spectrum and unphased spectrum, respectively. The green vertical lines are boundary lines of the highest signal in the spectrum. The green colored area represents the area of between zero and first deciles of two parts of spectrum.

The normalized averaged summed negative area, NSMnegPenalty, is the summation of negative intensity in the spectrum phasing with (PH0, PH1). The summed negative area is proposed to find the spectrum phased with the optimal phase angles with least negative intensity. A phase-corrected spectrum should be the spectrum with least data point with negative intensity, with least summation of red area in Figure 2.10. NSMnegPenalty is calculated with the following equation:

$$\text{NSMnegPenalty} = \frac{\sum_{i \in \text{Neg}} |\text{Re}|}{\max(\text{Re}) \times |\text{Neg}|} \qquad (3)$$

$$\text{Neg} = \{ i \mid \text{Re}_i < 0 \}$$

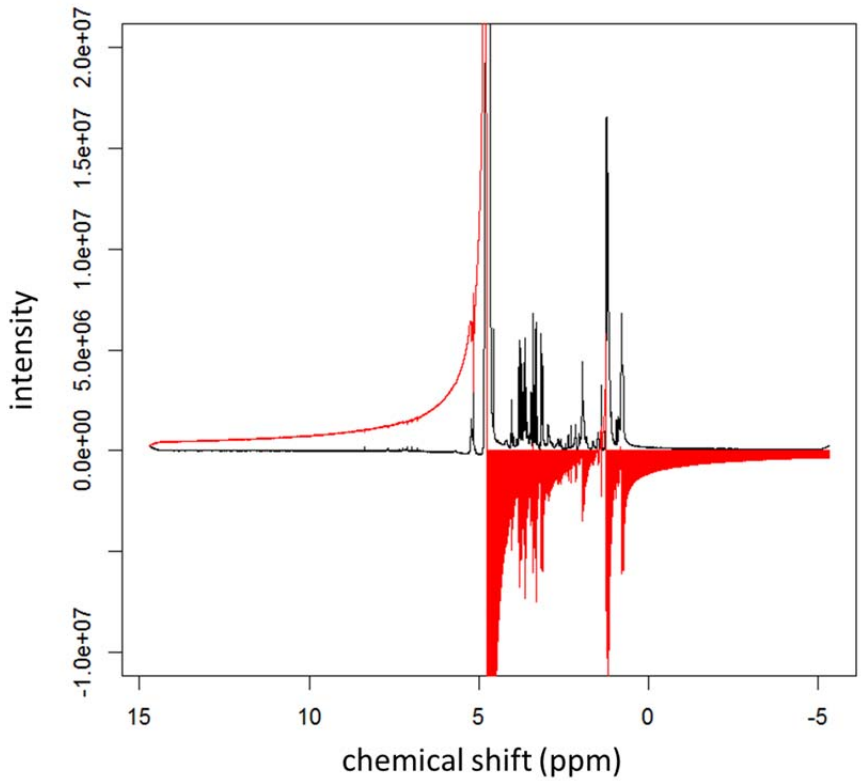


Figure 2.10: The illustration of how the normalized averaged summed negative area is calculated. The black and red lines represent the phased spectrum and unphased spectrum, respectively. The red colored area represents the summed area of negative intensity of the unphased spectrum.

The proposed scoring function is composed of the difference of baseline shoulder (sdBLdiff) and the normalized averaged summed negative area (NSMnegPenalty) with appropriate weighting on these proposed scores and the proposed scoring function is described in following equation:

$$\text{ScoringFunction} = \text{sdBLdiff} + 10000 \times \text{NSMnegPenalty} \tag{4}$$

2.3.4 *Performance Evaluation of PHASION on Phase Correction*

In general, there is no objective measurement of how ill-phased the spectrum is, so here we try to propose a new measurement for spectrum distortion. To evaluate the performance of phase correction on synthesized spiked spectra, the objective

measurement is defined as normalized SSE, normalized sum of squared error, of original spectrum and phased spectrum. The larger value is, the stronger the distortion is. The normalized SSE is calculated with the following equation:

$$\text{normalized SSE} = \frac{\sum_i (O_i - NP_i)^2}{\sum_i O_i^2} \quad (5)$$

where O_i is the real part of *i*th data in the original spectrum, NP_i is the real part of *i*th data point in the phased spectrum with intensity normalized to the maximum intensity of original spectrum.

To evaluate the performance of phase correction on synthesized spectra, the objective measurement are the proposed scoring function and the normalized SSE of original spectrum and phase-corrected spectrum is used for objective measurement of phase correction. The selected NMR phase correction softwares to be comprared with our algorithm are ACD NMR Spectroscopy Softwares (Advanced Chemistry Development, Inc., ACD/Laboratories,Canada) and Chenomx (Chenomx Inc.,Canada).

To evaluate the performance of phase correction on spectra of complex metabolomics sample, the only available objective measurement is the proposed scoring function. The normalized SSE is not used because we can't define phase-corrected spectra to be compared with phased spectrum by different algorithms.

2.4 Result and Discussion

2.4.1 *Convergence of Nelder-Mead Optimization*

Figure 2.11 shows how Nelder-Mead optimization on searching the optimal phase over PH0-PH1 plane with different initial phase angles. The optimizer with our proposed objective score did not converge to global optimal but trap in local optimal in

most cases, that is, the initial phase angles are important for whether Nelder-Mead simplex optimizer would converge to the optimal phase angles. In Figure 2.11(a). The spectrum phased with (0, 0) as initial phase angles converge to (0, 0) for the score distribution around (0, 0) almost stay the same. The score distribution over PH0-PH1 is plotted in Figure 2.11(b).

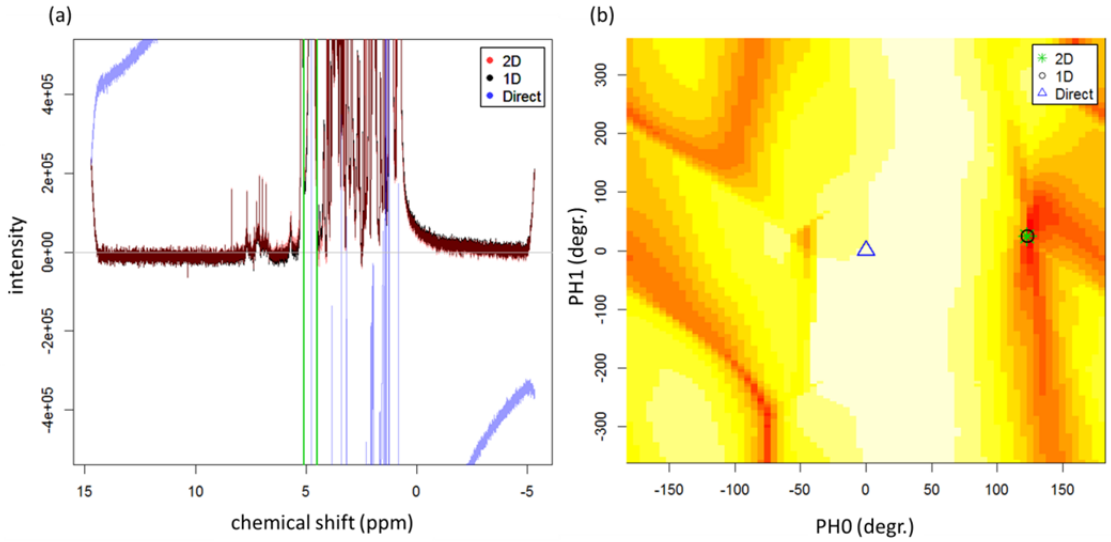


Figure 2.11: Different optimization searching approaches for selecting the optimal phase angle. The green vertical lines are boundary lines of the highest signal in the spectrum. (a) The spectrum phased with the optimal phase angles searching with initial phase angles assigned with (ph0*, ph1*), (ph0*, 0) and (0, 0) are colored in red, black and blue, respectively. (b) The score distribution over (PH0, PH1). The optimal phase angles searching with different initial phase angles assigned with (ph0*, ph1*), (ph0*, 0) and (0, 0) are labeled as green star, black circle and blue triangle, respectively. Yellow and red colored block represent the score of given (ph0, ph1) is high and low respectively.

