ANALYSIS OF SELDI MASS SPECTRA FOR BIOMARKER DISCOVERY AND CANCER CLASSIFICATION

By
YAPING CHENG

A thesis submitted to
The University of Birmingham
for the Degree of
DOCTOR OF PHILOSOPHY

CRUK Cancer Studies
Medical School
University of Birmingham
June 2009

Abstract

The thesis focused on data analysis of surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF MS) for biomarker discovery and cancer classification. It investigated quantitative measures of reproducibility and found that SELDI protein profiles are affected by sample storage and processing procedure. Two new peak alignment algorithms were proposed, one of which achieved the best performance when compared to the existing methods. The assumption of normality of SELDI protein profiles, on which the standard statistical methods are based, was examined. Normality tests and multiple testing procedures revealed that SELDI protein profiles do not follow normal distributions, implying that it may be reliable to use non-parametric methods for detecting disease-associated proteins. A new normalisation algorithm was proposed, and was shown to give a better improvement of normality compared with the existing methods. An integrated algorithm to discover proteomic biomarkers for cancer diagnosis was proposed and applied to two published SELDI data sets. The results demonstrated that the receiver operating characteristic (ROC) curve method may be more reliable to determine the discriminatory powers of the identified biomarkers compared to Wilcoxon test. The methods for proteomic biomarker discovery presented here may be generalisable and applicable to other mass spectrometry and genomics approaches.

Declaration

This thesis has prepared in accordance with the regulations for the degree of Doctor of Philosophy. It is composed by myself, and has not been submitted in any previous application for any degree. The work described in this thesis has been undertaken by myself, except where otherwise stated. I have contributed to six publications (Cheng et al.,2006; Cheng et al.,2008; Ward et al.,2006a; Ward et al.,2006d; Ward et al.,2006c; Ward et al.,2006b) (Appendix 8-1).

Dedication

This thesis is dedicated to my parents.

Acknowledgements

I would like to thank my supervisors, Dr. Wenbin Wei and Professor Philip J. Johnson, for offering me a great environment, enabling the work to take place, for the support and supervision throughout my study here and for suggestions on my thesis writing up.

A special thanks to Dr. Stephen Nyangoma for his statistical advices, suggestions, and comments on the thesis, and for many fruitful discussions I had with him. I should highlight that although he was not my official supervisor, he diligently guided me through this work.

I would like to thank all members in the Biomarker Discovery and Proteomics group in which I have been working for last three years, especially to Dr. Douglas Ward, Dr. Cindy Billingham, Dr. Aiman Alzetani and Dr. Ashley Martine, for their fruitful Monday discussions and suggestions about the sample collection, experiment design and data analysis. At the same time, I would like to thank Mrs Sue Johnson for her help and arrangement for my study here. I am grateful to all the members in the Cancer Studies division for their help and assistance in many aspects.

I would like to thank my husband, Dechao, and my daughter, Weijie, for their understanding and encouragement; otherwise this thesis would not have happened.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDIXES

ABBREVIATIONS

AFP	α -Fetoprotein
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under ROC Curve
CA125	Cancer Antigen 125
CA15-3	Cancer Antigen 15-3
CA19-9	Cancer Antigen 19-9
CA27-29	Cancer Antigen 27-29
CC	Correlation Coefficient
CDF	Cumulative distribution function
CEA	Carcinoembryonic Antigen
COV	Covariance
CT	Computed Tomography
CV	Coefficient of Variation
DNA	Deoxyribonucleic acid
DT	Decision Tree
Da	Dalton
EAM	Energy Absorbing Molecule
FDR	False Discovery Rate
FFT	Fast Fourier Transform
FN	False Negative
FOM	Figure Of Merit
FP	False Positive
FWER	Family Wise Error Rate
HCC	Hepatocellular Carcinoma
hCG	Human Chroriogonadotropin
IQR	Inter-Quartile Range
kNN	k Nearest Neighbour
LC-MS	Liquid Chromatography Mass Spectrometry
LDI	Laser Desorption Ionisation
m/z	Mass/charge Ratio
MALDI-TOF MS	Matrix-assisted Laser Desorption Ionisation Time-of-flight Mass Spectrometry
MIC	Median Iron Current
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
NN	Neural Network
OOB	Out-of-bag
PABS	Peak Alignment by Beam Search
PABSD	Peak Alignment by Beam Search With Dynamic Maximum Sideway Movement Value
PACS	Peak Alignment by Cubic Splines

PAFFT	Peak Alignment by Fast Fourier Transform
PAGE	Polyacrylamide Gel Electrophoresis
PAPM	Peak Alignment by Peak Matching
PASS	Peak Alignment by Sequential Search
PCA	Principal Component Analysis
PSA	Prostate Specific Antigen
RF	Random Forest
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SEAC	Surface Enhanced Affinity Capture
SELDI-TOF-MS	Surface Enhanced Laser Desorption Ionisation Time-of-flight Mass Spectrometry
SEND	Surface Enhanced Neat Desorption
SVM	Support Vector Machine
SWA	Segment Wise Alignment
TIC	Total Iron Current
TN	True Negative
TOF	Time-of-flight
TP	True Positive
UK	United Kingdom
UV	Ultraviolet

1 INTRODUCTION

1.1 Strategy to combat cancer

The efforts to combat cancer have not been very successful. Firstly, although we have seen advances in molecular medicine, genomics, proteomics, and translational research, the mortality rates for the most common cancers have not been significantly reduced (Diamandis,2004b). One in four of all deaths in the United Kingdom (UK) are caused by cancer. There were 153,491 cancer deaths in the UK in 2005 (Cancer Research UK,2007; Gavin,2007; General Register Office for Scotland,2007; Office for National Statistics,2007). Cancer accounted for 13% (7.6 million out of 58 million) of deaths worldwide in 2005. Currently, more than 11 million people are detected with cancer each year. Cancers are still the main cause of morbidity and mortality (Alaoui-Jamali & Xu,2006; Cancerbackup,2007; Semmes et al.,2005). It is estimated that there will be 16 million new cases each year by 2020, 9 million people dying from cancer in 2015, and 11 million dying in 2030 (Cho,2007).

Secondly, the progress in the understanding of cancer progression has been very slow and frustrating due to the complex multifactorial nature and heterogeneity of the cancer syndrome (Alaoui-Jamali & Xu,2006). Although progress in bio-technology has made it possible to identify genetic patterns that can be used to diagnose cancer and to predict cancer progression and treatment response (Assikis et al.,2004; Balmain et al.,2003; Curtis et al.,2001; Forrest et al.,2005; Gonzalez et al.,2007; Hoheisel,2006; Listgarten et al.,2004; Miyaki et al.,2004; Risch,2000; Rise et al.,2004; Rodin & Boerwinkle,2005; Scharpf et al.,2007; Staunton et al.,2001; Tomita et al.,2004; Unneberg et al.,2005; van de Vijver et al.,2002; Yoon et al.,2003;

Ziauddin & Sabatini,2001), the complexity and diversity of cancer make it difficult to interpret the genetic patterns.

Some of the best available options to combat cancer include primary prevention, earlier diagnosis, and improved therapeutic interventions. We have been witnessing the development of new drugs against cancer that are based on rational instead of empirical designs. There is hope that some of these drugs will prove to be more effective in the clinic than older generations of medicines (Diamandis,2004b).

Many of the most prevalent human cancers can be prevented to a significant extent through medical interventions or life-style changes. For example, smoking control is very successful in prevention of lung cancer. Dietary changes can reduce the risk of developing large bowel cancer (Henderson et al.,1991). However, the mechanisms of cancer initiation and progression are still not well understood (Cavalierie & Rogan,2006; Diamandis,2004b; Yu,2002). Fortunately, progress in bio-technology provides an opportunity to identify biomarkers, which are substances used as indicators of biologic states (Biomarkers Definitions Working Group,2001; caBIG,2007; Dalmasso,2008). A combination of biotechnology and bioinformatics are indispensable tools in detecting biomarkers that may be useful for earlier detection of cancer. There is evidence that the survival rate exceeds 85% when the common cancers are diagnosed at early stage and are organ confined (Etzioni et al.,2003; Gloeckler Ries et al.,2003; Semmes et al.,2005). Clearly, there is an urgent need to unravel novel biomarkers for early detection of cancer.

1.2 Biomarker discovery in cancer research

Identification of cancer biomarkers is one of the most rapidly advancing fields in clinical diagnostics. Cancer biomarkers can be used to screen asymptomatic individuals in the population, assist diagnosis in suspected cases, monitor treatment efficacy and predict treatment response (Diamandis,2002; Editorial,2004; Ludwig & Weinstein,2005). A few clinically approved biomarkers (Table 1-1) are available for early detection or for successful monitoring of treatment and relapses. Early cancer detection is critical in cancer control. Many screening approaches have been studied in solid cancers[1]. Established screening tools for the early detection of cancer include mammography for breast cancer, colonoscopy for colorectal cancer, prostate specific antigen (PSA) test for prostate cancer, and pap smear for cervix cancer.

However, none of the biomarkers listed in Table 1-1 has adequate sensitivity, specificity, and predictive value for population screening (Diamandis,2004b). PSA is produced by normal prostate cells in small amounts. The high amount of PSA in serum may correlate to the existence of prostate cancer. However, there are reasons other than cancer that cause rise in PSA, such as infections within the prostate gland, and increased exercise with irritation of the affected area (PacificLife,2007; Zimmermann,2007). Cancer antigen 125 (CA-125) can be a biomarker of ovarian cancer risk or an indicator of malignancy. However, level of this biomarker can be high in subjects who have pancreatitis, liver or kidney disease (PacificLife,2007). Imaging techniques such as chest X-ray and spiral computed tomography (CT) are also used, but are limited to a tumour size detection limit of 0.5-1.0 cm (representing close to 10^9 cells).

[1] Solid cancers are defined as abnormal cellular growths in "solid" organs such as the breast or prostate, as opposed to leukaemia, a cancer affecting the blood, which is liquid.

Table 1-1 Common cancer biomarkers used in primary care

Biomarkers	Cancer type	Clinical use	Source
AFP	Hepatoma, testicular	Staging	(Alaoui-Jamali & Xu,2006; Diamandis,2004b)
CA125	Ovarian, cervical, uterine, fallopian tube	Disease monitoring	(Alaoui-Jamali & Xu,2006; Diamandis,2004b)
CA15-3	Breast	Disease monitoring	(Alaoui-Jamali & Xu,2006; Diamandis,2004b)
CA19-9	Gastrointestinal, pancreatic, stomach	Disease monitoring	(Alaoui-Jamali & Xu,2006; Diamandis,2004b)
CA27-29	Breast	Disease monitoring	(Alaoui-Jamali & Xu,2006)
CEA	Colorectal, breast, lung, pancreatic, medullary thyroid	Disease monitoring	(Alaoui-Jamali & Xu,2006; Hutchens & Yip,1993)
hCG	Testicular cancer, trophoblastic tumours testis, ovary	Screening	(Diamandis,2004b)
Epidermal growth Factor receptor	Colon, non-small cell lung cancer	Selection of therapy	(Alaoui-Jamali & Xu,2006)
Her2/Neu	Breast, ovarian	Disease monitoring, selection of therapy	(Alaoui-Jamali & Xu,2006)
Human chronic gonadotropin-β	Testicular, ovarian	Staging	(Alaoui-Jamali & Xu,2006)
Immunoglobulins	B cell dyscrasias	Diagnosis Disease monitoring	(Diamandis,2004b)
PSA	Prostate	Screening, disease monitoring	(Alaoui-Jamali & Xu,2006; Diamandis,2004b)
Steroid hormone receptors	Breast	Screening	(Diamandis,2004b)
Thyroglobulin	Thyroid	Disease monitoring	(Alaoui-Jamali & Xu,2006)

The targeting of some biomarkers have contributed to reduced mortality rates and increased overall survival for cancers (Alaoui-Jamali & Xu,2006; Diamandis,2004b; Hutchens & Yip,1993). For example, carcinoembryonic antigen (CEA) has been identified as a biomarker for colorectal cancer, pancreas cancer, lung cancer, and breast cancer. It has been used for monitoring disease. Human chronic gonadotropin-β has been identified as a biomarker for ovarian cancer and has a clinical use in cancer staging. Epidermal growth factor receptor has been identified as a biomarker for colorectal cancer and non-small cell lung cancer. It has been used in helping determine an appropriate therapy.

Serum biomarkers are produced by body organs or tumours. They can be suggestive of tumour activity when detected in high amounts in blood (PacificLife,2007). Several serum markers have been identified through the years but, with a few exceptions such as PSA for prostate cancers and alpha-fetoprotein (AFP) for hepatocellular carcinomas and germ cell tumour, most have failed general integration into routine clinical practice. It is therefore important to devise new methods that may provide sensitive and reliable diagnostic biomarkers for solid cancers (Finne et al.,2001; Menon & Jacobs,2002; Stephan et al.,2003; Tibshirani et al.,2004). The absence of selective biomarkers may hamper efforts to improve early detection and therapeutic management (Alaoui-Jamali & Xu,2006).

1.3 Potential biomarker discovery for cancer research

The advanced technologies in microarrays and mass spectrometry, and the completion of the sequencing of the human genome have sparked new interest in the area of cancer biomarkers. The microarray and mass spectrometry assays allow thousands of measurements to be carried out in short periods of time (Diamandis,2002; Negm et al.,2002). The sequencing of the human genome can provide fundamental structural information about human genes (Diamandis,2002; Editorial,2004).

A potential biomarker can be a fragment of deoxyribonucleic acid (DNA) sequence, ribonucleic acid (RNA), protein or peptide. These correspond to the different stages of information flow from genotype to phenotype. The DNA in human cells contains long chains of four chemical building blocks, namely adenine (A), thymine (T), cytosine (C), and guanine (G). A human cell has more than 3 billion of these chemical bases, strung together in 23 pairs of chromosomes (SNP Fact Sheet,2009). Single nucleotide polymorphisms (SNPs) are the most frequent form of DNA variation present in the human genome. Because of their abundance, even spacing, and stability across the genome – they are estimated to occur at one out of every 100-300 bases (SNP Fact Sheet,2009) – SNPs are considered to be excellent genetic markers and offer significant potential for cancer genetic research (HapMap,2007b; HapMap,2007a; Risch & Merikangas,1996; Schork et al.,1998). It is estimated that there are about 10-30 million of SNPs in human genome (SNP Fact Sheet,2009). Detection of genetic SNP variants that underlie cancer means sifting through hundreds of thousands of SNPs to identify a relevant subset of markers that could be further examined. The identification of SNPs associated with cancer from a huge amount of SNPs can be daunting (Breiman & Spector,1992; Bureau et al.,2005; Klein et al.,2005; Lunetta et al.,2004; Maraganore et al.,2005; Ozaki et al.,2002; Raychaudhuri et al.,2001; Toivonen et al.,2000) and may lead to computational problems and difficulties in interpreting results. However, many recent successes relating SNPs to disease have been achieved in genome-wide association studies. For example, two correlated common variants on chromosome 8q24 associated with prostate cancer in European and African populations have benn identified (Amundadottir et al.,2006; Gudmundsson et al.,2007). A locus on chromosome 8q24 close to SNPs rs10505477 and rs6983267 is thought to be associated with colorectal cancer susceptibility (Zanke et al.,2007). Five independent SNPs that exhibited strong and consistent evidence of association with breast cancer have been reported (Easton et al.,2007). The Wellcome Trust Case Control Consortium has identified 24 independent association signals for bipolar disorder, coronary artery disease, Crohn's disease, rheumatoid arthritis, type 1 diabetes and type 2 diabetes (The Wellcome Trust Case Control Consortium,2007).

Microarray technology has enabled us to quantify the transcript levels of tens of thousands of genes simultaneously and is another powerful approach towards understanding genome function (Chu et al.,2005; Kelly & Ghosh,2005; Orchekowski et al.,2005; Simon et al.,2003; Spellman et al.,1998). The human genome has been estimated to contain 20,000-45,000 genes (Cho,2007; Human Genome Project,2007). Microarray is a rapidly growing scientific field, primarily concerned with the identification of potential biomarkers. It has provided promising results on genetic basis that can help stage cancer, and predict cancer progression and monitor therapy response (Datta & Lara,2006; Datta et al.,2004; Notterman et al.,2001). However, comparative transcriptional profiling alone is unlikely to fully identify biomarkers that are associated with cancer phenotype (Alaoui-Jamali & Xu,2006). The investigation of RNA expression obtained from microarray data may be an indirect way to understand the aetiology of a disease (Datta & Lara,2006).

Although gene studies have received a lot of attention, it is the proteins that perform most life functions and even make up the majority of cellular structures (Human Genome Project,2007). The ultimate element to control all the functions of a cell is the protein. Proteomics is the large-scale study of protein function and expression. It may hold enormous potential for the early detection of cancer (Hartwell et al.,2006;

National Cancer Institute,2007a; Negm et al.,2002; Srinivas et al.,2002). The cancer proteome contains information on perhaps every biological process that takes place in cancer cells, cancer tissue microenvironment and caner cell-host interaction (Alaoui-Jamali & Xu,2006). Proteomic technologies have a great potential to identify biomarkers for cancer diagnosis, to monitor disease progression, and to identify therapeutic targets. The proteome contains not only the intrinsic genetic information of a cell, but also the impact of its immediate environment. A transformation of a healthy cell into a neoplastic cell may cause altered expression profile, differential protein modification and activities, and this in turn may affect cellular function. Therefore, investigation of the cancer proteome could be a starting point in identifying biomarkers (Alaoui-Jamali & Xu,2006; Bensmail & Haoudi,2003; Zolg,2007). It has been estimated that 20,000-45,000 human genes produce approximately 250,000 spliced variants of RNA, which are translated into more than one million proteins as a result of post-translational processing and modification (Cho,2007; Ozier et al.,2003). Therefore, there is a need to develop sound quantitative methods for measuring proteomic changes caused by a disease.

1.4 Proteomic biomarker discovery for cancer research

Current progress in proteomics has been mainly due to developments in mass spectrometry (Aebersold & Goodlett,2001; Alaoui-Jamali & Xu,2006). There are many mass spectrometry technologies, including matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS, henceforth it is called MALDI), surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF MS, it is called SELDI), and liquid chromatography mass spectrometry (LC-MS).

LC-MS is a powerful method for sensitive detection and quantification of proteins and peptides in complex biological fluids like serum (Nyangoma et al.,2007), which combines the physical separation capabilities of liquid chromatography with the mass analysis capabilities of mass spectrometry.

MALDI is an ionisation technique applied to mass spectrometry and was first introduced by Franz Hillenkamp (Karas et al.,1985; Laiko et al.,2000) in 1985. Franz Hillenkamp and colleagues found that amino acid alanine[2] could be ionised more easily if it was mixed with the amino acid tryptophan[3] (also called the matrix) and irradiated with a pulses 266 nm laser. The ionisation was triggered by a laser beam. A chemical matrix was used to protect biomolecules from being destroyed by direct laser beam and to facilitate vaporisation and ionisation. The first commercial MALDI instruments were introduced in the early 1990s (Karas & Bahr,1990). MALDI has been used to profile and quantify individual peptides or proteins from mammalian cells and tissue section (Aebersold & Goodlett,2001; Chaurand & Caprioli,2002; Cho,2007; Kachman et al.,2002; Kasthuri et al.,2006; McCormack et al.,1997; Poon,2007; Stoeckli et al.,2001; Wolters et al.,2001).

Although MALDI has potential for biomarker discovery, it encompasses a complex sample preparation procedure. Furthermore, interpretation of in-source fragmentation mass spectra has a limitation in the low mass region of the mass spectrum where the abundance of intense, low mass matrix ion signals typically dominates the MALDI spectra, making it difficult to identify components. This limits the use of MALDI in

[2] Alanine is an α-amino acid with the chemical formula HO2CCH(NH2)CH3

[3] Tryptophan is an amino acid essential in human nutrition. It is one of the 20 amino acids encoded by the genetic code.

profiling complex protein mixtures. Prefractionation of protein mixtures using chromatographic approaches has been used to address this limitation. SELDI integrates MALDI and chromatography technology and was first introduced by Hutchens & Yip (1993).

SELDI may succeed in discovering new diagnostic modalities for cancers if the methods are used properly (Kroczak et al.,2006; Kuwata et al.,1998; Merchant & Weinberger,2000; Plebani,2005). Compared to MALDI, SELDI is different in the construction of the sample targets, the design of the analyser and the software used to interpret the acquired data (Vorderwülbecke et al.,2007). SELDI employs 8- or 16-spot chip for analysing 8-16 protein samples per chip. Each spot contains a solid-phase chromatographic surface for binding proteins at a particular binding condition. After washing and adding the matrix chemical, the retained proteins or peptides are charged and detected as peaks in mass spectra (Poon,2007). The analyser used for SELDI is especially adapted to achieve high-sensitivity quantification. The ion source and detector are constructed to support efficient ion transmission and ion detection over a wide mass range. The precise positioning of the laser beam is controlled by software in automatic or manual mode. Software used in SELDI allows normalisation of the resulting spectra to their total ion current (TIC) for internal quantitative calibration. SELDI is discussed in more details in Chapter 2. In summary, by combining selective protein binding with quantitative mass detection, SELDI enables the comparative analysis of proteins in a fast and simple process (Vorderwülbecke et al.,2007). Advancements in SELDI may lead to the identification of potential serum/plasma biomarkers at ng/ml or even lower levels (Poon,2007). However, many researchers believe that SELDI has inferior performance. The resolution of SELDI is usually quite low, and the ion peaks of interest cannot be directly identified (Petricoin et al.,2006). Another disadvantage of SELDI is that protein sequences, and thus specific identifications, are not obtained, requiring further biochemical/mass spectrometry analysis to identify differentially-expressed proteins (Miyamae et al.,2005).

Opportunities also mean challenges. Unlike genomic studies, proteomic studies have technical challenges to acquire proteomic data. Other challenges include lacking of standardised methodologies, unknown reproducibility, and intra- or inter-individual cancer heterogeneity (Alaoui-Jamali & Xu,2006). As far as we know, there is no evidence to show that the results obtained with SELDI technique are robust and reproducible among laboratories. Before the technique can be used for clinical purposes, the method needs to be internally and externally validated.

This thesis will investigate the quantitative methods used for biomarker discovery for cancer diagnosis using SELDI. We focus on the issue of reproducibility assessment and study the influence of important pre-processing steps, including spectrum baseline correction, normalisation, transformation, and peak alignment on biomarker discovery and cancer classification. We propose three methods to assess the reproducibility of SELDI protein profiles. Two improvements to the current peak alignment methods, suitable for SELDI mass spectra alignment are proposed. We also propose one new normalisation algorithm for pre-processing SELDI data, and an integrated algorithm for biomarker discovery using SELDI.

1.5 Objectives of the thesis

The objectives of the thesis are to investigate quantitative methods used in proteomic biomarker discovery for cancer diagnosis. In particular, we focus on the following issues.

1. Review the use of SELDI technology in cancer diagnosis and statistical methods for SELDI data analysis.
2. Investigate the methods for assessing reproducibility of SELDI data.
3. Study the methods for pre-processing SELDI data, including spectrum normalisation, transformation, and peak alignment.
4. Study the normality of SELDI data for selecting appropriate methods to identify biomarkers from SELDI data.
5. Investigate the methods for feature selection and sample classification and propose an integrated framework for identifying biomarkers using SELDI data.

1.6 Layout of the thesis

The rest of this chapter is devoted to providing a context for the thesis. Chapter 2 reviews SELDI technology and the statistical methods for SELDI data analysis. A general procedure of applying SELDI to cancer biomarker discovery and cancer classification is presented and discussed, which include issues on experiment designs, reproducibility analysis, protein profile pre-processing, identification of putative biomarkers, and cancer classification. The problems in analysing SELDI protein profiles are identified, and the solutions for these problems are outlined.

Chapter 3 proposes three quantitative reproducibility measures using Euclidean distance, correlation coefficient, and the paired t-test to assess the reproducibility of SELDI protein profiles. The proposed methods are applied to SELDI mass spectra generated from identical samples in replicate experimental runs for assessing the reproducibility of SELDI protein profiles.

Chapter 4 proposes two new peak alignment algorithms. The performance of the peak alignment methods are assessed and compared with other existing peak alignment methods. Classifiers are also built to evaluate the effect of peak alignment on cancer classification. The prediction accuracy of the classifiers that are trained using mass spectra with and without peak alignment is estimated and compared.

Chapter 5 examines the assumption of the normality of SELDI protein profiles, on which standard statistical methods (such as t-tests) used for differential expression detection are based. Four statistical tests of normality are applied to investigate the normality of SELDI protein profiles. The effects of normalisation and data transformation on normality of SELDI data are investigated. Our proposed normalisation algorithm is compared with the standard methods. The cancer classification accuracies using SELDI data with and without logarithmic transformation are also evaluated.

Chapter 6 proposes an integrated algorithm for biomarker discovery for early detection of cancer using SELDI expression profiles, which consists of two major components: feature selection methods and machine learning techniques. Both univariate feature selection and multivariate feature selection approaches are assessed. Two published SELDI data sets have been used to demonstrate how the algorithm works.

Conclusions and suggestions of possible future work in this area are given in Chapter 7.

Finally, the programs used in this thesis are listed in appendixes.

1.7 Summary

Continuous efforts have been made to combat cancer for many years. Cancers are still the main cause of morbidity and mortality. Cancer initiation and progression have been largely unknown due to the complexity and diversity of cancer genome. Although cancer biomarkers are not very effective at screening the population and diagnosing cancer early, they are valuable in monitoring patients during and after therapy. There has been evidence to show that the availability of a few biomarkers have significantly contributed to reduced mortality rates and increased survival of cancers. Therefore, there is an urgent need to discover biomarkers for early cancer detection and for selection of the most effective treatment from currently available therapeutic agents.

Proteome contains both the intrinsic genetic information of cells and the impact of their immediate environments. Therefore, investigating the cancer proteome could be a starting point in cancer biomarker discovery. The current progress in proteomics has been mainly due to the rapid development in mass spectrometry. The next chapters will be devoted to investigate SELDI mass spectra and its applications in biomarker discovery for early cancer diagnosis.

2 SELDI TECHNOLOGY AND DATA ANALYSIS

2.1 Introduction

SELDI is a recently developed technology for proteomic research. It couples chromatographic separation of samples using reactive surfaces and high-throughput mass spectrometry analyses and provides a way to study protein profiles over a wide range of molecular weights (0-200kDa) in small biological specimens, such as serum. It has great potential for biomarker discovery and for protein-protein interaction studies. In this chapter, the SELDI technology and its applications will be reviewed. One approach to biomarker discovery is discussed. This includes the following components: experiment design, reproducibility analysis, mass spectrum pre-processing, and biomarker identification. The problems in each component are discussed. The solutions for these problems are outlined and to be addressed in the subsequent chapters.

2.2 SELDI technology

SELDI technology encompasses two major subsets of MS technology: surface-enhanced neat desorption (SEND) and surface-enhanced affinity-capture (SEAC) (Hutchens & Yip,1993; Poon,2007). Figure 2-1 outlines the SELDI technology. It consists of three major components: the ProteinChip Array, the ProteinChip Reader, and the ProteinChip Software.

The ProteinChip array distinguishes this technology from other mass spectrometry-based analytical systems and has a 10-mm wide x 80-mm long chip with eight or sixteen 2-mm spots comprised of a specific chromatographic surface. There are two major types of surfaces (Figure 2-2): chemical surfaces and biochemical surfaces. The chemically treated surfaces include anionic, cationic, hydrophobic, hydrophilic, immobilised-metal affinity metal. The pre-activated biochemical surfaces are available for covalently coupling antibody, receptor, DNA, enzyme, and so on.

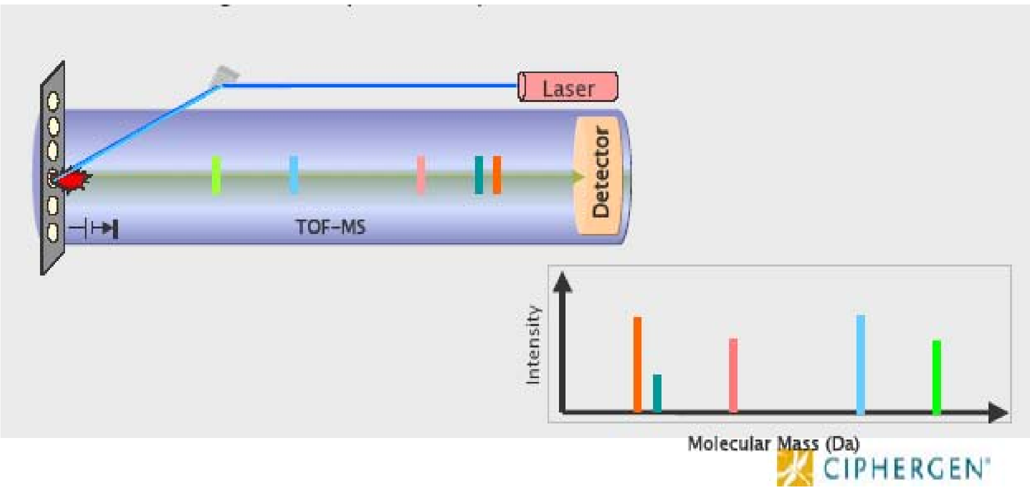
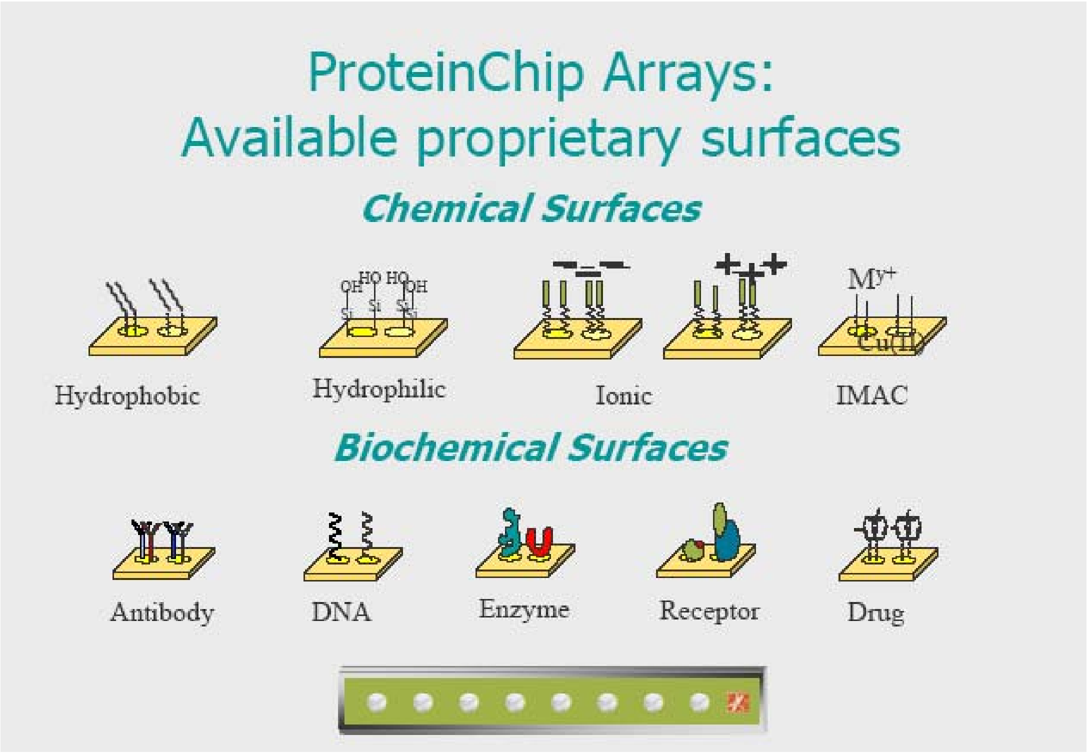


Figure 2-1 Diagram of SELDI technology

In the general application, a small volume of protein sample, such as biological fluid, is taken from people with or without cancer and applied on the ProteinChip array. A diluted sample is added to the pre-processed ProteinChip array and incubated (sample will be resolved into a subset of proteins with common properties). The ProteinChip is then washed to remove unbounded components, air-dried, and crystallised with an energy absorbing molecule (EMA) called “matrix”. The ProteinChip array usually binds a subset of proteins in the serum sample and the mass spectrum is recorded by the ProteinChip Reader.


ProteinChip Arrays:
Available proprietary surfaces

Chemical Surfaces

Biochemical Surfaces

Figure 2-2 ProteinChip Arrays.

The ProteinChip Reader is a laser desorption ionisation (LDI) time-of-flight (TOF) MS instrument equipped with a pulsed ultraviolet (UV) nitrogen laser source that uses state-of-the-art ion optic and laser optic technology. The bound proteins are hit with a laser in the ProteinChip Reader, causing the proteins to desorb and ionise when the matrix absorbs the energy produced at the wavelength of the nitrogen laser. This produces ionised protein molecules in the gas phase. A brief electric field is then applied to accelerate the ions down a flight tube, and a detector at the end of the tube records the time of flight. The small proteins fly faster and the large proteins fly more slowly. Therefore, the position of an individual protein in spectrum corresponds to its time-of-flight. The detected proteins are reported as a series of peaks (0-200kDa), with the mass/charge ratio (m/z) values of proteins and their corresponding intensities. The resulting data may be displayed in a two-dimensional plane, with m/z values displayed in the x-axis and the corresponding intensities (ion currents) on the y-axis (shown in the bottom part of Figure 2-1). The laser optics maximise ion extraction efficiency over the greatest possible sample area, and thus increases analytical sensitivity and reproducibility. The ProteinChip Reader's ion optics incorporates a four-stage, time-lag-focusing[4] ion lens assembly that provides precise, accurate molecular weight determination. Time-lag-focusing is used to increase the mass accuracy of the final output. Signal processing is accomplished by high-speed analog-to-digital converter, which is linked to a personal computer (Isaaq et al.,2002).

The ProteinChip Software controls all aspects of the ProteinChip Reader and facilitates data collection and processing. It has the function of loading ProteinChip array into the reader, calibrating the reader, analysing samples, normalising mass spectra, and presenting the spectra in the user-friendly manner. The quadratic equation is employed to convert TOF and its velocity through an ion chamber into m/z.

$$\frac{m/z}{Voltage} = a(TOF - t_0)^2 + b \qquad \textbf{(2.1)}$$

where, the parameters *a*, *b*, and t_0 can be estimated for known peptides/proteins. The equation generated by this process can then be used on mass spectra that are collected under the same instrument and conditions.

[4] Time-lag-focusing is an experimental technique in time-of-flight mass spectrometry in which improved mass resolution is obtained by using a controlled time delay between the initial pulse of ion formation and acceleration of the ions into the flight tube of the instrument.

2.3 Applications of SELDI technology in cancer studies

SELDI is currently one of the most rapid and sensitive proteomics analysis tools available. With further improvements in resolution and reproducibility, the unlimited type of biological material, less complicated protocol and the unique surface chemistries of the arrays make SELDI technology an attractive approach to protein profiling (Alterovitz et al.,2004; Austen et al.,2000; Cardone et al.,1998; Chernyak et al.,2001; Forde et al.,2002; Hampel et al.,2001; Henderson & Steele,2005; Hinshelwood et al.,1999; Jock et al.,2004; Li et al.,2002; Pawlik et al.,2006; Petricoin et al.,2002a; Petricoin & Liotta,2004; Poon et al.,2003; Reddy & Dalmasso,2003; Stoica et al.,2001; Tong et al.,2004; Yu et al.,2004; Zhang et al.,2004a; Zhao et al.,2004).

This versatile instrumentation is currently being used in projects ranging from the identification of disease biomarkers to the study of bio-molecular interactions. SELDI provides a means of evaluating the intensity of many proteins at the same time. The ProteinChip allows for protein profiling from a variety of complex biological materials such as serum, blood, plasma, intestinal fluid, urine, cell lysates, and cellular secretion products with less complicated sample preparation protocol. Some studies have used SELDI to characterise protein-protein interactions (Hinshelwood et al.,1999; Stoica et al.,2001), phosphorylation site mapping and glycoproteins characterisation (Cardone et al.,1998; Chernyak et al.,2001), and transcription factors (Forde et al.,2002). Others used SELDI to profile low molecular weight peptides (Sato et al.,2001) and proteins secreted by different cancer cell lines grown in serum free medium (Isaaq et al.,2002). SELDI has also been utilised to discover biomarkers for different types of cancers, such as ovarian (Petricoin et al.,2002a), bladder cancer (Zhang et al.,2004a), colon cancer (Yu et al.,2004; Zhao et al.,2004), hepatocellular carcinoma (HCC) (Poon et al.,2003), prostate cancer (Tong et al.,2004), breast cancer (Li et al.,2002), and renal cancer (Hampel et al.,2001; Rogers et al.,2003).

2.3.1 The use of SELDI in biomarker discovery

The concept of finding proteomic differences between cancer patients and normal subjects was based on an assumption that protein or protein fragments produced by cancer cells or their microenvironment may eventually enter into the general circulatory system. The biological fluid of interest is first interacted with a protein chip that incorporates some kind of affinity separation between "non-informative" and "informative" proteins. After washing, the informative proteins can be studied using SELDI. The biomarkers may be identified from the mass spectra of these protein or protein fragments and have a potential in early cancer diagnosis (Hutchens & Yip,1993). Table 2-1 shows the putative cancer biomarkers that have been identified using SELDI technology.

Many laboratories have demonstrated the feasibility of using mass spectrometry proteomic pattern analysis in the discovery of potential diagnostic markers and cancer classification (Adam et al.,2002; Cazares et al.,2002; Kozak et al.,2003; Li et al.,2002; Pang et al.,2006; Paweletz et al.,2001; Petricoin et al.,2002a; Petricoin et al.,2002b; Poon et al.,2003; Qu et al.,2002; Rosty et al.,2002; Vlahou et al.,2003b; Vlahou et al.,2001; Vlahou et al.,2003a; Vlahou et al.,2003c; Wadsworth et al.,2004; Won et al.,2003; Wulfkuhle et al.,2001; Xiao et al.,2003; Zhukov et al.,2003). The reported cancer classification accuracies are relatively high, with sensitivities varying from 83% to 100%, and specificities varying from 82% to 98% (Li et al.,2002; Petricoin et al.,2002a; Poon et al.,2003; Rogers et al.,2003; Tong et al.,2004; Yu et al.,2004; Zhang et al.,2004a; Zhao et al.,2004). These results show the potential of SELDI technology to discover significant patterns that may permit discrimination of cancer from normal samples.

Table 2-1 Putative cancer biomarkers identified using SELDI technology

Biomarker	Cancer type	Source
Apolipoprotein A1	Ovarian, pancreatic	(Alaoui-Jamali & Xu,2006; Kozak et al.,2005; Kozak et al.,2003; Zhang et al.,2004b)
a1-antitrypsin and a1-antichymotrypsin	Pancreatic	(Alaoui-Jamali & Xu,2006; Orchekowski et al.,2005; Yu et al.,2005)
Cytosolic ubiquitin	Breast cancer	(Cho,2007)
Ferritin light chain	Breast cancer	(Cho,2007)
Heptaglobin a-subunit	Ovarian, pancreatic, lung	(Alaoui-Jamali & Xu,2006; Ye et al.,2003)
Inter-alpha-trypsin inhibitor fragment	Ovarian, pancreatic	(Alaoui-Jamali & Xu,2006; Zhang et al.,2004b)
Osteopontin	Ovarian, prostate	(Alaoui-Jamali & Xu,2006; Khodavirdi et al.,2006)
Serum amyloid A	Nasopharyngeal, pancreatic, ovarian	(Alaoui-Jamali & Xu,2006; Moshkovskii et al.,2005; Orchekowski et al.,2005)
Transthyretin fragment	Ovarian	(Alaoui-Jamali & Xu,2006; Kozak et al.,2005)
Vitamin D-binding protein	Prostate, breast	(Alaoui-Jamali & Xu,2006; Corder et al.,1993; Pawlik et al.,2006)

Although SELDI technology has potential use in biomarker discovery and early detection of cancer, it also has many limitations. It is not clear whether the identified discriminant features were due to the inherent biological differences associated with cancer or to artefacts associated with the SELDI technology. The study by Diamandis (2004b) raised some open questions related to cancer diagnosis using SELDI technology, which include (a) whether the SELDI technologies are reproducible (Diamandis,2004a; Diamandis,2003b; Semmes et al.,2005), and (b) why the validated serum cancer markers (e.g. PSA, CA125, etc.) that could serve as positive controls have not been identified by SELDI technology. Several studies

(Adam et al.,2002; Banez et al.,2003; Diamandis,2003b; Gao et al.,2003; Petricoin et al.,2002b) concerned whether the use of SELDI can improve the early detection of cancer. The studies by Grizzle et al. (2003) and Semmes et al. (2005) suggested that these concerns should be addressed in an appropriately designed validation study.

The paper by Poon (2007) summarised the major limitations of SELDI in biomarker discovery, which include (a) conventional tumour markers (e.g. α-fetoprotein) have not been identified by SELDI, (b) The protein identities of the SELDI peaks cannot be obtained directly in the SELDI experiments, (c) The limited detection dynamic range causes difficulty in the identification of potential diagnostic proteins present at concentration below the μg/ml level in serum/plasma, (d) High susceptibility to the identification of false-significant biomarkers that are caused by systematic bias, (e) Quantitative proteomic profile is sensitive to small changes of experimental procedures and analytic variables, (f) Poor resolution for analysis of large proteins, particularly when resolving proteins with slight differences in post-translational modification (>20k Da), and (g) Lack of consensus assay protocols or standard operating procedures for SELDI experiments, leading to great variations in quantitative patterns and in the quality of the proteomic profiles across laboratories, and causing difficulty in cross-laboratory validation. These questions emphasise the need to evaluate the existing quantitative biomarkers discovery tools with the aim of deriving the optimal approach to biomarkers detection.

