ARTIFICIAL BEE COLONY WITH DIFFERENTIAL EVOLUTION ALGORITHM FOR FEATURE EXTRACTION AND SELECTION OF MASS SPECTROMETRY DATA

SYARIFAH ADILAH MOHAMED YUSOFF

UNIVERSITI SAINS MALAYSIA

2016

ARTIFICIAL BEE COLONY WITH DIFFERENTIAL EVOLUTION ALGORITHM FOR FEATURE EXTRACTION AND SELECTION OF MASS SPECTROMETRY DATA

by

SYARIFAH ADILAH MOHAMED YUSOFF

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

May 2016

ACKNOWLEDGEMENTS

Bismillahirahmanirahim walhamdulillah.....

In the name of Allah, the most Gracious and the most Merciful. All praise is due to Allah, the Cherisher and Sustainer of the worlds. Alhamdulillah, praise be to Allah for His glory and wills that gave me the strength and patience to complete this thesis in every possible way. First and foremost, I would like to express my sincere gratitude to my supervisor, Prof Dr Rosni Abdullah for her guidance and moral support. Much of this work would have been impossible without her charismatic supervision. Thank you to Dr. Ibrahim Venkat as the co-supervisor for his constructive comments throughout the study. My deepest love goes to my late mother, Wan Emas (1953-2009) that inspired my passion to continue to pursuing this journey. My father, Mohamed Yusoff for his never ending prayers and motivation. Not forgetting my heartfelt gratitude to my dearest husband, Dr. Hanapiah Abdullah for his concern, patience and understanding, and all my beloved kids, Faqih, Fahim, Fathiyyah and Fasiha Amni for everything special in my life. My deepest gratitude also goes to my siblings; Dr. Shafini, Fakrudin and Muhdi, and all my friends in USM and UiTM; Najihah, Ezedin, Anusha, Rahma, Hadri, Anisha, Ina, Ijat, Rafiza, Rihan, Rozi, K.Ina and many more for their continuous encouragement, prayer, precious love and support throughout this journey. Lastly, the memories remain for my late friends (UiTM's colleagues, 2010-16) TuanSya, K.Faidah, K.Lily, K.Koni, K.Husniah, Sopiah and Shafrah after a battle with cancer. Deeply, I have learned a lot of things throughout this journey, *'Life is just a passing moment, nothing is meant to stay'*. Thank you Allah.

TABLE OF CONTENTS

Page

CHAPTER 1 –INTRODUCTION

CHAPTER 2 –MASS SPECTROMETRY-BASED BIOMARKER DISCOVERY STEPS

CHAPTER 3 –LITERATURE REVIEW

CHAPTER 4 –RESEARCH METHODOLOGY

CHAPTER 5 –FEATURE EXTRACTION AND FEATURE SELECTION FOR MASS SPECTROMETRY-BASED BIOMARKER DISCOVERY

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony
Acc	Accuracy
ACO	Ant Colony Optimisation
CR	Crossover rate
CSV	Comma-separated value
DE	Differential Evolution
ESI	Electrospray Ionization
FN	False Negative
FP	False Positive
FS	Feature Selection
FT-MS	Fourier Transform-based Mass Spectrometry
GA	Genetic Algorithm
GC/MS	Gas Chromatography with Mass Spectrometry
MAD	Median Absolute Deviation
MALDI	Matrix-Assisted Lase Desorption/ionization
MCN	Maximum cycle
MS	Mass Spectrometry

MS/MS	Tandem Mass Spectrometry
M/Z	Mass-to-charge Ratio
PCA	Principle Component Analysis
PSO	Particle Swarm Optimisation
QAQC	Quality assurance, quality control
QSTAR-TOF	High Performance Hybrid Quadrupole Time-of-flight Mass Spectrometer
ROC	Receiver Operator Characteristic
Sen	Sensitivity
SELDI	Surface-Enhanced Laser Desorption Ionization
SNR	Signal-to-Noise ratio
Spec	Specificity
SVM	Support Vector Machine
TOF	Time-off Flight
TP	True Positive
XML	Extensible Markup Language

ALGORITMA KOLONI LEBAH TIRUAN BERSAMA EVOLUSI BERBEZA UNTUK CIRI PENYARINGAN DAN PILIHAN DATA JISIM SPEKTROMETRI

ABSTRAK

Kemajuan dalam teknik spektrometri jisim untuk kajian proteomik telah meningkatkan penemuan pengecaman-bio daripada corak kuantitatif proteomik. Pemprosesan data yang banyak untuk molekul yang terlibat boleh meningkat kepada siri puncak saling berkait dan bertindih di dalam spektrum jisim. Spektrum ini juga mengalami data berdimensi tinggi berbanding saiz sampel yang kecil. Beberapa kajian telah memperkenalkan teknik statistik dan pembelajaran mesin seperti Analisa Komponen Asas (*(PCA)*), Analisa Komponen Tak Bersandar (*(ICA)*) dan Analisa Riak Pekali (*wavelet-coefficient*) untuk mengekstrak data yang berpotensi. Namun, tiada satu pun daripada kaedah yang dibincangkan mengambil kira dengan serius masalah kelemahan data yang berdimensi tinggi benbanding saiz sample yang kecil. Kajian ini telah tertumpu kepada dua peringkat dalam analisa spektometri jisim. Pertama, kaedah ciri penyaringan iaitu akan menyaring puncak-puncak yang memberi inferens tentang maksud biologi bagi data tersebut. Anggaran pengecutan bagi kovarians telah di cadangkan untuk mengumpul *m/z windows* dan mengenalpasti pekali korelasi terbaik antara puncak-puncak bagi data spektometri jisim untuk ciri penyaringan. Kedua, kaedah ciri pemilihan yang mencari ciri-ciri terbaik berdasarkan keputusan yang paling tepat daripada model klasifikasi yang dijanakan. Suatu teknik pengkomputeran yang mimik kepada kemandirian dan proses semulajadi yang di kenali sebagai Koloni Lebah Tiruan (ABC) digabungkan dengan pengklasifikasi SVM telah di cadangkan sebagai ciri pemilihan dan kemudiannya dihibridkan dengan teknik Evolusi Berbeza (DE) sebagai algoritma deABC untuk mengembangkan lagi fungsi eksplorasi algoritma ABC yang asli. Kaedah yang dicadangkan ini telah diuji dengan data berdimensi tinggi daripada Spektroskopi jisim yang melibatkan set data kanser ovari, hati (HCC) dan Toksik Berasaskan Dadah (TOX) untuk menilai kuasa diskriminasi, ketepatan, kepekaan dan spesifikasi. Untuk kaedah penyaringan, keputusan ciri diskriminasi yang di hasilkan oleh anggaran pengecutan telah di uji dengan laporan kajian terdahulu dan menunjukkan keputusan yang lebih baik. Manakala untuk kaedah pilihan, perbandingan telah dibuat dengan algoritma *Particle Swarm Optimisation* (PSO), *Ant Colony Optimisation (ACO)* dan laporan kajian terdahulu sebagai ciri pemilihan. Algorithma deABC yang dicadangkan telah menunjukkan ketepatan bagi 98.44, 88.89 dan 93.75 peratus untuk set data kanser ovari, TOX dan hati (HCC) dan secara purata mengatasi prestasi PSO, ACO dan kajian yang sama yang telah dilaporkan.