The score around (PH0, PH1)=(0, 0) stays high score. No score change around this area is the main reason for Nelder-Mead simplex optimizer trapped at (0, 0) when finding optimal phase angles in the (PH0, PH1) plane of proposed scoring function. The spectrum phased with (ph0*, 0) and spectrum phased with (ph0*, ph1*) both return nicely phased spectrum can be seen in Figure 2.11 (red line and black line). The only major differences between these two results are the time consumption. The time

consumption for finding initial phase angles and find optimal phase angles with given initial phase angle (ph0*, 0) and (ph0*, ph1*) are 3 and148 seconds, respectively. The searching space for finding (ph0*, 0) is a sequence starts from -180 to 180 with setpsize=5 on PH0. The searching space for finding (ph0*, ph1*) is a sequence starts from -180 to 180 with setpsize=5 on PH0 and another sequence starts from -360 to 360 with step size =5.

2.4.2 *Comparison of Different Pre-processing Methods*

Figure 2.12 shows the phased spectra using different pivot points. Among these spectra, the most likely phase-corrected spectrum is the one phased with pivot point set at the largest signal in the spectrum, the water signal (Figure 2.12 (c)). The other spectrum phasing with pivot point set at chemical shift=0 (Figure 2.12 (b)) or randomly set (Figure 2.12 (a)) show half of spectrum with negative intensities.

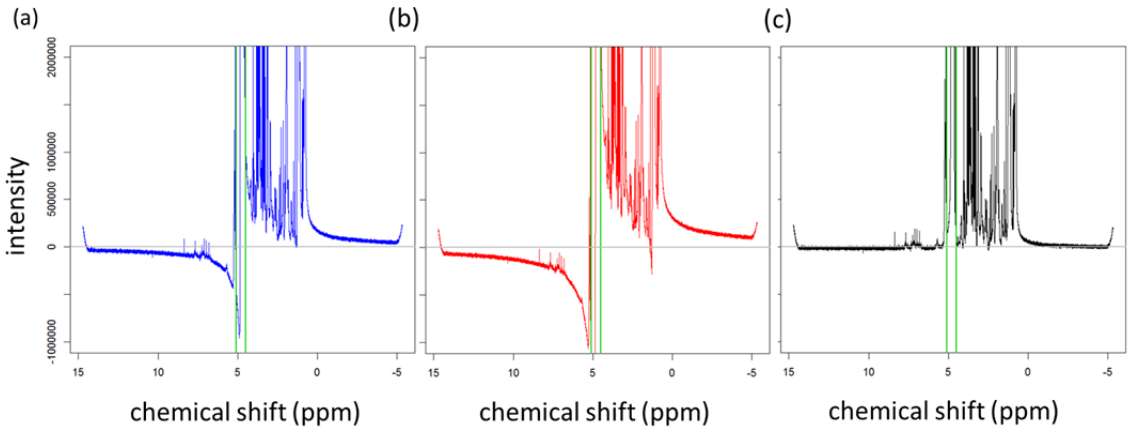


Figure 2.12: Effect of proposed data processing steps for phasing: phase optimization using different pivot points. The green vertical lines are boundary lines of the highest signal in the spectrum. (a) The phased spectrum with pivot point set at arbitrary point (ppm=2.49). (b) The phased spectrum with pivot point set at chemical shift=0 (ppm=0). (c) The phased spectrum with pivot point set at the highest signal in the spectrum (ppm=4.8).

Figure 2.13 shows the phased spectra using different scoring functions and examines how the segment of spectrum with largest signal would interfere with optimal

phase angle searching. The two spectra are phase-corrected by optimal phase angles of optimization with whole spectrum and largest signal removed spectrum. The difference between these two results is the phase-correction near the largest signal in the spectrum. The spectrum phased optimal phase angle with complete spectrum during optimization is distorted near the largest signal in the spectrum ranging from 4 to 7 ppm. In data processing, the water signal can be removed afterward so distortion at this area is acceptable, but in Figure 2.13 the segments of spectrum near water signal distorted are not acceptable.

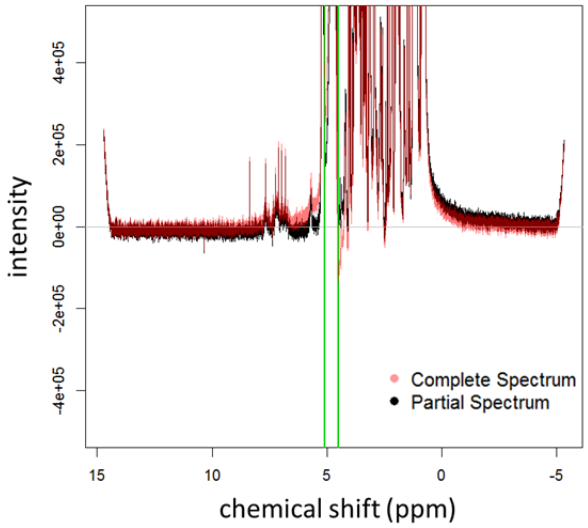


Figure 2.13: Effect of proposed data processing steps for phasing: phase optimization using partial spectrum. Two green vertical lines represent the location of water signal. The phased spectrum with complete spectrum and partial spectrum are colored in red and black, respectively.

Figure 2.14 shows the phased spectra using different weightings on penalty, NSMnegPenalty, in objective function to demonstrate how different penalty weightings affect the phased spectrum and how they affected. The results show the spectrum optimized with heavier penalty (Figure 2.14(c)) with relative flat baseline but the whole spectrum lifted up from zero to avoid heavy penalty from having negative intensity in phased spectrum. The one with no penalty (Figure 2.14(a)) comes with distorted

spectrum near the largest signal in the spectrum. The one with light penalty (Figure 2.14(b)), 10% of penalty, with larger difference between baseline of 1st part of spectrum and 2nd part of spectrum, when comparing with the spectrum phased with normal penalty.

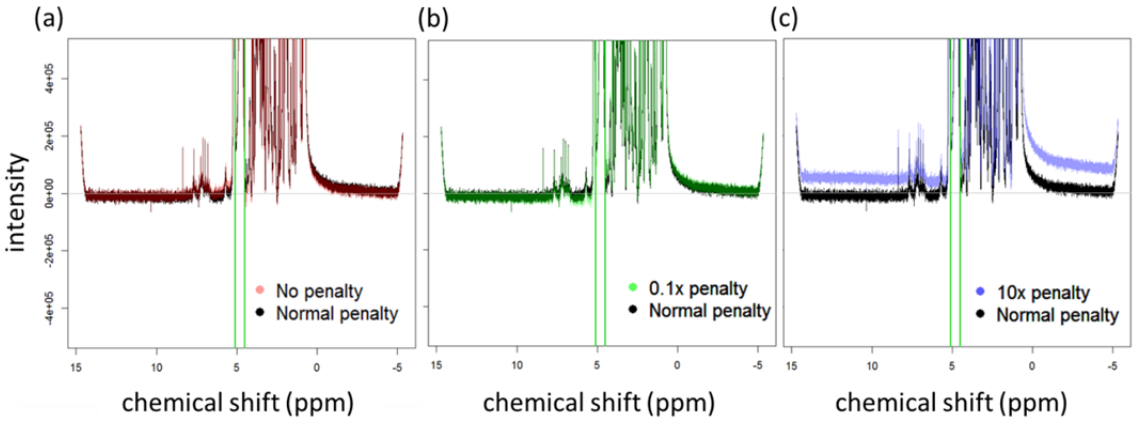


Figure 2.14: Effect of proposed data processing steps for phasing: phase optimization using different types of penalties. Two green vertical lines represent the location of water signal. (a) The phased spectrum with no penalty and normal penalty are colored in red and black, respectively. (b) The phased spectrum with 10% penalty and normal penalty are colored in green and black, respectively. (c) The phased spectrum with 10 times penalty and normal penalty are colored in blue and black, respectively.