2.4 Methods for SELDI data analysis

In SELDI, after washing and adding matrix chemicals, the retained proteins will be charged and recorded as mass spectra with intensities at different m/z values. Cancer biomarkers are proteins or peptides that are differentially expressed across disease conditions, which can be identified by comparing the intensity values between the cancer and the control groups. However, comparisons of large-scale multivariate proteomic patterns are subject to many challenging analytic issues, including experimental noise, systematic variations between experimental runs, and the huge number of measured features of which many are uncorrelated to cancer (Listgarten & Emili,2005). This section will review important mass spectrum acquisition and analytic methods that should be considered in comparative proteomic analysis, and outlines the procedures for proteomic biomarker discovery using SELDI as shown in Figure 2-3.

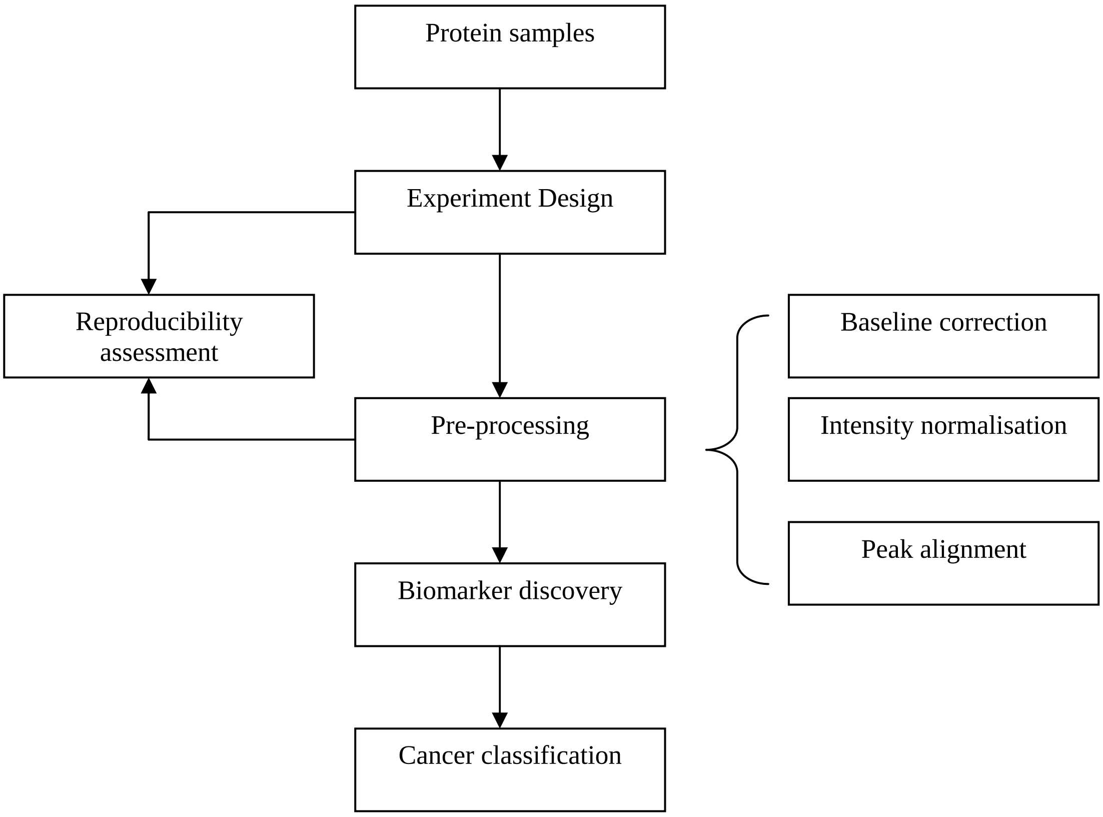


Figure 2-3 One approach for biomarker discovery using SELDI

2.4.1 SELDI experiment design

In SELDI experiments, the serum samples are added to a ProteinChip and resolved into subset of proteins with common properties. After washing, only a subset of the proteins and peptides present in the original sample is left on the chip to be examined. The subset of protein present in the final analysis depends on many factors, including the method used to extract proteins, the chip type used, the incubation and wash conditions, the type of matrix used and method of application, and the laser intensity settings.

Conducting preliminary experiments may be required to identify the best combination of the experimental conditions. The study by Cordingley et al. (2003) employed factorial experimental design to investigate factors which could affect mass spectrum quality and reproducibility during experiment process. The following factors were identified: (a) Urea/thiourea ratio in extraction buffer, (b) protease inhibitor, (c) time between crushing and homogenisation, (d) concentration and volume of samples, (e) length of washes, (f) number of washes, (g) time between last wash and addition of matrix, (h) age of matrix, (i) pipet type used for matrix deposition, and (j) number and volume of matrix depositions. The criteria defined to judge the performance of the mass spectra included the number of peaks detected, peak cleanness (i.e., sharp peaks that were well separated from each other, with smooth lines, with no secondary underlying peaks), and more reproducible traces (Cordingley et al.,2003).

By conducting experiments with different levels of these factors, many SELDI spectra were generated. These spectra were then analysed using a statistical algorithm to find the main effect factors. It was found that six factors, urea/thiourea ratio, protease inhibitor, time between crushing and homogenisation, length of washes, number of washes, and volume of matrix depositions had a significant impact on one or more of the response variables defined above, and that only two factors, urea/thiourea ratio, and volume of matrix depositions, had significant main effects on the number of peaks (Cordingley et al.,2003; Fung & Enderwick,2002).

In order to find structural features of cancer using SELDI, un-biasing the whole process is also extremely important because a small change in the way that the serum is collected between cancer patients and healthy subjects may contribute to artefacts (Petricoin, III & Liotta,2003). Other issues, such as sample handing (e.g. collection, and storage), laboratory environment (e.g. temperature, and humidity), ProteinReader calibration, and inadequate data pre-processing, might be the key factors to cause the systematic bias. It is vital to ensure that the same conditions are used in intra-assays[5] and inter-assays[6], as well as inter-laboratories.

2.4.2 Reproducibility analysis

Concern has recently been expressed as to whether SELDI based methods are reproducible. This is important as the identification of true biomarkers for cancer and the accurate classification of cancer patients from normal individuals largely depend on precision and consistency of measurements. Reproducibility refers to the ability of a SELDI experiment to be accurately reproduced. It measures the closeness of agreement between independent results obtained with the same method on identical

[5] Intra-assays refer to assays within a single assay run

[6] Inter-assays refer to assays over a number of different assay runs in one laboratory

test material but under different conditions, such as different operators, different apparatus, different laboratories, and after different intervals of time (IUPAC,2007).

The studies by Baggerly et al. (2004), Karsan et al. (2005) and Poon (2007) found that without a standardised experimental protocol, it is very difficult to compare the SELDI protein profiles generated from different laboratories, even though the same types of ProteinChip arrays, incubation buffer and washing solution are used. The study by Baggerly et al. (2004) assessed the reproducibility of SELDI protein profile by re-analysing three data sets of SELDI mass spectra derived from the serum of patients with ovarian cancer. The results of the analysis showed poor reproducibility. It was found that the separating feature sets were not reproducible across experiments. The features identified to discriminate well in one experiment did not generalise to other experiments. These results suggests that many of the structural features identified in the initial analyses could be due to artefacts of sampling process, rather than the underlying biological mechanisms of ovarian cancer (Baggerly et al.,2004).

Another example of lack of reproducibility of results from this technology was demonstrated in a re-analysis of three prostate cancer studies (Diamandis,2003a). The study by Diamandis (2003a) assessed the reproducibility of SELDI protein profiles by comparing three prostate cancer studies by Adam et al. (2002), Petricoin et al. (2002b), and Qu et al. (2002). The results showed that none of the peaks selected in the study by Petricoin et al. (2002b) was identified in the studies by either Adam et al. (2002) or Qu et al. (2002). Although the studies by Adam and Qu used the same chip for serum extraction and the same instrument to generate mass spectra, their distinguishing peaks were very different. This implies that the identified structural features may not be prostate cancer-related (Diamandis,2003b).

The reason for poor reproducibility of SELDI expression profiles is probably that the different centres did not adopt any standardised protocol and analysis procedure. The lack of agreement might be due to differential handing, processing of samples, changes in the type of ProteinChip array, mechanical adjustments to the mass spectrometer, or a shift to a different instrument or lab, etc. This indicates a need for careful experimental design, for varying experimental conditions, and for better methods of external calibration (Baggerly et al.,2004).

The study by Semmes et al. (2005) assessed interlaboratory reproducibility of SELDI protein profile by comparing the means, standard deviations, and coefficient of variations[7] of intensity, mass, signal and noise ratio, and resolution of 14 prostate cancer mass spectra and 14 non prostate cancer mass spectra generated from six different centres. The results showed that between laboratory reproducibility of SELDI serum profiling approached that of within-laboratory reproducibility when a standardised experimental protocol and quality control strategy were used (Poon,2007; Semmes et al.,2005). All these results suggest that standardised experimental protocols and analysis procedures are needed for biomarker discovery using SELDI technology.

Compared with the studies by Diamandis (2003a) and Baggerly et al. (2004), the study by Semmes et al. (2005) used a small number of identical serum samples across six different centres and the reproducibility analyses were based on only three

[7] Coefficient of variation is a normalised measure of dispersion of a probability distribution and defined as the ratio of the standard deviation to the mean.

intensity peaks. Although between-laboratory reproducibility of SELDI serum profiling was similar to that of within-laboratory reproducibility, further research on reproducibility of SELDI is needed. In the studies of both Diamandis (2003a) and Baggerly et al. (2004) reproducibility was measured by agreement in separating feature sets. It is not a quantitative measure, making it difficult to identify experimental conditions that may affect reproducibility of SELDI expression profiles. In order to assess the reproducibility of SELDI expression profiles and to facilitate identifying factors that affect the reproducibility of SELDI, it is necessary to develop quantitative reproducibility measures. This thesis revisits the issue of reproducibility and proposes three other methods for quantitatively assessing the reproducibility of SELDI protein profiles using Euclidean distance, correlation coefficient, and the paired t-test. The reproducibility of SELDI protein profiles will be studied in Chapter 3.

2.4.3 Mass spectrum pre-processing

The data generated from SELDI need pre-processing before further analysis. Due to matrix and interference of measurement procedure, chemical and/or electronic noise may be generated, and these may result in imprecise measurements of both m/z values and intensity values. For each SELDI mass spectrum i, we observe protein intensities $I_i(t_j)$ at t_j, $j = 1,2,...,T$. These protein intensities can be modelled by

$$I_i(t_j) = B_i(t_j) + \frac{1}{N_i} S_i(t_j) + \varepsilon_{ij} \qquad \textbf{(2.2)}$$

where, $B_i(t_j)$ is the baseline, representing a systematic artefact in the mass spectrum i, ε_{ij} is thought to originate from chemical and/or electronic noise in the SELDI detector, N_i is a normalisation factor, and $S_i(t_j)$ is the true signal and consists of a set of peaks, corresponding to proteins or peptides (Moriss et al.,2005).

The measurements of m/z values are imprecise by approximately 0.1-0.2% of the m/z value and the coefficient of variation of intensity measures is approximately 15-20% according to the manufacturer's specification of Ciphergen PBS IIc (Ciphergen Biosystems, UK). The variation of these values from spectrum to spectrum, and from experiment to experiment is well known. It is not infrequent that the coefficient of variation for intensity measures can be as high as 50–60% (Hong et al.,2005). Therefore, it is necessary to perform data pre-processing to remove noise and improve the quality of SELDI data before further analysis can be carried out. In fact, pre-processing is also a crucial factor in determining reproducibility of protein profiles from SELDI instruments. Data pre-processing includes baseline correction, intensity normalisation and transformation, peak alignment and peak picking.

Ciphergen ProteinChip Software 3.2 (Ciphergen Biosystems, UK), and PROcess library of BioConductor Project[8] (Bioconductor,2007b) are commonly used mass spectrum pre-processing packages. The study by Beyer et al. (2006) compared the performance of Ciphergen ProteinChip Software and PROcess for pre-processing SELDI mass spectra for rat liver proteome. Comparison of results showed that baseline correction and normalisation algorithms implemented in Ciphergen ProteinChip Software and PROcess had the same underlying concept and gave fairly similar results, but that the results differed after following the peak detection and

[8] An open source and open development software project for the analysis and comprehension of genomic data

alignment procedures. The concepts and methods for pre-processing SELDI data are discussed as follows.

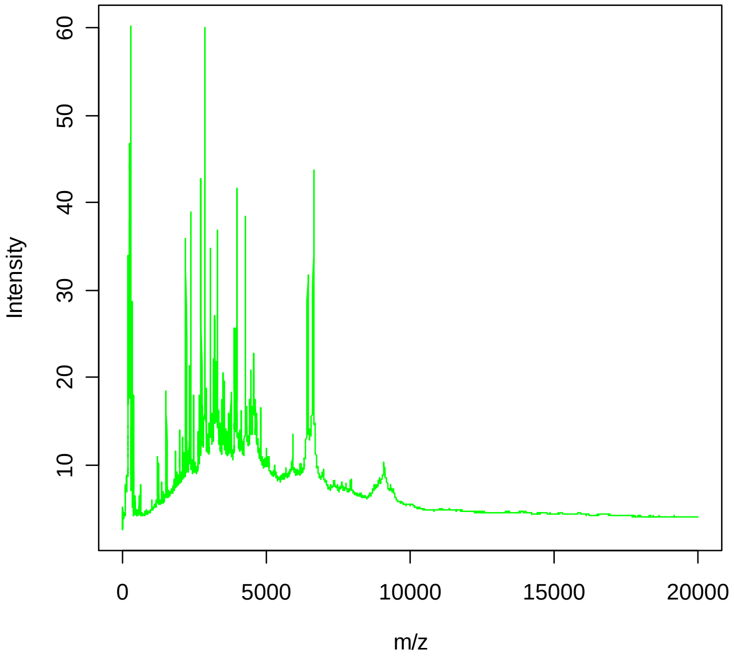


Figure 2-4 A raw spectrum (Gentleman et al.,2005).

2.4.3.1 Baseline correction

Baseline correction aims at removing or reducing the baseline artefact characteristics $B_i(t_j)$ due to matrix and interfering biochemical or physical processes of the measurement procedure in the experiment. The study by Malyarenko et al. (2005) introduced the idea that the source of the baseline is in the ion detector that gets saturated by the matrix molecules at low m/z values. This saturation usually decays slowly over time. Thus baseline is a vertical offset as shown in Figure 2-4, which significantly elevates the intensity of the peaks for lower m/z values between about 2,000 and 10,000 Da. At higher m/z, the baseline levels off to a plateau (Gentleman et al.,2005).

The baseline can be corrected by subtracting from a spectrum an estimate of chemical and/or electronic noise. The algorithm implemented in the Ciphergen ProteinChip Software fits a piece-wise convex-hull[9] that attempts to find the chemical and/or electronic noise and to correct the peak height (Beyer et al.,2006). The PROcess algorithm estimates the chemical and/or electronic noise using locally weighted regression by fitting a curve to the spectrum local minima. The traditional weight function used for local regression is the tri-cube weight function defined as follows.

$W(x) = (1 - |x|^3)^3$ if $|x| < 1$, otherwise 0 **(2.3)**

where, x is the scaled distance between each point to the point of estimation, which is calculated by scaling the distance by the maximum distance over all points in the local minima set.

At each point a polynomial is fitted to a subset of the data using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away. The value of the regression function for the point is then obtained by evaluating the local polynomial using the explanatory variable values for that data point. The spectra are then corrected by subtracting the baseline from the raw spectra (Beyer et al.,2006; Bioconductor,2007b; Gentleman et al.,2005). The implementation of the PROcess baseline correction is summarised as follows (Gentleman et al.,2005).

[9] The convex hull for a set of points X is the minimal convex set containing X

1. Segment the m/z range for each spectrum and find local minima for each interval.
2. Fit a local regression to the local minima for each spectrum.
3. Subtract the estimated baseline from each spectrum.

Figure 2-5 shows the result of a spectrum with baseline removed, where the middle curve is the baseline fitted by the weighted least squares. It can be seen that the process of baseline correction may introduce negative net intensity values. Empirically, using local minima in the regression tends to yield fewer negative values (Gentleman et al.,2005).

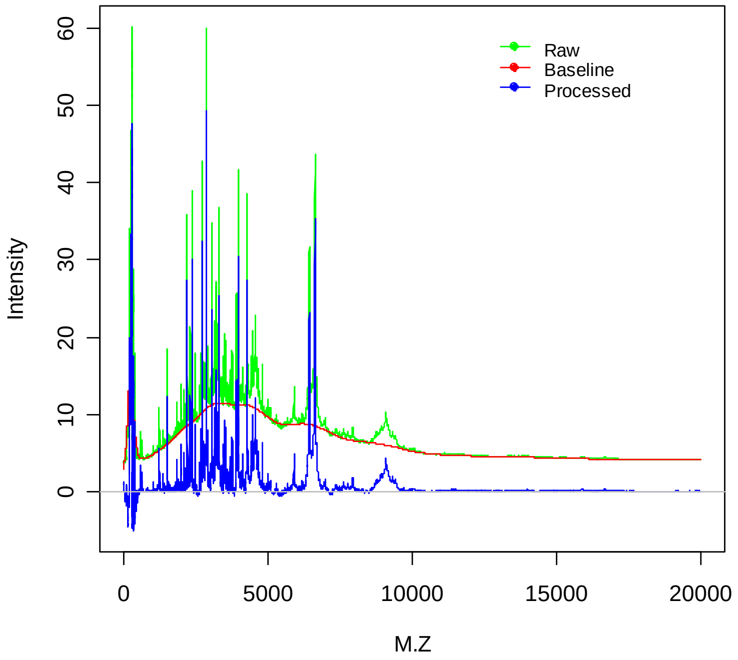


Figure 2-5 Baseline correction using PROcess (Gentleman et al.,2005).

2.4.3.2 Normalisation

Intensity normalisation aims at correcting for systematic differences in the total amount of protein desorbed from the sample plate. Due to variation in sample preparation and heterogeneity of samples on the spots of a chip, the detected intensities are only relative rather than absolute measurements of the proteins in the samples (Callister,2006). Therefore, prior to differential expression analysis, intensities need to be normalised. Normalisation techniques rescale the sample intensities by a normalisation factor N_i and enable the comparison of intensities across different spectra by eliminating the difference in the total amount of protein desorbed (Bolstad et al.,2003).

The commonly used normalisation methods are the global normalisations[10]. The assumption behind the global normalisation is that, on average, the number of proteins that are over expressed is approximately equal to the number of proteins under expressed and the number of proteins whose expression levels change is few relative to the total number of proteins (Sauve & Speed,2003). Such global normalisations assume that the sample intensities are all related by a constant coefficient. A common choice for this normalisation coefficient is the spectrum median or the mean. For mass spectrometric data, each protein concentration is measured by the area under curve (AUC) of its peak.

The software package PROcess in BioConductor (Bioconductor,2007a) uses the median AUC to normalise a set of mass spectra. It consists of the following phases. Firstly, for each spectrum, its AUC for a selected m/z range is calculated. The value

[10] Global normalisation applies a constant scaling factor to every measurement in a specified m/z range.

of AUC is proportional to the sum of the intensities in the selected m/z range if intensities at m/z values are taken at equally spaced time points (Gentleman et al.,2005). Secondly, the median AUC of a set of mass spectra to be normalised is obtained. Thirdly, each spectrum is scaled to the median AUC (Beyer et al.,2006; Bioconductor,2007b; Gentleman et al.,2005).

Ciphergen ProteinChip Software uses the mean AUC (or called TIC) to normalise a set of spectra. It consists of the following phases. Firstly, for each spectrum, the average intensity in a specified m/z range, which equals to TIC divided by the number of data points in the m/z range, is calculated. Secondly, the overall average intensity (mean AUC) across all the spectra to be normalised is calculated. Thirdly, each spectrum is normalised by scaling the intensities by the mean AUC (Beyer et al.,2006).

Figure 2-6 shows three spectra before normalisation with mocked-up trace data from Ciphergen ProteinChip Software 3.2 (Ciphergen Biosystems, UK). Figure 2-7 shows the spectra after normalisation. It can be seen that the systematic differences in the total amount of protein desorbed can be eliminated by the process of normalisation.

We propose a new normalisation algorithm and will compare it with the existing normalisation methods in Chapter 5.

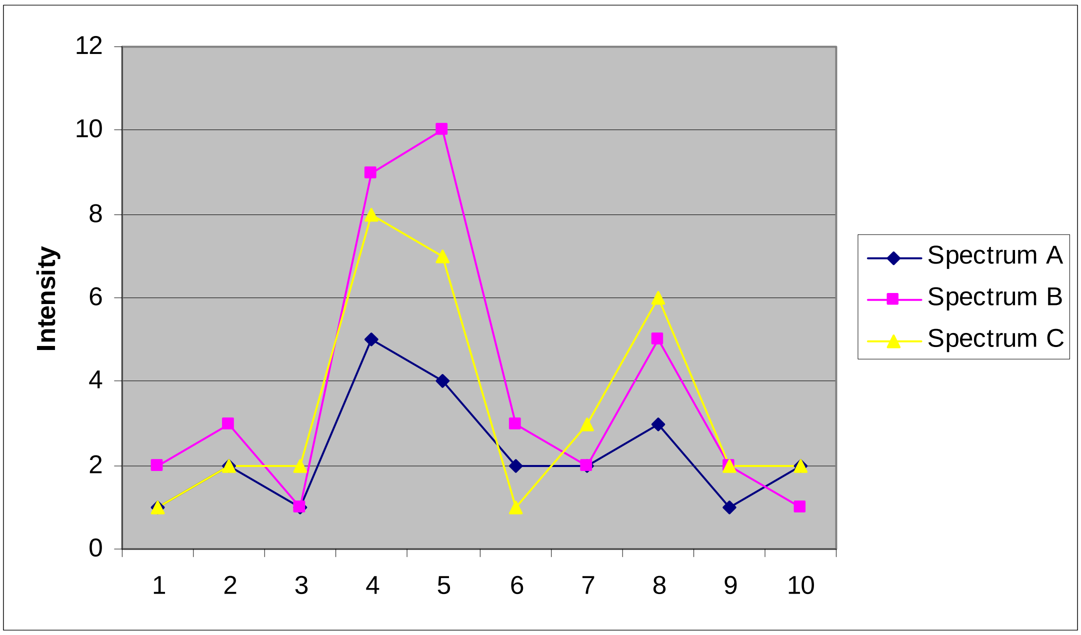


Figure 2-6 Three spectra before normalisation

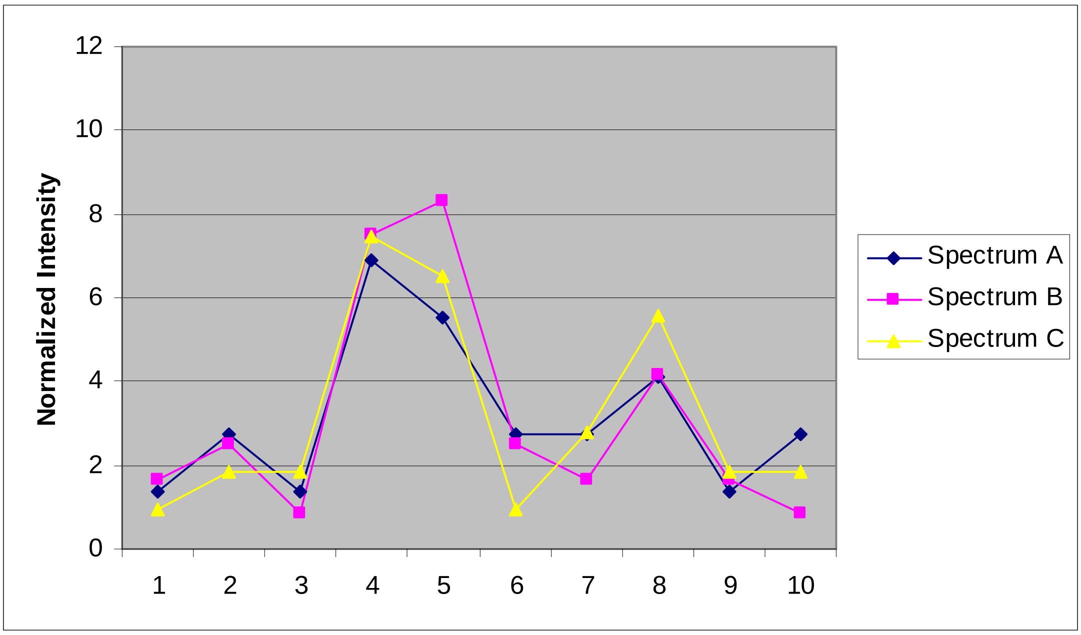


Figure 2-7 Three spectra after normalisation

2.4.3.3 Peak detection

A peak is a local maximum within a spectrum. Peak detection and quantification aim at identifying feature locations of the peaks in the true signal $S_i(t_j)$ across all spectra. Peak detection algorithms eliminate the intensities that are below a specified signal-to-noise (S/N) threshold guided for example by the magnitude of the signal to noise ratio[11]. Peak intensity (signal strength) is computed by determining the local maximum on a specified interval where the peak is detected (Beyer et al.,2006; Bioconductor,2007b). Figure 2-8 shows the detected peaks for the spectrum shown in Figure 2-5 using the PROcess package. Figure 2-9 shows the detected peaks in the zoomed region between 5,000 and 10,000 Da. The “local sigma” in Figure 2-8 and Figure 2-9 is a local estimate of the variation (Bioconductor,2007b).

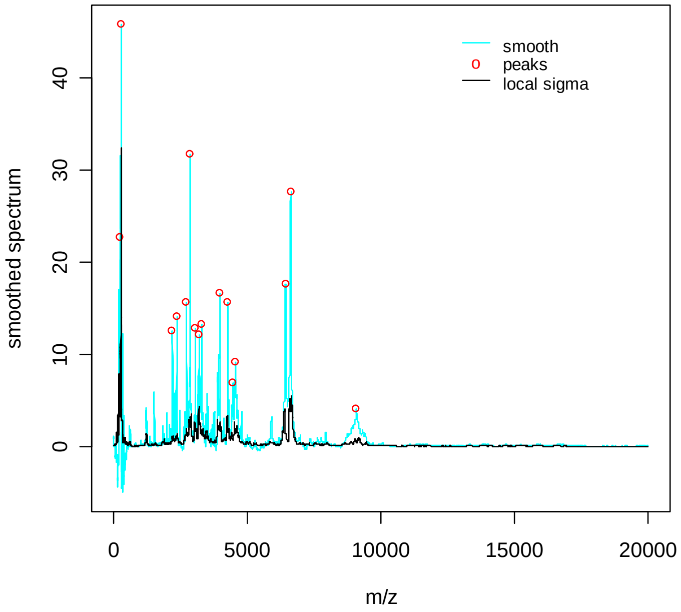


Figure 2-8 The detected peaks (Gentleman et al.,2005).

[11] The ratio of relevant or useful information (signal) to irrelevant information (noise)

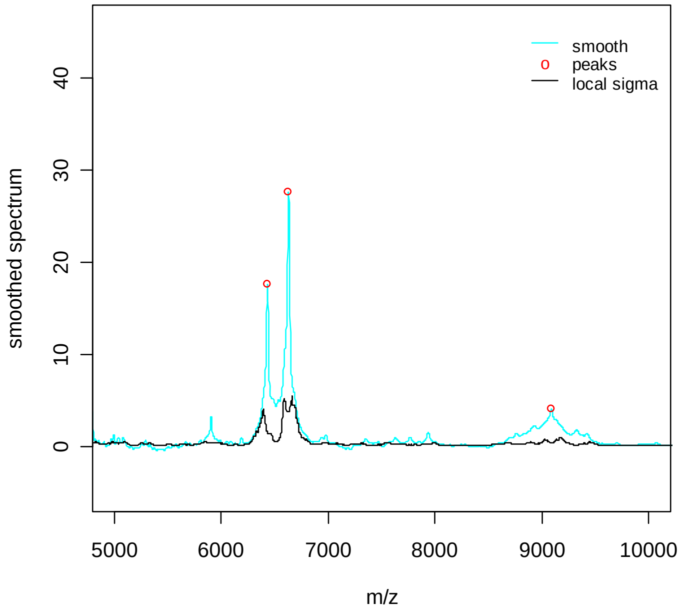


Figure 2-9 The detected peaks in a zoomed region (Gentleman et al.,2005).

2.4.3.4 Peak alignment

Because of the variations in the experimental conditions, such as machine drifts and temperature changes, undesirable variation may occur in SELDI mass spectra. This often results in differences in positions and shapes of peaks between spectra generated in single or multiple experiments. It has been known that the window of potential shift for a mass/charge point is approximately 0.1-0.2% of the mass/charge value of that point when using the model PBSII reader in external calibration. Figure 2-10 shows an example of peak shift. The unaligned peak is shifted to the left side of the reference peak. These technical variations might affect peak detection performance and can obscure biological differences between disease classes, which may affect the accuracy of biomarker discovery for cancer diagnosis.

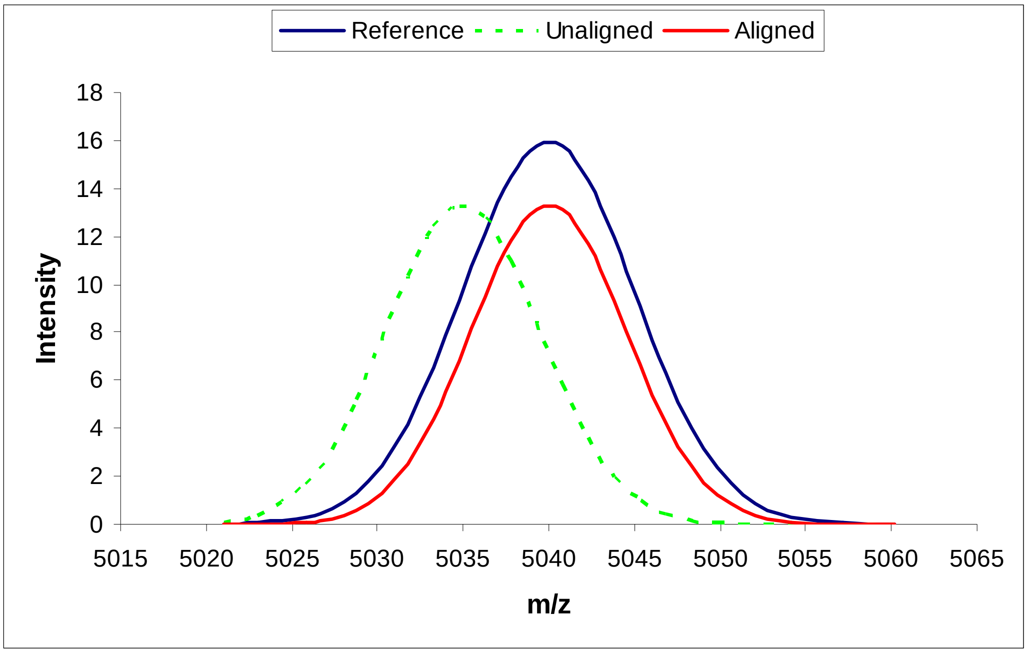


Figure 2-10 A reference peak and a sample peak before and after a right shift intended for aligning the sample peak.

Peak alignment aims at reducing the imprecise measurements of m/z points due to SELDI output shift within an experiment and between experiments and deciding which peaks in different samples correspond to the same protein. This can be achieved by aligning a mass spectrum to a reference spectrum by deleting or inserting data points in the shift regions. However, the two commonly used pre-processing packages, the Ciphergen ProteinChip Software and the PROcess, lack inbuilt functionalities for peak alignment.

Several peak alignment methods (Jeffries,2005; Wong et al.,2005b; Wong et al.,2005a) have been proposed for SELDI mass spectra data. These methods can be divided into two categories, namely peak-based alignment and segment-wise alignment (SWA). The former includes peak alignment by cubic splines (PACS)

(Jeffries,2005) and peak alignment by peak matching (PAPM) (Wong et al.,2005a). The latter was originally proposed by Forshed et al. (2003) for aligning nuclear magnetic resonance (NMR) spectra. In this scheme, the spectra are first divided into segments, whose features are then aligned in a piece-wise manner. Each segment is shifted sideway and stretched or shrunk by linear interpolation (warped) to fit a corresponding piece in the reference spectrum. The correspondence between segments is evaluated by the magnitudes of the correlation coefficient and the comparison with the largest correlation is declared a match. Many algorithms have been used to find optimal shift values for various MS data sets, which include peak alignment by beam search (PABS) (Forshed et al.,2003; Lee & Woodruff,2004) for NMR signal and peak alignment by fast Fourier transform (PAFFT) (Wong et al.,2005b) for chromatographic and spectral data sets.

The above existing peak alignment methods (Jeffries,2005; Lee & Woodruff,2004; Wong et al.,2005b; Wong et al.,2005a) assumed a constant horizontal shift during the process of peak alignment, which did not take the instrument resolution into account. For SELDI mass spectra, the accuracy in m/z positions is normally within 0.1-0.2% of the true m/z value, suggesting that the bigger the m/z value is, and the larger the horizontal shift would be.

In order to address this problem, we propose two new peak alignment methods and will compare their performance with that of the existing peak alignment methods in Chapter 4. The effect of peak alignment on cancer classification will also be assessed in Chapter 4.

2.4.4 Identification of putative biomarkers and cancer classification

2.4.4.1 Statistical methods for cancer classification

A number of methods have been proposed for cancer classification. These include logistic regression (Bhanot et al.,2006; Morra et al.,2007; Wada-Isoe et al.,2007), linear discriminant analysis, quadratic discriminant analysis (Datta & Lara,2006; Wagner et al.,2004; Wu et al.,2003), artificial neural network (ANN) (Ball et al.,2002; Bensmail & Haoudi,2003; Datta & Lara,2006; Poon et al.,2003; Rogers et al.,2003; Yu et al.,2004), random forest (RF) (Wu et al.,2003; Yu et al.,2004), support vector machine (SVM) (Yu et al.,2004; Zhang et al.,2006), and k-nearest neighbour (kNN) (Marchiori et al.,2005; Wu et al.,2003). Some commonly used methods for cancer classification are briefly described as follows.

2.4.4.1.1 Logistic regression

Logistic regression is a special form of generalised linear model and is usually used when the dependent variable is binary (Bhanot et al.,2006). It suits well if we want to classify cancer from normal subjects using SELDI protein profiles. In logistic regression, the dependent variable is a logit[12]; the regression equation is defined by

$$Logit(p) = \ln(\frac{p}{1-p}) = \alpha + \sum_{i=1}^{n} \beta_i I_i \quad \textbf{(2.4)}$$

where, p is the probability of cancer. Logistic regression applies maximum likelihood methods to a training data set to estimate the model parameters α and $\beta_i, i = 1,2,...,n$. For a new test sample, we calculate p from the regression equation

[12] A logit is the natural log of the odds, that is logit(p)=ln(p/(1-p))=ln(odds)

and determine the disease status by the proximity of this value to the two p values for cancer and non-cancer.

2.4.4.1.2 Linear and Quadratic discriminant analysis

Linear and quadratic discriminant analyses are based on a Bayesian classification rule (Datta & Lara,2006; Wu et al.,2003). Suppose that π_k denotes the prior probability of the class k, $k = 0,1$, representing normal subjects and cancer patients, respectively, and that $p(I \mid k)$ denotes the density of distribution of the observation I (SELDI protein profiles) for class k. We can then estimate the posterior distribution of class k given the observation of I by

$$p(k \mid I) = \frac{\pi_k p(I \mid k)}{p(I)} \propto \pi_k p(I \mid k) \quad \textbf{(2.5)}$$

The allocation rule is determined by choosing the class with maximal $p(k \mid I)$ (Wu et al.,2003). That is to say, classifying sample with SELDI protein profile I into class k such that the following holds.

$$y(I) = \arg\max p(k \mid I) = \arg\max(\ln(\pi_k) + \ln(p(I \mid k)) = \arg\max(\lambda_k(I)) \quad \textbf{(2.6)}$$

Suppose that $p(I \mid k)$ follows multivariate normal distribution with mean μ_k and covariance Σ_k, then we can derive the discriminant function $\lambda_k(I)$ as follows.

$$\lambda_k(I) = \ln(\pi_k) - \frac{1}{2}\ln(|\Sigma_k|) - \frac{1}{2}(I - \mu_k)^T \Sigma_k^{-1}(I - \mu_k) \quad \textbf{(2.7)}$$

It is a quadratic equation in I, therefore called quadratic discriminant analysis. The decision boundary between cancer and normal classes can then be determined by

$$\{I \mid \lambda_0(I) = \lambda_1(I)\} \quad \textbf{(2.8)}$$

When we assume the two classes have common covariance matrix, that is, $\Sigma_0 = \Sigma_1 = \Sigma$, then quadratic discriminant analysis becomes linear discriminant analysis. The linear discriminant function $\lambda_k(I)$ is determined as follows.

$$\lambda_k(I) = \ln(\pi_k) - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + I^T \Sigma^{-1} \mu_k \quad \textbf{(2.9)}$$

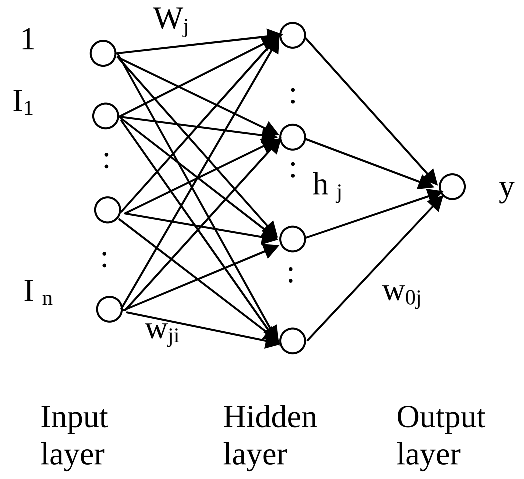


Figure 2-11 Diagram of a neural network

2.4.4.1.3 Artificial neural network

An artificial neural network is a classification or regression model and is represented by a network diagram, based on the concept of neurons in the human brain. It usually consists of at least one hidden layer of neurons (Bensmail & Haoudi,2003; Datta & Lara,2006) as shown in Figure 2-11. It has full interconnections from the input neurons to the hidden neurons and full interconnections from the hidden neurons to the output neurons. An artificial neural network model is trained using K $(n+1)$

patterns (I^k, y_k), $k = 1,2,...,K$, $I^k = (I_1^k, I_2^k, ..., I_n^k)$ are SELDI protein profiles, y_k is cancer status, for example, 1 for cancer and 0 for normal subject.

The output variable is cancer status and can be expressed as follows.

$$y = o(I; w) = g(\sum_{j=1}^{H} w_{0j} h_j) \qquad \textbf{(2.10)}$$

where, w_{0j} is the output weight from a hidden neuron j to an output neuron and g is an output function. The value of a hidden layer neuron h_j is given by

$$h_j = \sigma(\sum_{i=1}^{n} w_{ji} I_i + w_j), j = 1,2,...,H \qquad \textbf{(2.11)}$$

where, w_{ji} is the input weight from an input neuron i to a hidden neuron j, w_j is the threshold weight from an input neuron that has a constant value of 1 to a hidden neuron j, I_i are the intensity values at the input neurons, and H is the number of hidden neurons. σ is a sigmoid function and is defined by

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad \textbf{(2.12)}$$

The weights between neurons are obtained by minimising an error function defined by

$$\varepsilon = \frac{1}{2} \sum_{k=1}^{K} [o(I^k; w) - y_k]^2 \qquad \textbf{(2.13)}$$

For a new test sample, we calculate the output y_{new} of a trained NN model. If y_{new} is close to 1, then we classify it as cancer, otherwise normal subject.

2.4.4.2 Statistical methods for biomarker discovery

Identification of putative biomarkers from SELDI data is not a trivial task. Because a large array of data is generated from a single experiment, it is essential to utilise algorithms to identify proteomic patterns from large amount of data corresponding to a given phenotype from multiple samples (Bensmail & Haoudi,2003). Identification of proteomic fingerprints that underlie phenotypes means sifting through hundreds of SELDI protein profiles to identify a relevant subset of markers that could be further examined.

Many statistical methods have been employed to extract informative protein features using SELDI data sets. These methods can be divided into two categories: namely univariate feature selection method and multivariate feature selection method. The univariate feature selection method, also known as filter-based method, include the t-test (Bhanot et al.,2006; Datta & Lara,2006; De Torre et al.,2006), signal-to-noise (S/N) ratio (Bhanot et al.,2006), receiver operating characteristic (ROC) curve (Adam et al.,2002; Ball et al.,2002; Chen et al.,2002; Qu et al.,2002; Yu et al.,2004), Kolmogorov-Smirnov test (Levner,2005), and ANOVA F-statistic (Datta & Lara,2006; Pavlidis,2003; Wagner et al.,2004). The multivariate feature selection method, also known as wrapper-based method, include decision tree (Adam et al.,2002), genetic algorithms (Petricoin & Liotta,2004), self-organising-maps (Petricoin et al.,2002a), artificial neural networks (Ball et al.,2002), random forest (Izmirlian,2004) and support vector machine (Zhang et al.,2006). The univariate feature selection methods evaluate the discriminatory power of each feature individually, while multivariate feature selection methods evaluate the discriminatory power of each feature using a number of possible subset of features which may lead to a near optimal classifier. Some commonly used parametric test methods are briefly described as follows.