ARTIFICIAL BEE COLONY WITH DIFFERENTIAL EVOLUTION ALGORITHM FOR FEATURE EXTRACTION AND SELECTION OF MASS SPECTROMETRY DATA

ABSTRACT

The advancement in mass spectrometry technique for proteomic studies has proliferated the discovery of biomarkers from quantitative proteomics pattern. High-throughput data for a given molecule can give rise to a series of inter-related and overlapping peaks in a mass spectrum. The spectrum suffers from high dimensionality data relative to small sample size. Several studies have proposed statistical and machine learning techniques such as Principle Component Analysis (PCA), Independent Component Analysis (ICA) and wavelet-coefficient in order to extract the potential features. However, none of these methods take into account the huge number of features relative to small sample size. This study focused on two stages of mass spectrometry analysis. Firstly, feature extraction methods extract peaks as potential features to infer biological meaning of the data. Shrinkage estimation of covariance was proposed to assemble m/z windows and identify the correlation coefficient among peaks of mass spectrometry data for feature extraction. Secondly, feature selection techniques search parsimonious features through a learning model that exhibits the most accurate results. A computational technique that mimics survival and natural processing known as Artificial Bee Colony (ABC) integrated with linear SVM classifier was proposed for feature selection. Later, this was hybrid with Differential Evolution (DE) techniques (deABC) algorithm in order to expand the exploration of basic ABC. The proposed method was tested with several real-world high resolution mass spectrometry datasets which are ovarian cancer, liver (HCC) and Drug-induced toxicity (TOX) datasets to evaluate the discrimination power, accuracy, sensitivity and specificity. For feature extraction, the analysis was made with reported studies. The shrinkage estimation has performed better discriminative analysis on the similar features. For feature selection, the comparisons have been made with Particle Swarm Optimisation (PSO), Ant Colony Optimisation (ACO) algorithms and reported studies. The proposed feature selection deABC algorithm exhibited accuracy of 98.44, 88.89 and 93.75 percent on ovarian cancer, TOX and liver (HCC) datasets respectively and in average outperformed the PSO, ACO and similar reported study.

CHAPTER 1

INTRODUCTION

Bioinformatics can be viewed as the marriage of information technology with molecular biology and have covered algorithms, sequence representation, Markov modeling, neural network to predict protein secondary structure and other mathematical modeling method for analysis and storage of biological data. Discussions in bioinformatics frequently centered on two important biological molecules; (1) proteins and (2) nucleic acids (Ho, 2007). Proteins work accordingly to achieve a particular function in the cell, and they participate in every function of the cell. Meanwhile, nucleic acids are very large and complex organic molecules that include DNA and RNA to transmit genetic code from parents to offspring.

In proteins, many of them are enzymes that catalyse biochemical reactions, cell signaling, immune responses, cell adhesion and the metabolism. Hence, proteomics reveal large-scale studies of protein's structure, function, protein-protein interaction, and amounts expressed in living cells. In contrast to genomic, the proteomics study reflects more accurately on the dynamic state of cell, tissue or an organism (Amelina, 2011).

One of the major concerns in proteomics study is how their quantities, modifications and structures change in response to the need of the body or in disease. As an example, the cancer cells that often secrete specific proteins or fragments of proteins into the bloodstream and other body fluids such as urine, serum and saliva. Patterns of proteins called protein signature in these easily accessible body fluids could provide information about the risk, presence, and progression of disease (Jurisicova et al., 2008; Latterich et al., 2008; Paulovich et al., 2008). This knowledge ultimately could improve the diagnosis and prognosis cancer cases in the early stage or before symptoms are presented and customized treatment to the individual patient (therapy monitoring).

1.1 Background

In recent years, protein markers has shown great opportunity in diagnosis and prognosis of diseases. A study done by Farina (2014) has shown the rising trend in the proteomics cancer biomarker for the past 10 years. Proteomics biomarkers are based on the idea that the major workhorse of biological system, diseases and other malfunctions may be reflected by the proteomic level. As depicted in Figure 1.1, disturbances in proteome are caused by mutation, such as faulty post-translation modification, interference in protein-protein interaction, deleterious effects on pathways and networks and unnatural changes in protein expression.

Traditionally, identification and quantification of single protein biomarker is applied for biomarker discovery. However, due to complexity of biological pathway and heterogeneity between individual, single protein biomarker predictive utility might be limited (Catchpole, 2013). Alternatively, the panel of biomarkers are utilised to evaluate the activity of perturbation of biological system (Torrente et al., 2012; Sajic et al., 2015). The advancement in tandem mass spectrometry (MS/MS) through quantitative methods allow rapid identification and comparative quantification of several hundreds

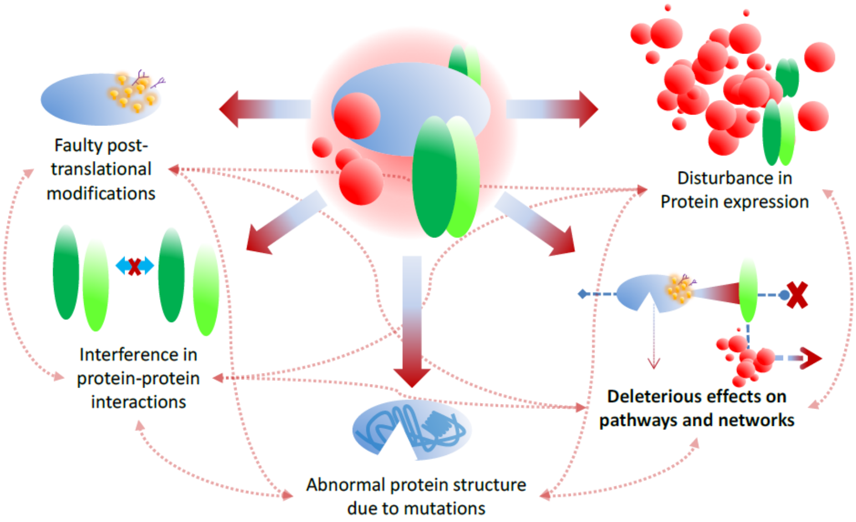

Faulty post-translational modifications
Disturbance in Protein expression
Interference in protein-protein interactions
Deleterious effects on pathways and networks
Abnormal protein structure due to mutations

Figure 1.1: Effects of disturbances in proteome (figure taken from Parviainen et al. (2014))

of proteins simultaneously (Pan et al., 2008).

