ALIGNMENT OF LC-MS DATA USING PEPTIDE FEATURES

A Thesis

by

XINCHENG TANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2011

Major Subject: Statistics

ALIGNMENT OF LC-MS DATA USING PEPTIDE FEATURES

A Thesis

by

XINCHENG TANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Alan R. Dabney
Committee Members,	Fred P. Dahm
	Scott Dindot
Head of Department,	Simon J. Sheather

December 2011

Major Subject: Statistics

ABSTRACT

Alignment of LC-MS Data Using Peptide Features. (December 2011)

Xincheng Tang, B.E., Wuhan University, Hubei, China

Chair of Advisory Committee: Dr. Alan R. Dabney

Integrated liquid-chromatography mass-spectrometry(LC-MS) is becoming a widely used approach for quantifying the protein composition of complex samples. In the last few years, this technology has been used to compare complex biological samples across multiple conditions. One challenge in the analysis of an LC-MS experiment is the alignment of peptide features across samples.

In this paper, we proposed a new method using the peptide internal information (both LC-MS and LC-MS/MS information) to align features from multiple LC-MS experiments. We defined Anchor points which are data elements that are highly confident we have identified and are shared by both samples. We chose one sample as template data set, find Anchor Points in this sample, then apply alignment to modify another sample and find Anchors in modified sample, these Anchors should line up with one another. One advantage of our method is that it allows statistical assessment of alignment performance. Use Anchor Points to perform alignment between samples, and labeling an objective performance in LC-MS.

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to my advisor, Dr. Alan R. Dabney, for his direction, suggestion, encouragement, patience and continuous support toward my professional development. He guided me to think about statistical research problems, how to do research and how to write a paper step by step.

I am very grateful to all my committee members, Dr. Fred P. Dahm, Dr. Scott Dindot, for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally I feel lucky to have my wife Xuan Wang, my parents by my side, thank you for your support and love.

NOMENCLATURE

GC-MS	Gas Chromatography Mass Spectrometry
LC-MS	Liquid Chromatography Mass Spectrometry
MALDI	Matrix Assisted Laser Desorption Ionization
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
M/Z	Mass-To-Charge Ratio
PNNL	Pacific Northwest National Laboratory
RT	Retention Time
SELDI	Surface Enhanced Laser Desorption Ionization
TOF	Time-Of-Flight

TABLE OF CONTENTS

Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

1. INTRODUCTION

1.1 Background

Mass spectrometry (MS) proteomics has become the tool of choice for identifying and quantifying the proteome of an organism, in recent years, tremendous improvement in instrument performance and computational tools are used. Several MS methods for interrogating the proteome have been developed: Surface Enhanced Laser Desorption Ionization(SELDI)[1], Matrix Assisted Laser Desorption Ionization (MALDI)[2] coupled with time-of-flight (TOF) or other instruments, and gas chromatography MS (GC-MS) or liquid chromatography MS (LC-MS).

In LC-MS-based proteomics, complex mixtures of proteins are first subjected to enzymatic cleavage, and the resulting peptide products are analyzed by using a mass spectrometer. In tandem mass spectrometry (denoted by MS/MS), fragmentation spectra are obtained for each subset of observed high-intensity peaks and compared to fragmentation spectra in a database, using software like SEQUEST[3], Mascot[4], or X!Tandem[5].

If the data used to be analyzed and interpreted at the protein level, then once a list has been constructed of the proteins believed to be present in the sample, the next task is to quantify the abundance of the proteins. Protein abundance information is contained in the set of peaks that correspond to the protein component peptides. Peak height or area is a function of the number of ions detected for a particular peptide, and is related to peptide abundance.

In LC-MS, each sample may have thousands of scans, each containing a mass spectrum. The mass spectrum for a single MS scan can be summarized by a plot of M/Z values versus peak intensities. These data contains signals that are specific to individual peptides. As a first step towards identifying and quantifying those

This thesis follows the style of Journal of Nuclear Materials.

peptides, features need to be identified in the data and distinguished from background noise, one simple method is to employ a filter on the signal-to-noise ratio of a peak relative to its local background. Each peptide gives an envelope of peaks due to a peptide constituent amino acids. The presence of a peptide can be characterized by the M/Z value corresponding to the peak arising from the most common isotope, referred to as the monoisotopic mass.

1.2 Experimental Procedure

A LC-MS-based proteomic experiment requires several steps of sample preparation (Figure 1.1), including cell lysis to break cells apart, protein separation to spread out the collection of protein into more homogenous groups, and protein digestion to break intact proteins into more manageable peptide components. Once this is complete, peptides are further separated, then ionized and introduced into the mass spectrometer.

A mass spectrometer measures the mass-to-charge ratio (M/Z) of ionized molecules, which is designed to carry out the distinct functions of ionization and mass analysis. The key components of a mass spectrometer are the ion source, mass analyzer, and ion detector (Figure 1.2). The ion source is responsible for assigning charge to each peptide. Mass analyzer measures the mass-to-charge (M/Z) ratio of each ion. The detector captures the ions and measures the intensity of each ion species. In terms of a mass spectrum, the mass analyzer is responsible for the M/Z information on the x-axis and the detector is responsible for the peak intensity information on the y-axis.