2.4.4.2.1 T-test

The t-test (also called Student's t-test) was introduced by William Sealy Gosset for monitoring the quality of beer brews. "Student" was his pen name. The t-test assumes that data in each group follows a normal distribution (Rice,1995). It is worthwhile to note the effect of sample size on the importance of the assumption. The central limit theorem states that the sum of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed. Let $X_1, X_2,..., X_n$ be a sequence of n independent and identically distributed random variables each having finite values of mean μ and variance σ^2. According to the central limit theorem, as the sample size n increases, the distribution of the sample average of these random variables approaches the normal distribution with a mean μ and variance ${\sigma^2}/{n}$ irrespective of the shape of the original distribution. That is to say, if sample size is large, then we should be more comfortable with using parametric statistics, such as t-test, as the central limit theorem guarantees the validity of the test even if the populations are non-normal. However, for small and moderate sample sizes, the validity of the test demands that the samples be drawn from normally distributed populations (Rice,1995).

We use *t* statistic to determine if the mean μ_p and μ_q of the intensity level of an m/z value across the samples in the two different groups (e.g. cancer and normal samples) are significantly different and it is defined as follows.

$$t = \frac{(\mu_p - \mu_q)}{\sqrt{\frac{\sigma_p{}^2}{n_p} + \frac{\sigma_q{}^2}{n_q}}} \qquad \textbf{(2.14)}$$

where, σ_p and σ_q are standard deviation of the intensity levels, n_p and n_q are the number of samples in each group. The null hypothesis of t-test is that $\mu_p = \mu_{q,}$ indicating that the mean of the feature values in one group is the same as that of the feature values in the other group (Bhanot et al.,2006). A large value of the *t* statistic at an m/z value corresponds to a small p-value. If the p-value is less than a cut-off (say 0.05), then the protein profile corresponding to this m/z value is a significant feature.

2.4.4.2.2 ANOVA

When there exist more than two different groups (e.g. cancer, benign and normal samples), ANOVA can be used to identify features from SELDI protein profiles. Suppose that I_{ikj} is the observed intensity value of the *j*th feature of the *k*th sample in the *i*th group, that *g* is the number of groups, and n_i is the number of samples in the *i*th group. The means of *j*th feature in *i*th group is defined as follows.

$$\bar{I}_{ij} = \frac{1}{n_i}\sum_{k=1}^{n_i} I_{ikj} \qquad \textbf{(2.15)}$$

The mean of *j*th feature across g groups is defined as follows.

$$\bar{I}_j = \frac{1}{\sum_{i=1}^{g} n_i}\sum_{i=1}^{g}\sum_{k=1}^{n_i} I_{ikj} \qquad \textbf{(2.16)}$$

The between group sum of squares for *j*th feature is defined as follows.

$$B_j = \sum_{i=1}^{g} (\bar{I}_{ij} - \bar{I}_j)^2 \quad \textbf{(2.17)}$$

The within group sum of squares is defined as follows.

$$W_j = \sum_{i=1}^{g} \sum_{k=1}^{n_i} (I_{ikj} - \bar{I}_{ij})^2 \quad \textbf{(2.18)}$$

The ANOVA F-statistic is defined as follows.

$$F_j = \frac{B_j / (g-1)}{W_j / (\sum_{i=1}^{g} n_i - g)} \quad \textbf{(2.19)}$$

A large value of the F_j statistic at an m/z value corresponds to a small p-value (Wagner et al.,2004). If the p-value is less than a cut-off (say 0.05), then the protein profile corresponding to this m/z value is a significant feature.

It is apparent that these feature selection approaches evaluate the discriminatory power of each protein one by one independently. That is, the statistical tests are applied to protein intensities at each protein m/z value to detect changes in expression profiles between different groups of samples. The proteins are then ranked according to the significance scores of SELDI protein profiles, such as p-value, *t* statistic or *F* statistic, from the most to the least informative. This ranking defines a series of protein sets as well as the order in which they are subsequently evaluated. More specifically, the first protein set to be evaluated is the best ranked protein; the second protein set the best two ranked proteins, and so on.

These feature selection approaches seem to provide a logical step in identifying proteomic biomarkers. However, there are three issues we need to consider: (a) how to deal with the problem of multiple hypothesis testing? (b) Are these statistical test methods valid for biomarker discovery using SELDI mass spectra? (c) Which method should be used to identify proteomic biomarkers?

2.4.4.3 Adjustment for multiple hypothesis testing

Because the statistical tests consider a single peak at a time, multiple hypotheses testing will lead to many false positives if the commonly used significance level is used. Several solutions have been proposed to filter false positives, which include family-wise error rate (FWER) (Bender & Lange,2001; Benjamini & Hochberg,1995b), Benjamini and Hochberg false discovery rate (FDR) (Bender & Lange,2001; Benjamini & Hochberg,1995a) and Bayesian approaches (Bender & Lange,2001).

FWER approach includes Bonferroni, Holm, and Westfall and Young permutation methods. Bonferroni method corrects the p-value of each variable by multiplying the number of variables in the test list. If the corrected p-value is still below a FWER, then the variable is considered to be statistically significant. Bonferroni method is simple. However, the price for this simplicity is low power. It was suggested that Bonferroni method should only be used in cases where the number of tests is quite small and the correlations among the test statistics are quite low (Bender & Lange,2001; Perneger,1998). Holm method ranks the p-values of individual variables from the smallest to the largest. The corrected p-value for the variable in the first rank is obtained by multiplying by the number of variables to its original p-value, the corrected p-value for the variable in the second rank multiplying the number of variables minus 1. If the corrected p-value is below the FWER, then the variable is considered to be statistically significant. The process is repeated until no variable is found to be significant. Unlike Bonferroni and Holm methods, Westfall and Young method permutes all the variables at the same time and creates a pseudo data set by dividing the data into artificial cancer and control groups. The p-values of variables are calculated on the pseudo-data set. The successive minima of the corrected p-values are retained and compared to the original p-values. The process is repeated a large number of times. The adjusted p-value is the proportion of resampled pseudo data sets where the minimum corrected p-value is less than the original p-value. If the adjusted p-value is below the FWER, the corresponding variable is considered to be statistically different.

The Benjamini and Hochberg FDR method ranks the p-values of variables from the smallest to the largest. The p-value of a variable is corrected by multiplying the total number of variables to its original variable and divided by its rank. If the corrected p-value is below an error rate, the corresponding variable is considered to be statistically different. In general, both FWER and FDR decrease the power to detect a single variable compared to unadjusted methods (Hunter et al.,2007).

Bayesian method differs from the above methods in estimating the marginal posterior probability that the alternative hypothesis is true, rather than controlling type I error rates. However, the values of these posterior probabilities depend on prior densities (Fayers et al.,1997), which are subjective and not well documented.

2.4.4.4 Are the parametric statistical test methods valid for biomarker discovery using SELDI mass spectra?

A typical SELDI spectrum may consist of up to 15,000 features. The high dimensional nature of this data has resulted in the popularity of parametric statistical test (e.g. t-test and ANOVA) as a method for identifying differentially expressed features in many high-throughput data sets, including SELDI and microarray data (Ball et al.,2002; Chen et al.,2002; Kerr et al.,2000; Shapiro et al.,1965; Smyth,2004; Wu et al.,2003). However, these tests assume that the data follow a normal or Gaussian distribution.

The assumption of normality has been adequately addressed for microarray data sets (Chen et al.,2005; Giles & Kipling,2003). However, as far as we know, it has not been studied in the case of SELDI data sets although some parametric methods, such as t-tests (Levner,2005; Pasinetti,2006; Wu et al.,2003), have been used to identify significant peaks from SELDI mass spectra. Therefore, it is necessary to assess the normality of SELDI protein profiles, in order to ascertain that the proteins detected to be differentially expressed more likely to be true positives.

We investigate the normality of SELDI protein profiles using skewness, kurtosis, Shapiro-Wilks test, Kolmogorov-Smirnov test, Cramér-von-Mises test, and Pearson χ^2 test, including implement of the goodness-of-fit testing. We also assess the effect of normalisation and mathematical transformation on the normality of SELDI protein profiles in Chapter 5.

2.4.4.5 Which method should be used to identify biomarkers from SELDI-TOF mass spectra?

Published studies have compared the performance of different feature selection and classification methods based on the accuracy of the classifiers which were built using different feature selection and classification algorithms (Levner,2005; Wu et al.,2003). The study by Wu et al. (2003) utilised different feature selection methods (such as t-test) and classification algorithms (including linear discriminant analysis and quadratic discriminant analysis) to analyse a MALDI ovarian cancer data set. It showed that different feature selection methods identified different sets of features, and that different algorithms gave different classification accuracies. The study by Levner (2005) examined the performance of classifiers trained using SELDI protein profiles identified by different feature selection methods (including t-test) based on five SELDI cancer data sets. Again it found that the classifiers trained with different feature sets gave different classification accuracies.

In general, it is more likely that different approaches will result in identification of different sets of proteins, and that different classification methods have different split performance. However, none of the previous studies addressed which approach should be used for identifying protein biomarkers from SELDI expression profiles. We therefore propose a general algorithm for biomarker discovery from SELDI protein profiles using a combination of feature selection methods and classification methods, and implement the algorithm using two published SELDI data sets in Chapter 6.

2.5 Discussion

SELDI technology appears to be a useful tool in biomarker discovery and early detection of cancer although its value has been limited at present by its inability to detect serum/plasma proteins at lower abundance (Poon,2007). But, SELDI is a new technology for proteomic studies. There is an urgent need to investigate how to make best use of the technology for biomarker discovery and cancer diagnosis with the current development of SELDI techniques.

Firstly, a standardised experimental protocol and operating procedure need to be established. The study by Poon (2007) demonstrated that the irreproducibility problem in MALDI analysis can be addressed when an unbiased acquisition strategy (Pang et al.,2004) was utilised. The intra-assay coefficient of variation of normalised protein peak intensity varied from 10% to 30%. The same author, Poon (2007) also reported that in SELDI data analysis, the unbiased acquisition strategy and use of hydrogel as a support for chromatographic materials could make the normalised intensity values of the protein or peptide peaks comparable (Coombes et al.,2003; Li et al.,2002; Pang et al.,2006; Poon et al.,2003; Poon et al.,2004; Poon et al.,2005). The intra-assay and inter-assay coefficients of variations for the normalised intensity of majority of the SELDI peaks varied from 5% to 25% (Ebert et al.,2004; Poon,2007; Wadsworth et al.,2004). The inter-laboratory coefficients of variation of the normalised SELDI peaks varied from 15% to 36% when a standardised experimental protocol and quality control strategy were used (Poon,2007; Semmes et al.,2005). This suggests that it is necessary to standardise the methodologies so that the results obtained with SELDI technology are robust and reproducible across laboratories.

The Human Proteome Organisation (Penfield,2007) is currently developing serum, plasma reference standards. The standard methodology of the SELDI technology will be essential for the quality-assurance instrument and the calibration of individual assays will make the results obtained from this technique robust and reproducible among laboratories. As the improvement of the resolution and reproducibility of the current instrument and introduction of sample loading robots, the SELDI technology should achieve more sensitivity and specificity. It has been argued that the rapid growth and development in SELDI technology and similar techniques will eventually overcome the limitations of SELDI technology in the near future (Poon,2007).

Secondly, raw SELDI data need to be pre-processed to reduce the bias caused by systematic variations in the experiments. Data pre-processing includes baseline subtraction, intensity normalisation, peak alignment and peak picking. The Ciphergen ProteinChip Software and PROcess package are two commonly used mass spectrum pre-processing packages. It has been shown that the baseline correction, and normalisation algorithms implemented in Ciphergen ProteinChip Software and PROcess packages gave similar performance. Although most previous studies (Li et al.,2002; Petricoin et al.,2002a; Poon et al.,2003; Rogers et al.,2003; Tong et al.,2004; Yu et al.,2004; Zhang et al.,2004a; Zhao et al.,2004) used Ciphergen ProteinChip Software to perform mass spectrum pre-processing, PROcess can also be used for the purpose of baseline subtraction and intensity normalisation. However, both packages only considered peak alignment during the process of peak picking. Therefore, further research on peak alignment is needed.

Thirdly, a variety of feature selection methods that have been successfully used in other fields, such as microarrays, are candidates for biomarker discovery using SELDI expression profiles. Some parametric methods, such as t-test, assume that the outcome measure follows a normal distribution. However, the studies by Wu et al. (2003) and Levner (2005) used the t-test to select significant peaks without testing the normality of SELDI expression profiles. The results were therefore questionable. Although there is a wide range of feature selection methods available, none of the previous studies gave a guideline in selecting methods for proteomic biomarker discovery. Further research on the normality of SELDI expression profiles and on identifying suitable methods for biomarker discovery are needed.

For cancer classification, some studies used non-parametric machine learning methods, such as neural networks. This is probably because that some classical statistical regression analyses, such as logistic regression, assume a functional form for the relation between SELDI expression profiles and the outcome measurement. However, precise information about the shape of the relation between SELDI expression profiles and the outcome measurement is lacking. Therefore, they often fail to deal adequately with the biological complexities and the multidimensional problem of variables selection.

Although analysis of single or small numbers of proteomic biomarkers is relatively simple task to conduct, the statistical analysis of information from hundreds of proteins is challenging. Multifactorial cancers are expected to be characterised by multiple proteins. Despite the small effects of some proteins, the effect of the combinations of these proteins and the interaction between protein-protein and protein-environment may be large enough to predispose a cancer. Therefore, it is necessary to handle not only a large number of proteins, but also the interactions between proteins and environmental factors. However, none of the previous studies investigated the influence of environmental factors on cancer classification accuracy. We will consider this issue in Chapter 7.

2.6 Summary

Although SELDI technology has shown a great potential in biomarker discovery and early detection of cancer, the validation and the clinical use of the technology is still being established. The value of SELDI technology has been limited by lack of consensus assay protocol and operating procedure. A general procedure and statistical methods for data mining SELDI mass spectra for biomarker discovery and cancer diagnosis are studied. Problems related to reproducibility analysis, peak alignment, normality test, and biomarker discovery have been identified. The solutions for these problems will be addressed in the next chapters.

2.7 Data sets used in the thesis

In the work that follows, we consider five SELDI data sets for analysis.

2.7.1 Colon cancer data set

A colon cancer data includes 15 samples taken from patient with colon cancer and 4 samples taken from normal individuals. All serum samples were applied on four chip surfaces: Q10, CM10, H50, and IMAC30 in 4 replicate experiments over a two-month period. The chips were analysed in a PBS IIc SELDI-TOF equipped with an autoloader (Ciphergen Biosystems, UK). Spectra were collected over 0-20 kDa (low range) and 0-200 kDa (high range). For the purpose of reproducibility analysis, the data set was divided into 4 subsets (Set 1, Set 2, Set3 and Set 4) according to the time when the sample were assayed. All the 19 samples were assayed in an identical fashion at each experimental run. The only difference between Set 1 and Set 2 is that the Set 2 was obtained from the samples which were left at room temperature about 4-5 hours after diluted 5-fold in 9 M urea, 50 mM Tris/HCl (pH 9.0) before diluted

10-fold dilution in the binding buffer prior to applied on bioprocessor. Set 3 and Set 4 were obtained from two different spots in one experimental run after 2 months later.

2.7.2 Pooled non-cancer data set

A pool of serum contains 26 samples taken from healthy individuals. The pooled sample was processed on seven distinct days at intervals over a 4-week period using identical handing techniques, the same equipment and personnel. 28 spectra were obtained from CM10 chip on a PBS IIc SELDI-TOF equipped with an autoloader (Ciphergen Biosystems, UK).

2.7.3 Lung cancer data set

A lung cancer data consists of 39 patients with histologically confirmed non-small cell lung cancer and 39 patients with no evidence of cancer disease. Sera were analysed in duplicate on CM10 ProteinChip arrays using a PBS IIc SELDI equipped with an autoloader (Ciphergen Biosystems, UK). Each spectrum was composed of peak intensities at 13429 points, which were corresponding to m/z values in the range of 0-20kDa.

2.7.4 Ovarian cancer data set 4-3-02

An ovarian data set 4-3-02**,** downloaded from the National Cancer Institutes of Clinical Proteomics Program web site (National Cancer Institute,2007b), consists of 100 samples from ovarian cancer patients and 100 samples from individuals without cancer. The spectra of these samples were generated by using WCX2 protein chip, and a Protein Biosystem II surface-enhanced laser desorption ionisation–time-of-flight mass spectrometer (Ciphergen Biosystems). Each spectrum was composed of peak intensities at 15154 points, which were corresponding to m/z values in the range of 0-20kDa.

2.7.5 Prostate cancer data set 7-3-02

A prostate cancer data set 7-3-02 was downloaded from the National Cancer Institutes of Clinical Proteomics Program web site (National Cancer Institute,2007b). It includes 69 samples from patients with prostate cancer and 63 samples from individuals with no evidence of disease. All spectra of these samples were generated using H4 protein chip and a Ciphergen PBS1 SELDI mass spectrometer. Each spectrum consists of 15,156 points and the baselines of spectra were subtracted.

3 REPRODUCIBILITY ANALYSIS

3.1 Introduction

In this chapter, we propose three quantitative measures using Euclidean distance, correlation coefficient, and the paired t-test to assess the reproducibility of SELDI protein profiles. The results of analysing a colon cancer data set have shown that the SELDI protein profiles of the identical samples have significant changes if they are left in ice for a period of 4-5 hours at room temperature. This suggests a new conjecture that protein profiles are affected by storage.

3.2 Materials and reproducibility assessment methods

3.2.1 Sample information

The colon cancer data set, described in Section 2.7.1, is used in this chapter to assess the reproducibility of SELDI expression profiles. Each spectrum consists of 13429 data points in the low mass range and 42426 data points in the high m/z range. All of the spectra were baseline-subtracted using Ciphergen ProteinChip Software, and normalised by total ion current, starting from the m/z value of 2000 for low range molecules (0-20kDa) and 20000 for high range molecules (20k-200kDa), respectively. We define “bioprocessor-to-bioprocessor” for the comparison of the difference between Set 1 and Set 2, “month-to-month” for Set 1 and Set 3, and “spot-to-spot” for Set3 and Set 4 in the following reproducibility analysis.

3.2.2 Reproducibility assessment methods

Let us consider a SELDI mass spectrum data set obtained from n samples. This spectral data can be put in a $m\times(n+1)$ data matrix, $(mz, I_1, I_2, ..., I_n)$, where, mz is a column vector denoting m measured m/z values and I's are the corresponding vectors of intensities.

The intensity profile for protein in the kth replicate, is denoted by $S_k = (mz_k, I_1^k, I_2^k, ..., I_n^k)$, where, I_j^k, k=1, 2, 3, 4, is the observed protein intensity, which is the sum of the true value of the intensity and a measuring error, i.e. $I_j^k = T_j + \varepsilon_j^k$, where, T_j is the underlying protein intensity, ε_j^k may be experimental error (or some contaminant) and j is the sample index. If SELDI mass spectra are reproducible, then we expect I_j^p to be close to I_j^q, where, q indexes replicate spectrum of spectrum p. That is, ε_j^p is close to ε_j^q.

This suggests a need to construct similarity measures for replicates of the same protein. Several approaches including, correlation coefficient, Euclidean distance and even conventional statistical methods such as paired t-test, may be considered.

3.2.2.1 Euclidean distance

A natural way of comparing measurements is to consider their differences. For vectors this means assessing combined differences between corresponding points across them. A classical tool for assessing the differences between two vectors is the Euclidean distance. For each sample j, the Euclidean distance between spectrum p and its replicate spectrum q is defined by

$$d_j^{pq} = \sqrt{\sum_{i=1}^{m}(I_j^p(i) - I_j^q(i))^2}\ ,\ j = 1,2,...,n \quad \textbf{(3.1)}$$

where, $I_j^p(i)$ represent the intensities of spectrum p of sample j, $i = 1,2,...,m$ denotes m measured m/z values, and $I_j^q(i)$ represent the intensities of the replicate spectrum q.

It measures how close the two spectra generated from an identical sample are. The smaller the Euclidean distance, the closer the two spectra, suggesting that the measurements of the SELDI protein profiles are more reproducible.

It should be noted that the study by Baggerly et al. (2004) also employed the concept of the Euclidean distance. But it was used in a different way, rather than directly measuring the reproducibility of SELDI protein profiles. In that study, the Euclidean distance was used to allocate samples into clusters. Each cluster was then labelled “cancer” or “normal” by majority vote, the fitness was defined in terms of the number of samples correctly classified. The fitness function was employed for selection of feature sets using a genetic algorithm[13]. The reproducibility was then assessed by agreement among different feature sets.

3.2.2.2 Correlation coefficient

Correlation coefficient is a commonly used measure of similarity. The reproducibility of SELDI mass spectra (i.e. the similarity between replicate spectra) can be measured by Pearson’s correlation coefficient, which is defined by

[13] A computer algorithm based on the mechanisms of biological natural selection, using populations of objects which can reproduce based on the biological concepts of survival of the fittest and mutation.

$$CC(I_j^p, I_j^q) = \frac{\mathrm{cov}(I_j^p, I_j^q)}{sd(I_j^p)sd(I_j^q)} \quad \textbf{(3.2)}$$

where, $\mathrm{cov}(I_j^p, I_j^q)$ is covariance between I_j^p and I_j^q, $sd(I_j^p)$ and $sd(I_j^q)$ are the standard deviations of I_j^p and I_j^q, respectively.

The numerator measures the shared variance between intensities I_j^p and I_j^q, while the denominator measures the observed variance (William,2007). The higher the correlation coefficient between I_j^p and I_j^q is, the higher the reproducibility of a mass spectrum.

The correlation is defined only when the two standard deviations are finite and non-zero. It has a range of [-1, 1]. However, in this context, a correlation of –1 would not be acceptable. The closer the correlation coefficient is to 1, the stronger the correlation between the replicate spectra. The values between 0 and 1 indicate the degree of dependence between the replicate spectra. The coefficient of 0 means that the replicate spectra are independent each other, suggesting that the measurements of the SELDI protein profiles are not reproducible.

3.2.2.3 The paired t-test

Another way of assessing similarities of replicate samples is to determine if the mean of the differences of corresponding measurements is zero. The paired t-test can be used for this purpose and it is defined by

$$t = \frac{E(I^p - I^q)}{SD(I^p - I^q)}\sqrt{N} \quad \textbf{(3.3)}$$

where, $E(I^p - I^q)$ is the mean difference of the intensity levels at an m/z value between spectra and their replicate spectra, $SD(I^p - I^q)$ is the standard deviation of these differences, and N is the sample size. The null hypothesis of the paired t-test is that the population mean of the differences between the two groups is zero, indicating that the measurements of the SELDI protein profiles are reproducible.

3.2.2.4 Linear mixed-effects models

Linear mixed-effects models are used to investigate the variations in SELDI protein profiles over different sample processing times. Like in the study by Banks et al. (2005), a separate model was fitted for each peak, chip type, and mass range. In these models, the response variable is the intensity value at each peak, and the independent variable is the sample processing time, which is a three-factor variable, having values of 0, 4-5 hours, and 2 months, with time 0 as baseline. A random-effect term is used to describe individual-specific effects, which takes the correlation between peak intensities measured on the same individual into account.

3.3 Results of reproducibility assessment

3.3.1 Euclidean distance

For each type of chip, the Euclidean distances of the identical samples between the spot-to-spot (between Set 3 and Set 4) groups, the bioprocessor-to-bioprocessor (between Set 1 and Set 2) groups, and the month-to-month (between Set 1 and Set 3) groups, were calculated through R (CRAN,2007). The same process was repeated for all four chips. The means and standard deviations of the Euclidean distances for the 19 samples were calculated and listed in Table 3-1. The Euclidean distance distributions of the 19 samples in the four types of chips with two m/z range (low: 0-

20k and high: 0-200k) are shown by the Box and Whisker plot depicted in Figure 3-1.

Table 3-1 Means and standard deviations of Euclidean distance between mass spectra obtained from identical samples in spot-to-spot, bioprocessor-to-bioprocessor, and month-to-month experiments.

Chip type	Mass range	spot-to-spot	bioprocessor-to-bioprocessor	month-to-month
IMAC30	Low	1.0749±0.3993	1.7756±0.8795 **	2.5099±0.5299 ***
	High	0.5311±0.2665	1.2496±0.6798 ***	1.5651±0.6712 ***
CM10	Low	1.4935±0.4624	2.0509±0.3831 ***	2.4399±0.4359 ***
	High	2.2200±0.9226	1.9612±0.7981	2.6537±0.9271
H50	Low	1.0749±0.3887	1.7757±0.8560 **	2.5099±0.5157 ***
	High	0.8609±0.5152	1.3929±1.0327	1.9656±0.3916 ***
Q10	Low	0.8640±0.3789	2.1695±0.8259 ***	2.5824±0.4907 ***
	High	1.6128±0.8375	2.2231±0.7257 *	4.4477±1.1324 ***
Significant levels: `***' <0.001, `**' <0.01, `*' <0.05.				

For spot-to-spot comparison (between Set 3 and Set 4), the means of Euclidean distances across four chips for both low and high mass ranges varied from 0.5311 to 2.2200, with a median of 1.0749. For bioprocessor-to-bioprocessor comparison (between Set 1 and Set 2), the means of Euclidean distances across the four chips vary from 1.2496 to 2.2231, with a median of 1.8685. For month-to-month comparison (between Set 1 and Set 3), the means of Euclidean distances across four chips vary from 1.5651 to 4.4477, with a median of 2.5099.

It is noteworthy that the mean of Euclidean distances increases from spot-to-spot comparison to bioprocessor-to-bioprocessor and month-to-month comparisons, with CM10 in the higher range setting a notable exception. The medians (across the four chips) of the means of Euclidean distances (19 samples) increase by 75% ((1.868-1.07)/1.07*100) for the bioprocessor-to-bioprocessor groups and by 135% ((2.510-1.07)/1.07*100) for the month-to-month groups, compared to the spot-to-spot groups.

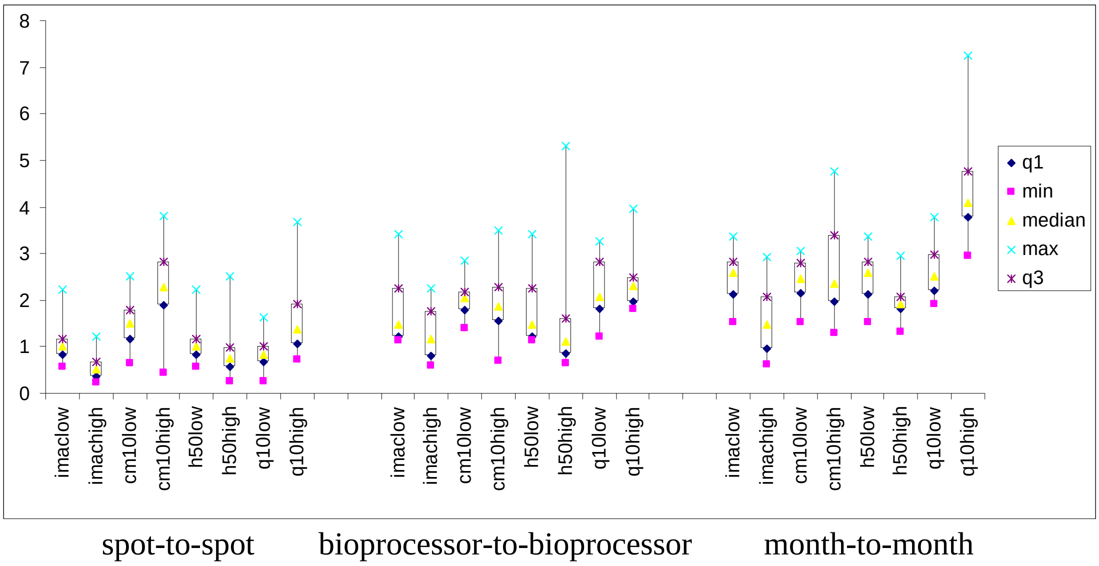


Figure 3-1 Box and Whisker plot of Euclidean distance between spot-spot, bioprocessor-bioprocessor, and month-2-month comparisons. q1 denoting for the first quartile; q3, the third quartile; min, the minimum of distances; max, the maximum of distances; median, the median of distances.

For each chip with both high and low mass ranges, we applied the paired t-tests to the 19 samples to investigate the difference of the Euclidean distances between the spot-to-spot and bioprocessor-to-bioprocessor groups, and between the spot-to-spot and month-to-month groups. The mean differences of Euclidean distances between the spot-to-spot and month-to-month groups were highly significant at p-value < 0.001 for all four chips in two different settings except CM10 in the higher m/z range. The mean differences of Euclidean distances between the spot-to-spot and bioprocessor-to-bioprocessor groups were significant at p-value < 0.001 for IMAC30 in the higher m/z range, and CM10 and Q10 in the lower m/z range, significant at p-value < 0.01 for IMAC30 and H50 in the lower m/z range, and significant at p-value < 0.05 for Q10 in the lower m/z range, while the mean differences of Euclidean distances between the spot-to-spot and bioprocessor-to-bioprocessor groups were not significant with p-values of > 0.05 for H50 and CM10 in the higher m/z range.

3.3.2 Correlation coefficients

For each type of the four chips, the corresponding correlation coefficients of identical samples between the spot-to-spot (between Set 3 and Set 4) groups, the bioprocessor-to-bioprocessor (between Set 1 and Set 2) groups, and the month-to-month (between Set 1 and Set 3) groups, were calculated by using R correlation coefficient functions (CRAN,2007). The same process was repeated for all four chips. The means and standard deviations of the correlation coefficients for the 19 samples were calculated and listed in Table 3-2. The distributions of the correlation coefficient of the 19 samples across four chips with two different range settings are shown in Figure 3-2.

Table 3-2 Means and standard deviations of correlation coefficients between mass spectra obtained from identical samples in spot-to-spot, bioprocessor-to-bioprocessor, and month-to-month experiments.

Chip type	**Mass range**	**spot-to-spot**	**bioprocessor-to-bioprocessor**	**month-to-month**
IMAC30	Low	0.9861±0.0123	0.9561±0.0417 **	0.9259±0.0305 ***
	High	0.9941±0.0041	0.9634±0.0359 **	0.9710±0.0210 ***
CM10	Low	0.9244±0.0435	0.8891±0.0364 *	0.8328±0.0519 ***
	High	0.6131±0.2284	0.8768±0.1105 **	0.7862±0.1983 *
H50	Low	0.9861±0.0119	0.9561±0.0405 **	0.9259±0.0297 ***
	High	0.9863±0.0112	0.9525±0.1442	0.9489±0.0205 ***
Q10	Low	0.9770±0.0183	0.8860±0.0726 ***	0.8367±0.0689 ***
	High	0.9874±0.0149	0.9706±0.0142 **	0.8877±0.0342 ***
Significant levels: `***' <0.001, `**' <0.01, `*' <0.05.				

For spot-to-spot comparison (between Set 3 and Set 4), the means of correlation coefficients across four chips for both low and high mass ranges vary from 0.6131 to 0.9941, with a median of 0.9861. For bioprocessor-to-bioprocessor comparison (between Set 1 and Set 2), the means of correlation coefficients across four chips vary from 0.8768 to 0.9706, with a median of 0.9543. For month-to-month comparison (between Set 1 and Set 3), the means of correlation coefficients across four chips vary from 0.7862 to 0.9710, with a median of 0.9068.

It is noteworthy that the mean of correlation coefficients decreases from spot-to-spot comparison to bioprocessor-to-bioprocessor and month-to-month comparisons, with CM10 in the higher range setting a notable exception. The medians (across the four chips) of the means of correlation coefficients (19 samples) decrease by 3% ((0.9543-0.9861)/0.9861*100) for the bioprocessor-to-bioprocessor groups and by 8% ((0.9068-0.9861)/0.9861*100) for the month-to-month groups, compared to the spot-to-spot groups.

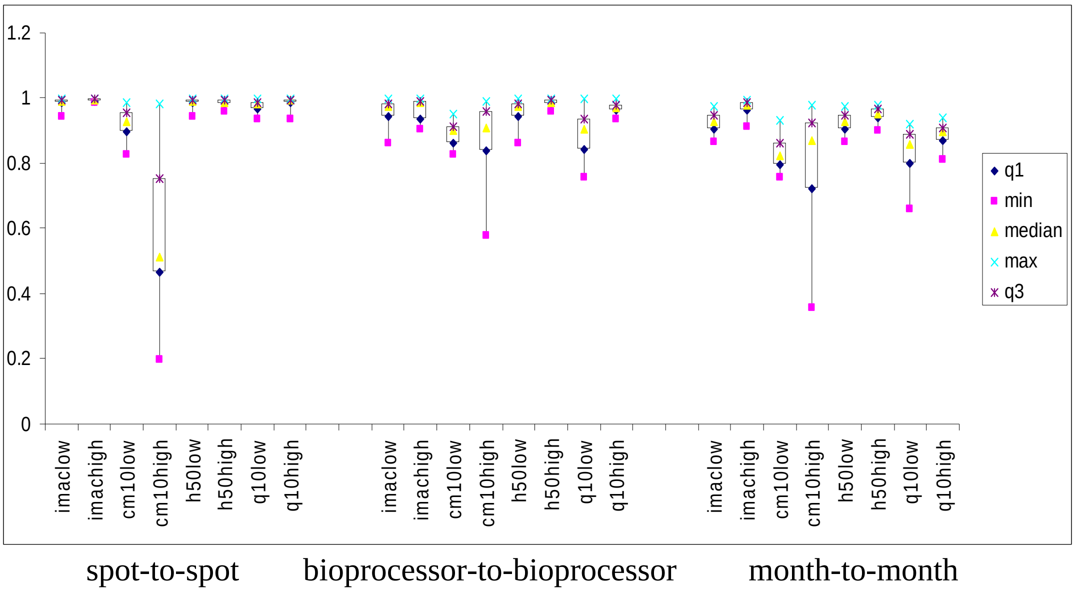


Figure 3-2 Box and Whisker plot of correlation coefficient between spot-spot, bioprocessor-bioprocessor, and month-2-month comparisons. q1 denoting for the first quartile; q3, the third quartile; min, the minimum of distances; max, the maximum of distances; median, the median of distances.

For each chip with both high and low mass ranges, we applied the paired t-tests to the 19 samples to investigate the difference of the correlation coefficients between the spot-to-spot and bioprocessor-to-bioprocessor groups, and between the spot-to-spot and month-to-month groups. The mean differences of the correlation coefficients between the spot-to-spot and month-to-month groups were highly significant at p-value < 0.001 for IMAC30, H50, Q10 in both range settings and CM10 in the lower range setting, significant at p-value < 0.05 for CM10 in the higher range setting. The mean differences of the correlation coefficients between the spot-to-spot and bioprocessor-to-bioprocessor groups were highly significant at p-value < 0.001 for Q10 in the lower range setting, significant at p-value < 0.01 for IMAC30 in both range settings, CM10 in the higher range setting, H50 in the lower range setting, and Q10 in the higher range setting, while the mean difference of the correlation coefficients between the spot-to-spot and bioprocessor-to-bioprocessor groups was not significant with a p-value of 0.322 for H50 in the higher range setting. These results are in accordance with those obtained from the Euclidean distance analyses.

3.3.3 The paired t-test

For each chip with both high and low mass ranges, the paired t-tests were performed between the spot-to-spot (between Set 3 and Set 4) groups, the bioprocessor-to-bioprocessor (between Set 1 and Set 2) groups, and the month-to-month (between Set 1 and Set 3) groups at each m/z point by calling the R function *t.test* (CRAN,2007). The percentages of the m/z points at which the p-values of the paired t-tests were less than 0.05 were counted for each of the 19 samples. The average percentages of m/z points at which the p-values were less than 0.05 were calculated across the 19 samples and listed in Table 3-3. The results of the paired t-tests showed that in average, 24% of m/z points in the spot-to-spot groups, 36% of m/z points in the bioprocessor-to-bioprocessor groups, and 62% of m/z points in the month-to-month groups were significantly different at a significance level of 0.05. This suggests that the reproducibility of SELDI mass spectra in the bioprocessor-to-bioprocessor, and the month-to-month groups became worse compared to that in the spot-to-spot groups.

Table 3-3 The average percentages of m/z points at which p-values are < 0.05 in paired t-tests between spot-to-spot, bioprocessor-to-bioprocessor, and month-to-month experiments.

Chip type	Range	spot-to-spot	bioprocessor-to-bioprocessor	month-to-month
		(%)	(%)	(%)
IMAC30	Low	22	42	52
	High	26	58	71
CM10	Low	59	45	55
	High	14	1	61
H50	Low	8	15	51
	High	48	10	74
Q10	Low	2	47	61
	High	11	69	67
Average		24	36	62

3.3.4 Linear mixed-effects models

The R (CRAN,2007) function lme() in library(nlme) (Pinheiro et al.,2009) is used to undertake the analyses of variations in SELDI protein profiles over different sample processing times. The number and percentage of peaks at which intensities have significant changes with time are shown in Table 3-4. The results of these linear mixed-effects models showed that in average, sample processing time 4-5 hours after baseline results in 29% of peaks at which intensities have significant changes (bioprocessor-to-bioprocessor), and sample processing time 2 months after baseline results in 48-49% of peaks at which intensities have significant changes (month-to-month). This implies that the reproducibility of SELDI mass spectra in the month-to-month groups became worse compared to that in the bioprocessor-to-bioprocessor groups. The indirect comparison of the numbers of peaks at which intensities have significant changes between baseline and 2 months later (the last two columns in Table 3-4) shows that the spot-to-spot groups may give a better reproducibility compared to the bioprocessor-to-bioprocessor, and the month-to-month groups. To estimate the changes of protein profiles between the spectra in Set 3 and their replicate spectra in Set 4 (for spot-to-spot comparison), the linear fixed-effects models were re-run, taking Set 3 as the baseline. The results show that there are 13% of peaks at which intensities have significant changes when comparing spectra in Set 3 and their replicate spectra in Set 4.

Table 3-4 The numbers and percentages of peaks at which intensities have significant changes over the sample processing times

Chip type	Range	Sample processing time		
		4-5 hours vs. baseline	**2 months vs. baseline**	**2 months vs. baseline**
		Set 2 vs. Set 1	**Set 3 vs. Set 1**	**Set 4 vs. Set 1**
		m/n (%)	m/n (%)	m/n (%)
IMAC30	Low	58/229 (25)	82/229 (36)	70/229 (31)
	High	38/105 (36)	52/105 (50)	52/105 (50)
CM10	Low	49/137 (36)	57/137 (42)	63/137 (46)
	High	25/92 (27)	35/92 (38)	61/92 (66)
H50	Low	8/92 (9)	42/92 (46)	43/92 (47)
	High	3/125 (2)	72/125 (58)	60/125 (48)
Q10	Low	51/96 (53)	59/96 (61)	59/96 (61)
	High	37/79 (46)	44/79 (56)	46/79 (58)
Average		269/922 (29)	443/922 (48)	454/922 (49)
n: the total number of peaks m: the number of peaks at which p<0.05/n				

3.4 Discussion

We proposed three quantitative reproducibility measures for SELDI expression profiles using Euclidean distance, correlation coefficient, and the p-value of the paired t-test. We then assessed the reproducibility of SELDI mass spectrum at each m/z point between the nineteen replicated identical samples assayed at different times in the same experimental conditions using the colon data set. Highly significant differences of SELDI protein profiles between the identical samples that were generated in different experimental runs over a short period or a long period of time have been found in this data set. The significant difference observed over a short period indicated that the mass spectra of the identical samples had significant changes if they were left in ice for a period of 4-5 hours at room temperature. This is consistent with the result reported in the study by Marshall et al. (2003), which found that when plasma was left sitting at room temperature for 4 or 8 hours, the MALDI data, as recorded by a SELDI instrument, changed significantly (Diamandis,2004b; Marshall et al.,2003). The bias of SELDI technology was also discussed in the studies by Baggerly et al. (2004) and Diamandis (2003a).

The reproducibility analysis results show that the number of m/z points, where the p-values of the paired t-tests were less than 0.05, increases from the spot-to-spot, bioprocessor-to-bioprocessor, and month-to-month comparisons in chip IMAC30, Q10, and H50 with low range setting, while opposite results were observed (the higher number of significant m/z points in the spot-to-spot comparison) in chip CM10 and H50 with higher setting. These observations were not in accordance with the results of Euclidean distances and correlation coefficients. This may be because that the number of significant m/z points from the paired t-tests only indicates that at how many specific m/z points the difference between assays were significant, while the Euclidean distance and correlation coefficient methods took the overall difference into account.

These data suggest that the difference between the identical samples over the replicate experiments was significantly different in terms of Euclidean distances and correlation coefficients. This difference was not due to inherent biological difference, but presumably due to artefacts associated with experimental conditions, such as time to assay, temperature, humidity, mechanical adjustments to the mass spectrometer itself, differential handing and/or processing of the samples. This indicates that careful experimental design is needed, a standard protocol should be drawn up to minimise the effect of irrelevant sources of variations (Baggerly et al.,2004). The proposed quantitative measures of reproducibility should be useful in helping identify important experimental conditions for improving the reproducibility of SELDI expression profiles.