Figure 2.15 shows the phased spectra whether the smile artifact would affect optimal phase angles searching. In the result we can see, there is difference in the part of spectrum near the largest signal in the spectrum. The phased spectrum optimization using spectrum with smile artifacts (Figure 2.15 red line) have stronger distortion near the largest signal when comparing with the spectrum optimization using spectrum without smile artifacts.

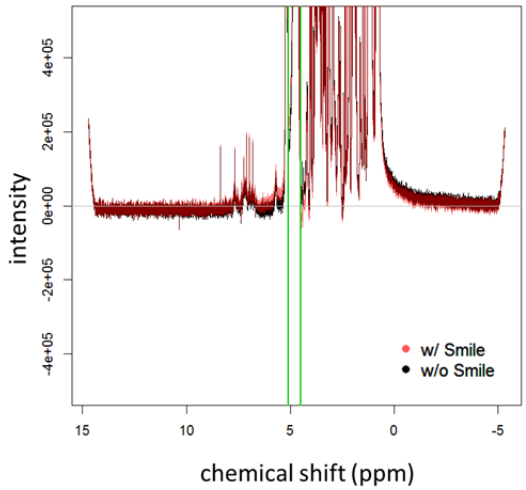


Figure 2.15: Effect of proposed data processing steps for phasing: phase optimization with the smile elimination. Two green vertical lines represent the location of water signal. The phased spectrums with smile artifact and without smile artifact are colored in red and black, respectively.

2.4.3 *Comparison of Performance on Synthesized Spectra with Gaussian Noise Introduced*

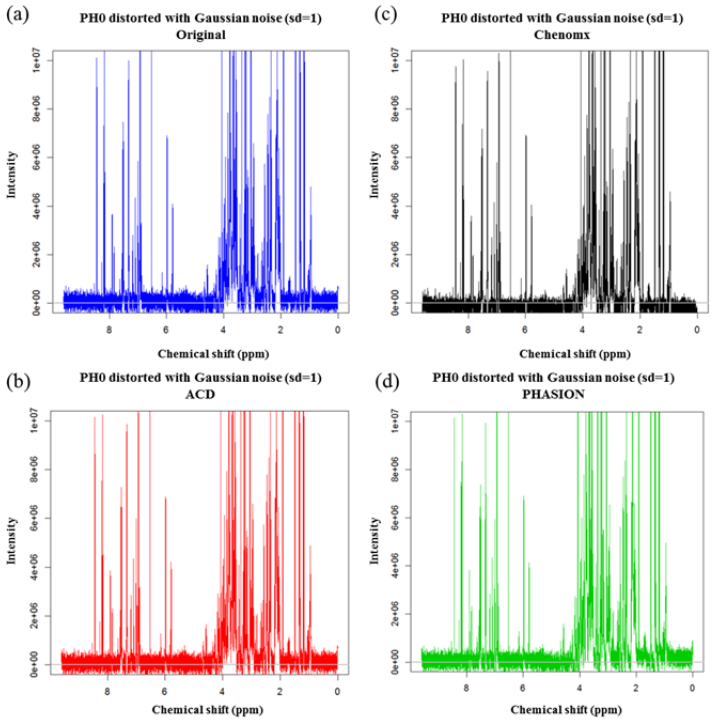


Figure 2.16: Evaluation of different autophasing algorithms on synthesized data. The synthesized data with the PH0 distorted with Gaussian noise (sd=1).

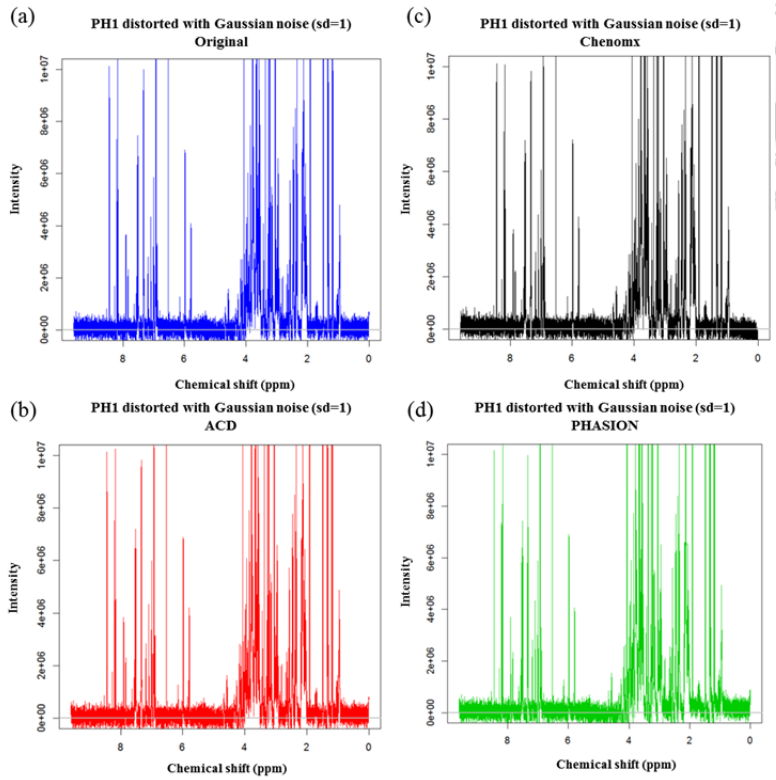


Figure 2.17: Evaluation of different autophasing algorithms on synthesized data. The synthesized data with the PH1 distorted with Gaussian noise (sd=1).

Figure 2.16 and Figure 2.17, we can hardly see the difference among the phased spectra and original spectrum in both the PH0 and the PH1 distorted and with Gaussian noise. Figure 2.18 and Figure 2.19 show the difference between these phased spectra and original spectrum in both the PH0 and the PH1 distorted and with Gaussian noise. The difference is shown in normalized SSE. In Figure 2.18, all phased spectra show large difference between phased spectrum and original spectrum at both ends of spectrum and the chemical shift range from 4.8 ppm to 5.7 ppm. The large difference happened at end of spectrum is phasing with inappropriate PH0. The large difference happened at chemical shift range from 4.8 ppm to 5.7 ppm is inappropriate pivot point is set.

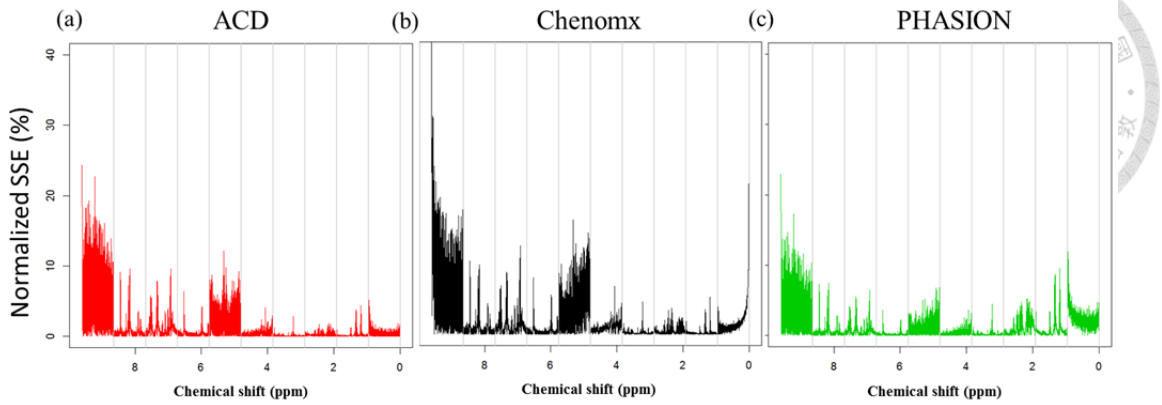


Figure 2.18: Evaluation of different autophasing algorithms on synthesized data introduced with the PH0 distorted with Gaussian noise (sd=1).