1.3 Existing Alignment Methods

The goal of alignment is to match corresponding peptide features in the M/Z vs scan plot(see Figure 1.2) from different experiments. A time warping method

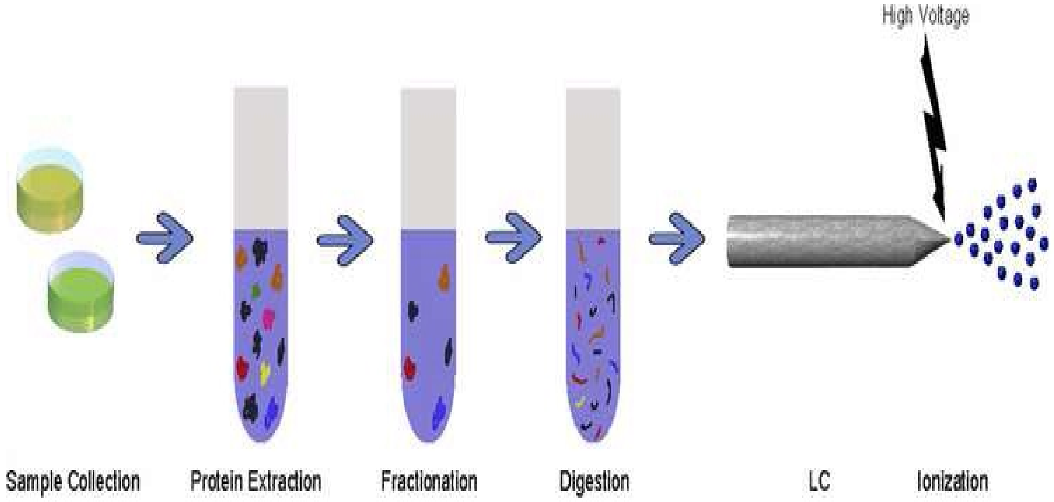

Fig. 1.1. Sample Preparation

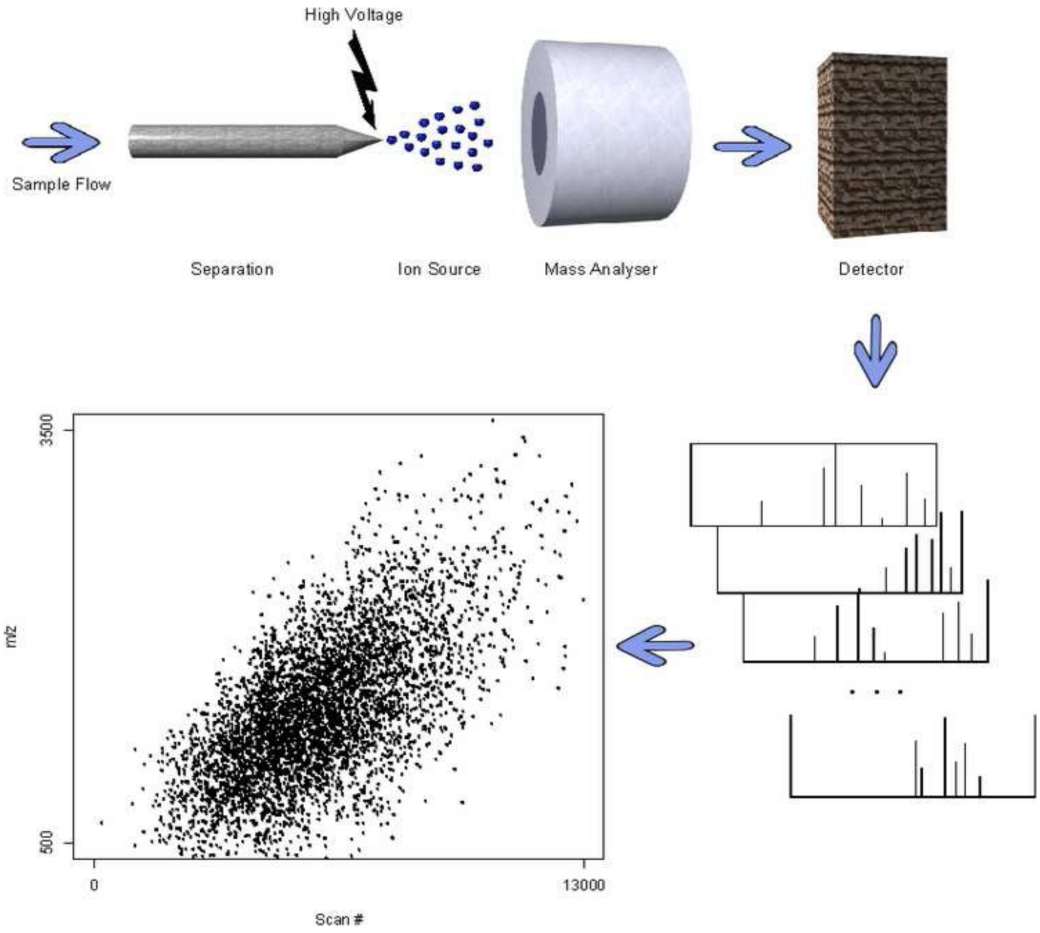

Fig. 1.2. Mass Spectrometry

based on raw spectrum for alignment of LC-MS data was introduced by Bylund and others[6],which is a modification of the original correlated optimized warping algorithm[7]. Wang and others [8], implemented a dynamic time warping algorithm allowing every RT point to be moved. However, the LC-MS data have added dimension of mass spectral information, so only mapping the retension time coordinates between two LC-MS files is not sufficient to provide alignment for individual peptides. Radulovic and others [9] performed alignment based on (M/Z, RT) values of detected features. Their method first divides the M/Z domain into several intervals and fitted different piece-wise linear time warping functions for each M/Z interval. After the time warping, they applied a "wobble" function to peak and allow peak to move (1-2% of total scan range) in order to match with the nearest adjacent peak in another file. Their method relies on the (M/Z, RT) values of detected peptide features, it fails to take advantage of other information in the raw image. Wang and others [10] proposed an alignment algorithm, PETAL, for LC-MS data. It uses both the raw spectrum data and the information of the detected peak features for peptide alignment.

In this paper, two Shewanella datasets are obtained from Pacific Northwest National Laboratory (PNNL) and they were analyzed by SEQUEST on different days. SEQUEST correlates uninterpreted tandem mass spectra of peptides with amino acid sequences from protein and nucleotide databases, which determines the amino acid sequence and thus the protein(s) and organism(s) that correspond to the mass spectrum being analyzed. Based on the SEQUEST output files, each sample has thousands of scans, and the M/Z, peak intensities, peptide information associated. It's obvious that there's some systematic error between the alignment of the two datasets.

In this study, we first applied some filter criteria to choose data points matched with high confidence in both samples, which are called "Anchor Points". We then use these "Anchor Points" in sample one as the baseline and modify the data points in sample two, to make the “Anchor Points” between the two samples aligned together, which means after alignment the “Anchor Points” in both samples show up at the same locations. The alignment algorithm is then applied on all the data points in sample two. Finally, statistical measurements of the performance of alignment are given on sample level and regional level.