3.5 Summary

The reproducibility of SELDI expression profiles has been investigated by applying the proposed quantitative measures of reproducibility to the colon cancer data set. The results showed that the differences of SELDI data between identical samples over the replicate experiments are statistically significant in terms of Euclidean distances and correlation coefficients. The mass spectra generated in different spots in one experimental run have a better reproducibility compared to the mass spectra generated in different experimental runs. The mass spectra generated in different experimental runs over a short period of time have a better reproducibility than the mass spectra generated in different experimental runs over a long period of time. The results of the linear fixed-effects models have confirmed these reproducibility changes for spectra generated in different experimental runs over different periods of time. The reproducibility analyses have found that the mass spectra of the identical samples had significant changes if they were left in ice for a period of 4-5 hours at room temperature. This suggests a new conjecture that protein profiles are affected by storage. The reproducibility analysis implies that apart from a standard experiment protocol should be adopted; there is also a need for further research on SELDI mass spectrum pre-processing. The next chapter will investigate peak alignment methods (one of the data pre-processing procedure) and their effects on the reproducibility of SELDI protein profiles and on the accuracy of cancer classification.

4 PEAK ALIGNMENT

4.1 Introduction

In this chapter, we propose two new peak alignment methods based on the SWA scheme, in which the maximum sideway shift value is set dynamically according to the characteristics of the SELDI mass spectra to allow the search algorithms to find the best shift in the search space of interest. The performance of the proposed peak alignment algorithms is assessed in terms of correlation coefficient, the number of peaks identified and coefficient of variation and compared with that of the existing peak alignment methods, peak alignment by beam search (Lee & Woodruff,2004), peak alignment by fast Fourier transform (Wong et al.,2005b), peak alignment by peak matching (Wong et al.,2005a) and peak alignment by cubic splines (Jeffries,2005). We then examine the extent to which different peak alignment methods improve the reproducibility of SELDI protein profiles.

The study by Wu et al. (2003) compared the performance of a number of statistical methods for the classification of ovarian cancer data sets generated by MALDI technology. The methods studied included linear discriminant analysis, quadratic discriminant analysis, bagging and boosting classification trees, k-nearest neighbour classifier, support vector machine, and random forest. In this analysis, the random forest model outperformed other statistical methods. We assess the effects of peak alignments on the discriminatory power of random forest classifiers in terms of the AUC.

It has demonstrated that one of the newly proposed peak alignment algorithms provides the best alignment criterion.

4.2 Materials and peak alignment methods

4.2.1 Data sets

The pooled non-cancer data set and lung cancer data set, described in Section 2.7.2 and Section 2.7.3, respectively, are used for peak alignment analysis. The spectra were background corrected and intensity normalised using the Ciphergen ProteinChip 3.2 software, and normalised to the total ion current in the range 2k to 20kDa. The reference spectrum was generated by averaging all spectra in each data set and is shown in Figure 4-1.

We also use the ovarian data Set 4-3-02 and prostate cancer data set 7-3-02, which were described in Section 2.7.4 and Section 2.7.5, respectively, to test the performance of the peak alignment methods.

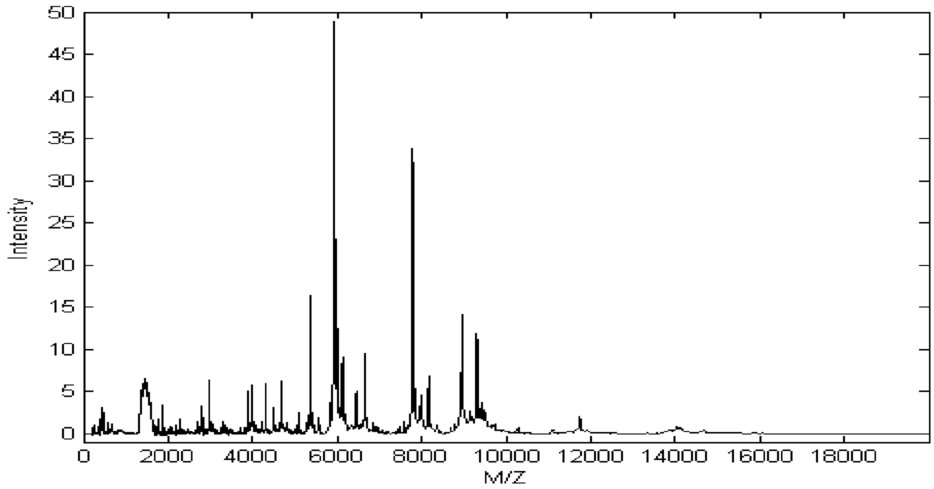


Figure 4-1 The reference spectrum generated by averaging 28 non-cancer spectra.

4.2.2 Peak alignment methods

4.2.2.1 The existing peak alignment methods

The four previously published peak alignment methods are described as follows.

4.2.2.1.1 Peak alignment by peak matching algorithm

The peak alignment by peak matching algorithm is a peak-based alignment scheme (Wong et al.,2005a). It inserts or deletes data points to shift regions in each spectrum to be aligned with the corresponding region in a reference spectrum, guided by the so-called reference points. The reference points are mainly automatically (or sometimes manually) selected peaks within each spectrum. The reference spectrum and the spectrum of interest are first divided into windows of specified sizes and then insertions (or deletions) are made, guided by the presence or absence of a match between reference peaks across the two spectra within a given window. The window size can be varied to ensure no matching peak is omitted within reasonable distance of the edge of a specified window. The rules for insertion or deletion are described as follows (Wong et al.,2005a). Suppose that there are *s* spectra with *d* data points. Each spectrum *m* is to be aligned with the corresponding region in a reference spectrum *r*, as marked by reference points, P_{im} and P_{jr}, where, i and j are points between 0 and d.

1. For each j in P_{jr}, find the closest matching P_{im}. If no match is found within a window of size w, which is specified by the user, then move to the next point $j+1$.

2. If P_{im} is found, but not aligned to P_{jr}, find the minima between, P_{im} and $P_{(i-1)m}$, ($\min_{-1}$), and, P_{im} and $P_{(i+1)m}$, ($\min_{+1}$), where insertions or deletions are to be made for alignment of P_{im} and P_{jr}
3. If $P_{im} >$ P_{jr} (for the value of the x-axis), then points are to be deleted from $\min_{-1}$ and points to be inserted at $\min_{+1}$. If $P_{im} < P_{jr}$, then the reverse applies.
4. Where points are inserted, the y-axis value for the inserted point is estimated by a least square quadratic polynomial fit to its adjacent w points.

4.2.2.1.2 Peak alignment by fast Fourier transform algorithm

Peak alignment by fast Fourier transform algorithm uses segment-wise alignment scheme and fast Fourier transform (FFT) cross-correlation to determine a shift range between a segment and its corresponding reference segment (Wong et al.,2005b). FFT is an algorithm for converting data from time to the frequency domain and often used for measuring correlation and time delay or shift between signals. It provides an accurate estimation of the shift between two signals. Suppose that we have two functions, one is a reference spectrum $r(x)$; the other is a sample spectrum $s(x)$ to be aligned. At any shift position u, their cross-correlation function $Corr(r,s)_u$ is defined by

$$Corr(r,s)_u = \int_{-\infty}^{\infty} r(x)s(x+u)dx \quad \textbf{(4.1)}$$

The Fourier transformation of $r(x)$ is given by

$$R(w) = \int_{-\infty}^{\infty} r(x)e^{2\pi iwx}dx \quad \textbf{(4.2)}$$

where, $R(w)$ is the Fourier transformed function in the inverse wavelength w domain. The reverse Fourier transformation is given by

$$r(x) = \int_{-\infty}^{\infty} R(w)e^{-2\pi iwx}dw \quad \textbf{(4.3)}$$

Similarly, the Fourier transformed function and the reverse Fourier transformation for $s(x)$ are given as follows.

$$S(w) = \int_{-\infty}^{\infty} s(x)e^{2\pi iwx}dx \quad \textbf{(4.4)}$$

$$s(x) = \int_{-\infty}^{\infty} S(w)e^{-2\pi iwx}dw \quad \textbf{(4.5)}$$

The cross-correlation can then be written as

$$Corr(r,s)_u = \int_{-\infty}^{\infty} R(w)S^*(w)e^{-2\pi iwx}dw \quad \textbf{(4.6)}$$

where, $S^*(w)$ represents the complex conjugate of the function. That is to say, we can calculate the cross-correlation by applying forward Fourier transformation to $r(x)$ and $s(x)$, then multiplying the transformed functions $R(w)$ and $S^*(w)$, and then performing reverse Fourier transformation of this product.

The implementation of this algorithm is described as follows (Wong et al.,2005b).

1. Calculate the cross-correlation function between a segment and its corresponding reference segment using FFT described above.
2. Find the optimal shift position u_{op} by $u_{op} = \max_u(Corr(r,s)_u)$
3. Shift the segment by that amount using the following formula.

$$s'(x) = \begin{cases} s(0) \quad x \in [0, u_{op} - 1], \quad s(x - u_{op}) \quad x \in [u_{op}, N-1] \quad if \quad u_{op} > 0 \\ s(x - u_{op}) \quad x \in [0, N + u_{op} - 1], \quad s(N-1) \quad x \in [N + u_{op}, N-1] \quad if \quad u_{op} < 0 \\ s(x) \quad x \in [0, N-1] \quad if \quad u_{op} = 0 \end{cases}$$

(4.7)

where, N is the size of the segment.

4.2.2.1.3 Peak alignment by cubic splines algorithm

The peak alignment by cubic splines algorithm is a peak-based alignment scheme that finds peak locations by fitting cubic splines to the ratio of peak locations (m/z values) of the spectrum of interest to its target value (Jeffries,2005).

A cubic spline is used to interpolate n data points, (t_1, y_1), (t_2, y_2), …, (t_n, y_n) with a piecewise cubic polynomial. The general form of the cubic spline is

$$S(t) = \begin{cases} S_0(t), \quad if \quad t \in [t_0, t_1] \\ S_1(t), \quad if \quad t \in [t_1, t_2] \\ \ldots \\ S_{n-1}(t), \quad if \quad t \in [t_{n-1}, t_n] \end{cases} \qquad \textbf{(4.8)}$$

In order to provide a smooth fit between data points, the spline function $y = S(t)$ should satisfy the interpolation conditions $S(t_i) = y_i$, and continuous first and second derivatives at interior data points $t_2,...,t_{n-1}$. The values of the second derivative at the endpoints can be arbitrary. A natural cubic spline sets second derivatives to zero at the endpoints t_0 and t_n. The value of y at any point t in $[t_1, t_n]$ can be interpolated by

$$S_i(t) = \frac{z_{i+1}(t-t_i)^3 + z_i(t_{i+1}-t)^3}{6h_i} + (\frac{y_{i+1}}{h_i} - \frac{h_i}{6}z_{i+1})(t-t_i) + (\frac{y_i}{h_i} - \frac{h_i}{6}z_i)(t_{i+1}-t) \quad \textbf{(4.9)}$$
$$h_i = t_{i+1} - t_i$$

where $z_i = S''(t_i)$ are the values of the second derivative at t_i

The coefficients of the cubic spline are determined by applying the interpolation conditions, continuity of first derivative and the endpoints conditions of second derivative to the above equations. For a natural cubic spline, it is to solve the following equations (De Boor,2001).

$$z_0 = 0$$
$$h_{i-1}z_{i-1} + 2(h_{i-1} + h_i)z_i + h_i z_{i+1} = 6(\frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}), i = 1,2,...,n-1 \quad \textbf{(4.10)}$$
$$z_n = 0$$

The peak alignment by cubic spline algorithm finds a postulated cubic spline and determines the optimal parameters that minimise the distances between the cubic spline and the ratio of peak locations of the spectrum of interest to its target value. The cubic spline transformation function is then used to recalibrate the peaks. Suppose that a single spectrum has a set of *N* peaks with associated m/z values (denoted by p_i) and a set of target m/z values (denoted by m_i). The algorithm is described as follows (Jeffries,2005).

1. Let u_i denote the ratio of the original mass value to its target value, that is, $u_i = \frac{p_i}{m_i}$.

2. For given { p_1, m_1, p_2, $m_2 \ldots$, p_N, m_N }, a cubic spline $f(p)$ is fitted such that the following error term is minimised.

$$E = \sum_{i=1}^{N} \{u_i - f(p_i)\}^2 + \lambda \int_{\alpha}^{\beta} \{f''(p)\}^2 dp \quad \textbf{(4.11)}$$

   where, α and β indicate the limits of the m/z range under consideration. λ is chosen by cross-validation (by successively leaving out one pair of peak location values) and determining what value of λ generally yields low estimated error, $\{u_i - f(p_i)\}^2$, for the omitted data.

3. Let M_{orig} denote one of the original *m/z* and I_{orig} denote its associated intensity. Then a recalibrated mass associated with the original intensity, I_{orig}, is calculated as follows.

$$M_{recal} = \frac{M_{orig}}{f(M_{orig})} \quad \textbf{(4.12)}$$

   where, $f(M_{orig})$ denotes the value of the spline function at the mass value of M_{orig}. These recalibrated masses are computed for every one of the original masses and are associated with the original intensities.

4. Linear interpolation of the recalibrated masses that are closest to the original mass is used to obtain a new intensity for the original mass.

A disadvantage of PACS is that it assumes that the target spectrum and the sample spectrum to be aligned have the same number of peaks.

4.2.2.1.4 Peak alignment by beam search algorithm

The peak alignment by beam search algorithm uses the SWA scheme and beam search algorithm. It aligns a sample spectrum to a reference spectrum by shifting and shrinking m/z values so that the correlation coefficient between the corresponding segments in the two spectra is maximised. The implementation of the algorithm is described as follows (Lee & Woodruff,2004).

1. Establish the set of promising solutions: $S = \{(0,0)\}$ and the search radius $r = 2 * ra / 3$, where, [-ra, ra] is a pre-specified search range.
2. For all $(i,s) \in S$, evaluate (i,s) and the solutions in its neighbourhood $(i \pm r, s \pm r)$, where, i is the shrinkage amount (if $i > 0$, then the segment is shrunk i points, otherwise it is stretched by $-i$), and s is the amount of shift (if $s > 0$, then right shift, otherwise left shift)
3. Select the *k* best solutions evaluated in step 2 to be the next set S.
4. Let $r = r / 3$. If $r < 1$ then report the member of S with the best correlation coefficient and stop the search, otherwise goto step 2

The beam search algorithm is a heuristic method in which a set of likely solutions is first created on the edge of a pre-determined search range [-ra, ra] (steps 1 and 2). These values are then pruned and the *k* most promising solutions are selected for further evaluation (step 3), where, *k* is a parametric input, called the beam width. For each search step, the search range is reduced until a stop criterion is met and the best solution is then reported (step 4).

4.2.2.2 Our proposed peak alignment methods

We propose two new alignment methods, the peak alignment by beam search with dynamic maximum sideway movement value (PABSD) and the peak alignment by sequential search (PASS). Both methods are based on the SWA scheme, but with different search algorithms. The PASS uses a sequential algorithm while the PABSD uses a beam search algorithm. In both algorithms, the maximum sideway movement value is set dynamically. These algorithms were implemented using MATLAB.

An important step in SWA method is the setting of the maximum range of sideway movement value. If this is too large, there could be inconsistencies in peak matching between spectral segments (Forshed et al.,2003). If it is too small, then the insufficient search range may lead to the failure of the search algorithm to find the best movement value to align spectral segments. Thus it is vital to set this parameter accurately before peak alignment can be done (Forshed et al.,2003). The current alignment algorithms that use the SWA scheme fix this parameter at a constant value (Forshed et al.,2003; Forshed et al.,2005; Lee & Woodruff,2004; Viant,2003; Wong et al.,2005b). However, in SELDI data sets, the shifts in m/z values are not constant and thus using a fixed parameter will not sufficiently align its features. Depending on the instrument resolution and mass calibration methods used, the accuracy in m/z positions is normally within 0.1-0.2% of the true m/z value. Therefore, the appropriate maximum range for sidewise movement should vary from segment to segment. To meet this challenge, the maximum sideway movement values are set dynamically. For a given segment, the maximum sideway movement value is set to be the minimum m/z value in the segment multiplied by the manufacture's specification error. The detailed description of the two proposed algorithms is given below.

4.2.2.2.1 PABSD method

The peak alignment by beam search with dynamic maximum sideway movement value algorithm is adapted from PABS. The difference between PABS and PABSD is that for each segment of interest, the maximum sideway movement value is not a constant. The implementation of the method is as follows.

1. Obtain minimum m/z value of the segment to be aligned, *minmz*.
2. Calculate the maximum sideway movement value: *msmv* = *minmz* *λ, where, λ is the instrument resolution (0.1%-0.2%).
3. Initialise *k* (1 or 2) temporary best shifts (s_i, i=1,…, k), and calculate the search radius: r = 2**msmv*/3
4. For all of the *k* temporary best shifts, calculate the correlation coefficient between the segments to be aligned in three shifts (s_i, $s_i \pm r$) and select best *k* shifts for next iteration of search. Let $r = r /3$.
5. Search is stopped when search radius r <1 and the best shift is reported.

4.2.2.2.2 PASS method

This method was developed based on “hill-climbing” search algorithm. It calculates the correlation coefficient between a reference segment and the segment of interest, and the results are compared with the goal in an uphill direction. The procedure is repeated at each data point until the maximum correlation coefficient is found and the corresponding shift value is the optimal solution. The implementation of the method is as follows.

1. Obtain minimum m/z value of the segment to be aligned, *minmz*.

2. Calculate the maximum sideway movement value: *msmv* = <u>*minmz*</u> *λ, where, λ is the instrument resolution (0.1%-0.2%).
3. Initialise search set **Ω** = (-1, 0, 1)
4. Calculate the correlation coefficient between the segments to be aligned in three shifts in the search set **Ω**, and find the maximum correlation coefficient *maxCC*
5. If the maximum correlation coefficient achieves at 0, then set **Ω** = ( )
6. If the maximum correlation coefficient achieves at 1, and if *msmv* ≥ 2, then set **Ω** = (2,…, *msmv*), otherwise set **Ω** = ( )
7. If the maximum correlation coefficient achieves at -1, and if *msmv* ≥ 2, then set **Ω** = (-2,…, -*msmv*), otherwise set **Ω** = ( )
8. If **Ω** = ( ), then stop searching, the shift corresponding to *maxCC* is reported, otherwise,
9. Calculate the correlation coefficient *newCC* between the segments to be aligned in the next shift in the search shift set **Ω**
10. If *newCC* > *maxCC,* then go to 9, otherwise,
11. Stop searching, the shift corresponding to *maxCC* is reported.

4.2.3 Evaluation and comparison of peak alignment methods

In the analysis of mass spectrometry data sets, an important pre-processing step is peak alignment. However, there is always a risk of introducing errors in data when performing peak alignment (Forshed et al.,2003). Therefore, the importance of checking the data for errors after peak alignment cannot be overemphasised. Several statistical methods, including correlation coefficient, coefficient of variation and classification methods, are used to evaluate the performance of the two new peak alignment algorithms and the four published methods.

4.2.3.1 Effect of peak alignment on reproducibility

Ideally, all spectra from the pooled serum should be similar (Jeffries,2005). Thus they would be expected to be associated with lower coefficient of variation and higher correlation coefficients between the spectra of interest and the reference spectrum. Both the correlation coefficient and the coefficient of variation are used to examine the effect of peak alignment on reproducibility of the SELDI data.

4.2.3.2 Effect of peak alignment on cancer classification

Applying peaks alignment to mass spectra will reduce technical variations and improve the accuracy of matching proteins across samples. To assess the performance of peak alignment methods on cancer classification, models are built using RF machine learning algorithm with both un-aligned and the aligned MS data generated by various alignment algorithms. The examination was implemented in the following steps.

1. *Detection of peaks*: Although valuable information might exist in any part of the spectrum, because of high measurement variation in SELDI data, peaks are the most suitable value to identify biomarker and classify cancer patients from normal individuals (Bhanot et al.,2006). Thus all the analyses are based on peaks extracted with the BioConductor software package PROcess (Gentleman,2005). To make the results comparable, the same parameter settings were applied to both aligned and un-aligned spectra in selection of peaks. The lower bound for signal/noise ratio is set to 2, the minimum intensity to1 and total area under the curve is taken as 0.01.

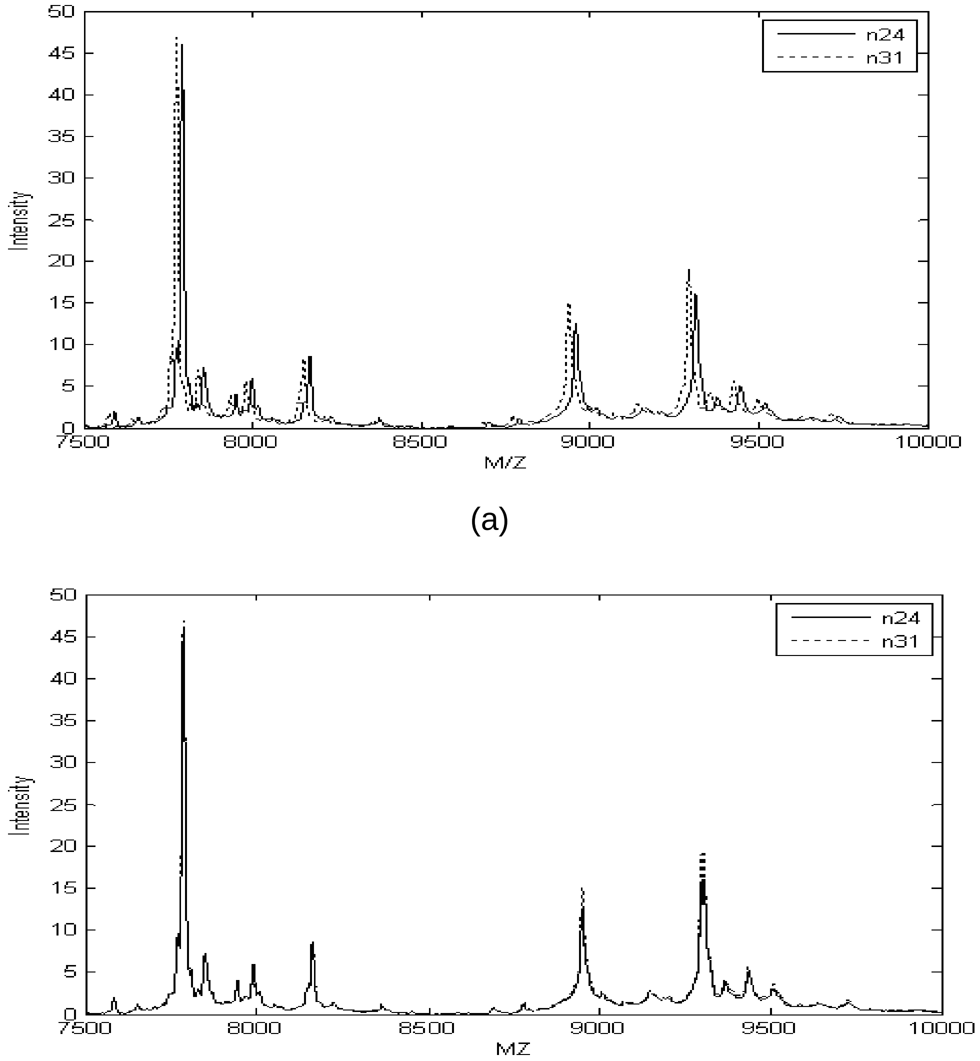
2. *Identification of significantly different peaks*: In order to assess the effect of peak alignment on cancer classification, the RF machine learning algorithm is used to build classification models for both aligned and un-aligned lung cancer data sets. Due to the large number of variables relative to the sample size, an important issue in analysing such data is to extract disease-associated biomarkers using the limited number of samples guided by the critical aim of minimising the number of false positives (Li et al.,2002). Peaks are ranked and selected using Wilcoxon test at the 5% significance level (10% FDR) and are then used to build classifiers to discriminate cancer from normal samples using RF technique.

3. *Cancer classification using RF:* A random forest model consists of multiple decision trees, grown in parallel. The R software package called ***randomForest*** is used**,** which is based on an algorithm proposed by Breiman (Breiman,2001) that combines two powerful machine learning techniques, bagging and random feature selection. The classification procedure involves the following steps: firstly, data is divided into training and test sets. Secondly, bootstrap samples are drawn from this training set. For each bootstrap sample, a tree is built to predict the *out-of-bag* (OOB) samples, which are not present in the bootstrap sample. When constructing a tree, the best split from a randomly selected subset of input variables is used at each node splitting. For each tree grown, about two-thirds of the samples are selected at random and used to train model and the rest of the samples are treated as OOB. To classify an input vector in the random forest, the vector is submitted as an input to each of the trees in the forest and the classification is determined by the majority vote.

The performance of the resulting models is evaluated using the AUC of the ROC curve. The McNemar test (Agresti,1990) is used to assess the significance of the difference between classification models for the aligned and un-aligned spectra.

4.3 Results of peak alignment

The maximum range of sideway movement was set dynamically from segment to segment. For the lung cancer data set, it was set to the minimum m/z value in that segment multiplied by 0.2% (manufacture's specification error). Figure 4-2 shows a section of SELDI mass spectra from 7.5kDa to 10kDa before and after peak alignment by PABSD (the reference spectrum is shown in Figure 4-1). Two spectral profiles n24 (generated on day 1) and n31 (generated on day 2) had differences in peak height and peak positions. It can be seen that the peaks were not well aligned originally, but clearly aligned after the PABSD algorithm was applied. Figure 4-2 also shows that the peaks keep their original shape and intensities after peak alignment. This is because the peak alignment only corrects the horizontal shifts (eliminates noise) and reserves the spectral composition, the signals of interest.



(a)

(b)

Figure 4-2 Two spectral segments generated from a pooled non-cancer sample on different days, n24 generated on day 1 and n31 day 2. (a) before peak alignment and (b) after peak alignment.

4.3.1 The parameter settings for peak alignment methods

For the proposed methods, the number of segments was set to 40 and the maximum sideway movement value was set dynamically. The beam width was set to 1 for both PABS and PABSD methods. For the PAFFT algorithm, the minimum segment size was set to 500 data points and the maximum shift was set to 20 data points for both the PAFFT and PAPM algorithms. For PACS, the individual calibration equations for each spectrum were generated for the pooled non-cancer data set in order to achieve higher alignment performance during the calibration and peak alignments. The parameters used for peak alignment are summarised in Table 4-1.

Table 4-1 Parameter setting for each peak alignment methods.

Alignment methods	Number of segment	Segment size	Maximum sideway movement	Window size
PABSD	40	Dynamically	Dynamically	-
PASS	40	Dynamically	Dynamically	-
PABS	40	Dynamically	20	-
PAFFT	-	500 (minimum)	20	-
PAPM	-	-	-	20
PACS	-	-	-	Dynamically

4.3.2 The effect of peak alignment on reproducibility

The pooled non-cancer data set was used for the reproducibility analysis. Table 4-2 shows the CVs and correlation coefficients for the 28 spectra after peak alignment using the six methods, the new two and the four published methods. The median CV was higher before peak alignment and reduced by aligning the peaks for all the methods but PAPM. The possible causes of poor performance of the PAPM method are that PAPM uses only local peak information to determine the amount of shift (insertion or deletion) and does not consider the correlation coefficient between the reference spectrum and the spectrum to be aligned during the process of peak alignment. These may introduce further errors into the spectrum after peak alignment, and thus gives poor performance, even compared to without peak alignment. In contrast, another peak-based method, PACS, determines the optimal parameters that minimise the distance between the cubic spline and the ratios of peak locations of the spectrum of interest to their target values, which takes information of all peaks into account. The SWA method, PAFFT, finds the optimal shift position by calculating the maximum cross-correlation function between a segment and its corresponding reference segment, which takes all data points in the spectra into account. Similarly, the SWA methods PABS, PABSD, and PASS all determine the shift position based on the maximal correlation coefficient between spectra, which take all data points in the spectra into account.

It is interesting to note that all inter-quartile ranges (IQR) of CV were in the similar range (from 0.485 to 0.508) but PAPM. This makes it comparable among the median CVs across the un-aligned and the aligned mass spectra using different peak alignment methods. The PASS method performed best. It had both the lowest CV (0.360) and the highest correlation coefficient (0.978), with the largest changes in median CV (from 0.411 to 0.360) and correlation coefficient (from 0.942 to 0.978) compared to those without peak alignment. These changes represent a reduction of 12% for the median CV, and an increase of about 4% in the correlation coefficient. This is consistent with what reported (4-7%) for PABS and PAFFT (Wong et al.,2005b).

The results of Wilcoxon tests show that compared to without peak alignment, all peak alignment methods, except PAPM, significantly reduce the median CV (p-value < 0.0001). However, the newly proposed methods with dynamic maximum sideway movement, PABSD and PASS, do not outperform PABS in terms of the median CV (p-value > 0.1). The results of Wilcoxon tests also show that none of the peak alignment methods significantly increases the correlation coefficient compared to without peak alignment (p-value > 0.05).

Table 4-2 CVs distribution and correlation coefficient under different peak alignment methods based on the pooled non-cancer data set.

Alignment methods	Median (CV)	IQR (CV)	CV % change compared to no alignment	Correlation coefficient (CC) ±SD	CC % change compared to no alignment
No alignment	0.411	0.486	-	0.942±0.049	-
PABSD	0.361	0.485	-12.2	0.976±0.011	3.6
PASS	0.360	0.489	-12.4	0.978±0.010	3.8
PABS	0.365	0.488	-11.2	0.976±0.012	3.6
PAFFT	0.366	0.508	-10.9	0.978±0.010	3.8
PAPM	0.506	0.599	23.1	0.882±0.092	-6.4
PACS	0.361	0.495	-12.2	0.977±0.013	3.7

The PASS method was also applied to the two published data sets (ovarian data set 4-3-02 and prostate data set 7-3-02) (National Cancer Institute,2007b). The average of the correlation coefficients between the reference spectrum and spectra of interest increased from 0.940 to 0.942 for the ovarian data set and from 0.798 to 0.801 for the prostate data set as shown in Table 4-3.

Table 4-3 The average of correlation coefficients between a reference spectrum and a spectrum of interest before and after applying the PASS peak alignment (Instrument resolution 0.2%)

Data sets	Before alignment	After alignment
Ovarian data set 4-3-02	0.940	0.942
Prostate data set 7-3-02	0.798	0.801

4.3.3 The effect of peak alignment on the cancer classification

The six peak alignment methods were applied to the lung cancer data set and the performance of cancer classification was evaluated in terms of the classification accuracy based on the AUC values. The results are shown in Table 4-4 and Figure 4-3. From Table 4-4, it can be seen that the number of significantly differently peaks ($p < 0.05$) remains relatively the same after the peak alignment, for the majority of the peak alignment methods (PABSD, PASS, PABS, PAFFT and PACS), but there was a huge increase (from 34 to 63) for the PAPM method. The classification performance of the RF models using six peak-aligned data sets and one data set without peak alignment are summarised using a bar-chart.

Table 4-4 The number of peaks with p-values of Wilcoxon tests less than 0.05 after applying different peak alignment methods to the lung cancer data set.

Alignment methods	**None**	**PABSD**	**PASS**	**PABS**	**PAFFT**	**PAPM**	**PACS**
Number of peaks	34	34	34	34	34	63	35

Figure 4-3 summarises the AUC values of the models. The RF models trained using the data set aligned by PASS and PAFFT methods gave the highest AUC values (0.84) compared to the RF model trained using raw data set (0.82), representing a 2% improvement, which is not statistically significant (*p*-value = 0.48, McNemar test). The RF model trained using data set aligned by PAPM method gave the lowest AUC value (0.81), which becomes worse than the model trained using the raw data set. Overall, we can see that the performance of the RF models were increased after the spectra were aligned by PABSD, PASS, PABS, PAFFT and PACS, but not by PAPM. These results are consistent with what was found after these peak alignment methods were applied to the pooled non-cancer sample data set.

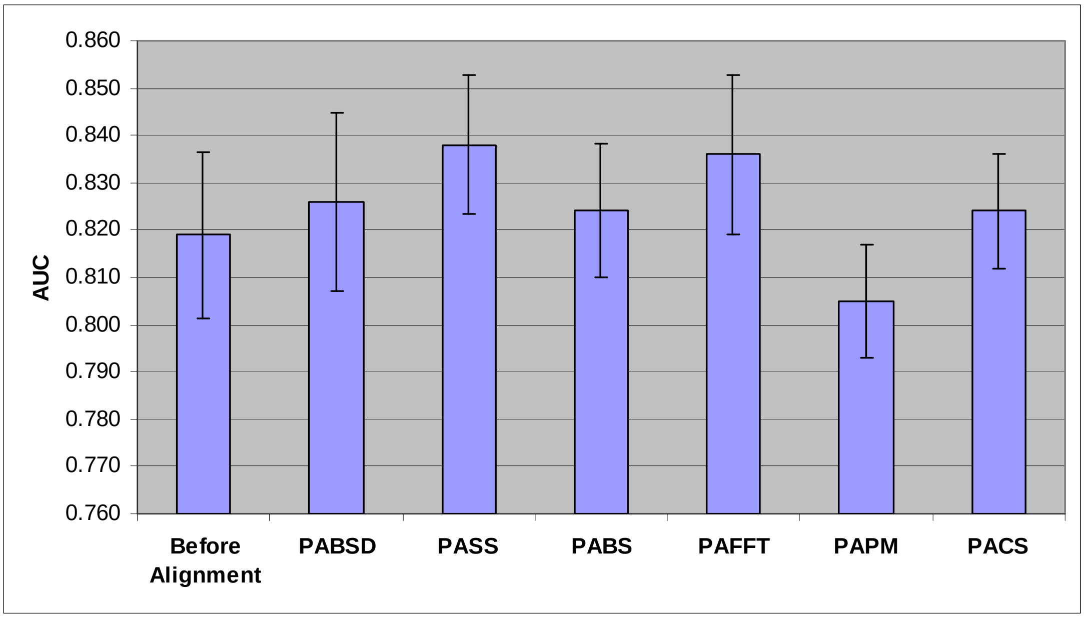


Figure 4-3 Summary of the classification performance of the RF models based on AUC.

4.4 Discussion

The performance of six peak alignment approaches has been examined and compared in terms of the reproducibility of the protein profiles and the AUC values of the classification models. All methods studied, apart from PAPM, led to a small improvement in cancer classification performance. The AUC values increased up to 2% compared to without peak alignment, which is not significantly different with a p-value of 0.48 (McNemar test). All peak alignment methods, except PAPM, increased the correlation coefficient up to 4% compared to without peak alignment, but these increases are not statistically significant (p-value > 0.05). The peak alignment methods, apart from PAPM, also resulted in reductions of 11 to 12% for the median CV compared to without peak alignment. These reductions are statistically significant (p-value < 0.0001). However, the proposed peak alignment methods with dynamic maximum sideway movement, PABSD and PASS, only made a small improvement in the reproducibility of MS data compared to PABS method, which is not a significant difference with p-value > 0.1. These analysis results suggest that PABSD and PASS methods with dynamic maximum sideway movement do not show superior performance compared to the existing SWA method PABS with fixed maximum sideway movement. This may be due to the fact that the data set used for these analyses were generated on one machine under a stable environment within a short period of time. It should be interesting to examine the effect of peak alignment on the improvement of the reproducibility of SELDI data, which are generated on circumstances when mass spectra are obtained from different machines or centres or within a long period of time. However, no data set generated on such circumstances can be found. We artificially introduced noise into some spectra in the existing pooled non-cancer data set. Suppose that 14 spectra (selected at random) were generated by centre 1, and the remaining 14 spectra were generated by centre 2. We kept the spectra generated by centre 1 unchanged, while added noise into the spectra generated by centre 2. For each spectrum generated in centre 2, 40 segments were partitioned. At each segment, 50% of data points were shifted left by 10 points, and 10 data points were inserted at the end of shifts. Table 4-5 shows the effect of peak alignment on the reproducibility of mass spectra in the simulated data set.

Table 4-5 CVs distribution and correlation coefficient under different peak alignment methods based on the pooled non-cancer data set.

Alignment methods	Median (CV)	IQR (CV)	CV % change compared to no alignment	Correlation coefficient (CC) ±SD	CC % change compared to no alignment
No alignment	0.563	0.594	-	0.798±0.139	-
PABSD	0.502	0.572	-10.8	0.956±0.011	28.0
PASS	0.504	0.575	-10.5	0.958±0.010	28.4
PABS	0.514	0.574	-8.7	0.956±0.011	28.0
PAFFT	0.519	0.596	-7.8	0.958±0.010	28.4
PAPM	0.642	0.721	14.0	0.747±0.011	-9.0
PACS	0.504	0.582	-10.5	0.955±0.012	27.8

Again, apart from PAPM, all the peak methods significantly reduce (p-value < 0.0001) the median CV down to 0.502 compared to 0.563 without peak alignment, leading to a reduction of 11%. They also significantly increase the correlation coefficient up to 0.958 compared to 0.798 without peak alignment, leading to an increase of 28%. The proposed methods PABSD and PASS outperform PABS in terms of the median CV. The p-values of Wilcoxon tests are 0.041 for PABS versus PABSD, and 0.037 for PABS versus PASS, respectively. The analysis results from the simulation data set demonstrate the potential superior performance of the peak alignment methods with dynamic maximum sideway movement over that of the existing SWA method with fixed maximum sideway movement.

4.5 Summary

We have proposed two new peak alignment algorithms and evaluated the performance of the two algorithms and other four existing peak alignment methods. Five out of six peak alignment methods (including the newly proposed two algorithms) significantly reduce the median CV by more than 10% compared to the results of the un-aligned mass spectra, while keeping the IQR in the same range. The five peak alignment methods also increase correlation coefficient by more than 3%. The two proposed peak alignment methods PABSD and PASS with dynamic maximum sideway movement do not show superior performance compared to the existing SWA method PABS with fixed maximum sideway movement for the data set that were generated on one machine under a stable environment within a short period of time. However, they do outperform PABS when a simulated data set that was assumed to be generated in different centres with artificially introduced noise was used.

The effects of peak alignment on the classification accuracy have been evaluated by training RF models using the un-aligned and aligned mass spectra by the six peak alignment methods. The results show that five of the six RF models trained using data set aligned by the peak alignment methods improve the AUC values by up to 2% compared to the RF model trained using raw data set, which are not statistically significant. The small improvement in cancer classification accuracy may be partly due to the fact that the relatively small number of mass spectra generated on one machine under a stable environment within a short period of time was used in this study. The effects of peak alignment methods on the performance of the classification models might not be detected. Further research is needed to assess the effects of peak alignment on biomarker discovery and cancer classification using large data sets obtained from different machines at different centres.

5 NORMALITY ANALYSIS

5.1 Introduction

In order to evaluate whether it is valid to apply the parametric statistical tests (such as t-test) to identify proteomic biomarkers, we investigate the normality of SELDI protein profiles using skewness, kurtosis, Shapiro-Wilks test, Kolmogorov-Smirnov test, Cramér-von-Mises test, and Pearson χ^2 test. We propose a new normalisation method and compare it with the existing normalisation methods in the Ciphergen Biosystems Software and PROcess package by evaluating how normality of protein profiles is affected by the normalisation methods. We also explore the role of data transformation on the normality of SELDI protein profiles.

It has been conjectured that the level of a gene expression and the pre-processing of the gene expression data in microarray are contributing factors to the lack of normality (Chen et al.,2005). In order to see if protein profiles exhibit similar behaviour, we examine the correlation between normality (as measured by the p-values from the normality test statistics, such as Shapiro-Wilks statistic) and protein expression levels.

The analysis results have shown that the newly proposed normalisation algorithm outperforms other existing methods. The rejection of the hypothesis of the normality of SELDI protein profiles by the goodness-of-fit tests implies that the SELDI mass spectra do not follow normal distributions. Therefore, it is unwise to employ the normal-theory based statistical methods to identify biomarkers from SELDI mass spectra.

5.2 Materials and statistical methods

5.2.1 Mass spectra

In this chapter, we use the ovarian data set 4-3-02, described in Section 2.7.4, to investigate the distribution of SELDI data at each data point and each peak, and to assess the effect of normalisation methods and data transformation methods on the distribution of the SELDI proteomic data. The data points with mass less than 2kDa were excluded because data points within this mass range were likely affected by matrix.

5.2.2 Intensity normalisation, transformation and normality tests

There are two ways in which intensities can be normalised. One is to standardise intensities, aiming at correcting for systematic differences in the total amount of protein desorbed from the sample plate, and another is to create a more normally distributed variable, aiming at improving the normality of intensities. We call the former intensity normalisation, and the latter intensity transformation.

5.2.2.1 Intensity normalisation

In this section we describe the two existing normalisation methods, mean AUC and median AUC, and propose a new ion current normalisation method, called median of ion current (MIC).

5.2.2.1.1 Mean AUC (TIC)

The implementation of the mean AUC normalisation method used in Ciphergen ProteinChip Software 3.2 (Ciphergen Biosystems, UK) consists of five steps. It is summarised as follows.

1. Calculate the TIC, denoted by s_i, $i = 1,2,...,n$, where, n is the number of spectra to be normalized and i is the index of a spectrum, s_i is the sum of intensities at all m/z values in the normalisation range for each spectrum.
2. Calculate the average ion current by dividing the TIC by the number of m/z data points $a_i = \frac{s_i}{n_i}$, $i = 1,2,...,n$
3. Calculate the normalisation coefficient across all selected spectra, $C = mean(a_i) = \frac{\sum_{i=1}^{n} a_i}{n}$
4. Calculate the normalisation factor for each spectrum by dividing the normalisation coefficient by the average ion current for each spectrum $N_i = \frac{C}{a_i}$
5. Scale each spectrum by multiplying intensities at each m/z data point by its normalisation factor.

5.2.2.1.2 Median AUC

The implementation of the median AUC normalisation method in the software package PROcess in BioConductor (Bioconductor,2007a) consists of four steps. It is summarised as follows.