The tandem mass spectrometry analysis produce huge number of peaks in high-dimensionality mode that are used to identify peptides. These peaks are generated from mass-to-charge ratio (m/z point) across the generated spectrum of MS data. Since the spectrum of high dimensionality data consist of huge number of m/z points as potential features, computational processing method is crucial for proper analysis and identification. Since a decade, machine learning technique has played important roles to generate reliable method for the complex analysis of raw data(Libbrecht and Noble, 2015). Generally, machine learning technique will process the raw high-dimensional data into tractable number of features using feature selection techniques. Most of the time, feature selection significantly plays as a vital key to balance the usage of few features as possible while maintaining high predictive and discriminative power (Yu et al., 2015). The balance is crucial since the goal is to produce highly accurate but small panels of biomarkers with potential clinical utility (Swan et al., 2013).

1.2 Motivation

High-resolution mass spectrometry generates extremely high-dimensional data of mass spectra (Sajic et al., 2015) consist of tens of thousands points of mass to charge ratio (m/z) of the substance. Each point might depict particular feature of protein or peptide. This huge number of features are relative to small number of samples. Some of the studies proposed particular statistical analysis to extract the data (Gibb and Strimmer, 2015), meanwhile some of the studies focusing on machine learning techniques such as dimensional reduction and wavelet analysis to extract and identified the parsimonious features (He et al., 2013; Neehar and Acharyya, 2013). However, none of these studies really address the dimensionality that focus to huge numbers of features relative to small sample size. This issue has been highlighted recently as crucial for biological data (Schäfer et al., 2005; Yao et al., 2008; Sanavia et al., 2012).

The high-resolution data produces much complex peaks due to two level of fragmentation compared to low-resolution data. Therefore, neighbourhood peaks in high-dimensionality data may infer similar proteins or peptides (He et al., 2009, 2013). Several studies have constructed features from high-resolution data as group of neighbourhood peaks known as peaks-bins or m/z windows. However, the proposed methods were machine dependent (Ressom et al., 2007) and the implementation of empirical statistical analysis are sensitive to huge numbers of features relative to small sample size (He et al., 2013). Therefore, a robust method to construct neigbourhood peaks across different platform of mass spectrometry instrument and without limit to particular dataset are desirable. The method should also take into account the weakness of empirical statistical analysis in evaluating the correlation of neighourhood peaks as a feature. Proper evaluation of neighbourhood peaks may also improve the discriminative characteristic of the features, thus capable to distinguish them between healthy and disease cases. Instead of that, according to Mostacci et al. (2010), the real 'biological' peaks are expected to sustain across samples. This is best known as reproducible issue in mass spectrometry data. Statistical analysis such as feature ranking concern only the correlation of the data, but does not evaluate the reproducible peaks across different samples. Therefore, the reproducible issue of peaks in mass spectrometry is also being highlighted in extracting the potential features (Zhang et al., 2010).

In the next phase of feature selection analysis, a classifier would accurately distinguish cancer and normal cases from entire thousands of features in spectra, but the classification model does not help in finding specific biomarkers. Therefore, small set of peaks are used to computationally predict markers with high accuracy (Ressom et al., 2007) and then are considered as panel of biomarkers. On the other hand, the feature selection method for biomarkers discovery are still open for improvement in terms of better accuracy for prediction (Swan et al., 2015). Furthermore, the challenge of biomarkers discovery also relies on robustness of the method that should be able to identify markers from different types of dataset. Hence, it is motivating to study on feature selection that is reliable in finding small set of marker over different types of cancer cases.

1.3 Research Questions and Objectives

The objectives of this study are identified by answering the following research questions:

1. How to assemble reliable and discriminative peaks-bins/ m/z windows from a list of detected peaks from high dimensional data but small sample sizes? The assumption is that strong correlation among peaks would represent a similar protein. (Objective 1)

2. How to consider stable reproducible peaks for feature extraction? (Objective 1)

3. Is it possible for bio-inspired optimisation to efficiently collaborate with standard classifier in order to select subset of features and generate high predictive and discriminative power? (Objective 2)

4. Are the proposed feature selection method with the integrated classifier, able to perform on different types of disease for biomarker discovery? (Objective 3)

The research objectives are:

1. To develop a feature extraction method that consider neighbourhood peaks and discriminative characteristics that are robust for huge number of features with relatively small sample sizes.

2. To develop a hybrid feature selection algorithm, built-in with SVM classifier that optimise searching of parsimonious features for biomarker discovery.

3. To evaluate the predictive and discriminative power performance of classification model for biomarker discovery from several diseases datasets.

1.4 Significance of This Study

The cancer rates keep rising over the past two decades and yet the cure factor is still low which raises alarming concern among the medical practitioner. According to Anderson (2010) even though several protein biomarkers have been introduced, the general amount of new clinical protein biomarkers have been low within the recent years. This is due to cost and effort to transform initial discovery to validated clinical solution (Amelina, 2011; Parviainen et al., 2014).

Searching proteins that indicate disease through mass spectrometry (MS) analysis has accelerated the discovery phenomena for biomarker identification. Computational methods proposed in comparing protein expression levels in normal cases with cancer cases sample lead to identification of potential biomarkers that can predict the degree of malignancy in tumors (Libbrecht and Noble, 2015). Therefore, computational technique is more reliable in reproducibility of prediction model of biomarkers in two ways; (1) same disease but different experiments and datasets; (2) different diseases. The results of the analysis provides valuable information about the efficacy of specific anti-cancer treatments or help to identify new molecular target for innovative therapeutic strategies (Parviainen et al., 2014). Further, those particular biomarkers can be used to evaluate the result of the treatments and monitor the long term recurrence of the disease on specific patient (Hathout, 2015).

1.5 Research Scope and Limitation

The research scope and limitation for this study are listed as follows:

1. Scope of data are label-free mass spectrometry from both MALDI and SELDI techniques and limited to high-resolution datasets.
2. Development of the feature selection studies are limit to wrapper approach.
3. Focusing on bio-inspired algorithms for feature selection.

1.6 Thesis Organisation

In general, the research is organised into 7 chapters.

Chapter 2 serves as introduction to fundamental aspects of mass spectrometry analysis that covers brief introduction to the instrument for proteomics analysis, data representation from the instruments, common mass spectrometry pipelines for biomarkers discovery and followed by a list of pre-processing methods applied.