In future study, we hope this alignment method can be applied to several samples of one organism, and as a guide to justify the points with same peptide information in different samples.

2. METHODS

2.1 Anchor Points

In our method, for different experiments, we have the raw data analyzed by SEQUEST. Based on the SEQUEST output files, each sample has thousands of scans, and the M/Z, peak intensities, peptide information of each scan, an example is given in Table 2.1. Define a point with high probability that its peptide shows up in SEQUEST output files as “Anchor Points”.

A sample record of data is given in Table 2.1

Table 2.1
An Example of Data Record Returned from SEQUEST.

ScanNum	MZ	PeakArea	PassFilt	PeakSignalToNoiseRatio
8239	826.4	4.98E+06	1	25.1
ChargeState	Xcorr	NumTrypticEnds	RankXc	Peptide
3	7.9171	2	1	K.LAYADGYVHA

Table 2.2
An Example of Anchor Points Located.

ScanNum	MZ	PeakArea	PassFilt	PeakSignalToNoiseRatio
8239	826.4	4.98E+06	1	25.1
ChargeState	Xcorr	NumTrypticEnds	RankXc	Peptide
3	7.9171	2	1	K.LAYADGYVHA
ScanNum	MZ	PeakArea	PassFilt	PeakSignalToNoiseRatio
7567	813.2	3.57E+06	1	19.2
ChargeState	Xcorr	NumTrypticEnds	RankXc	Peptide
3	7.5681	2	1	K.LAYADGYVHA

In LC-MS, we need to distinguish the peptide features from the background noise, the first step for doing this is MS peak detection. We employ a simple filter routine on the signal-to-noise ratio of a peak relative to its local background [11]. In our approach, in order to find peptides that exist in both samples with high confidence, three filtering criteria are applied. The first criterion is PassFilt equaling to 1, the second criterion is NumTrypticEnds equaling to 2 and the third criterion is SignalToNoiseRatio being greater than 10.

PassFilt is a score that does not come from by SEQUEST but is calculated from syn-fht summary generator using Xcorr, DelCN, RankXc and the number of tryptic termini. NumTrypticEnds (number of tryptic cleavage sites) is the number of termini that conforms to the expected cleavage behavior of trypsin (i.e. C-terminal to R and K). Note that K-P and R-P do not qualify as tryptic cleavages because of the proline rule. However, the protein N-terminus and protein C-terminus do count as tryptic cleavage sites. Values can be 0, 1, or 2 with 2 = fully tryptic; 1 = partially tryptic; 0 = Non tryptic. Any points in both sample satisfied these three criteria are called "Anchor Points". Table 2.2 is an example of the Anchor Points located after applying the three filtering criteria. In this example, we can find the points in both sample with same peptide information, which is K.LAYADGYVHA.