1. Calculate the sum of the intensities for each spectrum for a pre-specified m/z range, and denoted as s_i, $i = 1,2,...,n$, where, n is the number of spectra to be normalized and i is the index of a spectrum
2. Calculate the normalisation coefficient $C = median(s_i | i = 1,2,...,n)$

3. Calculate the normalisation factor for each spectrum $N_i = \frac{C}{s_i}$

4. Scale each spectrum by multiplying the intensities values at each data point by its normalisation factor.

5.2.2.1.3 MIC

The MIC algorithm is proposed by adapting the mean AUC and median AUC methods. The difference between the mean AUC and the MIC is that the former calculates the average of TIC over all m/z values for each spectrum and takes the mean of the average of TICs across all spectra to be normalised as the normalisation coefficient while the latter calculates the median of the ion currents over all m/z values for each spectrum and takes the median of the medians of the ion currents across all spectra as the normalisation coefficient. The difference between the median AUC and the MIC is that the former calculates the sum of intensities over all m/z values for each spectrum and takes the median of the sums of intensities across all spectra as the normalisation coefficient while the latter calculates the median of the ion currents over all m/z values for each spectrum and takes the median of the medians of the ion currents across all spectra as the normalisation coefficient.

For each spectrum selected for normalisation, the MIC method calculates the median intensity in a specified m/z range. It then calculates the median intensity, called normalisation coefficient, across all the spectra selected for normalisation. The algorithm then normalises each spectrum by scaling the intensities by its normalisation factor, which is equal to the normalisation coefficient divided by the median ion current. The implementation of the MIC algorithm is summarised as follows.

1. Calculate the MIC, denoted by md_i, this is the median intensity from all m/z values in the normalisation range for each spectrum, $i = 1,2,...,n$, where, n is the number of spectra to be normalized and i is the index of a spectrum.
2. Calculate the normalisation coefficient across all selected spectra, $C = median(md_i \big| i = 1,2,...,n)$
3. For each spectrum calculate the normalisation factor N_i, by dividing the normalisation coefficient by the median ion current for this spectrum, obtaining $N_i = \dfrac{C}{md_i}$
4. Scale each spectrum by multiplying intensities at each m/z data point by its normalisation factor.

The MIC normalisation method and the two existing normalisation methods are used to normalise the spectra in the ovarian SELDI data. The performance of these normalisation methods is then compared.

5.2.2.2 Intensity transformation

Data transformations are the commonly used tools for improving the normality of data and for removing or reducing systematic bias in the data (Osborne,2002). We study three types of transformation approaches to investigate the effect of transformation on the normality of SELDI data: the logarithmic transformation, the Box-Cox family transformation, and a variance stabilising transformation based on the arsinh function. The performance of these transformations is assessed by applying all transformation functions to the normalised and un-normalised protein profiles in the data set.

5.2.2.2.1 Box-Cox family transformation

The Box-Cox family transformation is defined as follows.

$$T(Y) = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda}, & if \quad \lambda \neq 0 \\ \ln(Y), & if \quad \lambda = 0 \end{cases} \qquad \textbf{(5.1)}$$

The parameter λ in formula enables the function to mimic many standard transformations. For example, $\lambda = 0.5$ is equivalent to the square root transformation, and $\lambda = 2$ results in the squared transformation.

5.2.2.2.2 Logarithmic transformation

Logarithms to base10 have been used in a number of published studies to transform MS data prior to statistical analysis (Baggerly et al.,2003; Zhang et al.,2006). The logarithmic transformation to base e is a special case of Box-Cox ($\lambda = 0$) transformation. Changing the base of the logarithm has the effect of multiplying by a constant and so will not affect the normality test results. However, the study by Osborne (2002) investigated the effect of the base of logarithms on the efficacy of transformations and demonstrated that a lower base (base 2) logarithmic transformation served resolution of the data. Therefore, apart from using logarithmic transformations to bases e (Box-Cox ($\lambda = 0$)) transformation, we investigate the effect of logarithmic transformations to bases 2 on the SELDI data in this thesis. To avoid numerical redundancies arising from taking logarithms of negative intensities, we added one plus the absolute value of the smallest negative intensity observed before taking logarithms. The function *box.cox* in ***car*** package of R (Wu,2007) is used to perform the Box-Cox transformation. The λ values are studied in the range between -2 and 2 in steps of 0.5.

5.2.2.2.3 Arsinh function transformation

The variant arsinh function with a form of $f(x) = \log_2(x + \sqrt{x^2 + 1}) - \log_2(2)$ was used in analysing SELDI data (Beyer et al.,2006). For large intensities, this transformation becomes equivalent to the logarithmic transformations to bases 2 as $\lim_{x \to \infty}(f(x) - \log_2(x)) = 0$. However, unlike the logarithmic transformations, it does not have a singularity at zero, and continues to be smooth in the range of small intensities. Furthermore, it stabilises the variance and is well-defined even for negative intensity values introduced into the data by baseline corrections (Beyer et al.,2006; Huber et al.,2002).

5.2.2.3 Normality tests

5.2.2.3.1 Gaussian (normal) distributions

Gaussian (normal) distributions are a family of distributions that are characterised by a bell-shaped, symmetric curve, with values more abundant around the mean and progressively fewer observations towards the tail. This distribution is one of the most important probability density functions because many sample populations from random events tend to approximate to a normal distribution. The normal distribution is defined by two parameters: the mean μ and the standard deviation σ and has a density function as follows.

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad \textbf{(5.2)}$$

A standard normal variate $Z = \frac{X - \mu}{\sigma}$ has μ = 0 and σ^2 =1, and its corresponding density function is given by

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} \qquad \textbf{(5.3)}$$

A plot of this density function is shown in Figure 5-1. The shape of this curve can assist in determining whether or not a variable follows a normal distribution. The probability density function has notable properties, including (1) symmetry about mean μ, (2) the mode and median both equal to the mean μ, and (3) the inflection points[14] of the curve occur one standard deviation away from the mean, that is $\mu - \sigma$ and $\mu + \sigma$ (Dekking et al.,2007).

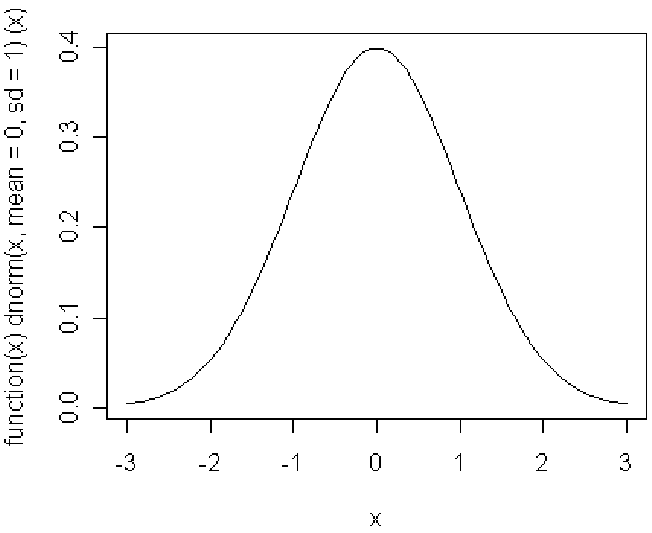


Figure 5-1 A standardised normal distribution

[14] An inflection point is a point on a curve at which the curvature changes sign.

The protein expression profile can be viewed as a random variable X . Suppose that it has mean μ, and variance σ^2, and that this protein profile is observed across n spectra, then the variance is defined by

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{n} \qquad \textbf{(5.4)}$$

A distribution is symmetric if it looks the same to the left side and right side of the centre point (Dekking et al.,2007). One measure that can be used to see whether a distribution is symmetric is skewness. More precisely, skewness is a measure of the asymmetry of a distribution. The skewness of a distribution with a mean of μ and a standard deviation of σ is defined by

$$sk_p = \frac{\sum (X_i - \mu)^3}{n\sigma^3} \qquad \textbf{(5.5)}$$

A value of skewness close to zero indicates that data are symmetrically distributed around the mean, otherwise it is skewed. A negative value of skewness indicates that data are skewed left, implying that the left tail is long relative to the right tail. A positive value of skewness indicates that data are skewed right, implying that the right tail is long relative to the left tail (Dekking et al.,2007) as shown in Figure 5-2.

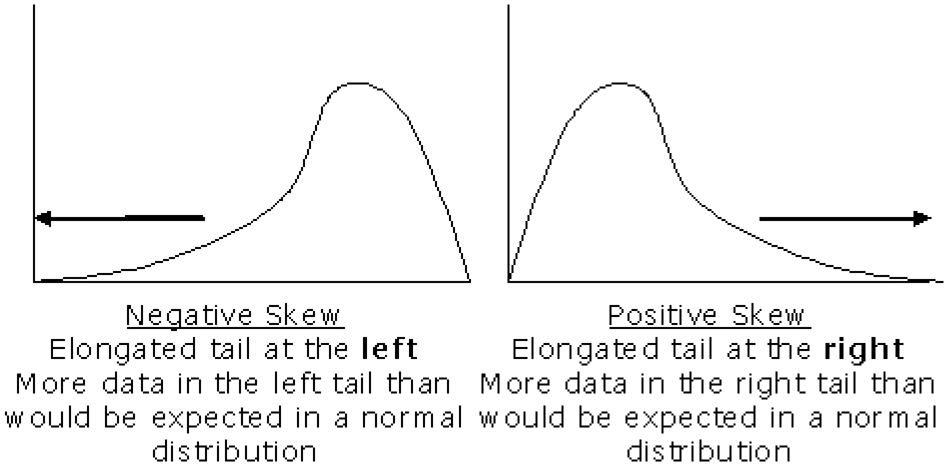

Negative Skew
Elongated tail at the **left**
More data in the left tail than would be expected in a normal distribution
Positive Skew
Elongated tail at the **right**
More data in the right tail than would be expected in a normal distribution

Figure 5-2 Positive skew and negative sknew.

Kurtosis is a measure that can be used to see whether the data are peaked or flat relative to a normal distribution. A data set with high kurtosis tends to have a distinct peak near the mean, declines rather rapidly, and has heavy tails. By contrast, a data set with low kurtosis tends to have a flat top, rather than a sharp peak near the mean (Dekking et al.,2007). The kurtosis of a distribution with a mean of μ and a standard deviation of σ is defined by

$$k = \frac{\sum (X_i - \mu)^4}{n\sigma^4} - 3 \qquad \textbf{(5.6)}$$

The "minus 3" at the end of kurtosis formula is a correction factor to make the kurtosis of the normal distribution equal to zero (Dekking et al.,2007; Joanes & Gill,1998). A value of kurtosis that is less than zero indicates a platykurtic (flat) distribution. A value of kurtosis that is great than zero indicates a leptokurtic (peaked) distribution. If the value of kurtosis is close to zero then the intensity distribution is normal or mesokurtic.

The population mean, standard deviation, skewness and kurtosis are usually unknown. They can be estimated by sample mean, sample standard deviation, sample skewness, and sample kurtosis. The sample mean, standard deviation, skewness, and kurtosis of the proteomic intensities are defined as follows.

$$\overline{X} = \frac{\sum X_i}{n} \qquad \textbf{(5.7)}$$

$$s = \frac{\sum (X_i - \overline{X})^2}{n-1} \qquad \textbf{(5.8)}$$

$$sk = \frac{\sum (X_i - \overline{X})^3}{(n-1)s^3} \qquad \textbf{(5.9)}$$

$$k = \frac{\sum (X_i - \overline{X})^4}{(n-1)s^4} - 3 \qquad \textbf{(5.10)}$$

The parametric methods, such as the t-test, have been used to identify significant peaks from SELDI mass spectra (Levner,2005; Pasinetti,2006; Wu et al.,2003). These tests make an assumption that SELDI data follows a normal distribution. To investigate the validity of these results, we carried out a large-scale SELDI normality study using the Shapiro-Wilks, Kolmogorov-Smirnov, Cramér-von-Mises, and Pearson χ^2 tests.

5.2.2.3.2 Shapiro-Wilks test

The Shapiro-Wilks test is a procedure for testing that a random sample comes from a normal distribution (Bai & Cheng,2003; Royston,1993; Royston,1995; Shapiro et al.,1965). Suppose that $I_1 < I_2 < ... < I_n$ is a vector of ordered random observations to be tested for normality, clearly if $\{I_j\}$ are a normal sample, then I_j can be expressed by $I_j = \mu + \sigma . x_j$, where, μ and σ are unknown mean and standard deviation of a normal distribution, and $x_1 \leq x_2 \leq ... x_n$ is an ordered random sample of size n from a normal distribution with mean 0 and variance 1. The Shapiro-Wilks W test statistic is defined by

$$W = \frac{(\sum_{j=1}^{n} a_j I_j)^2}{\sum_{j=1}^{n} (I_j - \bar{I})^2} \qquad \textbf{(5.11)}$$

where, $\bar{I} = \frac{1}{n} \sum_{j=1}^{n} I_j$, and $a_j, j = 1,2,...,n$ are a vector of weights and are determined by

$$(a_1, a_2, ..., a_n) = \frac{m'V^{-1}}{\sqrt{(m'V^{-1}V^{-1}m)}} \qquad \textbf{(5.12)}$$

where, $m' = (m_1, m_2, ..., m_n)$ denotes the vector of the expected values of the standard normal ordered statistics $x_1 \leq x_2 \leq ... x_n$ and that $V = (v_{ij}) = \text{cov}(x_i, x_j)$ denotes the covariance matrix (Royston,1993; Royston,1995; Shapiro et al.,1965).

A small value of W indicates a departure from normality and is the evidence of rejection of the null hypothesis (Shapiro et al.,1965).

5.2.2.3.3 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test provides a means of testing whether a set of observations are from a completely specified distribution (Birnbaum,1952; Dallal & Wilkison,1986; Lilliefors,1967; Massey,1951). Given a sample of n observations, the

Kolmogorov-Smirnov statistic is defined by $D = Max_I \left| S_n(I) - F^*(I) \right|$, where, $S_n(I)$ is the sample cumulative distribution function, $F^*(I)$ is the expected cumulative normal distribution with mean $\mu = \bar{I}$ (the sample mean), and standard deviation $\sigma = s$ (the sample standard deviation) (Dallal & Wilkison,1986; Lilliefors,1967; Massey,1951).

Let I be a random variable with the continuous distribution function defined by $S_n(i) = \Pr ob(I \le i)$, and $I_1 < I_2 < ... < I_n$ be a sample size of n for I, $S_n(i)$ can then be defined by the following step function (Birnbaum,1952; Juergen,2007a).

$$S_n(i) = \begin{cases} 0, \quad for \quad i < I_1 \\ \frac{j}{n}, \quad for \quad I_j \le i < I_{j+1}, j = 1,2,...,n-1 \\ ... \\ 1, \quad for \quad i \ge I_n \end{cases} \qquad \textbf{(5.13)}$$

$F^*(i)$ can be expressed by $F^*(i) = p(j) = \Phi([I_j - \bar{I}]/s)$, where Φ is the cumulative distribution function of the standard normal distribution. The Kolmogorov-Smirnov statistic D can then be re-expressed by

$$\begin{aligned} &D = \max\{D^+, D^-\} \\ &D^+ = \max\{\frac{j}{n} - p(j) \mid j = 1,2,...,n\} \\ &D^- = \max\{p(j) - \frac{j-1}{n} \mid j = 1,2,...,n\} \\ &p(j) = \Phi([I_j - \bar{I}]/s) \end{aligned} \qquad \textbf{(5.14)}$$

If the p-value of the Kolmogorov-Smirnov test is greater than 0.1, then it is computed from the following modified statistic as the test is only reliable when the p-value is smaller than 0.1 (Dallal & Wilkison,1986; Juergen,2007a).

$$Z = D(\sqrt{n} - 0.01 + 0.85/\sqrt{n}) \quad \textbf{(5.15)}$$

5.2.2.3.4 Cramér-von-Mises test

The Cramér-von-Mises test is used to judge the goodness of fit of one probability distribution compared to another probability distribution. The Cramér-von-Mises statistic is defined by

$$\begin{aligned} W^2 &= \int_{-\infty}^{\infty} [S_n(i) - F^*(i)]^2 dF^*(i) \\ &= \frac{1}{12n} + \sum_{j=1}^{n} (F^*(I_j) - \frac{2j-1}{2n})^2 \quad \textbf{(5.16)} \\ &= \frac{1}{12n} + \sum_{j=1}^{n} (p(j) - \frac{2j-1}{2n})^2 \end{aligned}$$

where, $S_n(i) = \Pr ob(I \le i)$ is the sample cumulative distribution function, $F^*(i) = p(j) = \Phi([I_j - \bar{I}]/s)$ is the expected cumulative normal distribution with mean $\mu = \bar{I}$ (the sample mean), and standard deviation $\sigma = s$ (the sample standard deviation), Φ is the cumulative distribution function of the standard normal distribution (Schmid & Trede,1996; Xiao et al.,2006).

If the value of the Cramér-von-Mises statistic w^2 is larger than the tabulated value we can reject the hypothesis that the data come from a normal distribution (Xiao et al.,2006).

5.2.2.3.5 Pearson χ^2 test

The Pearson χ^2 test is used to evaluate whether a sample comes from a distribution of a given form, by counting the number of observations falling into specified cells (Chernoff & Lehmann,1954; Stephens,1986). For the chi-square goodness-of-fit computation, the observed data are divided into k bins and the test statistic is defined as follows.

$$\chi^2 = \sum_{b=1}^{k} (O_b - E_b)^2 / E_b \quad \textbf{(5.17)}$$

where, O_b is the observed frequency for bin b, and E_b is the expected frequency for bin b and is defined by

$$E_b = n(F^*(I_b^u) - F^*(I_b^l)) \quad \textbf{(5.18)}$$

where, $F^*(I)$ is the expected cumulative normal distribution, n is the sample size, I_b^u is the upper limit, and I_b^l is the lower limit for bin b.

The null hypothesis of normality is rejected if $\chi^2 > \chi^2_{(\alpha,k-c)}$, where, α is a significant level, k is the number of classes, and c is the number of estimated parameters plus 1 (Chernoff & Lehmann,1954; Juergen,2007b). In the case of normality test, $c = 3$ as there are two parameters (the mean and the standard deviation) need to be estimated.

We applied the four statistical tests to both the normalised and un-normalised ovarian data sets to investigate the normality of the proteomic data. Some published studies also used peaks rather than data points to identify differentially expressed proteins and classify disease status (Bhanot et al.,2006; Yasui et al.,2003). Therefore, the normality tests were applied to both peaks and data points in this chapter. We extracted peaks using the BioConductor software package (PROcess) in R (Bioconductor,2007a). To make the results comparable, the same parameter settings were applied to both normalised and un-normalised spectra in peak selections. The lower bound for signal/noise ratio was set to 2, the minimum intensity 1 and the total area under the curve 0.01. The ovarian data set 4-3-02 contains samples of SELDI protein profiles generated by each of 10361 m/z values. A total of 160 peaks were detected by the PROcess peak picking program.

All the statistical tests were implemented in R statistical software (CRAN,2007) through the packages *nortest* and *stats*. A p-value significance threshold of 0.05 was used in all tests.

5.2.2.4 Multiple goodness-of-fit tests

These normality tests were actually performed at hundreds or thousands of peaks or data points, which involve multiple tests. In order to make a decision on whether SELDI data follows a normal distribution, we need to assess the goodness-of-fit testing. The conventional multiple testing adjustment methods, such as FWER and FDR, cannot be used for this purpose because they were designed for controlling FWER and FDR. The study by Chen et al. (2005) developed a special procedure of multiple testing adjustment for microarray data. The testing procedure involving m genes can be thought of as a sequence of m random trials with stochastically dependent outcomes (Chen et al.,2005). Let p be the probability of successful rejection of the null hypothesis in a given trial and let X be the number of successes in m trials, and let α be the significance level. Under the complete null hypothesis, the expected number of rejections is equal to $m\alpha$. The proposed resampling procedure is to test the hypothesis $H_0 : E(X) = m\alpha$ versus the alternative hypothesis $H_1 : E(X) > m\alpha$. It used the delete-d-jackknife[15] subsampling to estimate the sampling distribution of the shift test statistic $T = X - E(X)$. The cumulative distribution function $H(x)$ of the statistic T was estimated by

$$\hat{H}_{JACK}(x) = \frac{1}{N}\sum_{j=1}^{N} I\{\sqrt{\frac{r}{d}(T_{r,s_j} - T)} \le x\} \quad \textbf{(5.19)}$$

where, N was the number of subsamples, I was the indicator function, having a value of 1 if $\sqrt{\frac{r}{d}(T_{r,s_j} - T)} \le x$, and 0 otherwise. s_j was the collection of the subsamples of size d, which was drawn without replacement from $\{1,2,...,n\}$. $T_{r,s} = T_r(X_i, i \in s^c)$, where, s was a subset of $\{1,2,...,n\}$ with size d, s^c was the complement of s, d was an integer depending on n, $1 \le d \le n$, and $r = n - d$. The jackknife estimator $\hat{H}_{JACK}(x)$ was evaluated at the observed statistic value $x = X - m\alpha$ under the complete null hypothesis to produce an overall p-value associated with multiple tests (Chen et al.,2005).

In order to adjust for multiplicity of the tests, we adapted the method proposed by Chen et al. (2005) to evaluate the global null hypothesis of a normal distribution for SELDI data. Let $\alpha = 0.05$ be the significance level. Under the complete null hypothesis, the expected number of rejections is equal to 518 (10361*0.05) for data points and 8 (160*0.05) for peaks. The proposed resampling procedure is to test the

[15] Instead of leaving out one observation at a time, we leave out *d* observations. Therefore, the size of a delete-*d* jackknife sample is (*n* - *d*), and there are C_n^d jackknife samples.

hypothesis $H_0 : E(X) = 518$ versus the alternative hypothesis $H_1 : E(X) > 518$ for data points, and $H_0 : E(X) = 8$ versus the alternative hypothesis $H_1 : E(X) > 8$ for peaks.

We evaluate the utility of the four goodness-of-fit test statistics in describing the normality of SELDI data with or without pre-processing using various transformation and normalisation methods.

5.2.2.5 Distortion of expression profiles from normality

Apart from formal quantitative analysis of the normality of SELDI protein profiles using the statistical tests, visual inspection was also used to investigate the extent of the deviation of the protein profiles from normality. We used frequency histograms and Q-Q plots as aids to visual investigation of the normality of the protein profiles. Both graphs provide simple and effective visualisation of the data distribution and of any deviation from normality. For a perfect normal distribution, a Q-Q plot would show data points in a straight line near the centre with a positive gradient and a frequency histogram would show a bell-shaped frequency distribution which is symmetrical about the mean. The *graphics* and *stats* package in R are used to perform this analysis.

5.2.2.6 Effect of transformation on identification of differentially expressed protein profiles

Differentially expressed proteins can be identified by comparing the intensity for each protein using different statistical methods. These proteins could be potential biomarkers. The RF algorithm provides a routine for identifying the biomarkers (as described in Section 6.2.2.2.1). We assessed the effect of transformations listed in the previous section on the identification of biomarkers using this routine. Venn diagrams were used to explore the protein profiles detected to be differentially expressed for both the transformed and the un-transformed data

5.2.2.7 Effect of transformation on sample clustering

For many high-throughput assays the technical variance in data tends to rise with the measurement level (e.g. the intensity levels in SELDI) (Purohit et al.,2004). For the same percentage of error, a bigger value of the intensity means a bigger absolute error. The logarithmic transformation of the measured values is a method commonly used to convert multiplicative error into additive error and therefore stabilises the variance of the data (Thygesen & Zwinderman,2004). The question raised here is whether data transformation improves the normality of the data and whether it also leads to improvement in the performance of sample clustering. The principal component analysis (PCA) and RF were applied to the ovarian cancer data set to examine whether the logarithmic transformation improve the sample separation performance.

The PCA method constructs new variables known as the principle components that are linear combinations of the proteomic profiles. A plot of the projection of the sample positions along the first and second principals (also known as the scores plot) shows the relationship between the samples. Samples clustered together on the plot indicate similar protein profiles. RF classifier is a collection of many decision trees based on distribution free statistics and the output of RF is the majority votes of the trees, which have been described in Section 4.2.3.2. The function *princomp* of *MASS* package and ***randomForest*** from R (Wu,2007) are used to perform these analyses.

5.2.2.8 Correlation between protein expression level and normality

In order to investigate whether the level of a protein profile is related to their normality, a scatter plot is generated to show the relationship between p-values by the normality tests and the median values of protein intensity across all samples for the mean AUC normalised SELDI data set.

5.3 Results of normalisation, transformation and normality tests

5.3.1 Normality tests

The outcomes of the four statistical tests for normality using the un-normalised and normalised data sets under different transformation functions are summarised in Table 5-1, Table 5-2, Table 5-3, Table 5-4, Table 5-5, Table 5-6, Table 5-7, and Table 5-8, respectively. Table 5-1, Table 5-2, Table 5-3, and Table 5-4 summarise the percentage of rejections of the null hypothesis of normality for the 160 peaks detected in the ovarian cancer data set by PROcess peak detection software and Table 5-5, Table 5-6, Table 5-7, and Table 5-8 show the results for the 10361 data points. For all data transformations considered, a large proportion of protein profiles were not normally distributed, that is, had p-value less than 0.05. A large number of rejections were detected in the cancer group for both the normalised and un-normalised data sets under various data transformations. The asterisk in Table 5-1, Table 5-2, Table 5-3, Table 5-4, Table 5-5, Table 5-6, Table 5-7, and Table 5-8 show transformations and tests that resulted in the lowest percentage of rejections of the null hypothesis of normality in the control and the cancer groups.

Table 5-1 shows that among the four normality test results, logarithmic transformations give the best improvement of normality for the cancer protein profiles, whereas, the Box-Cox transformation (λ=-0.5) gives the best improvement for the protein profiles from the control group.

Table 5-2 shows that when the mean AUC normalisation method is used, the logarithmic transformations, including log to base 2 and natural logarithm (Box-Cox (λ=0)), in general give the best improvement of normality for the control group, whereas, Box-Cox (λ=-1, λ=-2) transformations give the best improvement for the cancer protein profiles under this normalisation. When the data is normalised by the median AUC, it turns out that the logarithmic transformations give the best improvement of normality for the protein profiles in the control group, whereas, Box-Cox (λ=-1, λ=-2) transformations appear to improve the normality for the cancer profiles as shown in Table 5-3.

Table 5-4, which displays the results of the normality analysis for the MIC normalised data, shows that the logarithmic transformations give the best improvement of normality of protein profiles for both the control and the cancer groups. Table 5-5 shows that logarithmic and arsinh transformations give the best improvement of normality for the cancer mass spectra and Box-Cox (λ=-0.5) gives the best improvement for the control mass spectra. Table 5-6 shows that logarithmic and arsinh transformations give the best improvement of normality for the control mass spectra and non-transformation gives the best improvement for cancer mass spectra. Table 5-7 shows that logarithmic transformations give the best improvement of normality for the control mass spectra and non-transformation gives the best improvement for the cancer mass spectra. Table 5-8 shows that logarithmic transformations give the best improvement of normality for the both control and cancer mass spectra. In summary, these results indicate that the logarithmic transformations give the best improvement of normality of the SELDI data. Table 5-9 summarised the percentage of rejections of the null hypothesis of the normal distribution under the logarithmic transformations with and without normalisation based on the data points and peaks.

Comparing the average number of rejections by the four normality tests between the data point data set and peak data set, we can see that peak selection can improve the normality of SELDI protein profiles generated from the control samples, but this is not true for cancer samples. A look at the numbers of rejections across the four normality tests for each of the normalisation methods (Table 5-9) indicates that normalisation has some effect on the normality of protein profiles. The MIC normalisation method appears to results in lower rejection rates across statistical tests and tissue types (cancer and normal controls) compared to other normalisation methods. However, the generality of this statement needs to be verified by large scale studies. On average, rejection rates appear to be higher for the cancer samples compared to the normal control samples. The average numbers of rejections by the four normality tests (from peaks data set) are 70% for the un-normalised data set, 90% for the mean AUC normalised data set, 90% for the median AUC normalised data set, and 69% for the MIC normalised data set. These figures are much lower for the normal control samples.

In summary, the normalisation methods in PROcess (the median AUC) and Ciphergen ProteinChip Software (the mean AUC) tend to give similar results. The MIC normalisation followed by the logarithmic transformation appears to give lower number of rejections and hence, the best performance in improving the normality of the SELDI protein profiles.

5.3.2 Distortion of SELDI expression profiles from normality

Table 5-9 shows that the rates of normality rejections vary from 55% to 95% by the Shapiro-Wilks test, from 38% to 93% by the Cramér-von-Mises test, from 30% to 87% by the Kolmogorov-Smirnov test and from 25% to 86% by the Pearson χ^2 test. Thus for this data set, the Shapiro-wilks normality test appears to be the most conservative, while the Pearson χ^2 normality test tends to be liberal. These are summaries of multiple normality tests on the 10361 data points and 160 proteins detected for the ovarian cancer data set, and we need to adjust for this multiplicity in order to make inference about their significance. Application of the global multiple testing procedure (Chen et al.,2005) to these test results, gave very small p-values (p-value ≈ 0) for each of them, indicating the overall rejection of the normal distribution of the SELDI protein profiles.

Table 5-10 shows deviation from normality, described by their skewness and kurtosis statistics, of six protein profiles with small p-values obtained by the Shapiro-Wilks tests on the control and the cancer samples, respectively. For all the peaks listed in that table, the values of the skewness statistics show greater deviations from 0 before logarithmic transformation, indicating that the distributions are skewed to the left with long right tails. The values of kurtosis for these peaks are also significantly greater than 0, indicating the distributions are leptokurtic, with low spread out peaks with data concentred in the tails. However, the values of skewness and kurtosis are much reduced toward zero after logarithmic transformation. To visualise these results, we constructed frequency histograms and the Q-Q normal plots. The frequency histograms (Figure 5-3 (a)) confirm that the selected proteins are characterised by having right-skewed and heavy-tailed distributions. The Q-Q plots (Figure 5-3 (b)) show that some data points are far away from a straight line near the centre, suggesting that positive skew was evident in the data distribution before and after logarithmic transformation. The frequency histograms (Figure 5-3 (c)) and Q-Q plots (Figure 5-3 (d)) show that there is improvement in normality of the protein profiles after logarithmic transformation. The Shapiro-Wilks test still show evidence of lack normality for these peaks (p-value < 0.05), but with decreased significance.

Table 5-1 The percentage of rejections of the null hypothesis for 160 peaks (un-normalised data).

Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)
Control	Raw data	100	99	97	98
	Log2	56	44	33	25
	Arsinh	62	51	41	29
	Box-Cox(-2)	97	92	86	89
	Box-Cox(-1)	59	50	37	34
	Box-Cox(-0.5)*	34*	23*	18*	14*
	Box-Cox(0)	56	44	33	25
	Box-Cox(0.5)	97	92	86	71
	Box-Cox(1)	100	99	97	98
	Box-Cox(2)	100	80	100	100
Cancer	Raw data	93	88	89	85
	Log2*	86*	69*	67*	57*
	Arsinh	87	70	67	59
	Box-Cox(-2)	82	69	69	65
	Box-Cox(-1)	82	77	69	58
	Box-Cox(-0.5)	86	80	67	60
	Box-Cox(0)*	86*	69*	67*	57*
	Box-Cox(0.5)	85	80	73	80
	Box-Cox(1)	93	88	89	85
	Box-Cox(2)	99	88	99	99

Asterisk indicates the lowest percentage of rejection of the null hypothesis in the control group and the cancer group under different transformations.
Note: the test results for raw data are equivalent to those for the Box-Cox (λ=1) transformed data.

Table 5-2 The percentage of rejections of the null hypothesis for 160 peaks (mean AUC normalised data).

Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)
Control	Raw data	96	81	73	75
	Log2*	56*	39*	32*	29*
	Arsinh	57	40	32	29
	Box-Cox(-2)	95	88	85	85
	Box-Cox(-1)	72	69	65	52
	Box-Cox(-0.5)	52	50	49	32
	Box-Cox(0)*	56*	39*	32*	29*
	Box-Cox(0.5)	75	54	39	45
	Box-Cox(1)	96	81	73	75
	Box-Cox(2)	100	86	98	96
Cancer	Raw data	96	92	91	91
	Log2	95	93	87	86
	Arsinh	95	94	87	86
	Box-Cox(-2)*	91*	84*	82*	80*
	Box-Cox(-1)*	89*	83*	84*	80*
	Box-Cox(-0.5)	95	88	85	86
	Box-Cox(0)	95	93	87	86
	Box-Cox(0.5)	95	94	92	89
	Box-Cox(1)	96	92	91	91
	Box-Cox(2)	98	84	94	91

Asterisk indicates the lowest percentage of rejection of the null hypothesis in the control group and the cancer group under different transformations.
Note: the test results for raw data are equivalent to those for the Box-Cox (λ=1) transformed data.

Table 5-3 The percentage of rejections of the null hypothesis for 160 peaks (median AUC normalised data).

Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)
Control	Raw data	95	80	72	73
	Log2*	55*	38*	30*	28*
	Arsinh	55	39	30	30
	Box-Cox(-2)	94	88	84	82
	Box-Cox(-1)	70	70	66	51
	Box-Cox(-0.5)	51	50	45	30
	Box-Cox(0)*	55*	38*	30*	28*
	Box-Cox(0.5)	74	51	39	45
	Box-Cox(1)	95	80	72	73
	Box-Cox(2)	100	87	98	95
Cancer	Raw data	95	91	91	90
	Log2	95	93	86	85
	Arsinh	95	94	87	85
	Box-Cox(-2)*	91*	82*	82*	79*
	Box-Cox(-1)*	89*	83*	86*	80*
	Box-Cox(-0.5)	95	87	84	85
	Box-Cox(0)	95	93	86	85
	Box-Cox(0.5)	95	94	91	89
	Box-Cox(1)	95	91	91	90
	Box-Cox(2)	98	84	93	90

Asterisk indicates the lowest percentage of rejection of the null hypothesis in the control group and the cancer group under different transformations.
Note: the test results for raw data are equivalent to those for the Box-Cox (λ=1) transformed data.

Table 5-4 The percentage of rejections of the null hypothesis for 160 peaks (MIC normalised data).

Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)
Control	Raw data	100	90	80	81
	Log2*	55*	41*	39*	31*
	Arsinh	56	41	39	31
	Box-Cox(-2)	94	86	88	87
	Box-Cox(-1)	78	80	76	59
	Box-Cox(-0.5)	58	52	44	43
	Box-Cox(0)*	55*	41*	39*	31*
	Box-Cox(0.5)	77	55	50	50
	Box-Cox(1)	100	90	80	81
	Box-Cox(2)	100	86	99	99
Cancer	Raw data	98	86	86	86
	Log2*	83*	65*	58*	70*
	Arsinh	83	66	58	70
	Box-Cox(-2)	78	73	69	66
	Box-Cox(-1)	84	77	64	54
	Box-Cox(-0.5)	81	68	55	58
	Box-Cox(0)*	83*	65*	58*	70*
	Box-Cox(0.5)	83	81	68	76
	Box-Cox(1)	98	86	86	86
	Box-Cox(2)	100	88	99	99

Asterisk indicates the lowest percentage of rejection of the null hypothesis in the control group and the cancer group under different transformations.
Note: the test results for raw data are equivalent to those for the Box-Cox (λ=1) transformed data.

Table 5-5 The percentage of rejections of the null hypothesis for 10361 data points (un-normalised data).

Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)
Control	Raw data	96	89	92	87
	Log2	80	74	70	63
	Arsinh	81	74	71	64
	Box-Cox(-2)	86	79	76	69
	Box-Cox(-1)	78	71	66	59
	Box-Cox(-0.5)*	73*	66*	64*	58*
	Box-Cox(0)	80	74	70	63
	Box-Cox(0.5)	93	86	85	76
	Box-Cox(1)	96	89	92	87
	Box-Cox(2)	97	77	95	92
Cancer	Raw data	86	79	76	67
	Log2*	76*	68*	64*	55*
	Arsinh	76	69	64	56
	Box-Cox(-2)	79	68	67	61
	Box-Cox(-1)	80	74	68	61
	Box-Cox(-0.5)	80	75	68	59
	Box-Cox(0)*	76*	68*	64*	55*
	Box-Cox(0.5)	74	69	66	59
	Box-Cox(1)	86	79	76	67
	Box-Cox(2)	93	87	90	86

Asterisk indicates the lowest percentage of rejection of the null hypothesis in the control group and the cancer group under different transformations.
Note: the test results for raw data are equivalent to those for the Box-Cox (λ=1) transformed data.

Table 5-6 The percentage of rejections of the null hypothesis for 10361 data points (mean AUC normalised data).

Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)
Control	Raw data	88	76	79	72
	Log2*	80*	70*	70*	63*
	Arsinh*	80*	70*	70*	63*
	Box-Cox(-2)	87	81	80	71
	Box-Cox(-1)	85	77	76	64
	Box-Cox(-0.5)	82	72	71	62
	Box-Cox(0)	80	70	70	63
	Box-Cox(0.5)	83	72	73	67
	Box-Cox(1)	88	76	79	72
	Box-Cox(2)	97	84	92	84
Cancer	Raw data*	89*	85*	86*	78*
	Log2	92	89	85	77
	Arsinh	92	89	85	77
	Box-Cox(-2)	94	91	89	81
	Box-Cox(-1)	93	90	87	79
	Box-Cox(-0.5)	92	89	86	78
	Box-Cox(0)	92	89	85	77
	Box-Cox(0.5)	91	86	85	77
	Box-Cox(1)	89	85	86	78
	Box-Cox(2)	92	86	88	81

Asterisk indicates the lowest percentage of rejection of the null hypothesis in the control group and the cancer group under different transformations.
Note: the test results for raw data are equivalent to those for the Box-Cox (λ=1) transformed data.

Table 5-7 The percentage of rejections of the null hypothesis for 10361 data points (median AUC normalised data).

Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)
Control	Raw data	88	76	79	72
	Log2*	80*	70*	70*	63*
	Arsinh	80	70	70	64
	Box-Cox(-2)	87	81	80	71
	Box-Cox(-1)	85	77	76	64
	Box-Cox(-0.5)	82	72	71	62
	Box-Cox(0)*	80*	70*	70*	63*
	Box-Cox(0.5)	83	72	73	67
	Box-Cox(1)	88	76	79	72
	Box-Cox(2)	97	84	92	84
Cancer	Raw data*	89*	85*	86*	78*
	Log2	92	89	85	77
	Arsinh	92	89	85	77
	Box-Cox(-2)	94	91	89	81
	Box-Cox(-1)	93	90	87	79
	Box-Cox(-0.5)	92	89	86	78
	Box-Cox(0)	92	89	85	77
	Box-Cox(0.5)	90	86	85	77
	Box-Cox(1)	89	85	86	78
	Box-Cox(2)	92	86	88	81

Asterisk indicates the lowest percentage of rejection of the null hypothesis in the control group and the cancer group under different transformations.

Note: the test results for raw data are equivalent to those for the Box-Cox (λ=1) transformed data.

Table 5-8 The percentage of rejections of the null hypothesis for 10361 data points (MIC normalised data).

Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)
Control	Raw	96	92	87	76
	Log2*	79*	72*	67*	57*
	Arsinh	79	72	67	58
	Box-Cox(-2)	85	80	75	64
	Box-Cox(-1)	83	77	70	59
	Box-Cox(-0.5)	79	74	67	56
	Box-Cox(0)*	79*	72*	67*	57*
	Box-Cox(0.5)	90	83	76	66
	Box-Cox(1)	96	92	87	76
	Box-Cox(2)	97	93	93	84
Cancer	Raw	88	76	70	63
	Log2*	76*	68*	65*	58*
	Arsinh	76	68	65	58
	Box-Cox(-2)	84	77	72	64
	Box-Cox(-1)	84	78	73	63
	Box-Cox(-0.5)	83	75	70	60
	Box-Cox(0)*	76*	68*	65*	58*
	Box-Cox(0.5)	76	69	66	59
	Box-Cox(1)	88	76	70	63
	Box-Cox(2)	93	86	86	79

Asterisk indicates the lowest percentage of rejection of the null hypothesis in the control group and the cancer group under different transformations.
Note: the test results for raw data are equivalent to those for the Box-Cox (λ=1) transformed data.

Table 5-9 Summarisation of the percentage of rejection of the null hypothesis of the normal distribution under the logarithmic transformation with and without normalisation based on the data points and peaks data sets

Data set	Group	Methods	Shapiro-Wilks (%)	Cramér-von-Mises (%)	Kolmogorov-Smirnov (%)	Pearson χ^2 (%)	Avg. (%)
Peaks	Control	Un-normalised	56	44	33	25	40
		Mean AUC normalised*	56*	39*	32*	29*	39
		Median AUC normalised*	55*	38*	30*	28*	38
		MIC normalised*	55*	41*	39*	31*	42
	Cancer	Un-normalised*	86*	69*	67*	57*	70
		Mean AUC normalised	95	93	87	86	90
		Median AUC normalised	95	93	87	86	90
		MIC normalised*	83*	65*	58*	70*	69
Points	Control	Un-normalised	80	74	70	63	72
		Mean AUC normalised*	80*	70*	70*	63*	71
		Median AUC normalised*	80*	70*	70*	63*	71
		MIC normalised*	79*	72*	67*	57*	69
	Cancer	Un-normalised*	76*	68*	64*	55*	66
		Mean AUC normalised	92	89	85	77	86
		Median AUC normalised	92	89	85	77	86
		MIC normalised*	76*	68*	65*	58*	67
Asterisks indicate that of different transformation methods the logarithmic transformation methods achieve the lowest percentage of rejections. Avg stands for average.							