Chapter 3 discusses the literature reviews on both feature extraction and feature selection in the domain of mass spectrometry analysis for biomarkers discovery. The literatures start with general concept of feature extraction and selection in optimisation. Further, the discussion focuses to the methods that applied to biomarkers discovery in mass spectrometry data.

Chapter 4 gives an insight of the methodology used in this research. The whole activities involved in this study on each phases are described and visualised. The phases start with formatting of raw data, pre-processing, feature extraction, feature selection and classification approaches. In addition, proposed methods on both feature extraction and selection are also highlighted in terms of the way data of mass spectrometry is mapped to the method.

Chapter 5 provides the extension of development of the proposed methods which are mainly about constructing m/z windows using shrinkage estimation as potential features and reproducible technique for feature extraction. While on feature selection, details mechanism such as adaptation and modification of the ACB algorithm as feature selection and further hybrid with Differential Evolution Algorithm are elaborated.

Chapter 6 covers the analysis and comparison of the analysis based on classification predictive results and discriminative analysis. The analysis mainly based on accuracy, sensitivity and specificity performance outputted from classification process, the ROC representation and discriminative analysis on three different datasets.

Chapter 7 concludes the main finding of this research and proposed for future works.

CHAPTER 2

MASS SPECTROMETRY-BASED BIOMARKER DISCOVERY STEPS

Mass spectrometry analysis is a complex process and contains complex mixture of proteins data, therefore fundamental concept highlighted in this chapter intends to provide clear picture of the basis in mass spectrometry data. Hence, the chapter begins with an overview of MS, introduction to the instrument and the concept of tandem mass spectrometry (MS/MS) as high resolution MS data. Next, the data representation for the raw data is elaborated and followed by discussion on MS's pipeline for biomarker discovery. Lastly, several sub-task applied for the pre-treatment or pre-processing analysis are presented.

2.1 Mass Spectrometry Overview

Mass spectrometry is a powerful analytical technique used for the analysis of large molecules. It is used to identify and quantify unknown compounds, determine molecular masses of large biological samples, elucidate their structural and quantitative information, and investigate intermolecular re-actions. These properties hold high significance for an analytical chemist or a life scientist in order to understand the behaviour of bio molecules that control biological systems and in turn, control our body. Mass spectrometry provides valuable information to a wide range of professionals such as chemist, biologist, astronomers and physician. For example, it is used to detect and identify the use of steroids in athletes, monitor the breath of patients by anesthesiolo-

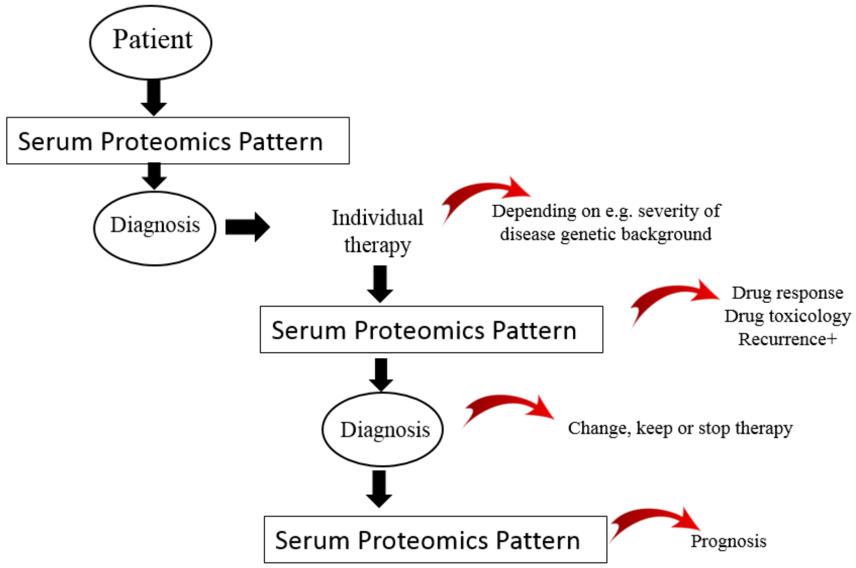


Figure 2.1: Systematic identification and characterization of protein pattern from body fluid for diagnosis and prognosis markers (figure taken from Seibert et al. (2004)).

gists during surgery, determine the composition of molecular species found in space, and determine how drugs are used by the body (Aebersold and Mann, 2003). In addition, the MS technology offers a helping hand in the systematic identification and characterization of protein for diagnostic and prognostic markers in tissue, blood serum and other body fluids as depicted in Figure 2.1.

There are wide ranges of mass spectrometry instruments used for identification and analysis in biotechnology itself. For example, Gas Chromatography with Mass Spectrometry (GC/MS) used by chemists to identify structural features of compound; Matrix-assisted laser desorption/ionization (MALDI) and Surface-Enhanced Laser Desorption Ionization (SELDI) used by pharmacists to identify proteomics patterns of proteins and peptides for various applications include biomarker analysis (Seibert et al., 2004). Both MALDI and SELDI techniques have created a beautiful insight towards high-throughput proteomics analysis to the researchers across multi-disciplines. It works on the principle that different molecules have different masses. Thus, once a substance is injected to the instrument, the constituent can be separated according to their masses.

2.2 Principle in Mass Spectrometry (MS)

Mass spectrometry (MS) is producing ions of the analytical compounds and separating ions according to their mass-to-charge-ratio (m/z). An important enhancement and capabilities of mass spectrometry is currently being used in tandem with chromatographic separation techniques. The two types of chromatography techniques adopted are; (1) Gas chromatography that separates compound chromatographically using gas in mobile phase; (2) Liquid chromatography that uses liquid in mobile phase which usually contains a mixture of water and organic solvents. Mass spectrometer is split into two main classes, the first class performs single mass spectrometry and the second one performs tandem mass spectrometry (MS/MS). This study is only focused on tandem mass spectrometry and will be explained in the next subsection. The measurement of mass spectrometer instruments consist of three major components; ionization source, the mass analyzer and the detector.

1. **Ionization Source**

   As mentioned above the compound under analysis has to be ionized before the mass can be measured and the ionization process is done in ionization source. When dealing with peptides and protein, ionization is commonly achieved by the addition of protons and the molecules. This addition also increases the mass of the molecule by the nominal mass of 1 Da per charge (per proton). Sources which cause only limited fragmentation are called soft ionization sources, as opposed to hard ionization sources, in which components typically fragment upon ionization. Soft ionization sources are used for peptides and proteins, and if fragmentation is desired afterward (as in MS/MS) other post-source method are used to achieve fragmentation. The most common uses of soft ionization in proteomics are Electrospray Ionization (ESI) and Matrix-Assisted Laser Desorption Ionization (MALDI). Anyhow, focusing to biomarker analysis MALDI and SELDI are most applicable ionization method (Ahmed, 2008).