In Figure 2.19, all phased spectra show similar result we observed in Figure 2.18, the large difference between phased spectrum and original spectrum at both ends of spectrum and the chemical shift range from 4.8 ppm to 5.7 ppm. However, in Figure 2.19(c), PHASION phased spectrum show smaller difference between PHASION phased spectrum and original spectrum than the other spectra phased with different algorithms. The PHASION is generally good at finding optimal PH0 for spectrum phasing and the optimal PH0 can be easier found when PH0 is not interfered by the noise existed spectrum.

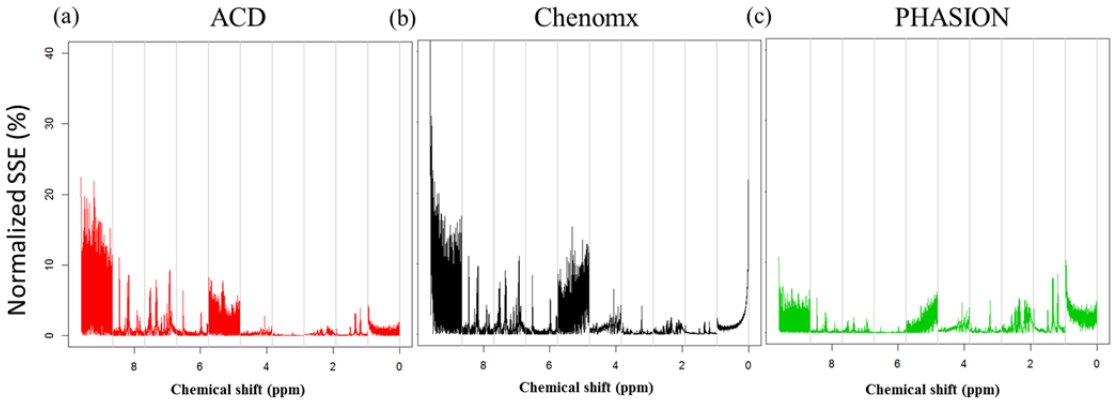


Figure 2.19: Evaluation of different autophasing algorithms on synthesized data introduced with the PH1 distorted with Gaussian noise (sd=1).

In Figure 2.20 the normalized SSE is shown with average and standard deviation of 35 synthesized data for each chemical shift segments. In Figure 2.20(a), ACD (red line) and PHASION (green line) show similar result except PHASION shows gradually increasing difference between normalized SSE of ACD and normalized SSE of PHASION at chemical shift range from 0 ppm to 3 ppm. In Figure 2.20(b), ACD (red line) and PHASION (green line) show similar result except PHASION shows gradually increasing difference between normalized SSE of ACD and normalized SSE of PHASION at chemical shift range from 4.8 ppm to 9.6 ppm. Except for Chenomx and ACD, PHASION show difference phase correction on data with different types of noise.

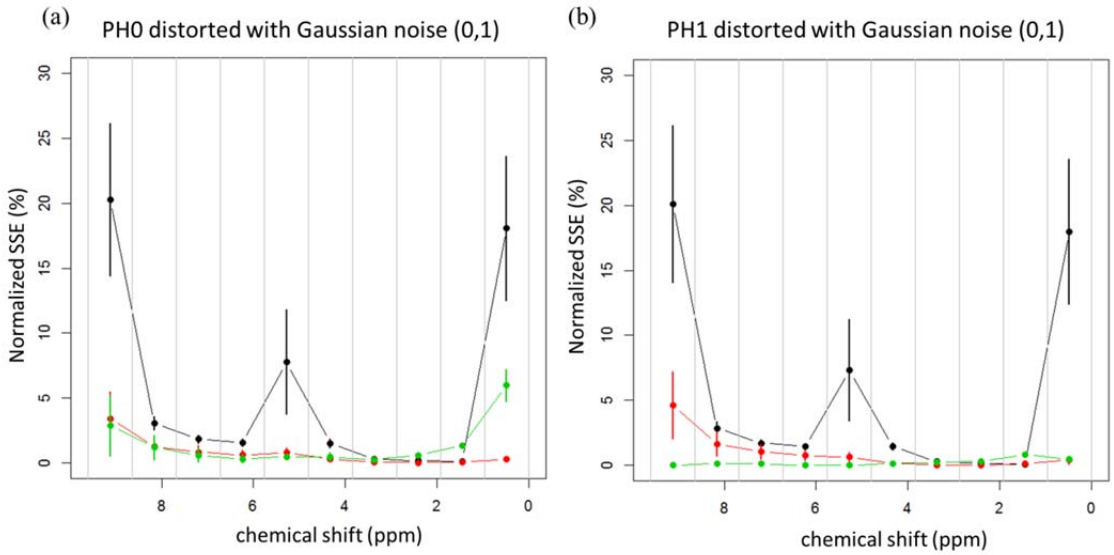


Figure 2.20: Evaluation of different autophasing algorithms on synthesized data with normalized SSE with average over 35 synthesized data on each chemical shift range. The bar represents the standard deviation of 35 synthesized data. The results of the synthesized data introduced with the PH0 distorted with Gaussian noise (sd=1) are shown in left panel. The results of the synthesized data introduced with the PH1 distorted with Gaussian noise (sd=1) are shown in right panel. The result of ACD, Chenomx and PHASION phased result is colored in red, black and green, respectively.

In Figure 2.21, the comparison of autophasing 35 synthesized data with introduced Gaussian noise in phase error. Proposed score for optimal phase angles searching used in PHASION show different result. In the result of PH0 distorted with Gaussian noise data set, the proposed score show similar result in SSE (Chenoomx > ACD >

PHASION). However, in the result of PH1 distorted with Gaussian noise data set, ACD is lower than PHASION in SSE, but ACD is higher than PHASION in proposed score. The proper explanation is the baseline would affect PHASION on finding optimal phase angles

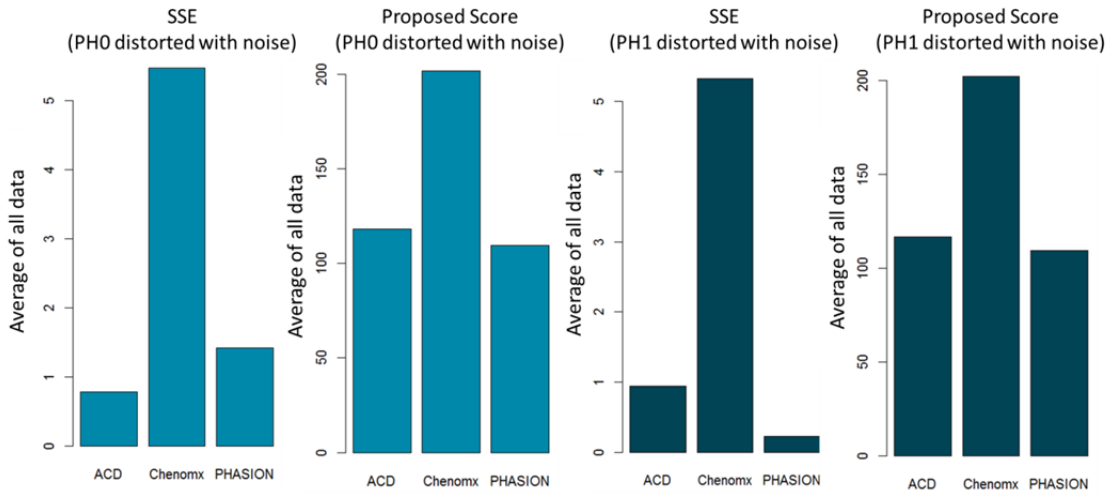


Figure 2.21: Evaluation of different autophasing algorithms on synthesized data with normalized SSE and Proposed Score. The results of the synthesized data introduced with the PH0 distorted with Gaussian noise (sd=1) are shown in left panel. The results of the synthesized data introduced with the PH1 distorted with Gaussian noise (sd=1) are shown in right panel.