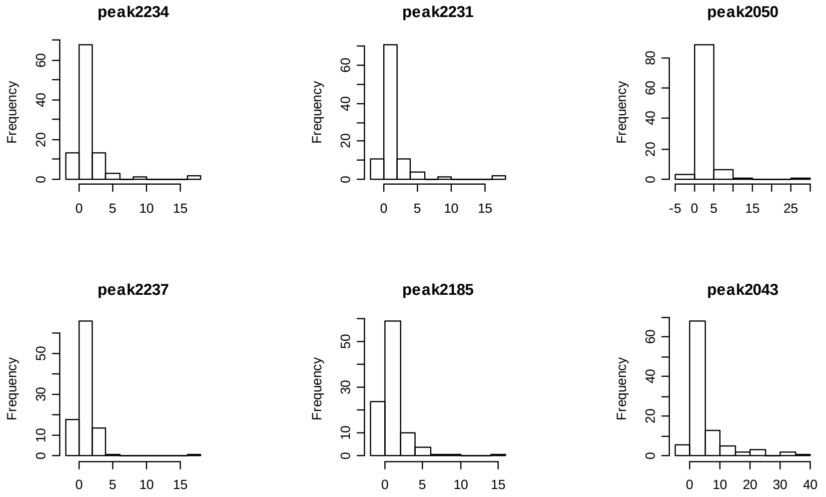
Table 5-10 The skewness and kurtosis of peaks with small p-values before logarithmic transformation (Shapiro-Wilks test)

	LOG	**Peaks (m/z)**	**M2234**	**M2231**	**M2050**	**M2237**	**M2185**	**M2043**
Control	Before	P-value	1.76E-16	4.53E-16	8.03E-16	8.80E-16	1.62E-14	3.16E-14
		Skewness	4.2	4.0	4.7	5.3	3.5	2.8
		Kurtosis	20.6	19.0	30.2	38.4	16.5	8.1
	After	P-value	6.75E-10	2.57E-9	3.22E-7	2.54E-7	1.46E-8	8.54E-6
		Skewness	2.0	1.9	1.6	1.6	1.2	1.1
		Kurtosis	5.7	4.8	4.3	5.4	3.5	0.8
Cancer		**Peaks (m/z)**	**M2183**	**M2114**	**M2050**	**M2175**	**M2185**	**M3298**
	Before	P-value	1.34E-16	1.81E-16	5.05E-16	5.86E-16	1.03E-15	1.12E-15
		Skewness	4.4	4.0	4.6	4.4	4.7	3.6
		Kurtosis	22.7	18.4	27.2	23.4	28.9	13.9
	After	P-value	2.46E-14	1.90E-13	4.17E-11	5.21E-9	1.1E-12	1.09E-11
		Skewness	3.5	3.0	2.5	3.3	3.3	2.4
		Kurtosis	15.0	10.4	8.4	14.6	15.5	6.6

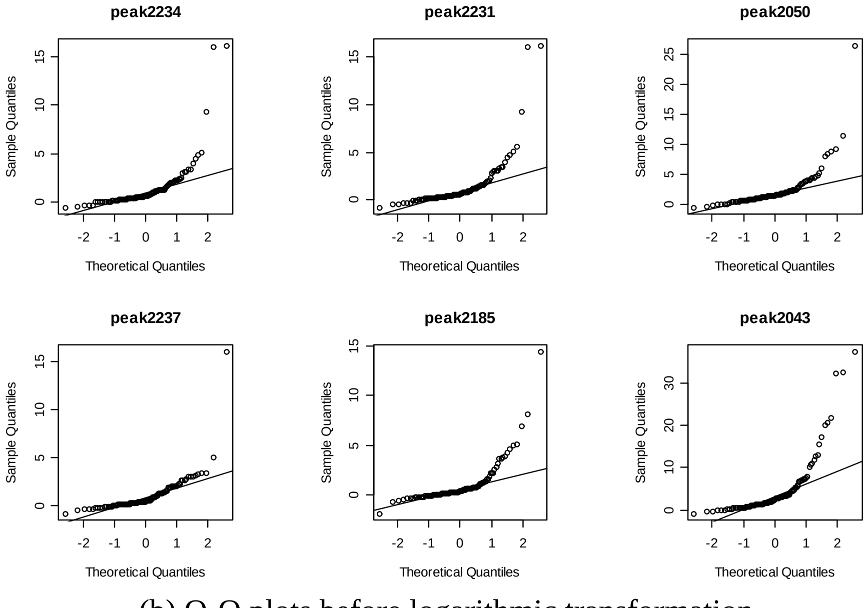
5.3.3 Effect of transformation on identification of differentially expressed protein profiles

The Venn diagrams in Figure 5-4 show the relationship between the significant protein profiles identified after logarithmic transformation to base 2 and those identified before the transformation. The Venn diagram in the right panel of Figure 5-4 (a) shows a huge overlap (11 out of 12 proteins) between the top 12 significant protein profiles identified with and without logarithmic transformation. The left panel of Figure 5-4 (a) shows the scatter plots of two sets of significant protein profiles identified using the data sets with and without logarithmic transformation. Each set of significant protein profiles are ranked by the p-values of Wilcoxon tests.

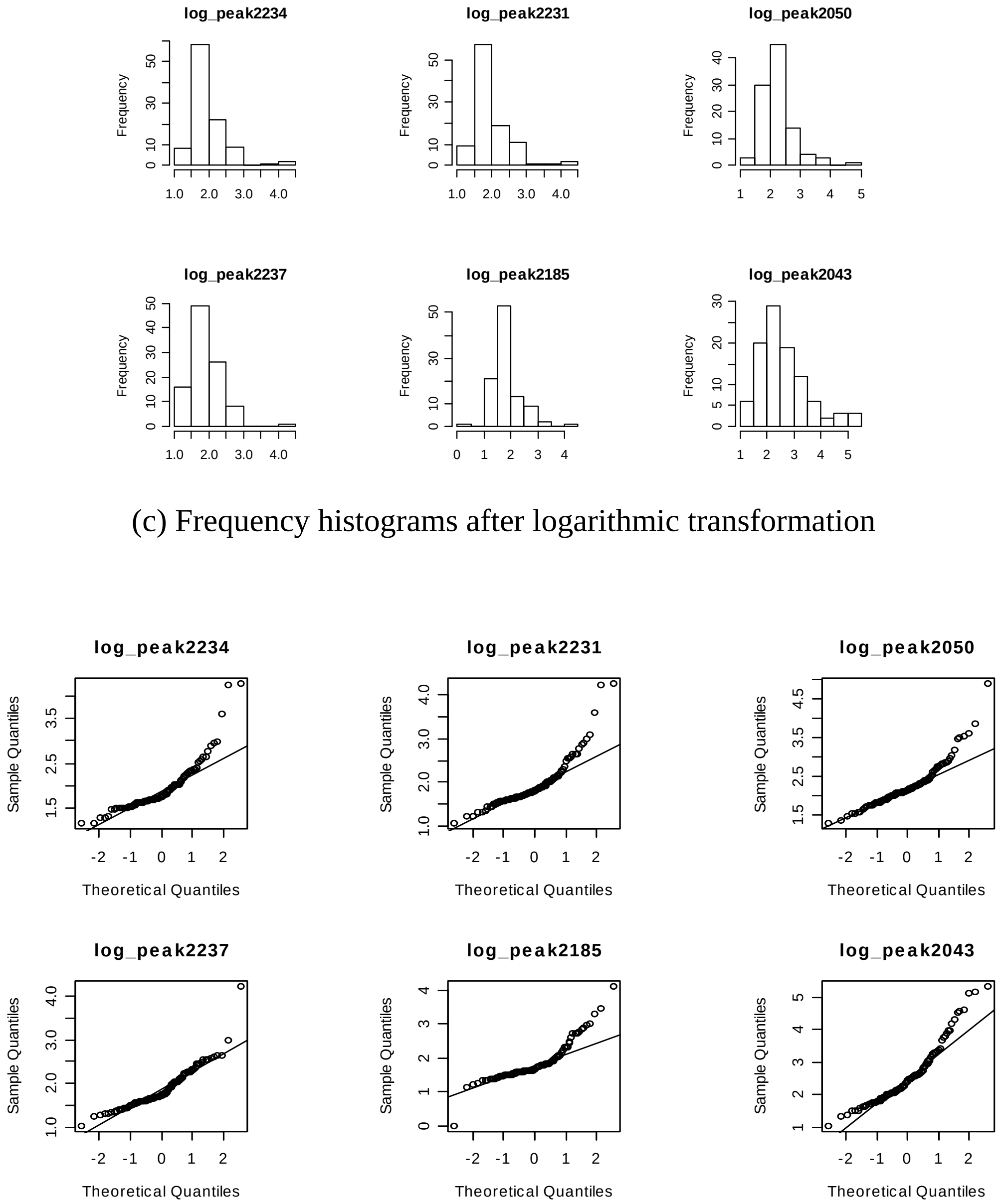
The lack of linear relationship between the two sets of significant protein profiles identified using the data sets with and without logarithmic transformation suggests that transformation affects the discriminatory power of the protein profiles.



(a) Frequency histograms before logarithmic transformation



(b) Q-Q plots before logarithmic transformation



(c) Frequency histograms after logarithmic transformation

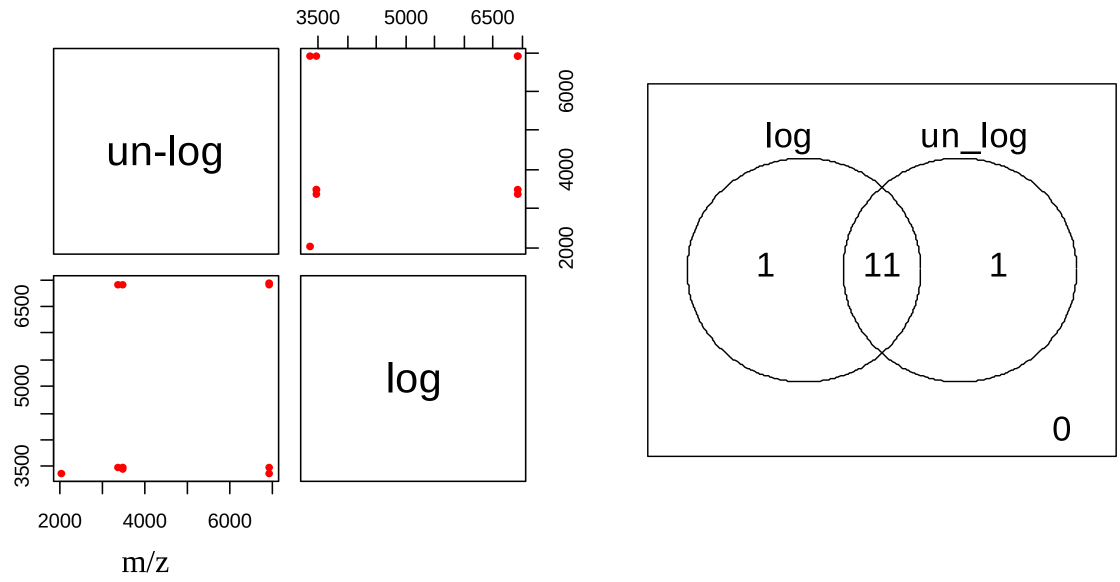
(d) Q-Q plots after logarithmic transformation

Figure 5-3 Frequency histograms and Q-Q plots of proteins with small p-values in control group before and after logarithmic transformation (MIC normalised). (a) Frequency histograms before logarithmic transformation. (b) Q-Q plots before logarithmic transformation. (c) Frequency histograms after logarithmic transformation. (d) Q-Q plots after logarithmic transformation.

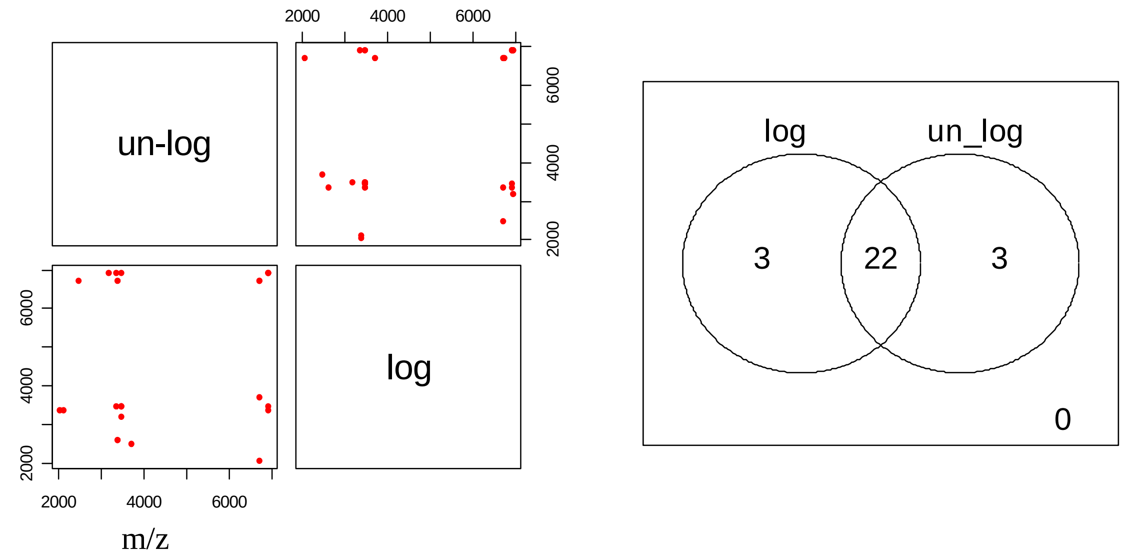
The right panel of Figure 5-4 (b) also shows a large overlap (22 out of 25 proteins) between the top 25 significant proteins identified before and after logarithmic transformation. And again we see that there is no linear relationship between the two sets of significant protein profiles identified using the data sets with and without logarithmic transformation (the left panel of Figure 5-4 (b)). The right panel of Figure 5-4 (c) shows a huge overlap of 45 out of the top 50 significant proteins identified with and without logarithmic transformation (or an overlap of 90%). The scatter plots (the left panels of Figure 5-4 (c)) show that the significant protein profiles have different discriminatory power before and after logarithmic transformation. These results indicated that the logarithmic transformation affects the discriminatory power of the protein profiles. It should be of interest to investigate whether the logarithmic transformation affects the accuracy of cancer classification.

5.3.4 RF classification

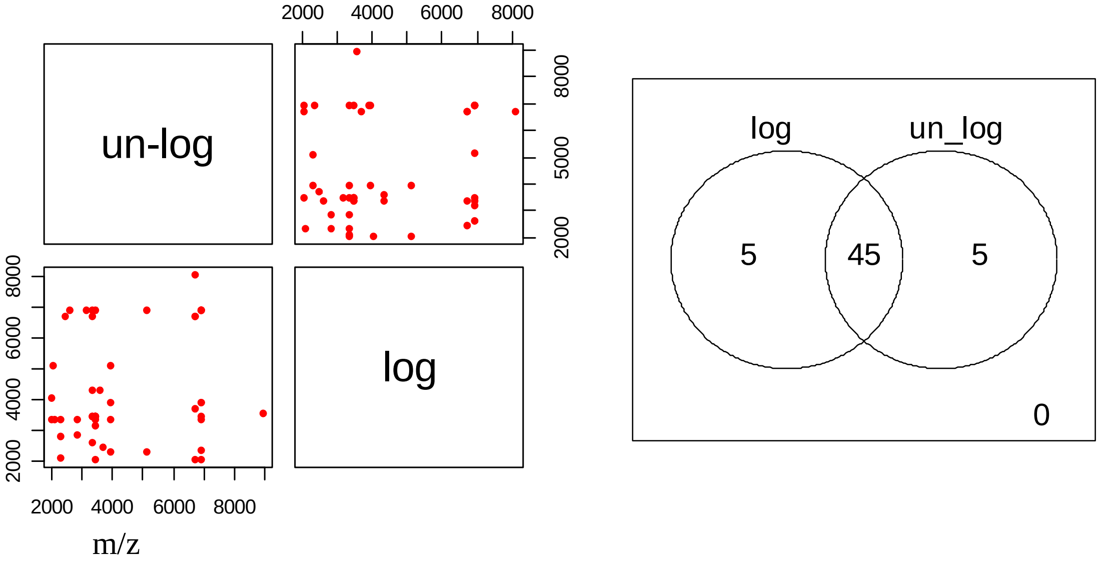
We now investigate how classification is affected by the logarithmic transformation to base 2. Table 5-11 lists the performance of cancer classification using the RF machine-learning algorithm (described in Section 4.2.3.2) based on the significant proteins identified from the raw data set and the logarithm-transformed data set. The misclassification rate of the RF model trained using the top 12 significant protein profiles reduces from 19% to 16%. For the top 25 protein profiles, the misclassification rate reduces from 12% to 11%, while using the top 50 protein profiles, the misclassification rate reduces from 15% to 14%. These results indicate that logarithmic transformation slightly increases the accuracy of cancer classification.



(a) The top 12 proteins



(b) The top 25 proteins



(c) The top 50 proteins

Figure 5-4 Concordance of the significant proteins identified using data set with and without logarithmic transformation. (a) top 12 proteins, (b) top 25 proteins and (c) top 50 proteins.

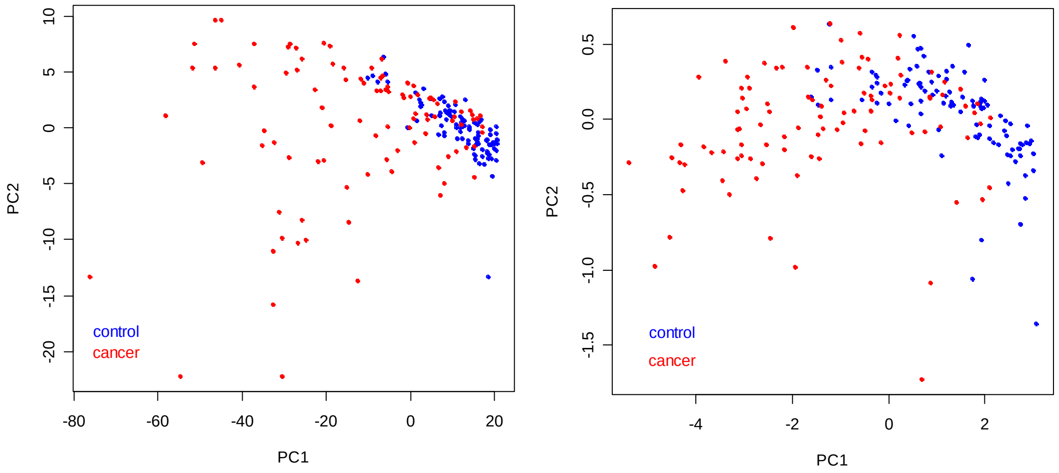
Table 5-11 Classification performance using RF machine learning algorithm

Data set	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**Misclassification rate (%)**
Top 12 proteins	87	75	78	19
Top 12 proteins (log2)	86	82	83	16
Top 25 proteins	86	91	91	12
Top 25 proteins (log2)	86	92	91	11
Top 50 proteins	84	87	87	15
Top 50 proteins (log2)	85	88	88	14

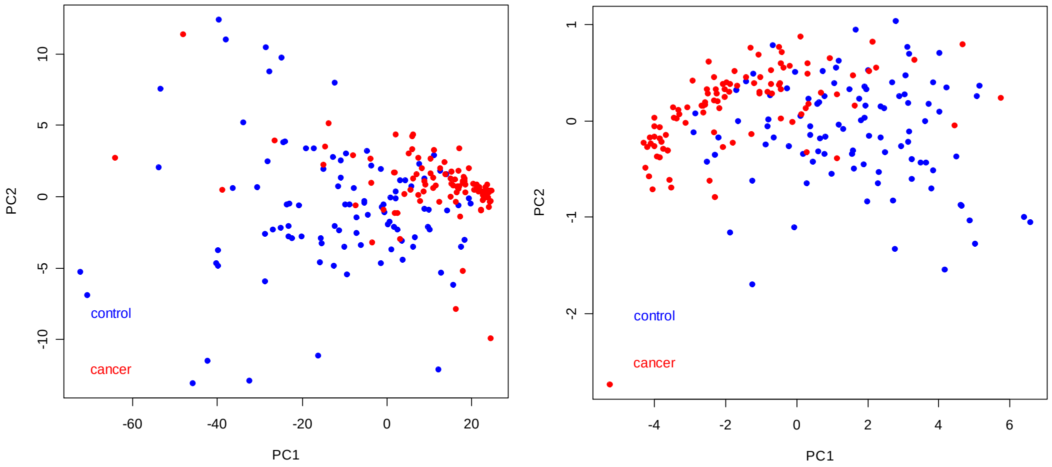
Top 12 proteins means that a data set consists of the top 12 protein profiles.
Top 12 proteins (log2) means that a data set consists of the top 12 protein profiles that were logarithm- transformed.

5.3.5 PCA analysis

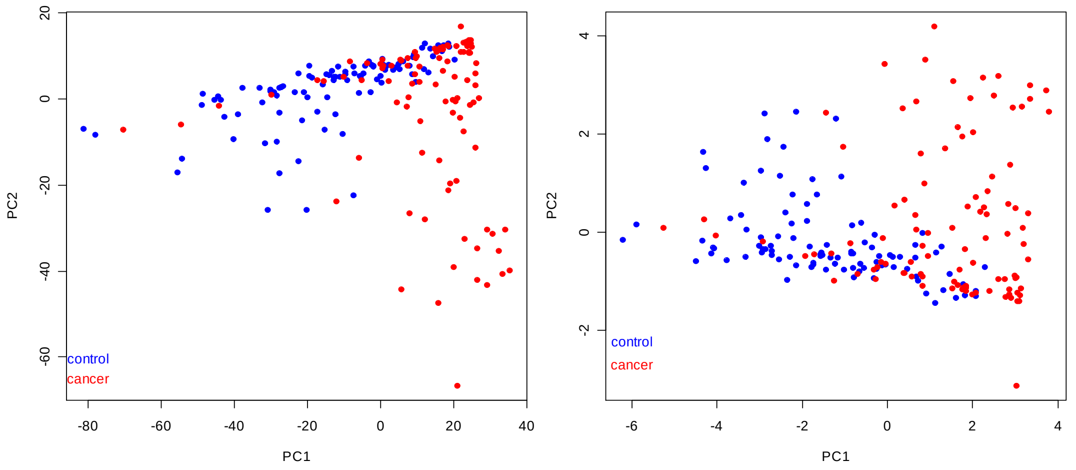
Figure 5-5 shows the results of PCA analysis using the significant proteins identified by the RF non-parametric methods based on the data sets with and without logarithmic transformation to base 2. The score plots generated using the top 12, 25, and 50 significant protein profiles before and after logarithmic transformation are shown in Figure 5-5(a), Figure 5-5(b), and Figure 5-5(c), respectively. These results visually show that the logarithmic transformation probably does not affect the performance of clustering.



(a) Top 12 significant proteins.



(b) Top 25 significant proteins.

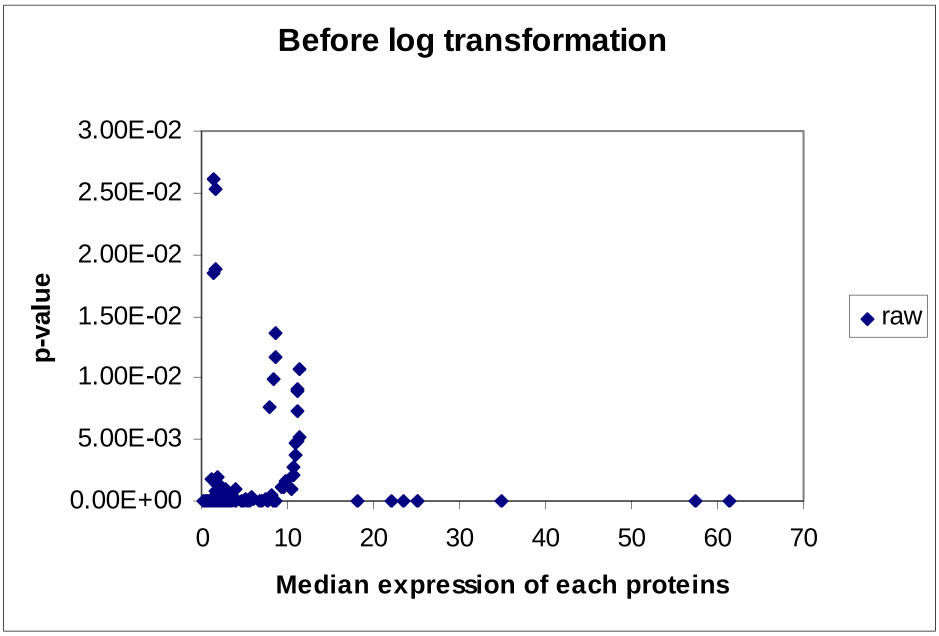


(c) 50 significant proteins.

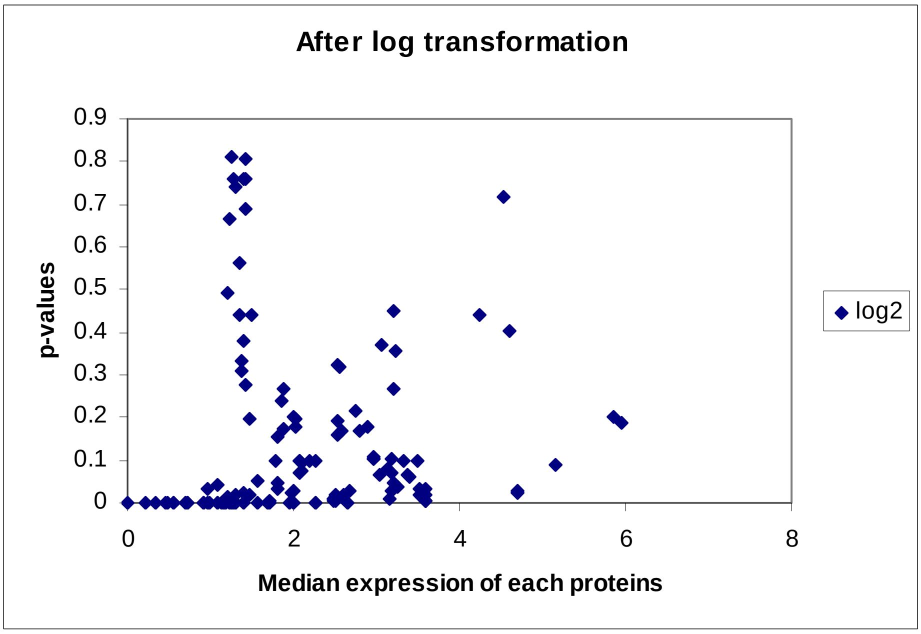
Figure 5-5 Scores plot of PCA (Left panel: before logarithmic transform; Right panel: after logarithmic transformation)

5.3.6 Correlation between protein expression level and the normality

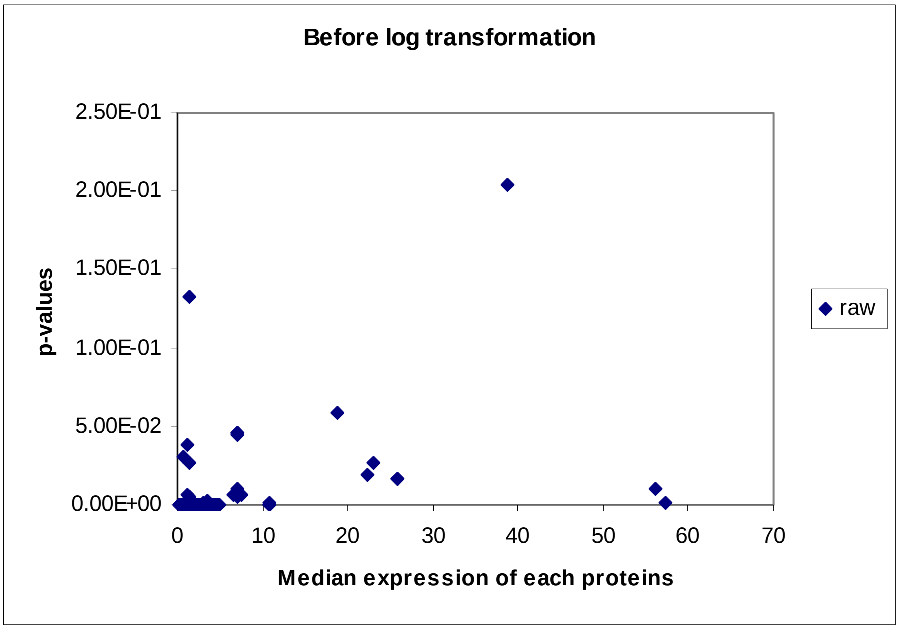
The relationship between the p-values of normality tests and the protein expression levels is shown in Figure 5-6 ((a) and (b) for the control group, and (c) and (d) for the cancer group). The x-axis represents the median protein intensities across all samples in the data set and the y-axis represents the corresponding p-values obtained from the Shapiro-Wilks tests. For the control group, the correlation coefficient between the p-values of normality tests and the protein expression levels is -0.001 with a p-value of 0.51 (Pearson’s product moment correlation coefficient test) for the raw spectra, and 0.007 with a p-value of 0.47 for the log-transformed spectra. This suggests that no correlation between the two variables is observed for the control group. For the cancer group, the correlation coefficient between the p-values of normality tests and the protein expression levels is 0.38 with a p-value of $1.79\text{x}10^{-6}$ for the raw spectra, and 0.29 with a p-value of 0.0003 for the log-transformed spectra. The observed correlation between the two variables implies that the low-expressed proteins from



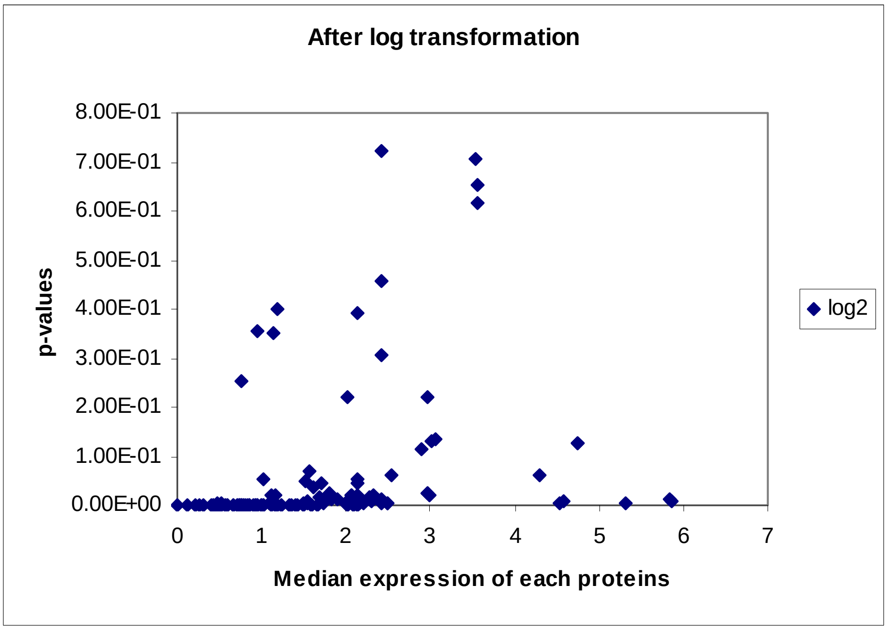
(a) Control group



(b) Control group



(c) Cancer group



(d) Cancer group

Figure 5-6 Scatter-plot of p-values (the Shapiro-Wilks test) and protein median expression level. (a) and (b) for the control group and (c) and (d) for the cancer group.

the samples of cancer patients do not follow normal distributions. This is consistent with what was observed in microarray data (Giles & Kipling,2003). However, the reason that the low-expressed proteins do not follow normal distributions is unknown. The results also show that the logarithmic transformation seems to reduce the correlation between SELDI protein profiles and the rejections of normality, especially as shown in Figure 5-6 (c) and (d).

5.4 Discussion

The outcomes of the four normality tests have shown that the hypothesis of the normality is rejected for a substantial number of protein expression profiles. Multiple testing procedures that combine the individual test results into a simultaneous hypothesis test confirm lack of normality, for both the logarithm-transformed and un-logarithm-transformed data sets. The standard mathematical transformation method for both proteomics and genomics data sets is the logarithm. We have extended this idea to include the use of arsinh and Box-Cox family transformations and have demonstrated that they have a role in improving the normality of the SELDI data. Our conclusions are based on the analysis of the publicly available ovarian cancer data set and it should be interesting to carry out large-scale simulation studies to generalise our results. Since goodness-of-fit tests are usually under-powered with small sample sizes, it is of critical importance to work with a data set that includes a large number of subjects (Chen et al.,2005). The ovarian cancer data set included 100 cancer samples and 100 control samples. Such a large sample size is rarely used in most experiments, and thus we feel that this sample size is large enough to lead to rather general conclusions.

It has been shown that generalised logarithmic transformations can stabilise the variance in NMR signal and microarray gene expression data (Purohit et al.,2004; Rocke & Durbin,2003). However, there is no report of its use in normality studies and this thesis explored that possibility. The results have shown that the standard mathematical transformation functions, such as logarithm and Box-Cox family transformation, may improve the normality for only some protein profiles. The results have indicated that the underlying structure of the high-throughput SELDI data cannot in general be modelled by a normal distribution even after data transformation. This result is consistent with what is described in the study by Giles et al. (2003). Giles stated that for many of the large-scale cancer studies, the nature of the underlying biology would make it surprising if significant deviations from normality were not found (Giles & Kipling,2003). Therefore, testing for normality is needed and warranted in every specific application. As large sets of SELDI protein expression data become more readily available, non-parametric testing would have the power to circumvent the problems raised by data that do not follow a normal distribution (Giles & Kipling,2003). Although it is quite common to apply logarithmic transformation to SELDI data set to stabilise variance (Coombes et al.,2005; Listgarten & Emili,2005), the data transformation can also alter the fundamental nature of the data (Osborne,2002). The results derived from this chapter have indicated that logarithmic transformation functions cannot convert all variables in a data set into normally distributed. Therefore, it may be unwise to apply the standard transformation function to SELDI data set and then employ normal-theory based statistical methods, such as the t-test.

5.5 Summary

The normality issues of SELDI mass spectrum have been assessed using the Shapiro-Wilks, Kolmogorov-smirnov, Cramér-von Mises and Pearson χ^2 statistical tests. The results from these four tests suggest that a large number of proteins reject the hypothesis of the normality of SELDI protein profiles. In addition, the goodness-of-fit tests reject the null hypothesis of normality, implying that SELDI mass spectra do not follow normal distributions. The normalisation results suggested that our proposed normalisation algorithm outperforms other existing methods. The transformation studies have shown that the logarithmic transformations yield the highest improvement of normality of SELDI protein profiles compared to other mathematical transformations. But the normalisation and mathematical transformations still cannot convert SELDI protein profiles into normally distributed. We think that it may be more reliable to use non-parametric methods, such as Wilcoxon test and the ROC curve method, as they overcome the many shortcomings of their parametric counterparts for high-throughput SELDI data analysis for detecting disease-associated proteins.

6 BIOMARKER DISCOVERY FOR CANCER CLASSIFICATION

6.1 Introduction

The identification of validated biomarkers correlating strongly to disease progression would not only classify the cancerous and non-cancerous tissues according to their molecular profile, but could also focus attention upon a relatively small number of molecules that might warrant further biochemical/molecular characterisation to assess their suitability as potential therapeutic targets (Bensmail & Haoudi,2003). Due to the complexity and diversity of proteomic patterns, it is essential to utilise algorithms to identify the underlying proteomic biomarkers from high dimensional data associated with cancer from multiple samples.

However, different feature sets are usually identified when different feature selection methods are applied because a set of features is heavily dependent on the mathematical framework of the algorithm at hand (Listgarten & Emili,2005). Similarly, the classification accuracy usually depends on the classification methods used (Datta & Lara,2006; Levner,2005; Listgarten & Emili,2005; Wu et al.,2003).

In this chapter, we propose an integrated algorithm for biomarker discovery from SELDI protein profiles using a combination of multiple feature selection methods and classification methods. Firstly, we used three feature selection methods, namely Wilcoxon test, the ROC curve and RF methods to produce three sets of significant proteomic features and to investigate how these feature selection methods affect identification of proteomic biomarkers. Secondly, we construct classification models using three machine learning algorithms, namely random forest, support vector machine, and k-nearest neighbour to investigate the performance of the models built on different feature sets, including the union and intersection of the different feature sets. Thirdly, we determine an optimal set of proteomic biomarkers from SELDI expression profiles.

6.2 Materials and biomarker discovery strategy

6.2.1 Data sets

The prostate cancer data set 7-3-02 (described in Section 2.7.4) and ovarian cancer data set 4-3-02 (described in Section 2.7.5) were used in this chapter to evaluate different feature selection methods for identifying biomarkers. The protein features with mass less than 2kDa were excluded for both data sets because data points within this mass range were likely affected by matrix. The spectra were normalised using the MIC algorithm, which was described in Section 5.2.2.1.3. The peaks of SELDI mass spectra were picked using PROcess software through R for both data sets.

6.2.2 Biomarker discovery strategy

Although a variety of methods can be used for biomarker discovery and cancer classification, there is evidence to show that different methods have identified different sets of biomarkers and give different classification accuracies (Geurts et al.,2005; Levner,2005; Wu et al.,2003). In order to address the question of which method should be used, in this section, we propose an integrated algorithm to identify biomarkers using SELDI protein profiles.

6.2.2.1 Univariate feature selection methods

Univariate feature selection approaches identify significant protein features by evaluating the discriminatory power at each protein m/z value. As demonstrated in Section 5.3.2, SELDI protein profiles do not follow normal distributions. Therefore, it is invalid to employ the parametric statistical tests (such as t-test) to detect changes in SELDI expression profiles between different groups of samples (e.g. control versus cancer). We utilise non-parametric univariate feature selection methods, such as Wilcoxon test and the ROC curve method. The criterion for identifying significant protein features using the non-parametric approaches is to rank the protein profiles according to the p-values of the Wilcoxon tests, or the AUC values of the ROC curve, from the most to the least informative.

6.2.2.1.1 Wilcoxon test

(1) Wilcoxon test: Wilcoxon test is also known as "Wilcoxon_Mann-Whitney" test. It is one of the best-known non-parametric significance tests. It was proposed initially by Wilcoxon in 1945 for equal sample sizes and extended by Mann and Whitney in 1947 to arbitrary sample sizes. The Wilcoxon test examines whether the samples in two groups come from the same distribution (Dekking et al.,2007). Suppose that a sample of n_x observations $\{I_x^1, I_x^2, ..., I_x^{n_x}\}$ is from one group x and that a sample of n_y observations $\{I_y^1, I_y^2, ..., I_y^{n_y}\}$ is from another group y. The Mann-Whitney test is based on a comparison of every observation I_x^i in the first sample with every observation I_y^j in the other sample. If the samples have the same median then each I_x^i has an equal chance of being greater or smaller than each I_y^j.

Therefore, under the null hypothesis $H_0 : P(I_x^i > I_y^j) = \frac{1}{2}$, under the alternative hypothesis $H_1 : P(I_x^i > I_y^j) \neq \frac{1}{2}$.

We count the number of times that a I_x^i from the group x is greater than a I_y^j from the group y and denote this value as U_x. Similarly, the number of times a I_x^i from the group x is smaller than a I_y^j from the group y is denoted by U_y. Under the null hypothesis we would expect U_x and U_y to be approximately equal (Shier,2004). The procedure for carrying out the Wilcoxon test is described as follows.

1) Arrange all the observations into a single ranked series. That is, rank all the observations without regard to which group they are in.
2) Add up the ranks in group x and denoted by R_x.
3) U_x is given by $U_x = R_x - \frac{n_x(n_x+1)}{2}$, n_x is the number of observations in the group x
4) R_y is calculated by $R_y = \frac{(n_x+n_y)(n_x+n_y+1)}{2} - R_x$, n_y is the number of observations in the group y
5) U_y is given by $U_y = R_y - \frac{n_y(n_y+1)}{2}$
6) U is given by $U = \min(U_x, U_y)$
7) Use statistical tables for the Mann-Whitney U test to find the probability of observing value of U or lower. If the test is one-sided, this is your p-value; if the test is a two-sided test, double this probability to obtain the p-value

(Shier,2004). If the number of observations is such that $n_x n_y$ is large enough (>20), a normal approximation $z = \frac{U - \mu_U}{\sigma_U}$ can be used (Dekking et al.,2007; Shier,2004), where, $\mu_U = n_x n_y / 2$, $\sigma_U = \sqrt{\frac{n_x n_y (n_x + n_y + 1)}{12}}$

(2) Multiple hypotheses testing: As discussed in Chapter 2, FWER and FDR are the two commonly used frequentist multiple hypotheses testing correction methods. FWER is a very conservative correction method and results in greatly diminished power to detect significantly differently expressed variables. The p-value used for a cut-off is computed by dividing the standard p-value (e.g. 0.05) by the number of tests. FDR (Baggio & Prodi,2005; Benjamini & Yekutieli,2005) is the expected value of the proportion of false positives among rejected null hypotheses. Compared to FWER control, FDR control is less conservative and provides increased power. It generates a good balance between discovery of statistically significant variables and limitation of false positive occurrences. In this case, the p-value used for a cut-off is computed as follows.

1. Ranking p-values of individual proteins in ascending order: $p_1 < p_2 < \ldots < p_k < \ldots < p_m$ (*m* is the total number of proteins)
2. Find the largest *k* so that $p_k \leq$ FDR*(k/m);
3. Set the cut-off $c = p_k$, assuming that the k proteins with the lowest p-values reject the null hypotheses.