2. **Mass Analyzer**

   Mass analyzer separates ionized peptides according to mass-to-charge ratio (m/z). This is achieved by the generation of electric or magnetic fields that separate the ions based on trajectories, velocity or direction. Thus, mass analyzer is the central of technology with the key parameters such as sensitivity, resolution, mass accuracy and the ability to generate valuable information of mass spectra from peptide fragmentation (MS/MS spectra) Aebersold and Mann (2003) . At present, there are four basic types of mass analyser which have been used in proteomics research; ion trap, time-of-flight (TOF), quadrupole and Fourier Transform ion cyclotron (FT-MS) analyser. They are different in design and performance, each with its own strength and weakness. Figure 2.2 shows how these four types of mass analyzer incorporate in mass spectrometer.

3. **Detector**

   Detector registers the relatives'number of ions at each *m/z* values and plots the spectrum as abundance intensities in Y-axis versus mass-to-charge (m/z) ratio of its ions in X-axis.

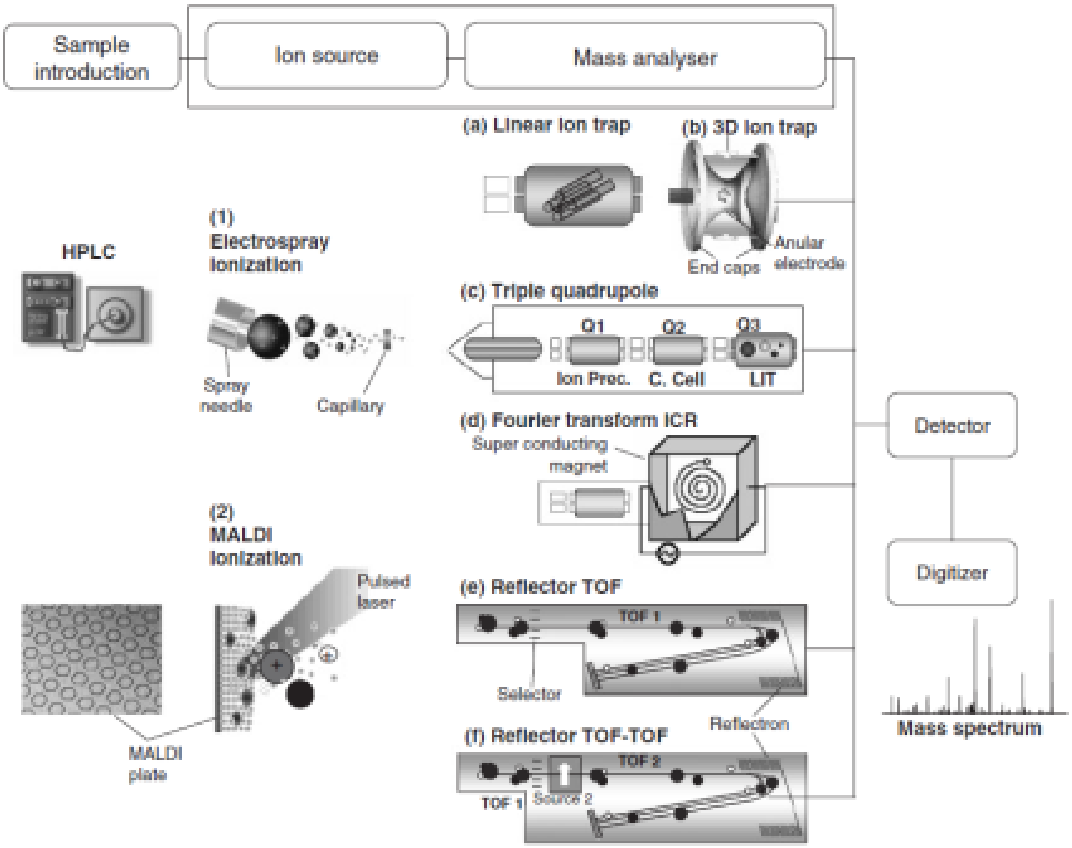


Figure 2.2: Main component of a mass spectrometer (Figure taken from Cañas et al. (2006)). Sample introduction device, ionization source for ion generation, mass analyzer for ion separation, and ion detector to transform analogue signals into digital signals and record a mass spectrum. Common ionization sources for proteomic research are ESI and MALDI. Widespread mass analyzer are ion traps (a) Linear, (b) Three-dimensional; (c) Triple quadroples; (d) Fourier transform cyclotron; and (e) Time-of-flight (TOF). Usually ion trap and quadrupole analyzer are coupled to ESI ion sources, whereas TOF analyzers are usually combined with MALDI ion source.

2.3 Tandem Mass Spectrometry (MS/MS)

Tandem mass spectrometry (MS/MS) is mainly used to produce structural information about a compound by fragmenting specific sample ions inside the mass spectrometer and identifying the resulting fragment ions (Parker and Borchers, 2014). Figure 2.3 shows the process of two level fragmentation from parent ion to produce daughter ion. For example, the parent ion is any specific protein existing in the sample, whereby daughter ion could be any peptides information that construct the protein. This information can then be pieced together to generate structural information regarding the intact molecule. Tandem mass spectrometry also enables specific compounds to be detected in complex mixtures on account of their specific and characteristic fragmentation patterns.

The advancement in tandem mass spectrometry (MS/MS) which produces high-resolution spectra embarks the in depth study of biomarkers through proteome profiling (Sajic et al., 2015). Through tandem mass spectrometry, high number of peaks are generated from a single spectrum of sample that represent peptides. Furthermore, they are much better reproducibility between and within machine runs (Conrads et al., 2003), thereby produce predicted model with higher sensitivity and accuracy. Moreover, the spectral resolution from low-level resolution or single fragmentation which produced only parents ions, have no ability to produce specific ions that are close in mass/charge, which can cause multiple specific discreet ions to coalesce into a single peak (Petricoin and Liotta, 2004). High-resolution mass spectrometry analysis remains to be seen as potential method for future clinical diagnostic platform.