2.2 Alignment Algorithm

"Anchor Points" found from both samples differ on "ScanNum", which represent retention time, so we need to find some algorithms to make these points in two samples aligned on "ScanNum" as well as M/Z and PeakArea. Let M_i be M/Z for peak i, S_i be the Scan Number for peak i, and P_i be the log intensity for peak i. The data is normalized to 0-1 range by dividing normalization factors of M/Z, scan number and log(Peak Area), denoted as MN, SN, and PN, which are the max M/Z, max Scan Number and max log(Peak Area) of the two samples. The reason why we do normalization is that due to the different scales of Scan Number,M/Z and Peak intensity values, the distance can not be equally measured, for example, the scan number is very large compared to the M/Z. So transforming the data into 0-1 range will give equal weight of all these three values.

With the pool of "Anchor Points" found between sample one and sample two, we are able to locate the five nearest anchor points for peak i in the second sample, where the distance is defined by Euclidean metric considering the three dimensions of normalized M/Z, Scan Number and log(PeakArea). With the defined range, let D_ij be the distance between peak i and Anchor Point j in the second sample two, $D_{ij} = ((M_i - M_j)/MN)^2 + ((S_i - S_j)/SN)^2 + (P_i - P_j)/PN)^2$

Let Δ_1 be the difference of averaged M/Z across the five nearest Anchor Points of peak i between the two samples, Δ_2 be the difference of averaged scan number and Δ_3 be the difference of averaged intensity. Then we use the differences to modify peak i in sample two by adding $(\Delta_1, \Delta_2, \Delta_3)$ to (M_i, S_i, P_i)

$\Delta_1 = \overline{M_{ij1}} - \overline{M_{ij2}}$

$\Delta_2 = \overline{S_{ij1}} - \overline{S_{ij2}}$

$\Delta_3 = \overline{P_{ij1}} - \overline{P_{ij2}}$

where $j = 1, 2, 3, 4, 5$

This algorithm uses both the raw spectrum data which are analyzed by SEQUEST and the information of the detected peak features for peptide alignment, we use all the information,such as Scan Number, intensity and M/Z to find the target points. Although in both samples,the points with same peptide information appear in different place due to the systematic bias, we assume that, for such point in sample one, the point with same peptide information in sample two should be not far away from the point in sample one.

3. REAL DATA EXAMPLE

In the Shewanella datasets, the normalizing factors of M/Z, Scan Number and Peak Area are as follows: $MN = 1519.48, SN = 10606, PN = 10.08476$

For comparison, both "Anchor Points" before and after alignment are plotted in Figure 3.1.

The distance of Anchors in both samples before alignment have systematic difference, and after alignment, the difference should be randomly distributed around the standard points (Anchor Points in sample one). We draw histograms to compare the distance of Anchors between both samples before and after alignment in Figure 3.2, Figure 3.3, and Figure 3.4, the Histogram shows that, after alignment, the distance differences between two samples are mostly around 0.

It is important to perform test to justify that after alignment, there is no systematic difference on the anchor points between two samples. As justification, perform Wilcoxon signed-rank test on the differences between two samples on the "Anchor points" before and after alignment. The null hypothesis is the true location shift is equal to 0 and the alternative hypothesis is the true location shift is not equal to 0.

We have the following results: Before alignment Wilcoxon signed-rank test on Scan Number of anchor point with continuity correction W = 537733, p-value =2.2e-16. After alignment Wilcoxon signed-rank test on Scan Number of anchor point with continuity correction W = 293222, p-value =0.0986.

Before alignment Wilcoxon signed-rank test on M/Z of anchor points with continuity correction W = 142212, p-value =0.1091. After alignment Wilcoxon signed-rank test on M/Z of anchor point with continuity correction W = 265884, p-value =0.9049.

Before alignment Wilcoxon signed-rank test on log(Peak Area) of anchor point with continuity correction W = 336362, p-value=1.655e-09. After alignment Wilcoxon