Figure 2.22 shows how PHASION find the optimal phase angles for the spectrum with baseline not centered at zero. The original spectrum with baseline lifted up from zero (green line). PHASION was proposed on assume the baseline of unphased spectrum should centered at zero and the phased spectrum using optimal phase angles should with baseline centered at zero. So in this case, multiple distorted segments of spectrum, baseline of spectrum goes down and suddenly goes different direction, vice versa, can be seen in PASION phased spectrum.

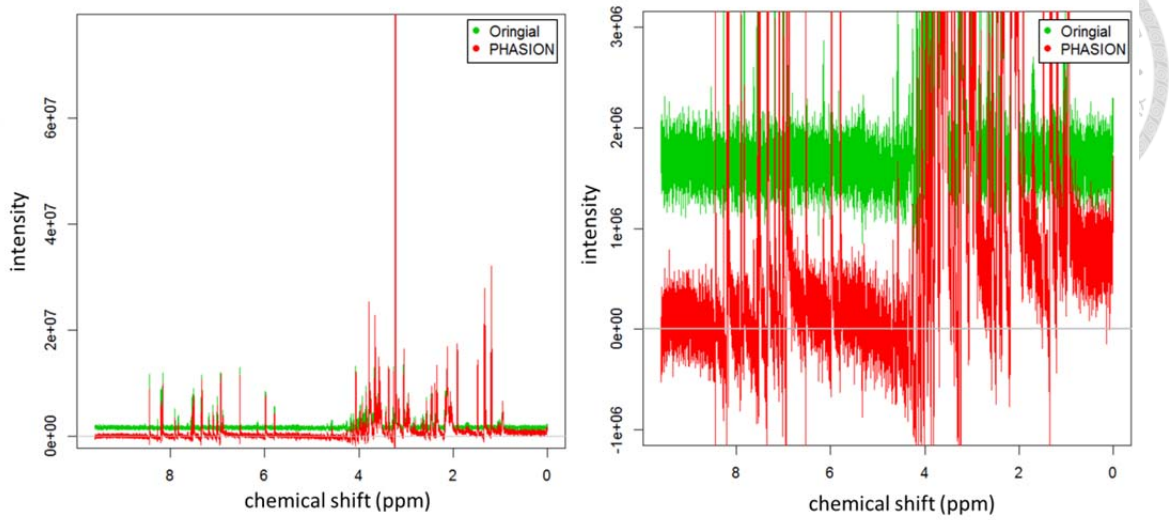


Figure 2.22: PHASION phasing spectrum with baseline not equals to zero

Figure 2.23 shows how PHASION find the optimal phase angles for the spectrum with baseline locate around zero. The original spectrum in moved to zero by subtracting a value for all data points in the spectrum. The PHASION phase spectrum now shows no distorted segments in spectrum which can be seen in Figure 2.24.

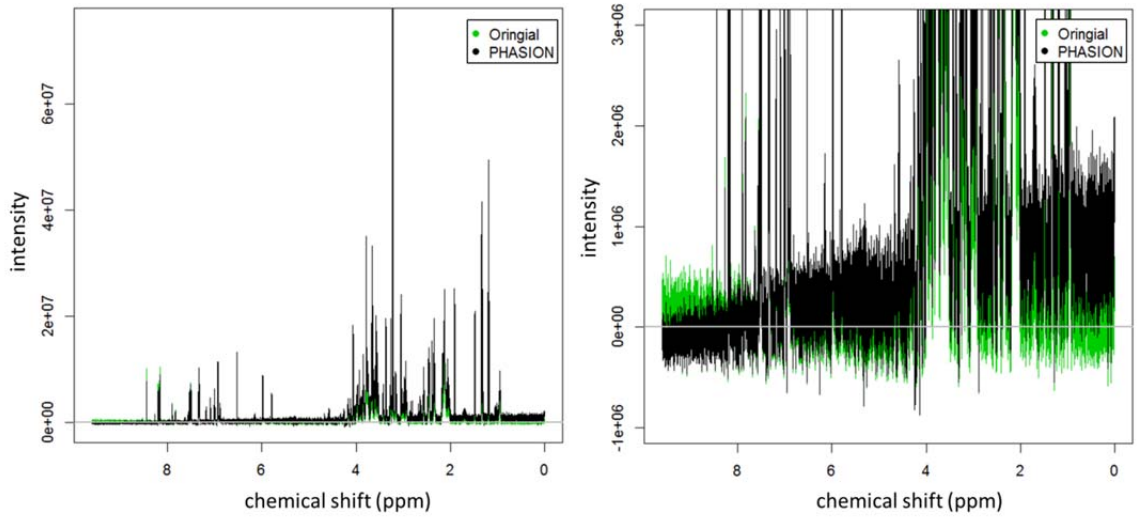


Figure 2.23: PHASION phasing with spectrum baseline equals to zero

2.4.4 *Comparison of Performance on Complex Metabolomic Plasma Samples*

In Figure 2.24 a-c, the *ACD* phased spectrum is similar to *PHASION* phased spectrum. The only difference between them is lifted baseline. The *Chenomx* phased spectrum return worst result with baseline below zero and strongly distorted baseline at water signal.

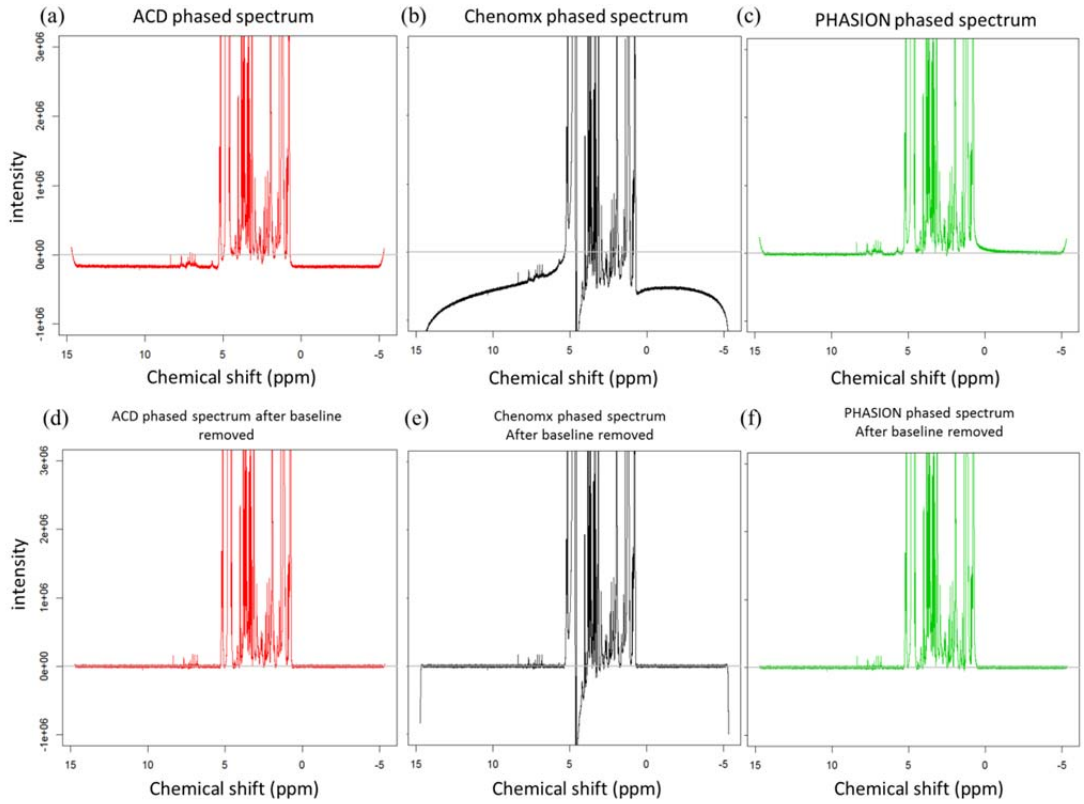


Figure 2.24: Comparison of different phasing methods on real sample. (a), (b) and (c) are the phased spectrum. (d), (e) and (f) are the baseline removed phased spectrum.