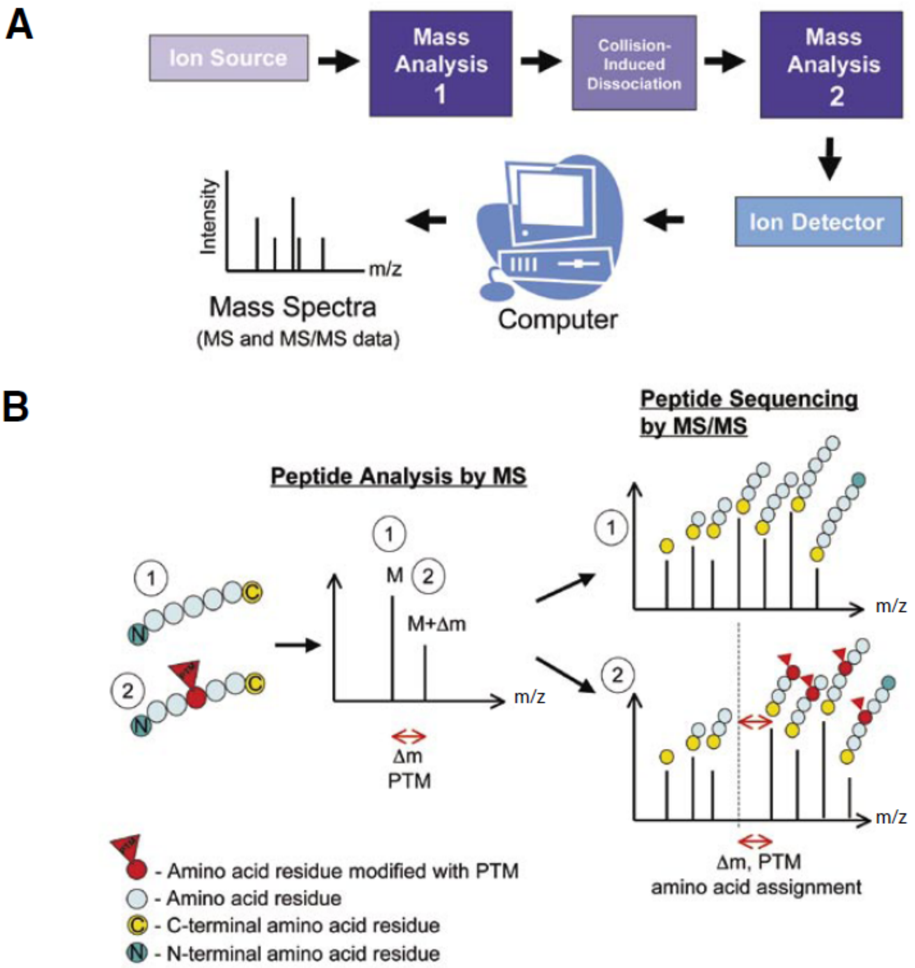
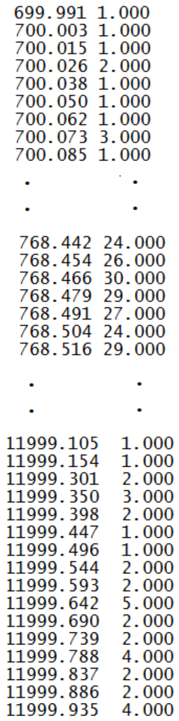


Figure 2.3: Process of fragmenting parent ions into daughter ions (MS/MS analysis). Figure taken from Larsen et al. (2006)



```
699.991 1.000
700.003 1.000
700.015 1.000
700.026 2.000
700.038 1.000
700.050 1.000
700.062 1.000
700.073 3.000
700.085 1.000
  .         .
  .         .

 768.442 24.000
 768.454 26.000
 768.466 30.000
 768.479 29.000
 768.491 27.000
 768.504 24.000
 768.516 29.000

  .         .
  .         .

11999.105  1.000
11999.154  1.000
11999.301  2.000
11999.350  3.000
11999.398  2.000
11999.447  1.000
11999.496  1.000
11999.544  2.000
11999.593  2.000
11999.642  5.000
11999.690  2.000
11999.739  2.000
11999.788  4.000
11999.837  2.000
11999.886  2.000
11999.935  4.000
```

Figure 2.4: Text file from a sample of raw data in mass spectrometry

2.4 Data Representation

The raw data for each sample of mass spectrometry either disease or normal case is commonly presented as CSV format or text files. The raw data is shown in Figure 2.4, where the first column represents mass to charge ratio (m/z) values of the spectrum and the second column represents relative signal intensity ion. The m/z is referred to distribution of ions by mass in unit of dalton (Da).

Both disease and normal samples from the raw data are then compiled into manageable form that combine all samples into a table of dataset which is easier to be analysed as represented in Table 2.1. Each sample is viewed as a spectrum composed of m/z on X-axis and intensity of particular m/z on Y-axis as portrayed in Figure 2.5.

Number of samples in different datasets might vary depending on how the data's are collected. For example, Figure 2.5 is depicted two spectrums from 216 samples

Table 2.1: Data representation

m/z ratio	Cancer sample1	Cancer sample2	. . .	Cancer sample*k*	Control samplek+1	.....	Control sample *n*
2000	0.1179	0.1735		0.3620	0.1727		0.0561
2000.384	0.1735	0.1619		0.2581	0.1238		0.0439
2000.786	0.2317	0.1883		0.1998	0.1078		0.0445
. .	. .	. .	. .	. .	. .	. .	
8787.623	0.0667	0.0425		0.0083	0.0241		0.0317
8788.428	0.0390	0.0345		0.0242	0.0275		0.0316
. .	. .	. .	. .	. .	. .		. .
11899.063	0.0256	0.0131		0.0562	0.0357		0.0369

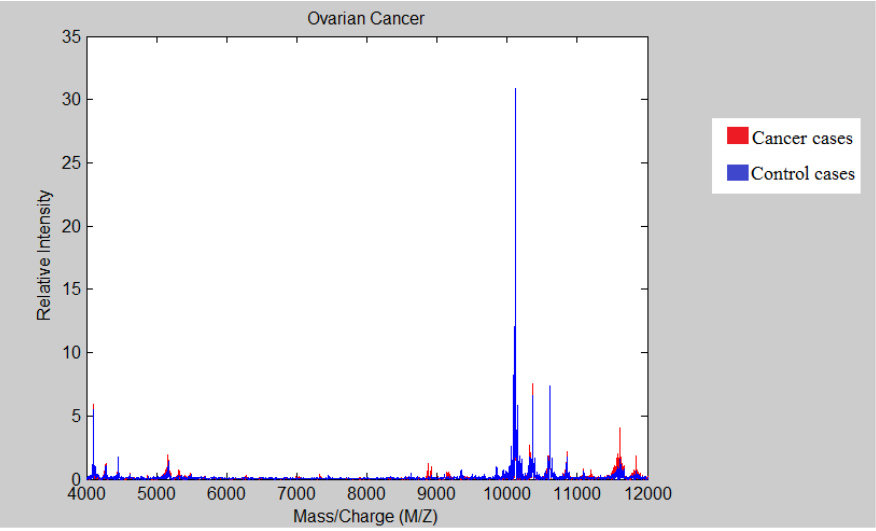


Figure 2.5: Sample of cancer and control spectrum from ovarian dataset

of Ovarian dataset that compose of 121 cancer cases and 95 control cases. From this figure, potential features for predicting biomarkers rely on each mass-to-charge ratio across 15,000 points on X-axis. Thus, tens of thousands potential features exhibit biological meaning from only 216 samples existing in dataset.