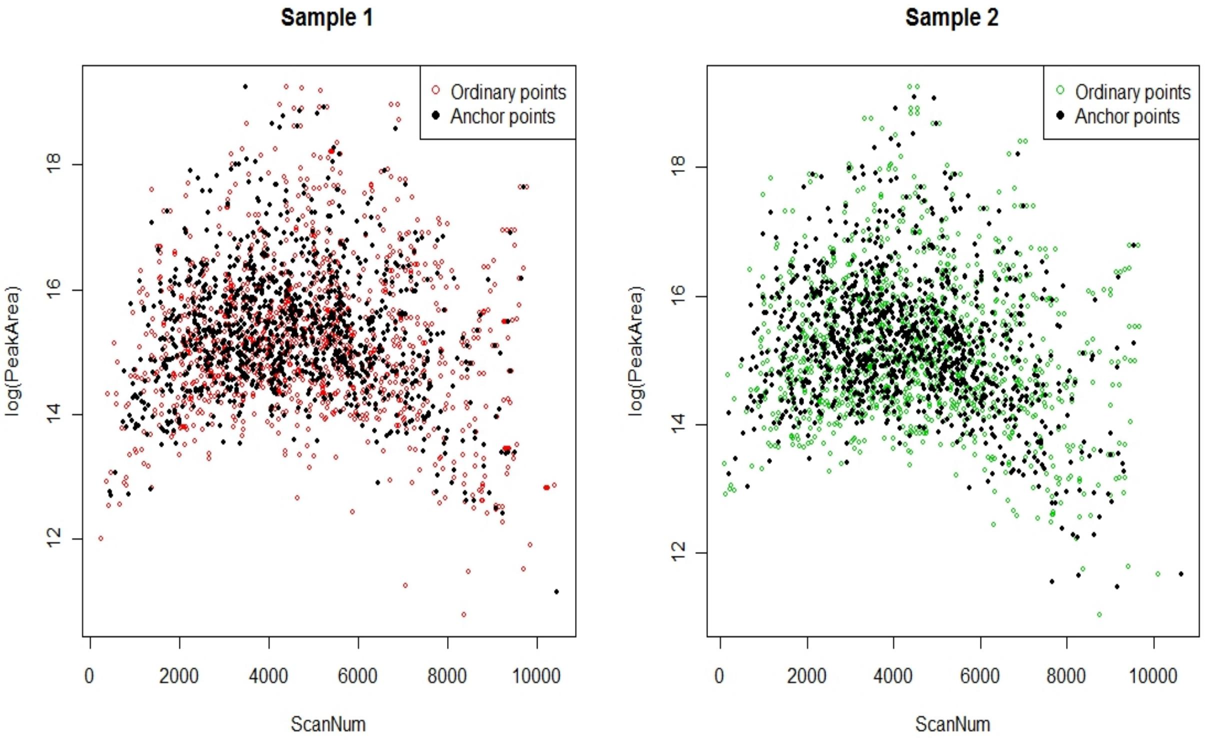

Fig. 3.1. Plot of Anchor Points Embedded in Both Samples. Left Panel is for Sample 1 and Right Panel for Sample 2.

signed-rank test on log(Peak Area) of anchor point with continuity correction W = 281910, p-value =0.6141.

So we conclude that the difference between ”Anchor points” after alignment in two samples has a common median 0, which indicates the method can be used as one alignment method to find or justify ”Anchor Points” in sample two.

All the data points in the two sampels before and after alignment are displayed in Figure 3.5, and Figure 3.6. It’s clear that the two samples mixed better in the scatter plots after alignment.

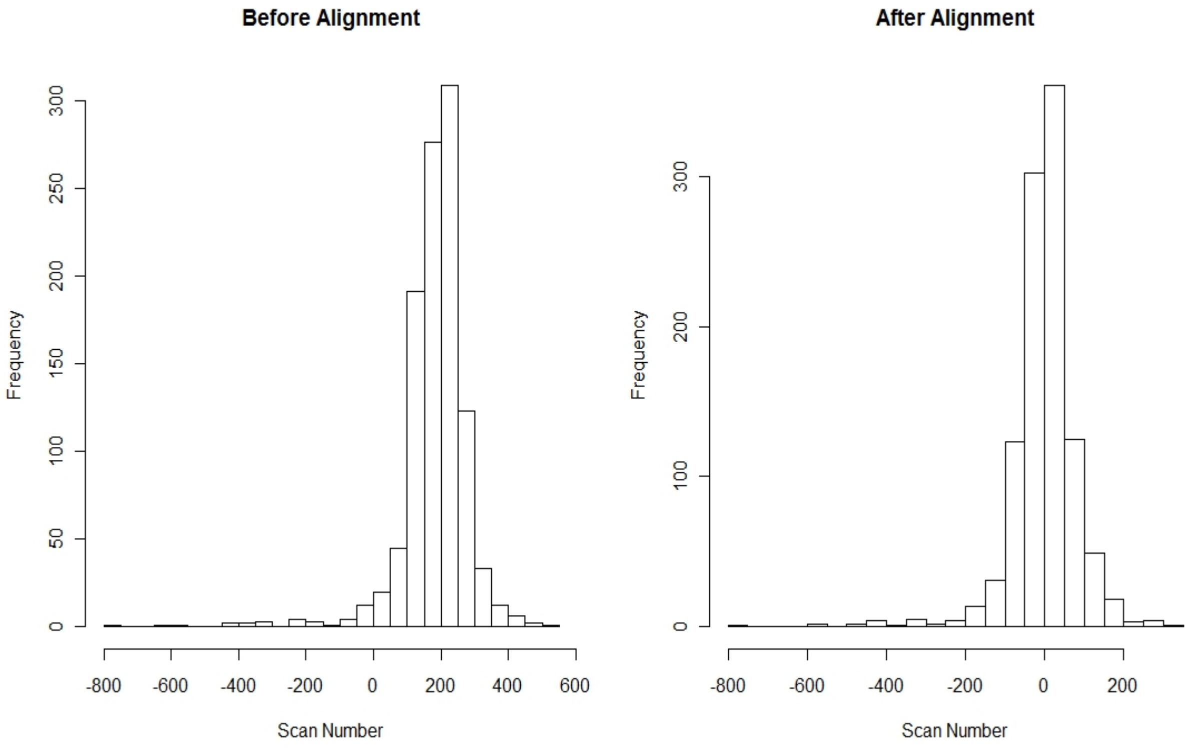

Fig. 3.2. Histograms of Scan Number of Anchor Points. Left Panel is for before Alignment and Tight Panel for after Alignment.