We try to see which one is with better baseline removed result, so we apply baseline removal[83] to these phased spectra. In the baseline removed spectra (Figure 2.24 d-f), the *ACD* phased spectrum and *PHASION* phased spectrum become comparable while the *Chenomx* phased spectrum is still with strong distortion at water signal, even though the water signal in general is ignored but the segment next to the water area is also strongly distorted.

In Figure 2.25, the comparison of autophasing on metabolomics complex plasma sample data, again, we can see our proposed algorithm shows better result in proposed scoring function. In general, the baseline of unphased spectrum is near zero, so the low proposed score with high normalized SSE scenario won't be see here, so we can treat proposed score as approximated normalized SSE and says PHASION can find optimal phase angles and return least distortions when comparing with ACD and Chenomx autophasing algorithm.

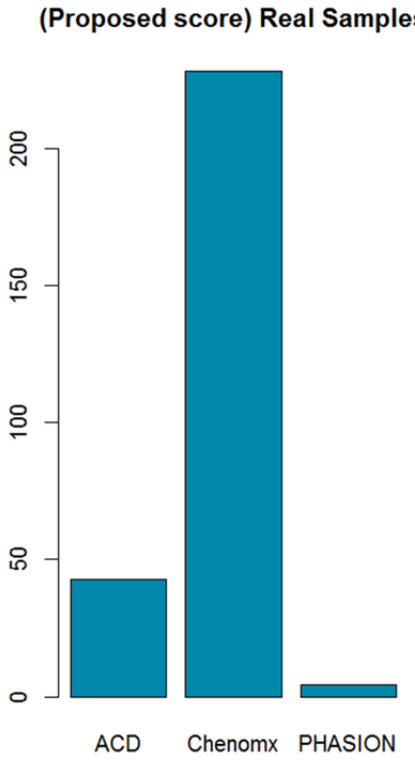


Figure 2.25: Evaluation of different autophasing algorithms on synthesized data with proposed score.

Comparing the results of phasing on synthesized data and complex samples, in most complex samples, the baseline is flat and closed to zero before spectrum phasing. Therefore, the noise in PH0 is less likely to be seen when phasing spectrum of complex sample.

2.5 Conclusion

The proposed autophasing algorithm, PHASION, is independent from any

commercial software and gives user more flexibility to processing data. It is hoped that the study will stimulate further study in metabolomics analyzing by NMR. PHASION achieves better phasing spectrum than the autophasing algorithms implemented in commonly used NMR data processing software.

PHASION can achieve by optimization using spectrum with pivot point set at largest signal in the spectrum, smile artifacts eliminated and the spectrum exclude largest signal.

PHASION can autophasing multiple files without any manual operation required. This is a major advantage over other two softwares. In ACD, users need to write macro script for batch processing and still need to select data import setting for each file before macro processing. In Chenomx, batch processing only available in commercial version. However PHASION is sensitive to spectrum with baseline not equals to zero, so move baseline to zero is needed in preprocessing stage. It is hoped that the study will stimulate further study in this field.

Table of Abbreviations

LC-MS	Liquid Chromatography-Mass Spectrometry
m/z	Mass-to-charge ratio
RT	Retention time
QC	Quality Control Sample
KDE	Kernel Density Estimation
^{1}H-NMR	Proton Nuclear Magnetic Resonance Spectrometry
FID	Free Induction Decay
PH0	Zero-order phase
PH1	First-order phase
SSE	Sum of squared error

Appendix

Table A-1 Information of 50 forensic drugs

No	Name	Formula	Cut off values (ng/mL)
1	morphine	C17H19NO3	200
2	norephedrine	C9H13NO	300
3	aminorex	C9H10N2O	300
4	pseudoephedrine	C10H15NO	300
5	nalorphine	C19H21NO3	300
6	methylephedrine	C11H17NO	300
7	dihydrocodeine	C18H23NO3	300
8	codeine	C18H21NO3	300
9	amphetamine	C9H13N	300
10	methamphetamine	C10H15N	300
11	MDA	C10H13NO2	500
12	MDMA	C11H15NO2	500
13	PMA	C10H15NO	300
14	PMMA	C11H17NO	300
15	MDEA	C12H17NO2	500
16	phentermine	C10H15N	300
17	norketamine	C12H14ClNO	100
18	ketamine	C13H16ClNO	100
19	tramadol	C16H25NO2	300
20	heroine	C21H23NO5	300
21	cocaine	C17H21NO4	300
22	methylphenidate	C14H19NO2	300
23	meperidine	C15H21NO2	200
24	2C-B	C10H14BrNO2	300
25	zolpidem	C19H21N3O	300
26	7-aminoflunitrazepam	C16H14FN3O	300
27	LSD	C20H25N3O	300
28	butorphanol	C21H29NO2	300
29	pentazocine	C19H27NO	200
30	PCP	C17H25N	300
31	meprobamate	C9H18N2O4	300
32	fentanyl	C22H28N2O	200
33	flurazepam	C21H23ClFN3O	300
34	midazolam	C18H13ClFN3	300
35	bromazepam	C14H10BrN3O	300
36	chlordiazepoxide	C16H14ClN3O	300
37	nitrazepam	C15H11N3O3	300
38	clonazepam	C15H10ClN3O3	300
39	methadone	C21H27NO	200
40	flunitrazepam	C16H12FN3O3	300
41	estazolam	C16H11ClN4	300
42	clobazem	C16H13ClN2O2	300
43	oxazepam	C15H11ClN2O2	300
44	alprazolam	C17H13ClN4	300
45	lorazepam	C15H10Cl2N2O2	300
46	temazepam	C16H13ClN2O2	300
47	lormetazepam	C16H12Cl2N2O2	300
48	nordiazepam	C15H11ClN2O	300
49	diazepam	C16H13ClN2O	300
50	prazepam	C19H17ClN2O	300

Reference

1. Andersen, J. S.; Lam, Y. W.; Leung, A. K. L.; Ong, S. E.; Lyon, C. E.; Lamond, A. I.; Mann, M. *Nature* **2005,** *433* (7021), 77-83.

2. Nesatyy, V. J.; Suter, M. J. F. *Environ. Sci. Technol.* **2007,** *41* (20), 6891-6900.

3. Oksman-Caldentey, K. M.; Inzé, D. *Trends in Plant Science* **2004,** *9* (9), 433-440.

4. Wilson, I. D.; Nicholson, J. K.; Castro-Perez, J.; Granger, J. H.; Johnson, K. A.; Smith, B. W.; Plumb, R. S. *J. Proteome Res.* **2005,** *4* (2), 591-598.

5. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat Biotech* **1999,** *17* (10), 994-999.

6. Ham, A. J. L.; Engelward, B. P.; Koc, H.; Sangaiah, R.; Meira, L. B.; Samson, L. D.; Swenberg, J. A. *DNA Repair* **2004,** *3* (3), 257-265.

7. Xiao, J. F.; Varghese, R. S.; Zhou, B.; Nezami Ranjbar, M. R.; Zhao, Y.; Tsai, T. H.; Di Poto, C.; Wang, J.; Goerlitz, D.; Luo, Y. *J. Proteome Res.* **2012,** *11* (12), 5914-5923.

8. Sasada, S.; Miyata, Y.; Tsutani, Y.; Tsuyama, N.; Masujima, T.; Hihara, J.; Okada, M. *Oncol. Rep.* **2013,** *29* (3), 925-931.

9. Armitage, E. G.; Barbas, C. *J. Pharm. Biomed. Anal.* **in press,** (0).

10. Zhang, T.; Wu, X.; Ke, C.; Yin, M.; Li, Z.; Fan, L.; Zhang, W.; Zhang, H.; Zhao, F.; Zhou, X. *J. Proteome Res.* **2012,** *12* (1), 505-512.