2.5 Mass Spectrometry Pipeline in Biomarker Discovery

Mass spectrometry analysis is varied in techniques, Figure 2.6 depicts the pipeline involved in the MS analysis for biomarker discovery which consider peaks as feature

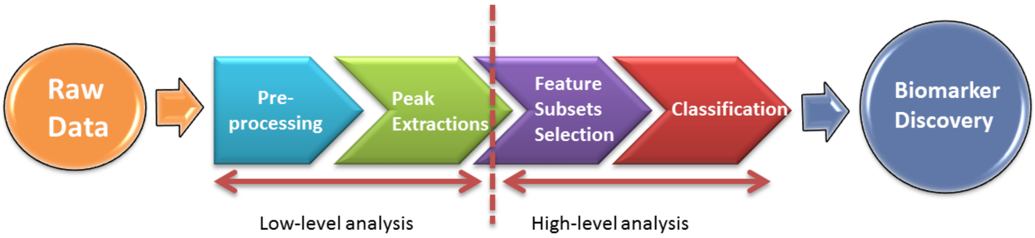


Figure 2.6: Mass Spectrometry pipeline

extraction method. In general, the whole process of biomarker discovery in mass spectrometry is classified into two levels which are; (1) low-level analysis which concerns with cleaning the raw data and finding potential features; and (2) high-level analysis which concerns on searching of parsimonious features and optimisation methods for biomarker discovery (Ressom et al., 2005; Armananzas et al., 2011).

2.6 Pre-processing Analysis

The raw data acquired from mass spectrometry equipment is not only affected by noise from the chemical source but also by variations and degradations encountered in preparation of the samples. This has been proved by several previous studies (Arneberg et al., 2007; Cruz-Marcelo et al., 2008; Wang et al., 2010) that proposes and compare several pre-treatment methods in order to standardize and maximize the quality of the raw data. However, considerable improvements in the reliability and accuracy of subsequent processes have been witnessed as a consequence of identifying and quantifying all the potential features present in the sample. This section discusses several pre-processing steps that can be performed in any order since there is no established or gold standard methods (Cruz-Marcelo et al., 2008). Some of the important pre-processing steps are as follows:

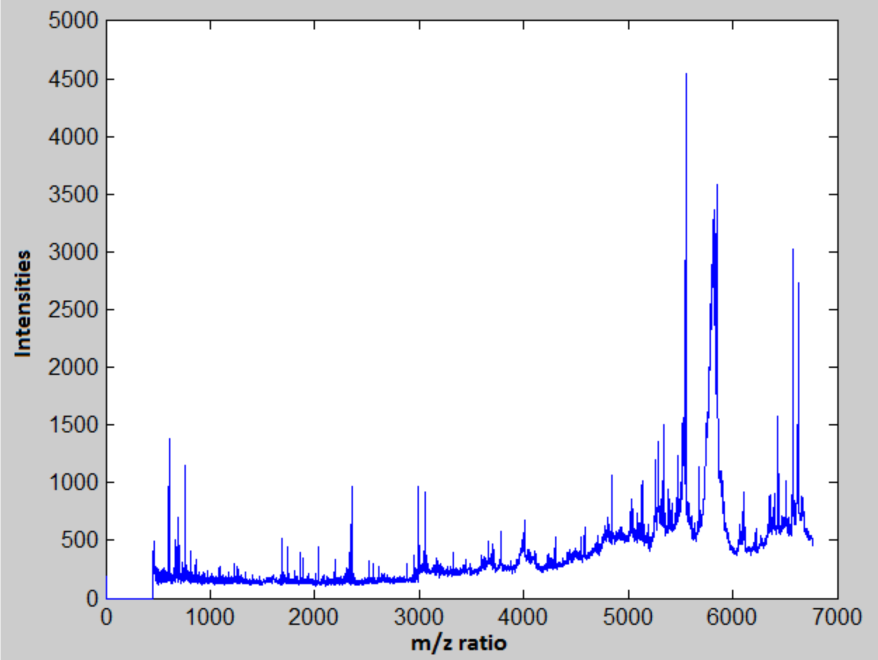


Figure 2.7: Standard baseline removal in ovarian cancer dataset: Before baseline removal

2.6.1 Baseline Removal

Baseline, or also known as background correction generally involves estimating and subtracting the baseline from the signal (Yang et al., 2009). Mass spectrometry raw data varies in their baseline across m/z values due to chemical contamination even for the same datasets. This chemical noise mainly occurs in matrix or is caused by detector overload during the experiment. The aim of baseline correction is to retain peak shape and flattens non-peaks. Several methods have been proposed to subtract the baseline across m/z axis. One of the common ways is to firstly adjust the variable baseline of a raw mass spectrum by estimating the baseline with multiple shifted windows of a specific width (Zhao and Davis, 2009). Secondly, a spline approximation algorithm is employed to regress the varying baseline to the window points. Alternative to this approach, a non-linear method such as top-hat morphological operator could be employed. A detailed discussion regarding baseline estimation and correction has been

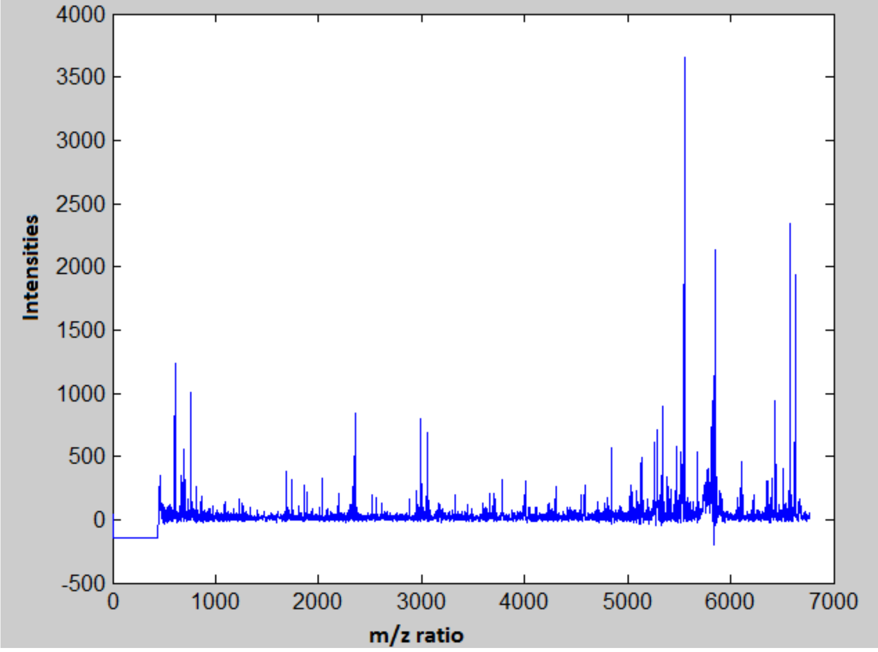


Figure 2.8: Standard baseline removal in ovarian cancer dataset: After baseline removal

discussed in Shin et al. (2008). Meanwhile, Figure 2.7 and 2.8 shows the result of before and after baseline removal for an ovarian cancer sample.