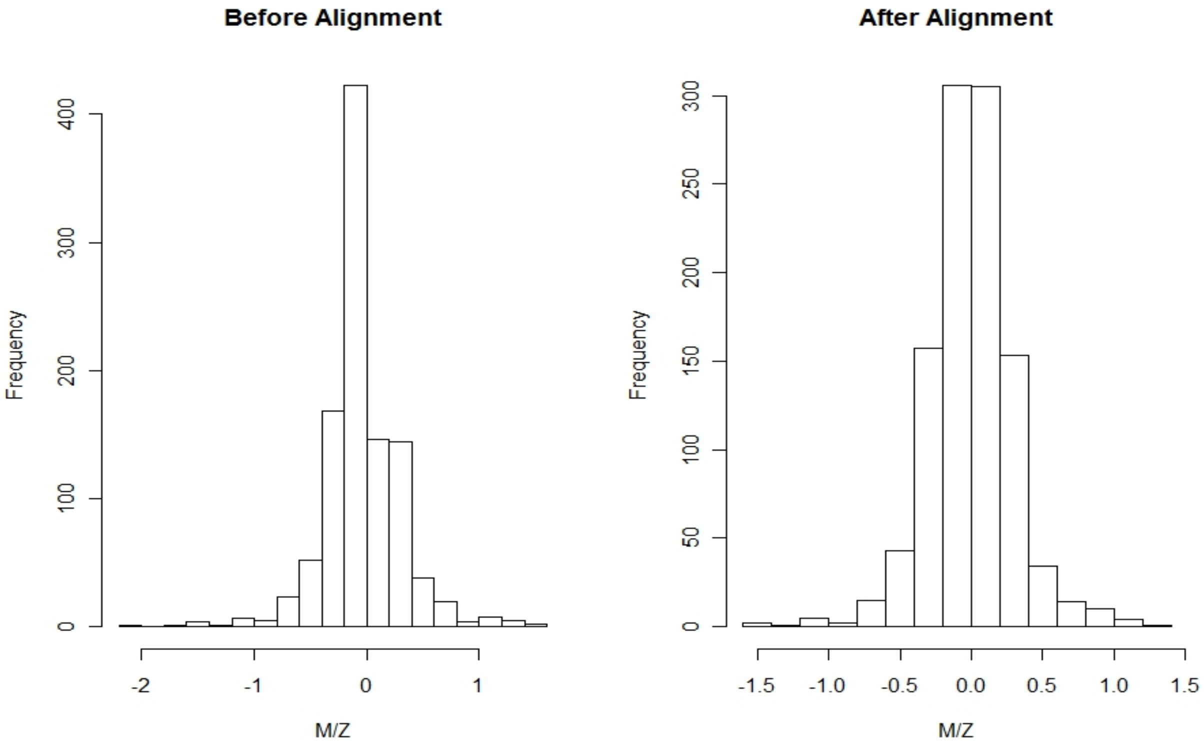

Fig. 3.3. Histograms of M/Z of Anchor Points. Left Panel is for before Alignment and Right Panel for after Alignment.