11. Weckwerth, W. *Bioanalysis* **2010,** *2* (4), 829-836.

12. Liberman, L. M.; Sozzani, R.; Benfey, P. N. *Curr. Opin. Plant Biol.* **2012,** *15* (2), 162-167.

13. Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M. *Bioinformatics* **2006,** *22* (15), 1902-1909.

14. Allwood, J. W.; De Vos, R. C. H.; Moing, A.; Deborde, C.; Erban, A.; Kopka, J.; Goodacre, R.; Hall, R. D. Chapter sixteen - Plant Metabolomics and Its Potential for Systems Biology Research: Background Concepts, Technology, and Methodology. In *Methods Enzymol.*, Daniel Jameson, M. V.; Hans, V. W., Eds. Academic Press: 2011; Vol. Volume 500, pp 299-336.

15. Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat Rev Drug Discov* **2002,** *1* (2), 153-161.

16. Werf, M. v.; Jellema, R.; Hankemeier, T. *J. Ind. Microbiol. Biotechnol.* **2005,** *32* (6), 234-252.

17. Watkins, S.; German, J. *Curr. Opin. Mol. Ther.* **2002,** *4* (3), 224.

18. Griffin, J. L. *Curr. Opin. Chem. Biol.* **2006,** *10* (4), 309-315.

19. Quinones, M. P.; Kaddurah-Daouk, R. *Neurobiol. Dis.* **2009,** *35* (2), 165-176.

20. Krumsiek, J.; Suhre, K.; Evans, A. M.; Mitchell, M. W.; Mohney, R. P.; Milburn, M. V.; Wägele, B.; Römisch-Margl, W.; Illig, T.; Adamski, J. *PLoS genetics* **2012,** *8* (10), e1003005.

21. Saghatelian, A.; Trauger, S. A.; Want, E. J.; Hawkins, E. G.; Siuzdak, G.; Cravatt, B. F. *Biochemistry* **2004,** *43* (45), 14332-14339.

22. Tang, Z.; Martin, M. V.; Guengerich, F. P. *Anal. Chem.* **2009,** *81* (8), 3071-3078.

23. Cho, J. Y.; Kang, D. W.; Ma, X.; Ahn, S. H.; Krausz, K. W.; Luecke, H.; Idle, J. R.; Gonzalez, F. J. *J. Lipid Res.* **2009,** *50* (5), 924-937.

24. Vinayavekhin, N.; Saghatelian, A. *ACS Chem. Biol.* **2009,** *4* (8), 617-623.

25. Mosier, A. C.; Justice, N. B.; Bowen, B. P.; Baran, R.; Thomas, B. C.; Northen, T. R.; Banfield, J. F. *mBio* **2013,** *4* (2).

26. Sreekumar, A.; Poisson, L. M.; Rajendiran, T. M.; Khan, A. P.; Cao, Q.; Yu, J.; Laxman, B.; Mehra, R.; Lonigro, R. J.; Li, Y.; Nyati, M. K.; Ahsan, A.;

Kalyana-Sundaram, S.; Han, B.; Cao, X.; Byun, J.; Omenn, G. S.; Ghosh, D.; Pennathur, S.; Alexander, D. C.; Berger, A.; Shuster, J. R.; Wei, J. T.; Varambally, S.; Beecher, C.; Chinnaiyan, A. M. *Nature* **2009,** *457* (7231), 910-914.

27. Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinformatics* **2008,** *9* (1), 504.

28. Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. *Mol. Cell. Proteomics* **2005,** *4* (12), 2010-2021.

29. Dolan, J. W. Retention Time Changes. http://www.chromatographyonline.com/lcgc/data/articlestandard//lcgceurope/042005/143655/article.pdf (accessed Oct 1).

30. Prince, J. T.; Marcotte, E. M. *Anal. Chem.* **2006,** *78* (17), 6140-6152.

31. Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006,** *78* (3), 779-787.

32. Jeong, J.; Zhang, X.; Shi, X.; Kim, S.; Shen, C. *BMC Bioinformatics* **2013,** *14* (1), 1-11.

33. Podwojski, K.; Fritsch, A.; Chamrad, D. C.; Paul, W.; Sitek, B.; Stühler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W. *Bioinformatics* **2009,** *25* (6), 758-764.

34. Krokhin, O.; Ying, S.; Cortens, J.; Ghosh, D.; Spicer, V.; Ens, W.; Standing, K.; Beavis, R.; Wilkins, J. *Anal. Chem.* **2006,** *78* (17), 6265-6269.

35. Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E. *Journal of chromatography. A* **2002,** *961* (2), 237-244.

36. Tomasi, G.; van den Berg, F.; Andersson, C. *Journal of Chemometrics* **2004,** *18* (5), 231-241.

37. Nordström, A.; O'Maille, G.; Qin, C.; Siuzdak, G. *Anal. Chem.* **2006,** *78* (10), 3289-3295.

38. Katajamaa, M.; Orešič, M. *J. Chromatogr. A* **2007,** *1158* (1–2), 318-328.

39. Fischler, M.; Bolles, R. *Commun. ACM* **1981,** *24* (6), 381-395.

40. Windig, W.; Phalp, J. M.; Payne, A. W. *Anal. Chem.* **1996,** *68* (20), 3602-3606.

41. Idborg-Björkman, H.; Edlund, P. O.; Kvalheim, O. M.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2003,** *75* (18), 4784-4792.

42. Shen, H.; Grung, B.; Kvalheim, O. M.; Eide, I. *Anal. Chim. Acta* **2001,** *446* (1–2), 311-326.

43. Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. *Nucleic Acids Res.* **2009,** *37* (suppl 2), W652-W660.

44. Andreev, V. P.; Rejtar, T.; Chen, H. S.; Moscovets, E. V.; Ivanov, A. R.; Karger, B. L. *Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics, Montreal* **2003**.

45. Pluskal, T.; Castillo, S.; Briones, A.; Oresic, M. *BMC Bioinformatics* **2010,** *11* (1), 395.

46. Yu, T.; Park, Y.; Johnson, J.; Jones, D. *Bioinformatics (Oxford, England)* **2009,** *25* (15), 1930-1936.

47. Reinert K Fau - Kohlbacher, O.; Kohlbacher, O. (1940-6029 (Electronic)).

48. Lommen, A.; Kools, H. *Metabolomics* **2012,** *8* (4), 719-726.

49. Li, X. J.; Yi, E. C.; Kemp, C. J.; Zhang, H.; Aebersold, R. *Mol. Cell. Proteomics* **2005,** *4* (9), 1328-1340.

50. Zhang, X.; Asara, J. M.; Adamec, J.; Ouzzani, M.; Elmagarmid, A. K. *Bioinformatics* **2005,** *21* (21), 4054-4059.

51. Mueller, L.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M. Y.; Vitek, O.; Aebersold, R.; Müller, M. *Proteomics* **2007,** *7* (19), 3470-3480.

52. Wang, J.; Lam, H. *Bioinformatics* **2013,** *29* (19), 2469-2476.

53. Nielsen, N. P. V.; Carstensen, J. M.; Smedsgaard, J. *J. Chromatogr. A* **1998,** *805* (1–2), 17-35.

54. Pravdova, V.; Walczak, B.; Massart, D. L. *Anal. Chim. Acta* **2002,** *456* (1), 77-92.

55. Boccard, J.; Veuthey, J. L.; Rudaz, S. *Journal of separation science* **2010,** *33* (3), 290-304.

56. Hoffmann, N.; Keck, M.; Neuweger, H.; Wilhelm, M.; Hogy, P.; Niehaus, K.; Stoye, J. *BMC Bioinformatics* **2012,** *13* (1), 214.

57. Ahmed, S.; Zhang, M.; Peng, L. GPMS: A Genetic Programming Based Approach to Multiple Alignment of Liquid Chromatography-Mass Spectrometry Data. In *Applications of Evolutionary Computation*, Esparcia-Alcázar, A. I.; Mora, A. M., Eds. Springer Berlin Heidelberg: 2014; Vol. 8602, pp 915-927.