The cut-off value for the smallest p-value in FDR control method is α/m, which is equivalent to Bonferroni adjusted cut-off value. It is known that the Bonferroni correction method is too conservative for most biological applications. FDR method allows some false positive results, but at a given number or proportion, it may be more conductive to scientific application (Birkner et al.,2006). Therefore, FDR control method is used in this chapter to select significantly differentially expressed proteins for comparing the univariate feature selection methods.

6.2.2.1.2 The ROC curve

(1) The ROC curve: The ROC curve was first used during the World War II for the analysis of radar signals. Since then, ROC analysis has been extensively used in medicine, radiology, psychology and other areas. It has been introduced in data mining and machine learning recently. The first application of ROC to machine learning was made by Spackman who investigated the area under the ROC curves in comparing and evaluating different classification algorithms (Spackman,1989). Several studies have since used the ROC curve to select features from SELDI protein profiles (Adam et al.,2002; Ball et al.,2002; Chen et al.,2002; Qu et al.,2002; Yu et al.,2004).

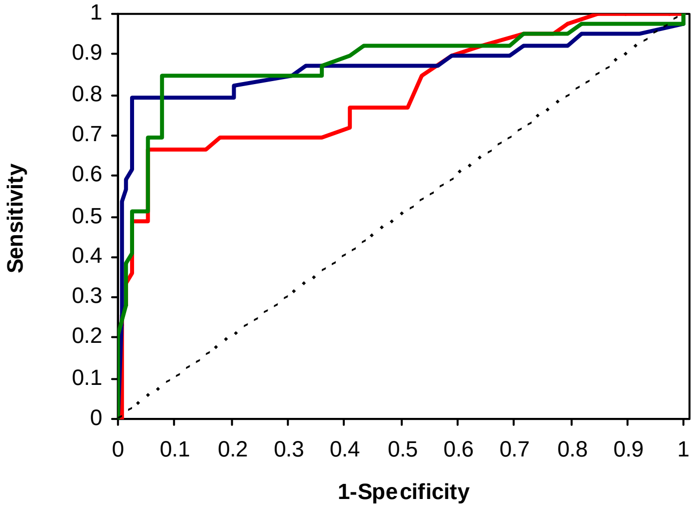


Figure 6-1 An example of the ROC curves.

The ROC curve is a plot of the sensitivity versus 1-specificity as showed in Figure 6-1. It displays the fraction of the true positives and the fraction of the false positives as discrimination thresholds vary in a diagnostic test. The advantage of ROC analysis includes the fact that it explicitly considers the tradeoffs in sensitivity and specificity. The AUC under the ROC curve actually represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. This probability of correct ranking curve is equivalent to the quantity estimated by the Wilcoxon statistic (Bradley,1997; Hanley & McNeil,1982).

(2) The AUC and Figure of merit (FOM): By calculating the AUC, an approximated measure of probability is determined for each model. FOM, which is used to assess the discrimination power of each model or test as a function of AUC, is assigned as shown in Table 6-1 (Swets,1988).

Table 6-1 The figure of merit

AUC	0.91-1.00	0.81-0.90	0.71-0.80	0.61-0.70	0.51-0.60
FOM	Excellent	Good	Fair	Poor	Failed

A model or test with the highest AUC will in general have the fewest false positive and false negative results, regardless of the final threshold. If the AUC reaches a value of 1, then complete separation between the two groups has been achieved. But if the AUC value approaches to 0.5 (as shown by the dashed line in Figure 6-1), then there is no discrimination.

6.2.2.2 Multivariate feature selection methods

Using the Wilcoxon test and the AUC to select significant protein profiles as biomarkers may have some limitations, as they do not take interactions among variables into account. More importantly, the univariate feature selection methods are not stable when sample sizes are relatively small (Kohavi,1997). Some classical statistical regression analyses, such as logistic regression, allow investigation of interactions between variables, and between variables and environmental factors, but may have a problem of the curse of dimensionality. The model becomes unstable as more main effects and interaction terms are added, in the sense that the variance of the parameter estimates becomes excessively large. Statistical parametric models, such as generalised linear models, assume a function for the relation between the input variables and the output variables. However, precise information about the shape of the relation between the input variables and the output variables is lacking. Therefore, they often fail to deal adequately with the biological complexities and the multidimensional problem of feature selection (Breiman,1992).

The non-parametric machine learning methods have potential to identify cancer susceptibility proteins without making any assumption about relationship between protein profiles and disease status. Several multivariate search strategies have been proposed in the literature, all involving combinatorial searches through the space of possible feature subsets (Duda et al.,2000; Kohavi,1997). The search can start with one feature, with all features or with a random subset of features. A forward selection algorithm chooses one feature as the starting point and then successively adds features from this point. A backward elimination algorithm chooses all features as the starting point, and successively removes what are considered irrelevant features. A random selection algorithm chooses randomly a number of features at the starting point, and iteratively adds relevant features and removes irrelevant features according to their discriminatory power. All the selection algorithms are based on heuristic search methods.

6.2.2.2.1 Random forest

The random forest algorithm combines bagging and random feature selection techniques. Bagging is to adaptively resample the original data so that the weights are increased for the most frequently misclassified samples. The random feature selection makes it possible to improve the predictive accuracy (Wu et al.,2003). A random forest model is a group of tree predictors $f(I;\theta_j),\ \ j=1,2,...,T$, where, I represents l observed protein profiles with associated random vector I and θ_t are independent and identically distributed random vectors. T represents the total number of trees. The training data set is assumed to be independently drawn from the joint distribution of (I,Y) and comprises $N(l+1)$ patterns (I_i, y_i), $i=1,2,...,N$ (sample size for each tree), Y is the observed cancer status. I_i represents the vector of intensities, y_i represents the true class.

A random forest classifier reports an importance measure for each variable based on the decrease mean accuracy or a decreased Gini distance criterion (Robnik-Sikonja,2004). These are internal estimates of the decrease in the classifier's overall accuracy if that particular variable is not used in building the classifier. For each protein feature, based on the classification performance in out-of-bag samples, random forest model calculates its margin as follows.

$$mg(I_i, y_i) = \frac{1}{T_i}\sum_{j=1}^{T} U(f(I_i;\theta_j) = y_i)t_{ij} - \max_{k\neq y_i}\{\frac{1}{T_i}\sum_{j=1}^{T} U(f(I_i;\theta_j) = k)t_{ij}\} \quad \textbf{(6.1)}$$

where, $T_i = \sum_{j=1}^{T} t_{ij}$, T_i represents the number of trees for which individual sample i is out-of-bag, j represents the index of trees, t_{ij} is an indicator with a value of 1 if individual i is out-of-bag for tree j , and $f(I_i;\theta_j)$ represents the vote for tree j .

$U(f(I_i;\theta_j)=y_i)$ denotes a indicator function, taking value of 1 if $f(I_i;\theta_j)=y_i$, and 0 otherwise.

In the context of a categorical response, the larger margin indicates the higher confidence of the forest prediction. If a feature is predictive of the response, its margin will be decreased when randomly permuting the values of this protein feature among the individuals in the out-of-bag samples. The decrease in margin, $W_M(P)$, is used as an index to measure the importance of the protein feature.

$$W_M(P)=\frac{1}{N}\sum_{i=1}^{N}[mg(I_i,y_i)-mg(I_i^{(P)},y_i)] \quad \textbf{(6.2)}$$

where, N is the total number of samples in a data set, and $W_M(P)$ is the importance index obtained by randomly permuting the values of a peak intensity.

With a specific threshold for the decrease in margin, a set of important features are determined. Thus, protein features with large importance measures are thought to have more discrimination power.

6.2.2.2.2 Concordance and discrepancy

We examine the concordance and discrepancy of the discriminatory feature sets extracted using different feature selection methods. The scatter plot matrices and Venn diagrams are generated using function *pairs* in *graphics* package and *vennDiagram* in *limma* package through R to implement these analyses.

The effects of different feature selection methods on classification accuracy are also assessed using three different machine-learning algorithms. This provides a way to identify biomarkers for SELDI data by combining findings from different methods.

6.2.2.3 An integrated algorithm for identifying proteomic biomarkers

We propose an algorithm for identifying proteomic biomarkers based on SELDI data (Figure 6-2), which is outlined as follows.

(1) Each feature selection method is applied to the same SELDI data set, and a set of proteomic biomarkers is generated, which is denoted by B_i, where, i is the ith feature selection method.

(2) The intersection set of the protein features identified using these individual feature selection methods is found and denoted by B_I, where, $B_I = \bigcap B_i$.

(3) The union set of the protein features identified using these individual feature selection methods are found and denoted by B_U, where, $B_U = \bigcup B_i$.

(4) Several classifiers are trained using the individual proteomic feature set B_i, B_I and B_U. The corresponding performance of the classifiers (e.g. misclassification rate) are recorded and denoted by R_j, where, j is the jth classifier. Note that the classifiers can be trained using one or more methods.

(5) The optimal set of proteomic biomarkers is determined and denoted by B_o, where, $B_o = \{B_j, j = i, I, U \left| \min_j (R_j) \right.\}$.

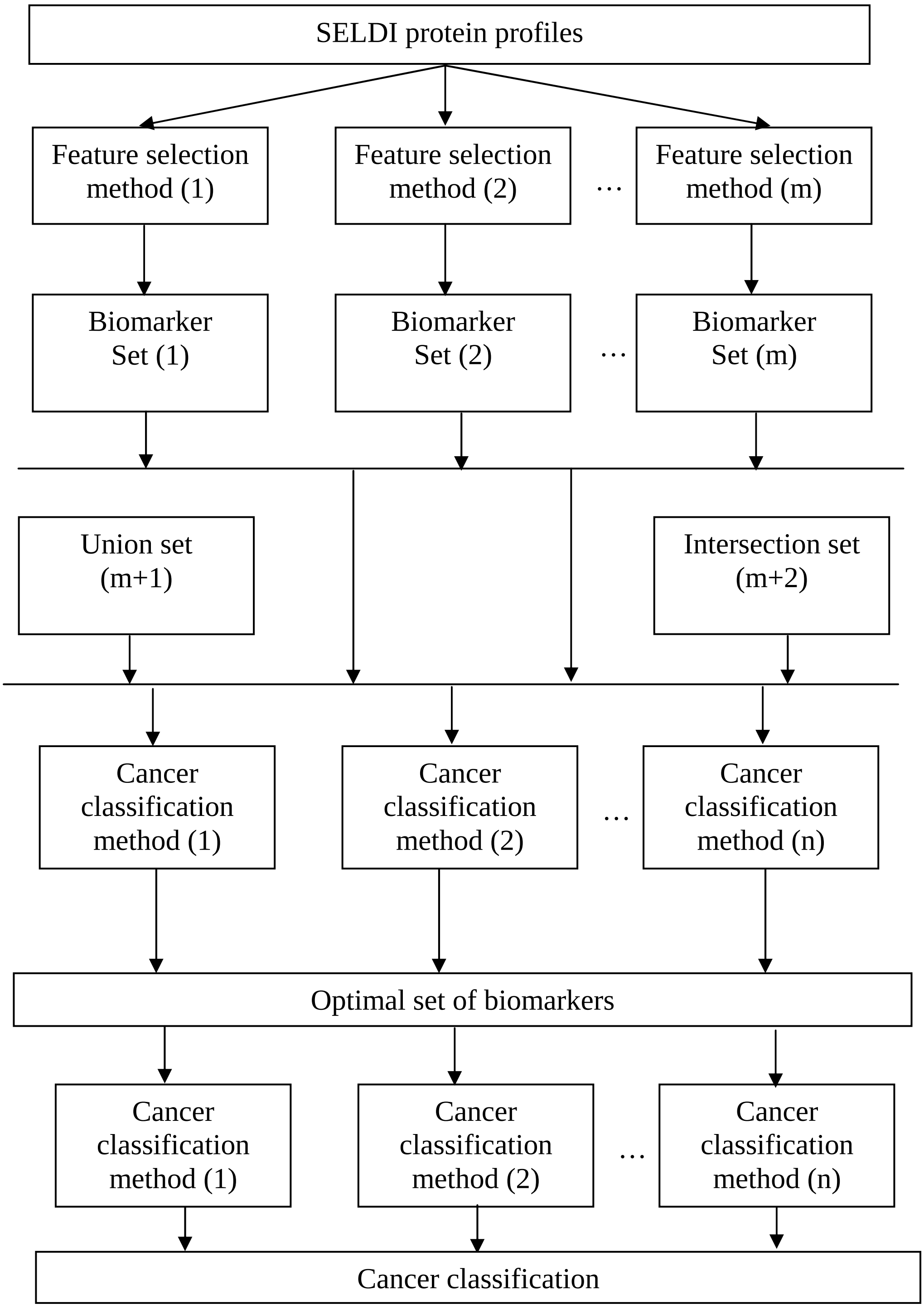

SELDI protein profiles
Feature selection method (1)
Feature selection method (2)
…
Feature selection method (m)
Biomarker Set (1)
Biomarker Set (2)
…
Biomarker Set (m)
Union set (m+1)
Intersection set (m+2)
Cancer classification method (1)
Cancer classification method (2)
…
Cancer classification method (n)
Optimal set of biomarkers
Cancer classification method (1)
Cancer classification method (2)
…
Cancer classification method (n)
Cancer classification

Figure 6-2 Diagram of biomarker discovery using SELDI

6.2.2.4 Cancer classification

The high-throughput SELDI mass spectrometry provides a platform to identify protein patterns for distinguishing cancer from normal. Due to the high dimensional data that is generated from a single experiment, it is essential to use mathematical tools to discover the boundaries or relationship between the groups (e.g. cancer versus control, treatment versus non-treatment) from SELDI protein profiles. Random forest, support vector machine, and k-nearest neighbour are used as examples in this chapter to classify cancer from normal subjects.

6.2.2.4.1 Random forest

The procedure to classify cancer using the random forest method has been described in 4.2.3.2. The software package *randomForest* in R is used to build classification models. The prediction error is evaluated based on out-of-bag estimation. To assess the error rate variation, the whole procedure is repeated 100 times and 500 trees are grown in each repeat.

6.2.2.4.2 Support vector machine

Support vector machine is a machine-learning algorithm. It maps the feature space into a hyperplane and performs regression in that space where all class-one points (for example cancer patients) are on one side and all class-two points (for example normal subjects) on the other side as shown in Figure 6-3, where the line Y=w*X+b is the ‘hyperplane’. However, such a hyperplane is not unique. In order to address this problem, SVM introduces a concept called margin (Cristianin & Shawe-Taylor,2000). The margin of the linear SVM is defined by $M = \frac{2}{|w|}$, which is the distance from the hyperplane to the nearest data point. The unique hyperplane can be determined by maximising the margin M.

Let $y_i = 1$ if the *i*th subject is a cancer patient and $y_i = -1$ if the *i*th subject is normal. In order to make correct classifications in a training data set, we need to satisfy the following equations: $y_i(wX_i + b) \geq 1$. Training a linear SVM classifier is equivalent to solving the following optimisation problem because the solution of maximising M is equivalent to the solution of minimising $|w|$.

$$\text{Minimise } \Phi(w) = \frac{1}{2}\|w\|^2$$
$$\text{Subject to } y_i(wx_i + b) \geq 1 \qquad \textbf{(6.3)}$$

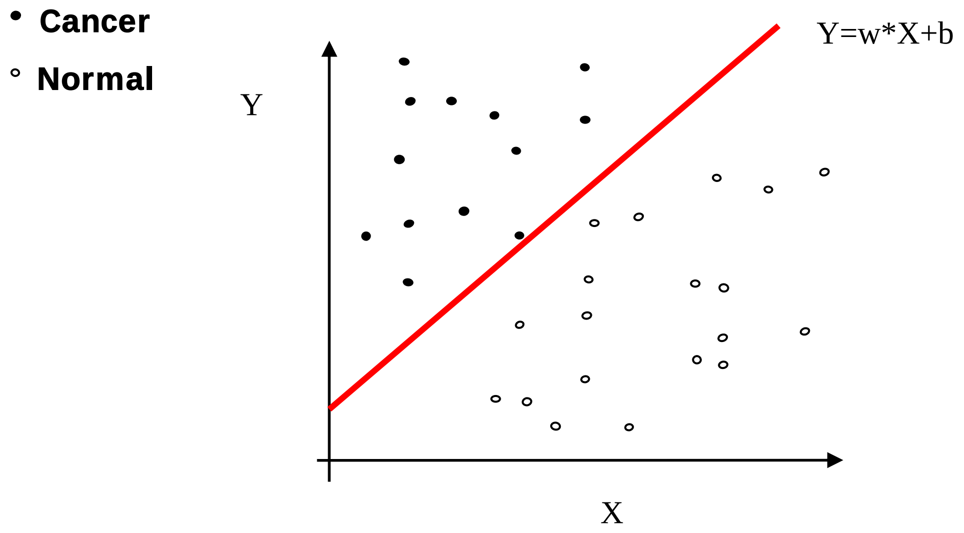


Figure 6-3 A linear SVM classifier

The optimisation problem requires that all data points are correctly classified, that is, the data set is separable. However, for a data set with noise as shown in Figure 6-4, one cannot train a SVM by solving the above optimisation problem. One possible solution is to use soft margin classification method by introducing slack variables, which measure the degree of misclassifications as shown in Figure 6-5.

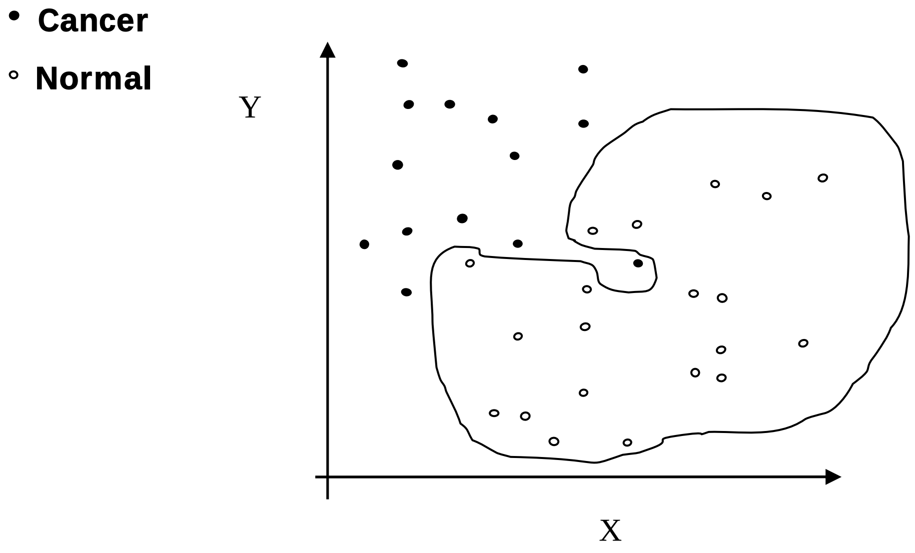


Figure 6-4 A problem in training a SVM classifier using a data set with noise

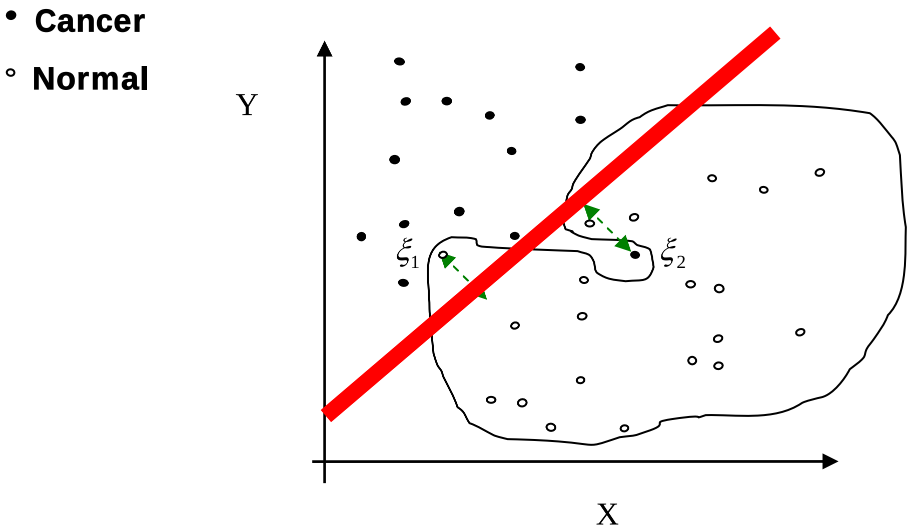


Figure 6-5 A linear SVM classifier with soft margin

Training a linear SVM classifier with soft margin is equivalent to solving the following optimisation problem defined by

$$\text{Minimise } \Phi(w) = \frac{1}{2}\|w\|^2 + C\sum \xi_i$$
$$\text{Subject to } y_i(wx_i + b) \geq 1 - \xi_i\,,\ \xi_i \geq 0 \qquad \textbf{(6.4)}$$

The solution of the above optimisation problem is a trade off between maximising the margin and minimising the error penalty (Cristianin & Shawe-Taylor,2000) and involves constructing a dual problem described as follows.

$$\text{Maximise } D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i{}^t x_j$$
$$\text{Subject to} \sum_{i=1}^{n}\alpha_i y_i = 0\,,\ \alpha_i \geq 0 \qquad \textbf{(6.5)}$$

From the solution, we can obtain $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, and $b = y_k - w^T x_k$ for any x_k such that $\alpha_k \neq 0$, implying that x_k is a support vector. The classification function is defined as follows.

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i x_i{}^T x + b) \qquad \textbf{(6.6)}$$

For data sets that are not linearly separable, non-linear SVM classifiers can be used. For example, the data set shown in Figure 6-6 (a) cannot be separated with one hyperplane. However, instead of looking at (x_i, y_i), we can consider $(x_i, y_i, x_i^2 + y_i^2)$, mapping source data into a higher-dimensional space as shown in Figure 6-6 (b), where we can separate the two classes.

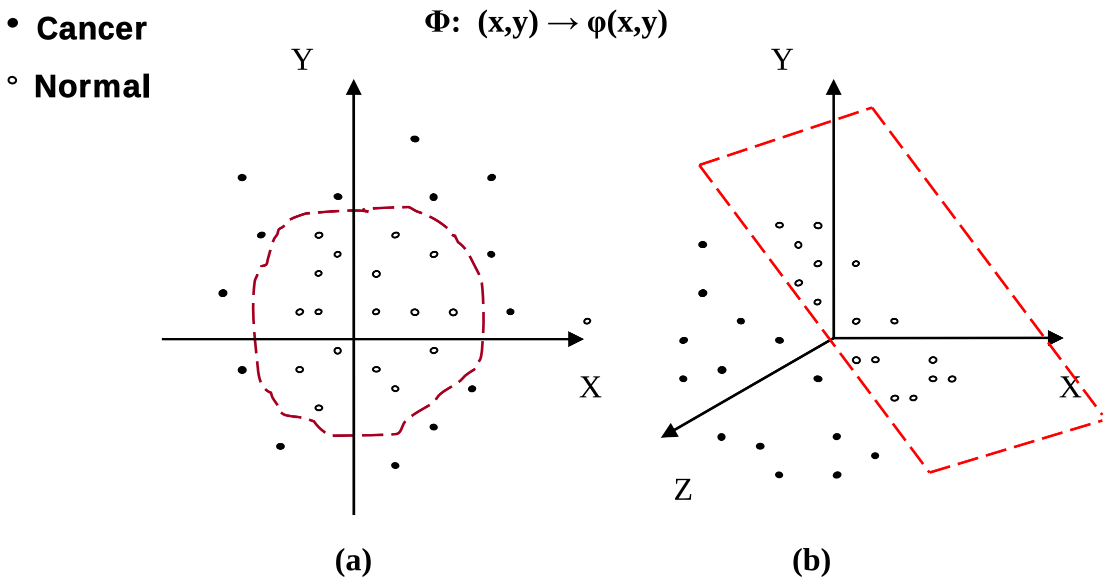


Figure 6-6 A non-linear SVM classifier.

A non-linear SVM classifier is a hyperplane in the transformed feature space, which is generated by kernel functions. The common non-linear kernels include polynomial, Gaussian, and sigmoid functions defined as follows, respectively.

$$\text{Polynomial: } K(x_i, x_j) = (1 + x_i{}^T x_j)^p \quad \textbf{(6.7)}$$

$$\text{Gaussian: } K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|}{2\sigma^2}} \quad \textbf{(6.8)}$$

$$\text{Sigmoid: } K(x_i, x_j) = \tanh(\beta_0 + \beta_1 x_i{}^T x_j) \quad \textbf{(6.9)}$$

The solution of the non-linear SVM involves constructing a dual problem described as follows.

$$\begin{aligned}&\text{Maximise } D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)\\ &\text{Subject to} \sum_{i=1}^{n} \alpha_i y_i = 0,\ \alpha_i \geq 0 \qquad \textbf{(6.10)}\end{aligned}$$

The corresponding classification function is defined by

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b) \quad \textbf{(6.11)}$$

The function *svm* in *e1071* package from *R* is used to implement the model construction. The Gaussian kernel is used. The classification error is estimated based on the 10-fold cross validation. The error variance is evaluated by running 100 times.

6.2.2.4.3 K-nearest neighbour

K-nearest neighbour is a simple distance-based method in cluster analysis. A kNN model partitions a training data set into a number of categories based on a distance measure between two points. We use the following Euclidean distance measure to partition a training data set into a number of categories.

$$d(I^x, I^y) = \sqrt{\sum_{i=1}^{m} (I_i^x - I_i^y)^2} \quad \textbf{(6.12)}$$

Where, I^x and I^y are the observed protein profiles from samples x and y.

The category centroids are fixed at random positions when the model is initialised. For a given set of test samples, the algorithm searches the *k* nearest neighbours in the training set. The sample classification is performed based on the sample status of its nearest neighbours in the training data set. A sample in the test data set is classified to cancer or control by votes from its nearest neighbours, which is weighted by the rank of the distance between its neighbour and the test sample (Bensmail & Haoudi,2003; Wu et al.,2003).

The function *knn.cv* in *class* from *R* is used to build the classification models. Leave-one-out cross validation scheme is used to estimate the prediction error and the error variance is calculated based on 100 runs.

6.3 Results of biomarker discovery

6.3.1 Feature selection

6.3.1.1 Wilcoxon test

The Wilcoxon test was applied to the SELDI protein profiles from the ovarian cancer data set and the prostate cancer data set. The cut-off of p-values was set to 10^{-7} for the prostate cancer data set and 10^{-15} for the ovarian cancer data set, which were calculated using the FDR correction method ($FDR = 5.0x10^{-7}$). A total of 16 out of 125 protein features with the p-values of the Wilcoxon tests less than 10^{-7} were extracted from the prostate cancer data set; and 26 out of 160 protein features with the p-values less than 10^{-15} were selected from the ovarian cancer data set as shown in Table 6-2.

6.3.1.2 ROC curve

Fourteen out of 125 protein features from the prostate cancer data set and 34 out of 160 protein features from the ovarian cancer data set were identified with the AUC values of greater or equal to 0.81 (fair separation performance). They are listed in Table 6-2.

6.3.1.3 Random forest

Eighteen out of 125 protein features from the prostate cancer data set and 24 out of 160 protein features from the ovarian cancer data set as shown in Table 6-2 were identified with the mean decrease accuracy of greater than 0.005.

Table 6-2 shows that the numbers of the protein features identified by the Wilcoxon test, ROC curve and RF methods are quite different for both the prostate cancer data set and the ovarian cancer data set. The ROC curve method produced the smallest feature set (14) for the prostate cancer data set compared with the other two feature selection methods. By contrast, the ROC curve method generated the biggest feature set (34) for the ovarian cancer data set, although the same cut-off value (AUC ≥ 0.81) was applied.

Table 6-2 The numbers of protein features identified by the Wilcoxon test, ROC curve and RF methods.

Data set	Wilcoxon test	ROC curve	Random forest
Prostate cancer	16	14	18
Ovarian cancer	26	34	24

Figure 6-7 shows the relationship between the feature sets extracted using the Wilcoxon test, ROC curve and RF methods. It can be seen that the three methods tend to identify some common discriminatory protein features. Thirteen common protein features from the prostate cancer data set and 19 common protein features from the ovarian cancer data set were identified by all the three methods. For the prostate cancer data set, a common protein feature was identified by using the Wilcoxon test and ROC curve method, but not by the RF method. For the ovarian cancer data set, a common set of seven protein features were identified by using the Wilcoxon test and ROC curve method, but not by the RF method. However, five protein features were identified by the RF method, but not by either the Wilcoxon test or the ROC curve method for each data set.

The protein features identified from the prostate cancer data set by the Wilcoxon test, ROC curve and RF methods are listed in Table 6-3. They are ranked according to their discrimination powers. The first column lists the features that were identified by the Wilcoxon tests and ranked (ascending) by their p-values. The features listed in the third column were identified by the ROC curve method and ranked (descending) by the AUC values. The features shown in the fifth column were identified by the RF method and ranked (descending) by the mean decreases in accuracy.

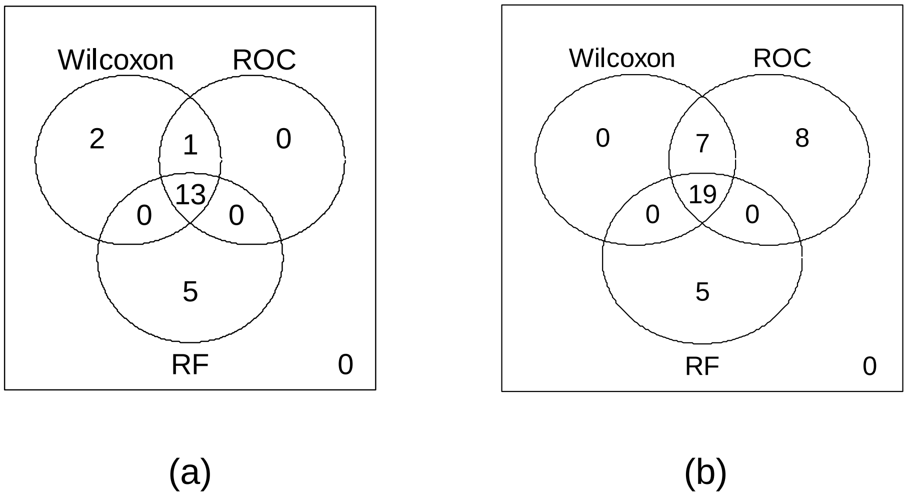


Figure 6-7 The relationship among the feature sets identified using the Wilcoxon test, ROC curve and RF methods. (a) for the prostate cancer data set and (b) for the ovarian cancer data set.

Sixteen, fourteen and eighteen protein features were identified by the Wilcoxon test, ROC curve and RF methods, respectively. The thirteen common features identified by the three methods are 3468.89, 3471.09, 3472.19, 3474.39, 3475.49, 3478.79, 6914.28, 6915.84, 6917.39, 6918.94, 6920.49, 6922.05, and 6925.15. The five features, 5250.73, 5257.49, 5260.20, 7633.59, and 7649.10 were identified only by the RF method. Two features, 4248.17 and 8469.88 were identified by the Wilcoxon test only. One feature 3478.79 was identified by both Wilcoxon test and ROC curve method.

Table 6-3 The protein features identified by the Wilcoxon test, ROC curve and RF methods from the prostate cancer data set.

Wilcoxon test		ROC curve		Random forest	
Feature (M/Z)	**P-value**	**Feature (M/Z)**	**AUC**	**Feature (M/Z)**	**Mean decrease accuracy**
6914.28	3.95E-14	6914.28	0.882	3468.89	0.020
6918.94	3.95E-14	6918.94	0.882	3471.09	0.017
6915.84	4.09E-14	6915.84	0.882	6915.84	0.017
6920.49	4.53E-14	6920.49	0.881	3472.19	0.017
6917.39	4.86E-14	6917.39	0.881	6917.39	0.015
3468.89	6.66E-14	3468.89	0.879	6914.28	0.015
3471.09	1.24E-13	3471.09	0.874	5257.49	0.015
6922.05	1.24E-13	6922.05	0.874	6918.94	0.015
3472.19	2.14E-13	3472.19	0.871	6920.49	0.015
6925.15	4.67E-13	6925.15	0.865	6922.05	0.013
3474.39	8.81E-12	3474.39	0.845	5250.73	0.011
3475.49	4.07E-11	3475.49	0.833	6925.15	0.011
3477.69	6.06E-11	3477.69	0.830	5260.2	0.008
3478.79	1.90E-10	3478.79	0.822	3478.79	0.006
8469.88	4.06E-09			7633.59	0.005
4248.17	1.49E-08			7649.1	0.005
				3475.49	0.005
				3474.39	0.005

Table 6-3 also demonstrates that the Wilcoxon test and ROC curve method yield the same discriminatory power for each of the 14 features 6914.28, 6918.94, 6915.84, 6920.49, 6917.39, 3468.89, 3471.09, 6922.05, 3472.19, 6925.15, 3474.39, 3475.49, 3477.69, and 3478.79 identified from the prostate cancer data set. The ranks of the 14 features identified by the ROC curve methods are exactly the same as those of the top 14 features identified by the Wilcoxon test. This confirms the fact that the AUC under the ROC curve is equivalent to the quantity estimated by the Wilcoxon test.

Table 6-4 The protein features identified by the Wilcoxon test, ROC curve and RF methods from the ovarian cancer data set.

Wilcoxon test		ROC curve		Random forest	
Feature (M/Z)	P-value	Feature (M/Z)	AUC	Feature (M/Z)	Mean decrease accuracy
3354.96	0	3468.89	0.888	3463.40	0.027
3362.00	0	3463.40	0.888	3354.96	0.025
3463.40	0	3471.09	0.881	6912.73	0.018
3468.89	0	3472.19	0.870	6917.39	0.017
3471.09	0	3473.29	0.867	3468.89	0.016
3472.19	0	6912.73	0.864	3471.09	0.016
3473.29	0	6709.40	0.861	6918.94	0.014
3474.39	0	6915.83	0.861	6915.83	0.013
3475.49	0	3354.96	0.861	3362.00	0.012
6709.40	0	3474.39	0.861	3472.19	0.011
6710.93	0	6917.39	0.860	6920.49	0.010
6712.46	0	6918.94	0.859	2040.27	0.010
6713.99	0	6710.93	0.858	6712.46	0.009
6912.73	0	6712.46	0.855	6709.40	0.009
6915.83	0	6920.49	0.853	6710.93	0.009
6917.39	0	6713.99	0.852	3473.29	0.009
6918.94	0	6922.05	0.846	3182.12	0.007
6920.49	0	3475.49	0.842	3699.61	0.006
6922.05	0	3362.00	0.841	3474.39	0.006
3367.40	2.22E-16	3368.49	0.836	3367.40	0.006
3368.49	2.22E-16	6715.52	0.836	3368.49	0.006
3370.65	2.22E-16	6923.60	0.836	6713.99	0.005
6715.52	2.22E-16	3370.65	0.835	2114.28	0.005
6923.60	2.22E-16	3367.40	0.835	2474.29	0.005
6925.15	4.44E-16	6925.15	0.832		
6717.05	8.88E-16	6717.05	0.830		
		3476.59	0.827		
		6926.70	0.823		
		6718.58	0.822		
		6928.26	0.819		
		3369.57	0.819		
		6929.81	0.812		
		6720.11	0.812		
		6721.64	0.810		

The protein features identified from the ovarian cancer data set by the Wilcoxon test, ROC curve and RF methods are listed in Table 6-4. They are ranked according to their discrimination powers. The first column lists the features that were identified by the Wilcoxon tests and ranked (ascending) by their p-values. The third column lists the features that were identified by the ROC curve method and ranked (descending) by the AUC values. The fifth column lists the features that were identified by the RF method and ranked (descending) by mean decreases in accuracy.

Twenty-six, thirty-four and twenty-four protein features were identified by the Wilcoxon test, ROC curve and RF methods, respectively, from the ovarian cancer data set. The nineteen common features identified by the three methods are 3354.96, 3362.00, 3367.40, 3368.49, 3463.40, 3468.89, 3471.09, 3472.19, 3473.29, 3474.39, 6709.40, 6710.93, 6712.46, 6713.99, 6912.73, 6915.83, 6917.39, 6918.94, and 6920.49.

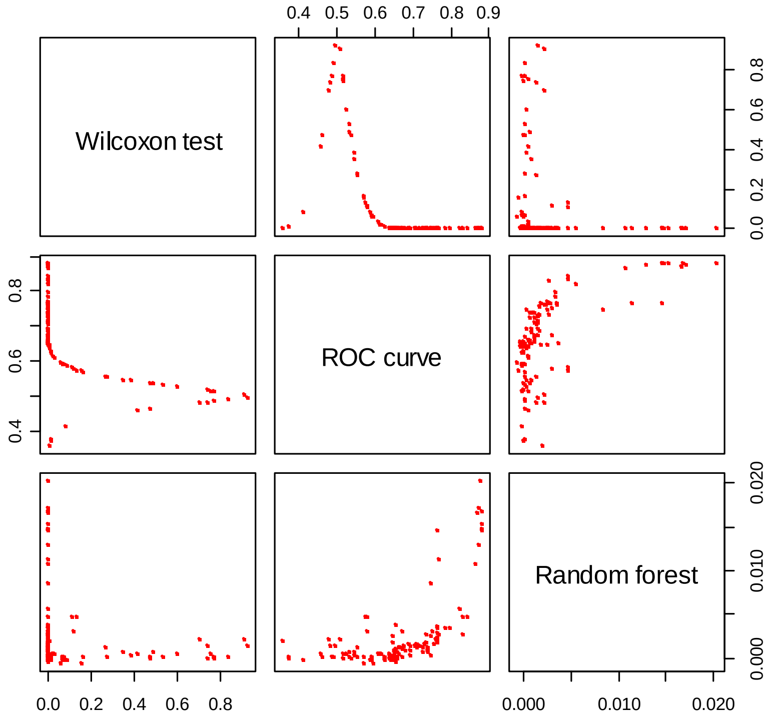
The five features, 2040.27, 2114.28, 2474.29, 3182.12, and 3699.61, were identified only by the RF method. Seven features, 3370.65, 3475.49, 6715.52, 6717.05, 6922.05, 6923.60, and 6925.15, were identified by both Wilcoxon test and ROC curve method. Eight features, 3369.57, 3476.59, 6718.58, 6720.11, 6721.64, 6926.70, 6928.26, and 6929.81, were identified by the ROC curve method only.

The results from Table 6-4 demonstrates that although the twenty-six features 3354.96, 3362, 3463.4, 3468.89, 3471.09, 3472.19, 3473.29, 3474.39, 3475.49, 6709.4, 6710.93, 6712.46, 6713.99, 6912.73, 6915.83, 6917.39, 6918.94, 6920.49, 6922.05, 3367.4, 3368.49, 3370.65, 6715.52, 6923.6, 6925.15, 6717.05 identified by the Wilcoxon test (listed in the first column) were also identified by the ROC curve method (listed in the third column), but the ranks of these features in the two feature sets identified by the Wilcoxon test and the ROC curve method are different. This indicates that if the p-values of the Wilcoxon test are too small (p-value = 0 in this study) to distinguish the discriminatory powers of the protein features, then the ROC curve method can be used for this purpose.

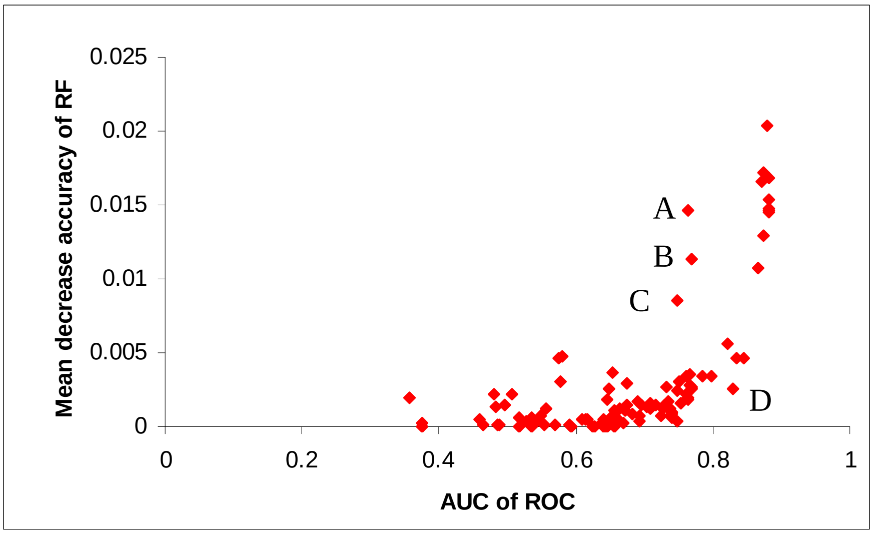
6.3.2 Concordance and discrepancy

Figure 6-8 plots the distributions of the discriminatory powers of the protein features identified by the Wilcoxon test, ROC curve and RF methods from the prostate cancer data set. Figure 6-8 (b) shows the zoomed distributions of the discriminatory powers of the protein features identified by the ROC curve and RF methods. It can be seen that some features can only be identified by one method, but not by the other. For example, the features A (5257.49), B (5250.73) and C (5260.20) were identified by the RF method, but not by the ROC curve method. By contrast, the feature D (3478.79) was identified by the ROC curve method, but not by the RF method.

Figure 6-9 plots the distributions of the discriminatory powers of the protein features identified by the Wilcoxon test, ROC curve and RF methods from the ovarian cancer data set. Figure 6-9 (b) shows the zoomed distributions of the discriminatory powers of the protein features identified by the Wilcoxon test and the RF method. Again the results show that some features can only be identified by one method, but not by the other. For example, the features E (6925.15) and F (6923.60) were identified by the Wilcoxon test, but not by the RF method. By contrast, the feature G (2474.29) was identified by the RF method, but not by the Wilcoxon test.