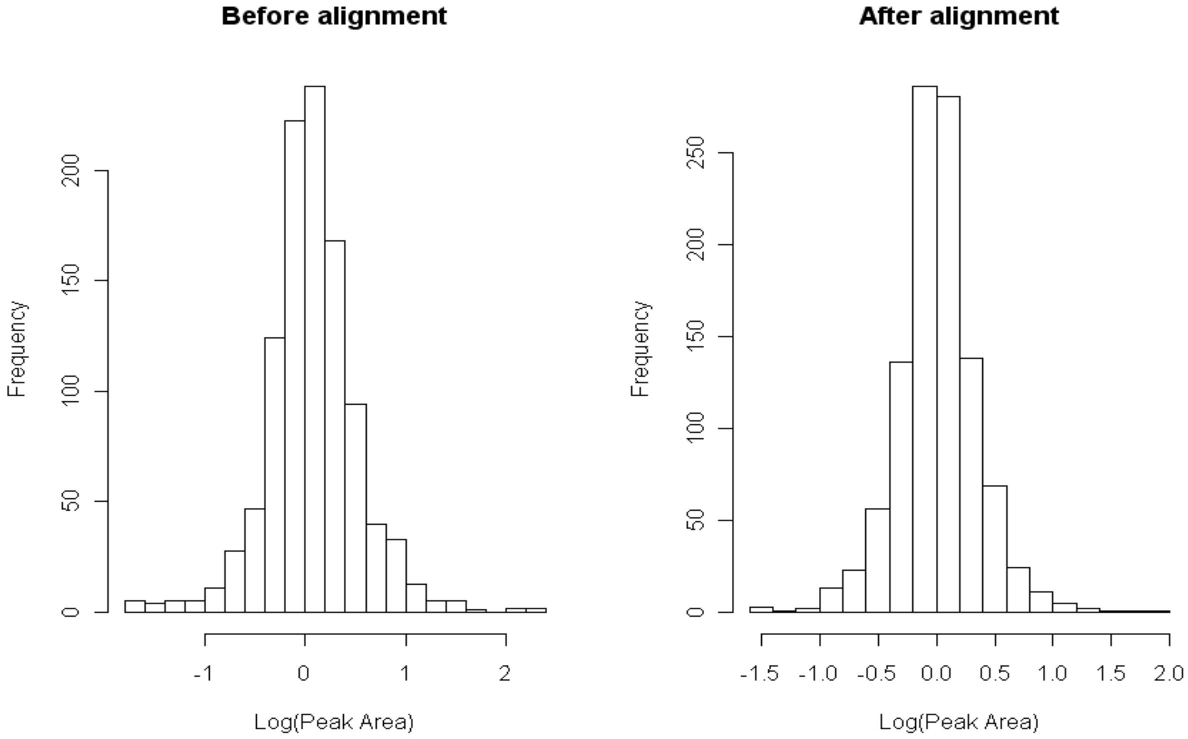

Fig. 3.4. Histograms of Log(Peak Area) of Anchor Points. Left Panel is for before Alignment and Right Panel for after Alignment.

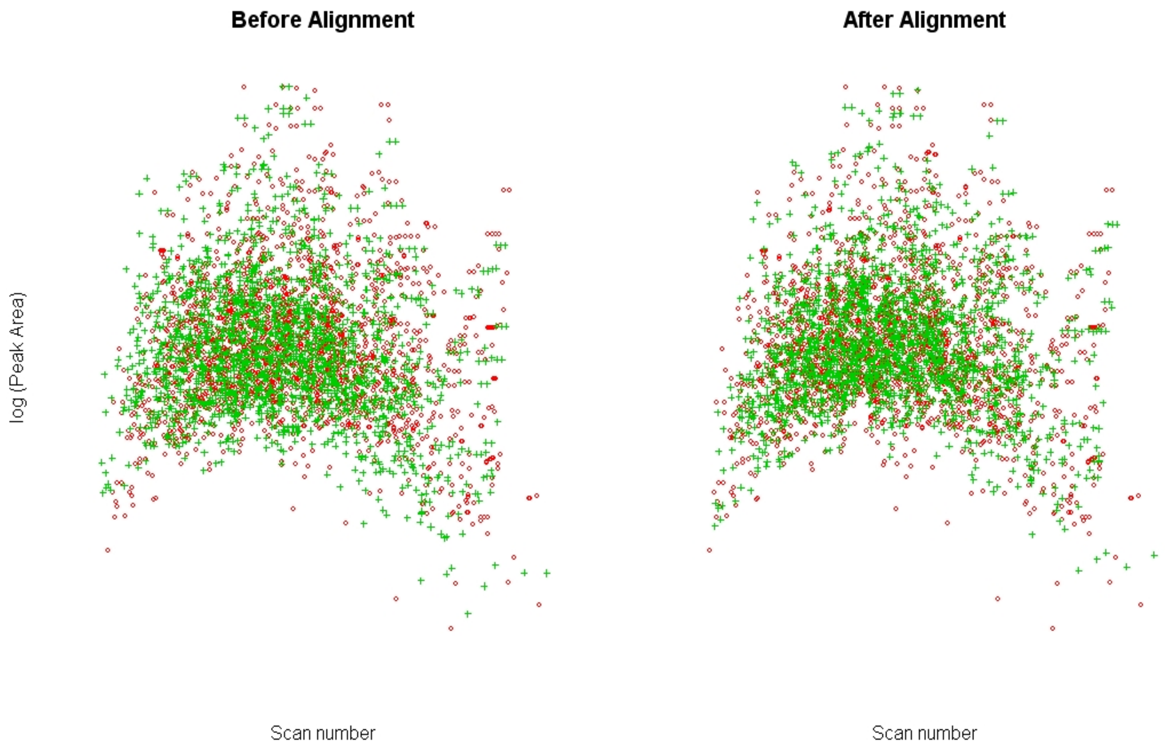

Fig. 3.5. Scatter Plot of Scan Number vs Log (Peak Area) on All Data Points. Left Panel is for before Alignment and Right Panel for after Alignment.