58. *MS Resolver*.

59. *MarkerLynx*.

60. Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Orešič, M. *Chemometrics and Intelligent Laboratory Systems* **2011,** *108* (1), 23-32.

61. Lange, E.; Tautenhahn, R.; Neumann, S.; Gropl, C. *BMC Bioinformatics* **2008,** *9* (1), 375.

62. Sandin, M.; Ali, A.; Hansson, K.; Månsson, O.; Andreasson, E.; Resjö, S.; Levander, F. *Mol. Cell. Proteomics* **2013,** *12* (5), 1407-1420.

63. Peters, S.; Velzen, E.; Janssen, H. G. *Anal Bioanal Chem* **2009,** *394* (5), 1273-1281.

64. Yu, T.; Park, Y.; Johnson, J. M.; Jones, D. P. *Bioinformatics* **2009,** *25* (15), 1930-1936.

65. Zhang, B.; Zhang, C.; Yi, X. *Pattern recognition* **2004,** *37* (1), 131-144.

66. Atkeson, C. G.; Moore, A. W.; Schaal, S. *Artif. Intell. Rev.* **1997,** *11* (1-5), 11-73.

67. Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. *Mol. Syst. Biol.* **2005,** *1*, 2005.0017-2005.0017.

68. Patti, G. J.; Tautenhahn, R.; Siuzdak, G. *Nat. Protocols* **2012,** *7* (3), 508-516.

69. R Core Team. **2014**.

70. Kazmi, S.; Ghosh, S.; Shin, D. G.; Hill, D.; Grant, D. *Metabolomics* **2006,** *2* (2), 75-83.

71. Duong, T. *Weatherburn Lecture Series for the Department of Mathematics and Statistics, University of Western Australia* **2001,** *24*.

72. Härdle, W. *{Applied Nonparametric Regression (Econometric Society Monographs)}*. Cambridge University Press: 1992.

73. Sheather, S. J. *Statistical Science* **2004,** *19* (4), 588-597.

74. Howe, F.; Barton, S.; Cudlip, S.; Stubbs, M.; Saunders, D.; Murphy, M.; Wilkins, P.; Opstad, K.; Doyle, V.; McLean, M. *Magn. Reson. Med.* **2003,** *49* (2), 223-232.

75. Tate, A. R.; Crabb, S.; Griffiths, J. R.; Howells, S. L.; Mazucco, R. A.; Rodrigues, L. M.; Watson, D. *Anticancer Res.* **1995,** *16* (3B), 1575-1579.

76. Griffiths, J. R.; McSheehy, P. M. J.; Robinson, S. P.; Troy, H.; Chung, Y. L.; Leek, R. D.; Williams, K. J.; Stratford, I. J.; Harris, A. L.; Stubbs, M. *Cancer Res.* **2002,** *62* (3), 688-695.

77. Griffiths, J. R.; Stubbs, M. *Adv. Enzyme Regul.* **2003,** *43*, 67-76.

78. Lenz, E. M.; Wilson, I. D. *J. Proteome Res.* **2007,** *6* (2), 443-458.

79. Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Nat. Protocols* **2007,** *2* (11), 2692-2703.

80. Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Expert review of molecular diagnostics* **2004,** *4* (2), 189-199.

81. Holmes, E.; Antti, H. *Analyst* **2002,** *127* (12), 1549-1557.

82. Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999,** *29* (11), 1181-1189.

83. Wang, K. C.; Wang, S. Y.; Kuo, C.H.; Tseng, Y. J. *Anal. Chem.* **2012,** *85* (2), 1231-1239.

84. Andrew Clayton, T.; Lindon, J. C.; Cloarec, O.; Antti, H.; Charuel, C.; Hanton, G.; Provost, J. P.; Le Net, J. L.; Baker, D.; Walley, R. J.; Everett, J. R.; Nicholson, J. K. *Nature* **2006,** *440* (7087), 1073-1077.

85. Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. *Anal. Chem.* **2006,** *78* (13), 4430-4442.

86. Nicholson, J. K.; Wilson, I. D. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1989,** *21* (4), 449-501.

87. Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2004,** *45* (1), 109-143.

88. Nicholson, J.; Wilson, I. D. High resolution nuclear magnetic resonance spectroscopy of biological samples as an aid to drug development. In *Progress in Drug Research/Fortschritte der Arzneimittelforschung/Progrès des recherches pharmaceutiques*, Birkhäuser Basel: 1987; pp 427-479.

89. Lindon, J. C.; Nicholson, J. K.; Wilson, I. D. *Journal of Chromatography B: Biomedical Sciences and Applications* **2000,** *748* (1), 233-258.

90. Robertson, D. G.; Reily, M. D.; Sigler, R. E.; Wells, D. F.; Paterson, D. A.; Braden, T. K. *Toxicol. Sci.* **2000,** *57* (2), 326-337.

91. Lindon, J.; Holmes, E.; Nicholson, J. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2001,** *39* (1), 1-40.

92. Wang, T.; Shao, K.; Chu, Q.; Ren, Y.; Mu, Y.; Qu, L.; He, J.; Jin, C.; Xia, B. *BMC Bioinformatics* **2009,** *10* (1), 83.

93. Ebel, A.; Dreher, W.; Leibfritz, D. *J. Magn. Reson.* **2006,** *182* (2), 330-338.

94. Smolinska, A. *Chemometrics and NMR spectroscopy for metabolomics analysis of neurological disorders*. [Sl: sn]: 2012.

95. de Brouwer, H. *J. Magn. Reson.* **2009,** *201* (2), 230-238.

96. Chen, L.; Weng, Z.; Goh, L.; Garland, M. *J. Magn. Reson.* **2002,** *158* (1–2), 164-168.

97. Craig, E. C.; Marshall, A. G. *Journal of Magnetic Resonance (1969)* **1988,** *76* (3), 458-475.

98. Neff, B. L.; Ackerman, J. L.; Waugh, J. S. *Journal of Magnetic Resonance (1969)* **1977,** *25* (2), 335-340.

99. Daubenfeld, J. M.; Boubel, J. C.; Delpuech, J. J.; Neff, B.; Escalier, J. C. *Journal of Magnetic Resonance (1969)* **1985,** *62* (2), 195-208.

100. Heuer, A. *Journal of Magnetic Resonance (1969)* **1991,** *91* (2), 241-253.

101. Brown, D. E.; Campbell, T. W.; Moore, R. N. *Journal of Magnetic Resonance (1969)* **1989,** *85* (1), 15-23.

102. Golotvin, S.; Williams, A. *J. Magn. Reson.* **2000,** *146* (1), 122-125.

103. Sterna, L. L.; Tong, V. P. *Fuel* **1991,** *70* (8), 941-945.

104. Džakula, Ž. *J. Magn. Reson.* **2000,** *146* (1), 20-32.

105. Nelder, J. A.; Mead, R. *The Computer Journal* **1965,** *7* (4), 308-313.

106. Morgan, S. L.; Deming, S. N. *Anal. Chem.* **1974,** *46* (9), 1170-1181.

107. Powell, M. J. D. *The Computer Journal* **1964,** *7* (2), 155-162.

108. Nocedal, J.; Wright, S. *Numerical Optimization*. Springer: 2000.

109. Wright, M. H. *Documenta Mathematica* **2010,** (7).

110. Lagarias, J.; Reeds, J.; Wright, M.; Wright, P. *SIAM Journal on Optimization* **1998,** *9* (1), 112-147.

111. King, R. W.; Williams, K. R. *J. Chem. Educ.* **1990,** *67* (4), A100.

112. Cobas, C. Bruker Smiles.

http://nmr-analysis.blogspot.tw/2010/05/bruker-smiles.html (accessed Nov,4).

113. Facey, G. Is My 1H NMR Spectrum Quantitative?

http://u-of-o-nmr-facility.blogspot.tw/2008/03/is-my-1-h-nmr-spectrum-quantitative.html (accessed Nov, 4).