2.6.2 Noise Removing

Noise filtering process is aimed to clean the raw spectra from noise produced by electronic (white noise) and chemical sources by eliminating peaks which fall below a pre-defined threshold. This process has been applied by various wavelet methods that attempt to not only clean the noise but also smoothing the observed signals. Frequently, the process of denoising and smoothing combine wavelet techniques with several types of filtering methods such as Savitzky-golay filter, Gaussian filter and Moving-average filter. More information about these methods are described by (Yang et al., 2009). Figure 2.9 shows the result of removing noise from a spectrum.

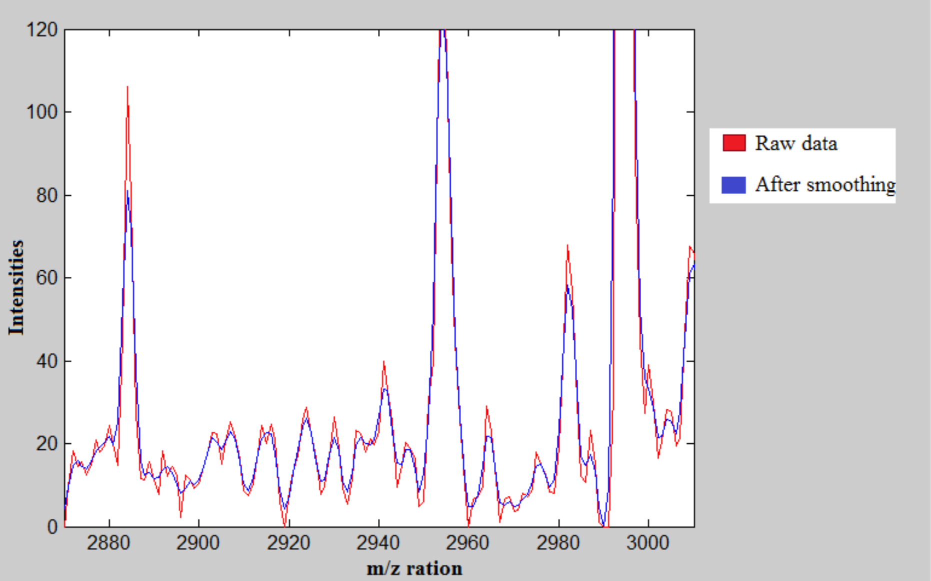


Figure 2.9: Noise removing and smoothing

2.6.3 Normalisation

Normalisation technique is applied to ensure that the sample falls on a specified range. Specific to mass spectrometry data, since the intensity of peaks in inter-spectrum are highly variable, normalisation is applied to increase the reliability of the data by using them in uniform way. One of the widely used methods is by dividing each spectrum by its Total Ion Current (TIC) and then rescaling the spectrum into relative intensity values below 100 *m/z*. This method and other comparison studies have been explained by Meuleman et al. (2008). Figure 2.10 depicts the normalisation result on a spectrum.

2.6.4 Alignment of Spectra

Due to large and high-throughput data from various spectra to be processed, alignment of spectra has been done to ensure that the similar peaks for every spectrum are correctly matched and reflected to the same protein intensities. It can overcome errors that could happen during the process of identifying peptide's signal with molecular weight. For example, biomarker discovery is to identify the location of peaks where peak inten-

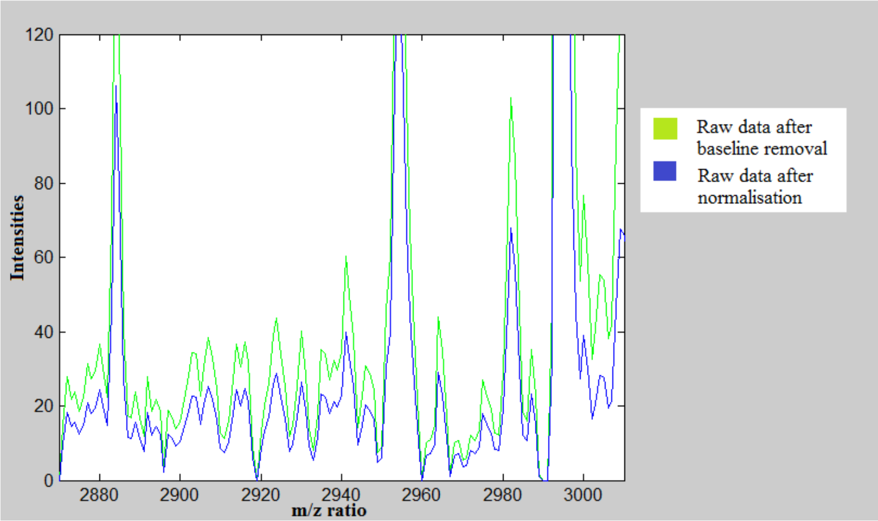


Figure 2.10: Normalise spectrum

sities (or absence /presence of peaks) have strong contrast between control and disease individuals. Hereby, alignment of peaks are important before the comparison of same peaks in different dataset is done.

Jeffries (2004) has proposed two algorithms to improve alignment among samples; the first algorithm works with SELDI data produced by Chipergen instrument; and the second is used for general format. The first algorithm is based on Chiphergen's conversion of time-of flight data to mass values via a quadratic equation. The second algorithm was created by assuming that the data was represented in a two column format; X-axis for *m/z* values and Y-axis for intensities values. The concept of alignment is based on fitting cubic splines to the data rather than quadratic equation. A detailed discussion on alignment process has been done by Wong et al. (2005).

2.7 Chapter Summary

This chapter identify the fundamental of mass spectrometry data from the basic instruments until pre-processing approaches. The study has analysed the pre-processing methods and highlighted the mass spectrometry pipeline for biomarker discovery pertaining to peaks extraction approach. The pipeline determines the importance of each phase that may impact the quality of biomarkers discovery analysis. Hence careful consideration for feature extraction and feature selection phases are the main concern of this study and the proper analysis are performed in the next chapter.