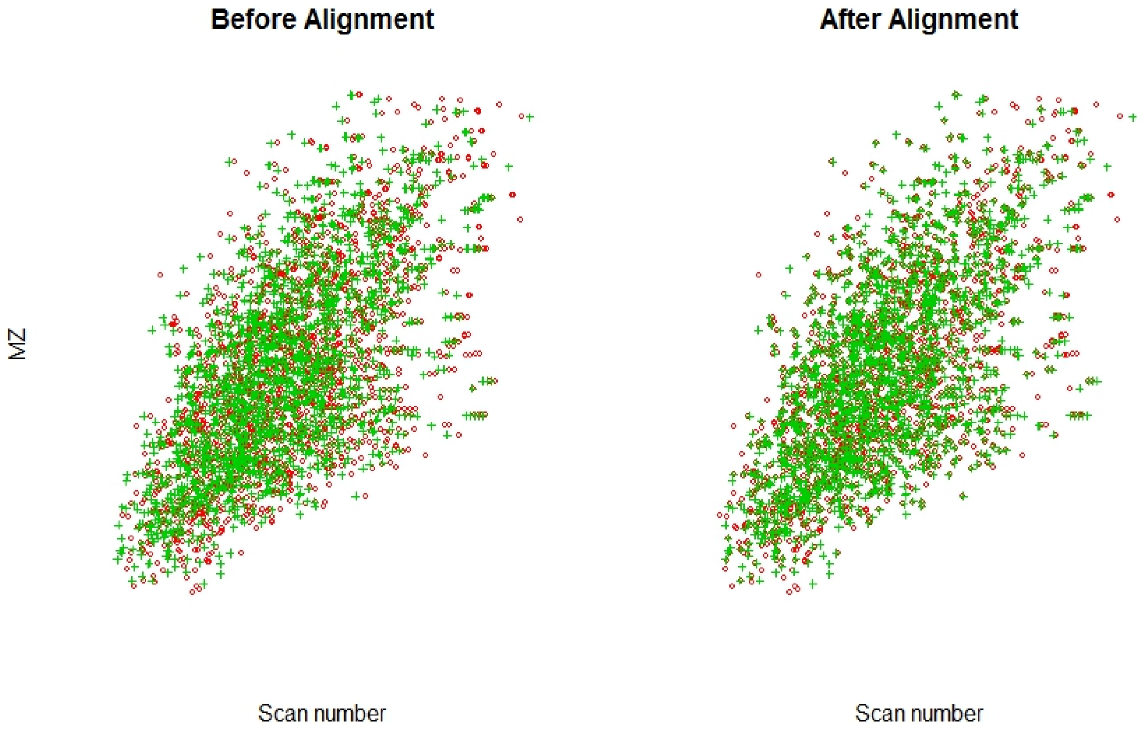

Fig. 3.6. Scatter Plot of Scan Number vs. M/Z on All Data Points. Left Panel is for before Alignment and Right Panel for after Alignment.

4. CONCLUSION

One advantage of our method is that it allows statistical assessment of alignment performance. We could statistically evaluate the performance of our methodology with other alignment algorithms on some dataset that has peptides identified with high confidence.

The statistical confidence measure of our method was given on sample level. We expanded it to region level and the future work would be developing peptide level statistical confidence measure and pass it to downstream quantitative analysis.

REFERENCES

[1] N. Tang, P. Tornatore, and S. R. Weinberger. Current Developments in SELDI Affinity Technology. Mass Spectrometry Reviews, 23(1) (2004) 33-34.

[2] M. Karas, D. Bachman, U. Bahr, and F. Hillenkamp. Matrix-Assisted Ultraviolet Laser Desorption of Non-Volatile Compounds. Int J Mass Spectrometry Ion Proc, 78 (1987) 53-68.

[3] J. K. Eng, A. L. Mccormack, and J. R. III. Yates. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in A Protein Database. J Am Soc Mass Spectrom, 5 (1987) 976-989.

[4] D. N. Perkins, D. J. C. Pappin, and D. M. Creasy. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. Electrophoresis, 20 (1987) 3551-3567.

[5] R. Craig and R. C. Beavis. TANDEM: Matching Proteins with Tandem Mass Spectra. Bioinformatics, 20(9) (2004) 1466-1467.

[6] D. Bylund, R. Danielsson, G. Malmquist, and K. E. Markides. Chromatographic Alignment by Warping and Dynamic Programming as A Pre-Processing Tool for Parafac Modelling of Liquid Chromatography-Mass Spectrometry Data. Journal of Chromatography, A 961 (2002) 237-244.

[7] N. P. Nielsen, J. M. Carstensen, and J. Smedsgaard. Aligning of Single and Multiple Wave Length Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping. Bioinformatics, A 805 (1987) 17-35.

[8] W. Wang, H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, M. Anderle, and C. Becker. Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards. Analytical Chemistry, 75 (2003) 4818-4826.

[9] D. Radulovic, S. Jelveh, S. Ryu, T. G. Hamilton, E. Foss, Y. Mao, and A. Emili. Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using

Liquid-Chromatography-Tandem Mass Spectrometry. Molecular & Cellular Proteomics, 3 (2004) 984-997.

[10] P. Wang, H. Tang, M. P. Fitzgibbon, M. Mcintosh, M. Coram, H. Zhang, E. Yi, and R. Aebersold. A Statistical Method for Chromatographic Alignment of LC-MS Data. Biostatistics, 82 (2007) 357-367.

[11] N. Jaitly, A. Mayampurath, K. Little_eld, J. N. Adkins, G. A. Anderson, and R. D. Smith. Decon2LS: An Open-Source Software Package for Automated Processing and Visualization of High Resolution Mass Spectrometry Data. Bioinformatics, 10(87) (2009) 1471-2105.

VITA

Xincheng Tang majored in mechanical engineering at Wuhan University(P.R.China), where he obtained his Bachelor of Engineering in 2006. He received his M.S. in statistics from Texas A&M University in December 2011. His research interests include Quantitative Proteomics, Clinical Trial Design.

He may be reached at:
Department of Statistics
Texas A&M University
3143 TAMU
College Station, TX 77843-3143.

and his email address is xtang@stat.tamu.edu.