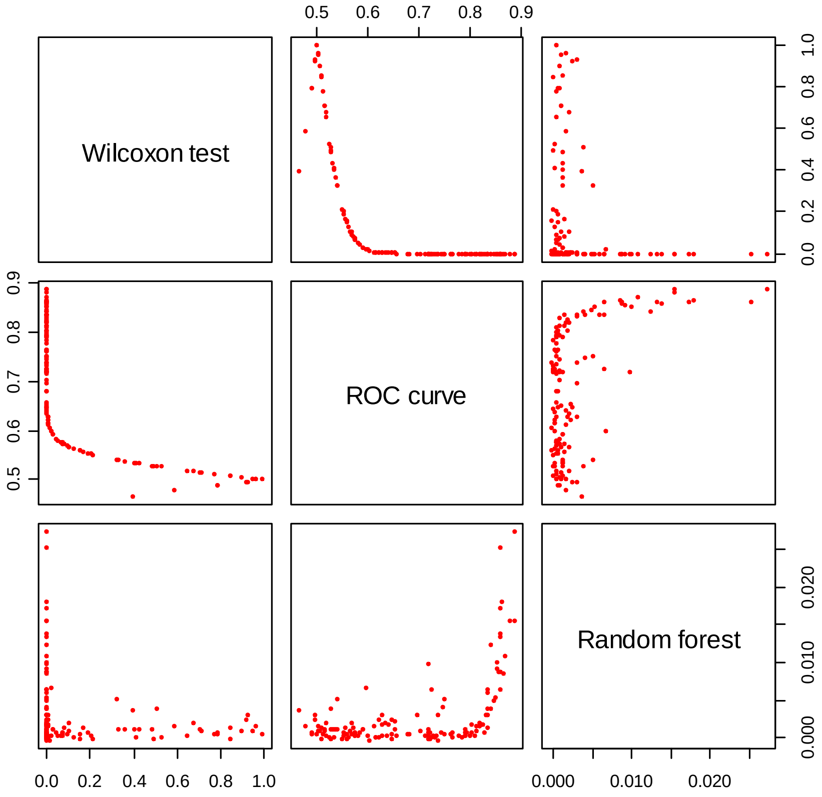


(a)

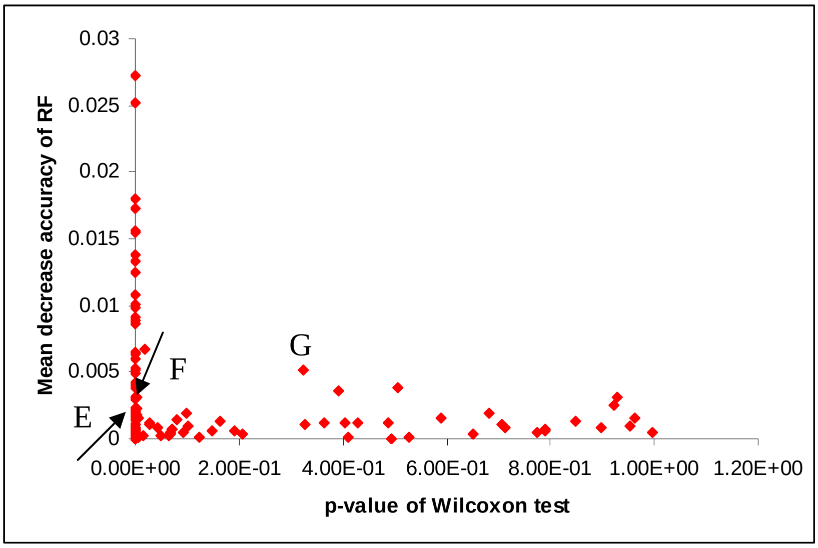


(b)

Figure 6-8 The distributions of the discriminatory powers of the protein features identified by the Wilcoxon test, ROC curve and RF methods from the prostate cancer data set.



(a)



(b)

Figure 6-9 The distributions of the discriminatory powers of the protein features identified by the Wilcoxon test, ROC curve and RF methods from the ovarian cancer data set.

6.3.3 Effect of different feature selection methods on the performance of cancer classification

Classification models were constructed using SVM, RF, and kNN machine learning algorithms based on the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods, the intersection set of the biomarkers, and the union set of the biomarkers for the prostate cancer data set and ovarian cancer data set, respectively. The means and standard deviations of misclassification rates of these classifiers are summarised in Table 6-5 and Table 6-7.

Table 6-5 The means and standard deviations of misclassification rates of the classifiers built using the different feature sets and classification approaches for the prostate cancer data set.

Feature set	**Misclassification rates (%)**		
	SVM	**RF**	**kNN**
Wilcoxon test	15.66 ± 1.37	17.11 ± 0.01	22.72 ± 0.00
ROC	17.85 ± 1.67	18.96 ± 0.01	2121 ± 0.00
RF	**9.87 ± 1.11***	14.15 ± 0.01	**16.67 ± 0.00***
Intersection	18.99 ± 1.52	21.84 ± 0.01	20.45 ± 0.00
Union	11.04 ± 1.10	**13.94 ± 0.01***	22.72 ± 0.00

Table 6-5 shows that the misclassification rates of the SVM classifiers vary from 10% to 19%. The best performance of the SVM classifiers was achieved with the proteomic biomarkers discovered by the RF method. The misclassification rates of the RF classifiers vary from 14% to 22%. The best performance of the RF classifiers was achieved with the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods. However, the misclassification rate of the RF classifier built using the set of proteomic biomarkers discovered by RF is only slightly lower compared with that of the RF classifier built using the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods. This suggests that the performance of RF classifiers may be relatively insensitive to changes in individual input variables. The misclassification rates of the kNN classifiers vary from 17% to 23%. The best performance of the kNN classifiers was achieved with the proteomic biomarkers discovered by the RF method. From these analyses, the optimal proteomic biomarkers have been selected from the prostate cancer data set and shown in Table 6-6. They are the set of protein features discovered by the RF method. The best classification accuracy is 90%.

Table 6-6 The optimal proteomic biomarkers identified from the prostate cancer data set

Feature (M/Z)	Identified by RF	Identified by ROC	Identified by Wilcoxon test
3468.89	√	√	√
3471.09	√	√	√
3472.19	√	√	√
3474.39	√	√	√
3475.49	√	√	√
3478.79	√	√	√
5250.73	√	X	X
5257.49	√	X	X
5260.20	√	X	X
6914.28	√	√	√
6915.84	√	√	√
6917.39	√	√	√
6918.94	√	√	√
6920.49	√	√	√
6922.05	√	√	√
6925.15	√	√	√
7633.59	√	X	X
7649.10	√	X	X
√ represents Yes; X represents No			

Table 6-7 shows that the misclassification rates of the SVM classifiers vary from 11% to 19%. The best performance of the SVM classifiers was achieved with the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods. The misclassification rates of the RF classifiers vary from 13% to 20%. The best performance of the RF classifiers was achieved with the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods.

However, the misclassification rate of the RF classifier built using the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods is only slightly lower compared with that of the RF classifier built using the set of proteomic biomarkers discovered by RF. Again, this suggests that the performance of RF classifiers may be relatively insensitive to changes in individual input variables. The misclassification rates of the kNN classifiers vary from 19% to 27%. The best performance of the kNN classifiers was achieved with the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods.

Table 6-7 The means and standard deviations of misclassification rates of the classifiers built using the different feature sets and classification approaches for the ovarian cancer data set.

Feature set	**Misclassification rates (%)**		
	SVM	**RF**	**kNN**
Wilcoxon test	18.75 ± 0.90	19.53 ± 0.01	26.75 ± 0.00
ROC	17.38 ± 1.05	18.80 ± 0.01	23.73 ± 0.00
RF	16.21 ± 1.21	13.09± 0.01	22.71 ± 0.00
Intersection	19.37 ± 0.83	20.42 ± 0.01	27.27 ± 0.00
Union	**10.61 ± 0.19 ***	**12.94 ± 0.01 ***	**19.00 ± 0.00 ***

From these analyses, the optimal proteomic biomarkers have been selected from the ovarian cancer data set and shown in Table 6-8. They are the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods. The best classification accuracy is 89%.

Table 6-8 The optimal proteomic biomarkers identified from the ovarian cancer data set

Feature (M/Z)	Identified by RF	Identified by ROC	Identified by Wilcoxon test
2040.27	√	X	X
2114.28	√	X	X
2474.29	√	X	X
3182.12	√	X	X
3354.96	√	√	√
3362.00	√	√	√
3367.40	√	√	√
3368.49	√	√	√
3369.57	X	√	X
3370.65	X	√	√
3463.40	√	√	√
3468.89	√	√	√
3471.09	√	√	√
3472.19	√	√	√
3473.29	√	√	√
3474.39	√	√	√
3475.49	X	√	√
3476.59	X	√	X
3699.61	√	X	X
6709.40	√	√	√
6710.93	√	√	√
6712.46	√	√	√
6713.99	√	√	√
6715.52	X	√	√
6717.05	X	√	√
6718.58	X	√	X
6720.11	X	√	X
6721.64	X	√	X
6912.73	√	√	√
6915.83	√	√	√
6917.39	√	√	√
6918.94	√	√	√
6920.49	√	√	√
6922.05	X	√	√
6923.60	X	√	√
6925.15	X	√	√
6926.70	X	√	X
6928.26	X	√	X
6929.81	X	√	X

√ represents Yes; X represents No

6.4 Discussion

SELDI technology is a rapid growing scientific field. The highly promising proteomic patterns generated by SELDI suggest that this technology could be used clinically for biomarker discovery, early detection of cancer, and the prediction of treatment response. However, identification of huge number of proteins from complex biological samples is still a challenge in the field of quantitative proteomics.

Several studies (Datta & Lara,2006; Levner,2005; Listgarten & Emili,2005; Marchiori et al.,2005; Wu et al.,2003) investigated the performance of different feature identification and classification methods and reported that better performance does not necessarily correspond to stability of the method and interpretability of results. Although there are a variety of feature selection methods available, including both univariate and multivariate methods, a robust feature selection method for the identification of proteomic biomarkers from SELDI expression profiles does not exist. In addition, there is no prior for deciding which feature selection methods should be used.

We have proposed an integrated algorithm to address this problem. The algorithm has been tested using three feature selection methods and three machine learning approaches for the two published SELDI data sets. The analysis results did show that different feature selection methods identified different sets of biomarkers, and that different machine learning approaches lead to different classification accuracy. The three feature selection methods selected a total of 39 biomarkers from 160 peaks for the ovarian cancer data set. Nineteen out of the 39 biomarkers were selected by all the three feature selection methods, suggesting that the common biomarkers identified by the three feature selection methods account for 49% for the ovarian cancer data set. The three feature selection methods selected a total of 18 biomarkers from 125 peaks for the prostate cancer data set. Fifteen out of the 18 biomarkers were selected by all the three feature selection methods, suggesting that the common biomarkers identified by the three feature selection methods account for 72% for the prostate cancer data set. The misclassification rates of the classification models built using SVM, RF, and k-NN varies from 11% to 27% for the ovarian cancer data set, and 10% to 23% for the prostate cancer data set. The results also show that the biomarkers that fail to be discovered by some feature selection methods could be identified by others. The discrepancies among the biomarker sets identified by different methods make it very difficult to decide which set of proteomic biomarkers should be reported and used for cancer diagnosis. By applying our proposed algorithm, the optimal set of proteomic biomarkers has been selected for both data sets. However, we should be cautious in interpreting the optimal set of proteomic biomarkers. Firstly, the integrated algorithm is based on the process of optimisation. The optimal set of proteomic biomarkers is selected if a classifier built with these proteomic biomarkers achieves the best classification performance. However, the second best classification performance may not be remarkably different from the best classification performance. That is to say, the identified optimal set of proteomic biomarkers may be un-unique. As showed in the results of applying the algorithm to the published SELDI data sets, this is especially the case for RF classifiers. Taking the classification performance of different types of classifiers into account can solve this problem. For example, for the ovarian cancer data set, if only RF classifiers were used it would be very difficult to determine the optimal set of proteomic biomarkers as the RF classifiers built using the set of proteomic biomarkers discovered by RF and the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods gave relatively similar classification accuracy. However, both SVM and kNN classifiers confirm that the set of proteomic biomarkers discovered by RF feature selection method is an optimal one. Similarly, for the prostate cancer data set, SVM and kNN classifiers confirm that the union of the proteomic biomarkers discovered by the Wilcoxon test, ROC curve and RF methods is an optimal set of proteomic biomarkers. Secondly, due to the instrument resolution, some proteomic biomarkers in the optimal set might represent the same protein. This needs further experiments to verify.

It is worthwhile to mention that although only three feature selection methods and three machine learning approaches have been applied to the proposed algorithm in this chapter, the algorithm for proteomic biomarker discovery for cancer diagnosis itself has no any limit on the use of feature selection and classification methods. In general, any existing feature selection method and data-mining tool as well as future developed methods will fit into the algorithm. The algorithm has an advantage that it does not exclude any method before exploring it. It might be reasonable to assume that it is wise to use the multivariate methods to analyse genomics and proteomics data as genes usually work in a collaborative way, rather than in an independent way. However, the study by Zhang et al. (2006) reported that although the multivariate SVM-based method outperformed the univariate method in terms of the performance of classification, univariate methods can reveal more of the differentially expressed features especially when they are correlated. This suggests that univariate methods are still useful in the identification of proteomic biomarkers from SELDI protein profiles.

However, lack of normality of SELDI protein profiles encourages the use of non-parametric approaches, such as Wilcoxon test and the ROC curve method, for biomarker discovery. The results have demonstrated and confirmed that the AUC values under the ROC curve is equivalent to the p-values of the Wilcoxon tests. But in a case when it is impossible to rank the protein features according to the p-values of the Wilcoxon test, the AUC values under the ROC curve can be used to distinguish the discriminatory powers of these protein features.

6.5 Summary

We have investigated the feature selection and cancer classification methods, including both univariate and multivariate approaches. We argue that both the univariate and multivariate methods should be useful in the identification of proteomic biomarkers from SELDI protein profiles. However, none of the previous studies provided a guideline in selecting methods for feature selection and cancer classification.

We have proposed an integrated algorithm for proteomic biomarker discovery from SELDI protein profiles. It has been applied to two published SELDI data sets to show how it works. In addition, the results have also demonstrated that the ROC curve method may be more reliable to detect proteomic biomarkers, especially to determine the discriminatory powers of the identified biomarkers, compared to the Wilcoxon test.

Although only three feature selection methods and three machine-learning approaches were used in this study, the proposed algorithm for proteomic biomarker discovery presented here may be generalisable and applicable to other mass spectrometry and genomics approaches. It should be interesting to further test the proposed algorithm using a large SELDI data set generated within a centre or across different centres.

7 CONCLUSIONS AND FUTURE WORK

This chapter draws together the main points made in the thesis. This is followed by suggestions for possible future work.

7.1 Conclusions

The conclusions drawn from this thesis consists of the following components:

(a) Cancer research and needs for identifying biomarkers,

(b) SELDI technology and data analysis

(c) Quantitative measures of reproducibility

(d) SELDI mass spectrum peak alignment

(e) Normalisation and normality test of SELDI protein profiles

(f) Detection of proteomic biomarkers for cancer diagnosis

(g) Cancer prognosis and prediction of treatment response using SELDI proteomic biomarkers

(h) Integration of environmental factors with SELDI proteomic biomarkers.

7.1.1 Cancer research and needs for identifying biomarkers

The progress in the understanding of cancer initiation and progression has been frustrating. The mortality rates for the most common cancers have not been significantly reduced. Some of the best available options to combat cancer include primary prevention, earlier diagnosis, and improved therapeutic interventions. The progress in biotechnology provides an opportunity to detect biomarkers, which could facilitate the understanding of the mechanisms of initiation, prevention, diagnosis and prognosis of cancer. Functional genomics, such as microarray technology, has attracted a great interest because it provides the potential ability to monitor the expression of the whole genome on a single chip. However, comparative transcriptional profiling alone is unlikely to fully identify biomarkers that are associated with cancer phenotype. The investigation of RNA expression obtained from microarray data may be an indirect way to understand the aetiology of cancer. Proteomics is the latest functional genomics technology and is really a main target of our interest to understand biological systems, detect biomarkers for cancer diagnosis, monitor disease progression, and identify therapeutic targets. The proteome contains not only the intrinsic genetic information of a cell, but also the impact of its immediate environment. A transformation of a healthy cell into a neoplastic cell may cause altered expression profile, differential protein modification and activities, and this in turn may affect cellular function. Therefore, investigation of the cancer proteome could be a starting point in the identification of biomarkers. Current progresses in proteomics, especially in SELDI, have demonstrated the potential for biomarker discovery and early cancer diagnosis. It is important to note that the protein identities of SELDI expression profiles are unknown. However, this might not prevent SELDI technology from becoming a clinical tool because each protein has a unique m/z value and the combination of specific SELDI protein profiles could form proteomic patterns.

7.1.2 SELDI technology and data analysis

We have reviewed SELDI technology and the statistical methods for SELDI data analysis. One approach to biomarker discovery for cancer classification has been presented and discussed. It includes issues on experiment designs, reproducibility analysis, SELDI protein profile pre-processing, identification of putative biomarkers, and cancer classification. We have identified the problems in analysing SELDI protein profiles, especially in reproducibility analysis, peak alignment, data normalisation and transformation, and selection of statistical methods for detection of proteomic biomarkers and cancer diagnosis.

7.1.3 Quantitative measures of reproducibility

We have proposed three quantitative measures of reproducibility using Euclidean distance, correlation coefficient, and the paired t-test. The reproducibility of SELDI protein profiles has been investigated using a colon cancer data set. The results showed that the differences of SELDI protein profiles between identical samples over replicate experiments are statistically significant in terms of Euclidean distances and correlation coefficients. The mass spectra generated in different spots in one experimental run have a better reproducibility compared to the mass spectra generated in different experimental runs. The mass spectra generated in different experimental runs over a short period of time have a better reproducibility than the mass spectra generated in different experimental runs over a long period of time. The reproducibility analyses have found that the mass spectra of the identical samples had significant changes if they were left in ice for a period of 4-5 hours at room temperature. This suggests a new conjecture that protein profiles are affected by sample storage and processing procedures. The reproducibility analysis results suggest that standardised experiment protocol and analysis procedure should be adopted. The proposed quantitative measures of reproducibility should be useful to identify important experimental conditions for improving the reproducibility of SELDI protein profiles.

7.1.4 SELDI mass spectrum peak alignment

We have proposed two peak alignment algorithms and compared them with another four existing peak alignment methods. Five out of six peak alignment methods (including the newly proposed two algorithms) reduce the median CV by more than 10% compared to the results of the un-aligned raw mass spectra, while keeping the IQR in the same range. The five peak alignment methods also increase correlation coefficient by more than 3%. One of the two proposed peak alignment methods achieves the best performance in terms of correlation coefficient and coefficient of variance. The effects of peak alignment on the classification accuracy have also been evaluated by training RF models using the un-aligned and aligned mass spectra using the six peak alignment methods. The results show that five of the six RF models trained using data set aligned by the peak alignment methods improve the AUC values by up to 2% compared to the RF model trained using raw data set.

These results demonstrate that the peak alignment algorithms have a positive effect on reducing systematic bias in producing SELDI mass spectra. A small improvement in cancer classification accuracy has been observed, which is not statistically significant. However, because the relatively small number of mass spectra generated on one machine under a stable environment within a short period of time was used in this study, the effects of peak alignment methods on the performance of the classification models might not be detected.

7.1.5 Normalisation and normality test of SELDI protein profiles

We have examined the assumption of normality of SELDI protein profiles, on which the standard statistical methods are based. The results from the Shapiro-Wilks, Kolmogorov-smirnov, Cramér-von Mises and Pearson χ^2 statistical tests suggest that a large number of protein profiles do not follow normal distribution. Furthermore, the goodness-of-fit tests reject the null hypothesis of normality, implying that SELDI protein profiles do not follow normal distributions. The normalisation analysis has shown that our proposed intensity normalisation algorithm outperforms other existing methods. The intensity transformation studies have shown that the logarithmic transformations yield the highest improvement of normality of SELDI protein profiles compared to other mathematical transformations. However, the intensity normalisation and transformations still cannot convert SELDI protein profiles into normally distributed data. Therefore, it may be unwise to apply the standard transformation function to SELDI data set and then employ normal-theory based statistical methods, such as the t-test. We think that it may be more reliable to use non-parametric methods, such as Wilcoxon test and the ROC curve method, as they overcome the many shortcomings of their parametric counterparts for high-throughput SELDI data analysis for detecting disease-associated proteins.

7.1.6 Detection of proteomic biomarkers for cancer diagnosis

We have investigated the feature selection and cancer classification methods, including both univariate and multivariate approaches. The advantage of the univariate feature selection methods is that they are simple. However, multiple testing on a large number of SELDI protein profiles will generate many false positives if the standard significance level is used. The commonly used multiple hypotheses testing correction methods have been studied to address this problem. The advantage of multivariate feature selection methods is that they take the interactions between SELDI protein profiles into account. However, they may reveal less of the differentially expressed features compared to the univariate methods if some SELDI protein profiles are correlated. That is to say, both the univariate and multivariate methods should be useful in the identification of proteomic biomarkers from SELDI protein profiles.

There is evidence to show that different feature selection methods usually discover different sets of proteomic biomarkers, and that different cancer classification algorithms often produce different classification accuracies. However, none of the previous studies provided a guideline in selecting methods for proteomic biomarker discovery and cancer classification.

We have proposed an integrated algorithm for proteomic biomarker discovery from SELDI protein profiles. It has been applied to two published SELDI data sets. The optimal sets of biomarkers were found for both data sets. In addition, the results have also demonstrated that the ROC curve method may be more reliable to detect proteomic biomarkers, especially to determine the discriminatory powers of the identified biomarkers, compared to the Wilcoxon test.

Although only three feature selection methods and three machine-learning approaches were used in this study, the proposed algorithm for proteomic biomarker discovery presented here may be generalisable and applicable to other mass spectrometry and genomics approaches. It should be interesting to further test the proposed algorithm using a large SELDI data set generated within a centre or across different centres.

7.1.7 Cancer prognosis and prediction of treatment response using SELDI proteomic biomarkers

The proposed algorithm for biomarker discovery and cancer classification using SELDI technology can be easily adapted for cancer prognosis and for predicting treatment response. The prediction of clinical response can be modelled as follows: $y = f(I_1, I_2, ..., I_m, r_1, r_2, ..., r_l)$, where, $I_i, i = 1,2,...,m$ are the identified proteomic biomarkers, $r_j, j = 1,2,...,l$ are the available treatment regimens, and y is a continuous response variable of interest, for example, the cancer survival time. The algorithm for cancer prognosis and the prediction of treatment response is outlined below.

(1) Each feature selection method is applied to the same SELDI data set, and a set of proteomic biomarkers is generated, which is denoted by, B_i, where, i is the ith feature selection method.

(2) The intersection set of the protein features identified using these individual feature selection methods is found and denoted by B_I, where, $B_I = \bigcap B_i$.

(3) The union set of the protein features identified using these individual feature selection methods are found and denoted by B_U, where, $B_U = \bigcup B_i$.

(4) Mathematical models are trained using each supervised machine learning algorithm and the individual proteomic feature set B_i, B_I and B_U. The corresponding performance of the models (e.g. the residual errors) are recorded and denoted by E_j, where j is the jth trained model. Note that the models can be trained using one method or combination of different methods.

(5) The optimal set of proteomic biomarkers is determined and denoted by B_o, where, $B_o = \{B_j, j = i, I, U \mid \min_j(E_j)\}$.

(6) Simulation analysis is carried out using the optimal set of proteomic biomarkers and the best trained model to search for the best available treatment regimen for a given cancer patient based on her/his proteomic patterns.

7.1.8 Integration of environmental factors with SELDI proteomic biomarkers

So far the thesis has focused on cancer biomarker discovery and on capturing the underlying relationship between the identified proteomic patterns and the cancer status or the therapeutic response. However, most common cancers and clinical quantitative traits are extremely complex and result from interactions between many proteins and various environmental factors. For example, the proteomic patterns may contain information that influences human's physical traits, one's likelihood of suffering from cancer, and the responses of one's body to substances that one encounters in the environment. It is therefore desirable to incorporate environmental factors in the classification and prediction models. To our knowledge, none of the published SELDI studies considered the interactions among proteins and between proteins and environmental factors, and investigated the influence of environmental factors on the performance of classification or prediction models.

Our proposed algorithms for proteomic biomarker discovery and for cancer prognosis can be readily extended to integrate environmental factors. For example, the prediction of clinical response with consideration of environmental factors can be modelled as follows: $y = f(I_1, I_2, ..., I_m, r_1, r_2, ..., r_l, e_1, e_2, ..., e_s)$, where, $I_i, i = 1,2,..., n$ are the identified proteomic biomarkers, $r_j, j = 1,2,..., l$ are the available treatment regimens, $e_t, t = 1,2,..., s$ are the environmental factors, and *y* is a response variable of interest.

7.2 Future work

- The reproducibility of SELDI technology needs further investigation within a centre and across different centres.
- The proposed peak alignment and normalisation algorithms, and the normality issues need to be further tested using a large set of mass spectra generated from different SELDI machines at different times.
- It needs further tests on the proposed algorithm for proteomic biomarker discovery, and cancer diagnosis and prognosis using other commonly used feature selection, classification and prediction methods, such as wavelet decomposition, multivariate adaptive regression spline, Bayesian network, and artificial neural network.

8 APPENDIXES

Appendix 8-1 Publications

Cheng, Y., Nyangoma, S. O., & Johnson, P. J. (2008) "Identifying Biomarkers from SELDI-TOF Protein Profiles: An Integrative Approach". In: **XXIVth International Biometric Conference,13-18th July, 2008, Ireland**.

Cheng, Y., Ward, D. G., Wen, W., Billingham, L. J., & Johnson, P. I. (2006) "Plasma proteome profiling of liver tumours in flatfish". In: **NCRI conference, 8-11th October, 2006, Birmingham**.

Ward, D. G., Cheng, Y., N'Kontchou, G. et al (2006a) Changes in the serum proteome associated with the development of hepatocellular carcinoma in hepatitis C-related cirrhosis. **Br.J.Cancer,** 94 (2): 287-292

Ward, D. G., Cheng, Y., N'Kontchou, G. et al (2006b) Preclinical and post-treatment changes in the HCC-associated serum proteome. **Br.J.Cancer,** 95 (10): 1379-1383

Ward, D. G., Suggett, N., Cheng, Y. et al (2006c) Identification of serum biomarkers for colon cancer by proteomic analysis. **Br.J Cancer,** 94 (12): 1898-1905

Ward, D. G., Wei, W., Cheng, Y. et al (2006d) Plasma proteome analysis reveals the geographical origin and liver tumor status of Dab (Limanda limanda) from UK marine waters. **Environ.Sci.Technol.,** 40 (12): 4031-4036

Appendix 8-2 Programs for reproducibility analysis

```
### To generate average of duplicate spectra

rm(list=ls())
aveInten <- function(multiple) {
   Nrow <- nrow(multiple)
   ave = matrix(nrow = Nrow, ncol = 1)
   for(i in 1:Nrow){
        ave[i] <- mean(as.numeric(multiple[i,2:3]))
   }
   Ave
}

## read in data
files<-dir("./PABS", pattern="*.csv", full.names=TRUE)
fn <- "blank"
flag <- 0

for(file in files) {
        data <- read.table(file, header = TRUE, sep = ",")
        fileN <-gsub("_d", "", file)
        if (fn != fileN){# meet this file firstly
           if (fn != "blank") {# save average file
              data2 <- aveInten(data1)
              data3 <- cbind(data1[,1], data2[,1])
              write.table(data3, file = basename(fn), sep = ",", append = FALSE,
col.names = FALSE, row.names = FALSE)
              fn <- "blank"
           }
           fn <- fileN
           data1 <- data
         } else if (fn == fileN) {# meet same file nth times
              data1 <- cbind(data1, data[,2])
         }
}
## last spectrum
data2 <- aveInten(data1)
data3 <- cbind(data1[,1], data2[,1])
write.table(data3, file = basename(fn), sep = ",", append = FALSE, col.names =
FALSE, row.names = FALSE)
```

```
### To merge single spectrum file into one spectra file
# files: list of files to be read in.
# mzStart: the M/Z value from which the intensity will be read in.
# heager: indicating whether the file contains the header in its first line.
#1 represents TRUE and 0 FALSE.
# output: one matric which contains all the data.

GetIntens <- function(files, mzStart, header) {
   flag <- 0
   for(fname in files) {
      f <- basename(fname)
      ff <- gsub("low good spectra ", "", f)
      fff <- gsub(".csv", "", ff)
      flag <- flag + 1
      if( header == 1) {
         single<-read.table(fname, header = TRUE, sep = ",", col.names = 
c("M/Z",fname))
      } else {
          single<-read.table(fname, header = FALSE, sep = ",", col.names = 
c("M/Z",fname))
      }
      single<- subset(single, single[1] > mzStart)
      if (flag == 1) {
        multiple <- single[,1]
        multiple <- cbind(multiple, single[,2])
        names <- fff
         } else {
       multiple <- cbind(multiple, single[,2])
       names <- c(names, fff)
      }
   }
   colnames(multiple) <- c("M/Z", names)
   Multiple
}
```

Appendix 8-3 Programs for peak alignment study

### To calculate correlation coefficients and Euclidean distance of the duplicate spectra

```
# spectrumA.
# spectrumB: duplicate spectrum of "spectrumA".
CorDisDuplicate <- function(spectrumA, spectrumB) {
       numSample <- ncol(spectrumA) - 1 #first column is m/z value
       cordis = matrix(nrow = numSample, ncol = 4)
       numRow <- nrow(spectrumA)
       for (i in 1:numSample) {
           j <- i+1
           cordis[i,1] <- colnames(spectrumA)[j]
           cordis[i,2] <- colnames(spectrumB)[j]
           cordis[i,3] <- cor(spectrumA[1:numRow,j], spectrumB[1:numRow,j])
           cordis[i,4] <- sqrt((spectrumA[1:numRow,j]-
                              spectrumB[1:numRow,j])%*%(spectrumA[1:numRow,j]-
                              spectrumB[1:numRow,j])/numRow)
      }
      colnames(cordis) <- c("ID", "ID_D", "Cor", "Dis")
      Cordis
}

### for control group before alignment
files <- dir("./data/befor_alignment/control", pattern = "(low good spectra)[ ][0-9]*[
             ](n).csv",full.names=TRUE)
filesD <- dir("./data/before_alignment/control", pattern = "(low good spectra)[ ][0-
9]*[
             ](n_d).csv",full.names=TRUE)
ins <- GetIntens(files,2000,1)
insD <- GetIntens(filesD,2000,1)
cordis <- CorDisDuplicate(ins, insD)
write.table(cordis, file = "result/CorDisControl.csv", sep = ",", append = FALSE,
                 col.names = TRUE, row.names = FALSE)
```

### To perform McNemar's chi-squared test.

```
### ref: http://www.medcalc.be/manual/mpage06-15.php

per <- matrix(c(31, 7, 11, 29), nr = 2, dimnames = list("pabsd" = c("cancer",
"control"),
"before" = c("cancer", "control")))
mcnemar.test(per)
```

### To perform spectrum alignment (MATLAB)

```
Clear
seldi_input_parameters  % Load parameters for analysis;

% Fast peak alignment of spectra using a beam search algorithm
% REFERENCE: G.-C. Lee and D.L. Woodruff, Beam Search for Peak Alignment
of NMR Signals
% submitted to Analytica Chimica Acta (2004).

% read in reference spectrum
refName = input('Enter directory name of reference spectrum: ', 's');
cd (refName);
MZ_Inten_ref = csvread('reference.csv',1); %data includes mz and intensity
MZ = MZ_Inten_ref(:,1); % get mz ratio value

% Read in spectra to be aligned
directoryName = input('Enter directory name of SELDI spectra to be aligned: ', 's');

% Spectra in MATLAB working directory.
cd (directoryName);
files = dir('*.csv');

% Preallocate some space for the data.
numSamples = numel(files);
numDataPoints = numel(MZ);
Intens = zeros(numDataPoints,numSamples);

% Loop over the files and read in the data.
for i = 1:numSamples
   MZ_Inten = csvread(files(i).name, 1);
   Intens(:,i) = MZ_Inten(:,2);
End
disp (' ');
cd ../ % for protecting orig data

% Start peak alignment based upon linear distribution
% Calculate segmentation positions.
ref_spec = MZ_Inten_ref(:,2);
raw_length = length (ref_spec);
start_point = 1;
segment = input('Enter number of segments [default = 50]: ');
if isempty(segment)
   segment = 50;
End
disp (' ');
```

```
si = raw_length;
s_index = round (raw_length / segment);
s = [s_index : s_index : raw_length];

% Call peak alignment algorithm
% calculate "ra" "rb"
% ra: maximum range of sideway movement
% rb: maximum range searched forward to find local minimum in ref and target
ra_state = input('Maximum sideway movement range dynamica or fixed? [dynamica
= 1 (default) and fixed = 2]: ');
if isempty(ra_state)
   ra_state = 1;
End
methods = input('Enter alignment methods. [PAES = 1 (default); PABS = 2;
PAFMCC = 3]: ');
if isempty(methods)
   methods = 1;
End
if methods == 2
   bw = input('Enter beam width 1 or 2. [beam width = 1 (default)]: ');
   if isempty(bw)
      bw = 1;
   End
else bw = 0;
End

for n = 1 : numSamples
   [spec_shift,optima,cc] =
seldi_fastpa(Intens(:,n),ref_spec,s,MZ(:,1),ra_state,rb,bw,fig,si, methods);
   %spec_shift = [files(i).name spec_shift']'
   Intens_aligned(:,n) = spec_shift;
   cc_matrix = [cc_matrix cc];
   shift_matrix = [shift_matrix optima(:,1)];
End

cc_av = mean (cc_matrix');
disp (['Average correlation coefficient for all alignments = ' num2str(cc_av)]);
cc_std = std (cc_matrix');
disp (['Std deviation of correlation coefficients for all alignments = '
num2str(cc_std)]);

%save the aligned spectra on directory named "data_aligned'
cd (directoryName);
cd ../;
if (exist('data_aligned')== 7)
      rmdir('data_aligned','s');
End
mkdir data_aligned;
```

```
cd data_aligned;
for i = 1:numSamples
   mz_intens_aligned = [MZ_Inten_ref(:,1) Intens_aligned(:,i)]; % plus mz value
   csvwrite(files(i).name, mz_intens_aligned);
End
cd ../
```

```
### To generate ROC curve and calculate AUC.
### Original author: Hemant Ishwaran. ishwaran@bio.ri.ccf.org
### Modified by Yaping Cheng. YXC466@bham.ac.uk

rocValue <- function(group,ordinal)
{
      group.uniq <- sort(unique(group))
      neg <- ordinal[group==group.uniq[1]]
      pos <- ordinal[group==group.uniq[2]]
      n.neg <- length(neg)
      n.pos <- length(pos)
      range.data <- range(neg,pos)
         #xtxt <- 'Test Results'

###   Find all distinct atoms generated by pos and neg
###   test values.
      atoms <- unique(sort(c(neg,pos)))
      atoms.neg <- unique(sort(neg))
      atoms.pos <- unique(sort(pos))
      n.atoms <- length(atoms)
      tp <- rep(0,n.atoms)
      fp <- rep(0,n.atoms)
      #cdf.neg <- rep(0,length(atoms.neg))
      #cdf.pos <- rep(0,length(atoms.pos))

###   Compute tp and fp values for each distinct atom
      for ( i in 1:n.atoms) {
         if (mean(atoms.pos) > mean(atoms.neg)) {
              tp[i] <- sum(pos >= atoms[i])/n.pos
              fp[i] <- sum(neg >= atoms[i])/n.neg
            } else {
              fp[i] <- sum(pos >= atoms[i])/n.pos
              tp[i] <- sum(neg >= atoms[i])/n.neg
            }
         }

###   Nice touch is to add end values for tp,fp values and cdf's
         tp <- c(1,tp,0)
```

```
            fp <- c(1,fp,0)
            delta <- (range.data[2]-range.data[1])/20
            atoms.neg <- c(atoms[1]-delta,atoms.neg,atoms[n.atoms]+delta)
            atoms.pos <- c(atoms[1]-delta,atoms.pos,atoms[n.atoms]+delta)

###   (1) Calculate area under curve using Wilcoxon U-statistic
###   (2) Estimate its standard error
            tp.mean <- (tp[1:(n.atoms+1)] + tp[2:(n.atoms+2)])/2
            fp.diff <- -diff(fp)
            area.U <- sum(fp.diff*tp.mean)
            area.S <- sqrt((area.U*(1 - area.U)+(n.pos - 1)*(area.U/(2 - area.U)
                         - area.U^2) + (n.neg - 1)*((2*area.U^2)/(1 + area.U)
                         - area.U^2))/(n.pos*n.neg))
        result <- list(fp = fp, tp = tp, areaU = area.U, areaS = area.S, Natoms = n.atoms)
        Result
}

rocPlot <- function(fp, tp, nbreaks=15, area.U, area.S, n.atoms,
minAtom.plot=25,mtxt="ROC plot for ",...)
{
            if(n.atoms>=minAtom.plot){
            plot(fp,tp,type="l",xlim=c(0.0,1.0),
                ylim=c(0.0,1.0),xlab="1-Specificity",
                ylab="Sensitivity",main=mtxt)
            text(0.7,0.1,paste('Area=',format(round(area.U,3))))
        }
            else{
                 plot(fp,tp,type="b",xlim=c(0.0,1.0),
                 ylim=c(0.0,1.0),xlab="1-Specificity",
                 ylab="Sensitivity",main="mtxt")
                 text(0.7,0.1,paste('Area=',format(round(area.U,3)),
                                  '+/-',format(round(area.S,3))))
            }
}

###  modified
rocPeak <- function (peaks) {
        nn <- nrow(peaks)
        ll <- ncol(peaks)
        auc_peak <- matrix(nrow = ll-2, ncol = 2)
        group <- as.numeric(peaks[,2])
        ### first column: sample ID; second column: group info.
        senSpe <- c()
        senSpeNames <- c()
        protein_names <- colnames(peaks)[3:ll]
        for (i in 3:ll) {
            j <- i-2
            auc_peak[j,1] <- protein_names[j]
            ordinal <- peaks[,i]
            rocV <- rocValue(group,ordinal)
            rocPlot(rocV$fp, rocV$tp,nbreaks=15, area.U=rocV$areaU,
area.S=rocV$areaS,n.atoms = rocV$Natoms,
minAtom.plot=25,mtxt=colnames(peaks)[i])
            auc_peak[j,2] <- rocV$areaU
            senSpe <- cbind(senSpe, rocV$fp, rocV$tp)
            senSpeNames <- c(senSpeNames, protein_names[j], protein_names[j])
        }
        colnames(auc_peak) <- c("Peak", "AUC")
        colnames(senSpe) <- senSpeNames
        result <- list(auc = auc_peak, ss = senSpe)
        Result
}
```

Appendix 8-4 Programs for normality test

```
### To perform goodnees-of-fit test for normality test
### Original autor: Linlin Chen (for microarray data)
### Modfied by Yaping Cheng (for proteomic SELDI data)
rm(list=ls())
### Read in the data set (row: protein  column: sample ID)
f <- dir("../data/peaks", pattern = "peak_median_cancer.csv", full.names=TRUE)
da <- read.table(f, header=TRUE, sep=",", row.name = 1)
da <- t(da)

### number of proteins
NN <- nrow(da)
### number of samples
LL <- ncol(da)
### number of resampling steps
SN <- 1000

### log transformation
ndata <- da
small <- min(ndata)
small <- 1-small
### replace values with "smallest" to make all datas >= 1
sdata <- ndata + small
da <- sdata

### to record the number of the rejection  for each sampling
yyn <- 1:SN
s <- 1:LL
size <- LL-10

library(nortest)
for(i in 1:SN)
{
    yyn[i] <- 0
    ### do the resampling, without replacement,
    ### size is LL-10, instead of LL
    ss <- sample(s, size)
    dd <- da[,ss]
     for(j in 1:NN) # for each gene
     {
         ### to perform the Pearson chi-square test for normality
         P <- pearson.test(dd[j,])$p.value
         if(p < 0.05)
             yyn[i]  <- yyn[i] +1
    }
}

r <- size
d <- 10
nd <- LL

### calculate t
### number of the rejections for each resampling
yy <- yyn
### 0.05*NN: the expected number of rejections under the complete null hypothesis.
### mean(yy): average number of rejection under each resampling
tt <- (mean(yy) - 0.05*NN)
aa <- sqrt(r/d)*(yy-mean(yy))
dd <- aa[aa > tt]
length(dd)
```

```
### Functions for normalisation, transformation and normality test

### row: m/z values
### column: samples

### Ciphergen: TIC
meanAUC <- function(data) {
    nc <- ncol(data)
    mea <- colMeans(data)
    nor_cor <- mean(mea)
    for(i in 1:nc) {
        data[, i] <- data[, i] * nor_cor/mea[i]
    }
    Data
}

### The newly proposed normalisation algorithm
medianAUC <- function(data) {
    nc <- ncol(data)
    medi <- matrix(nrow=1, ncol=nc)
    for(i in 1:nc) {
      medi[1,i] <- median(data[, i])
    }
    nor_cor <- median(medi)
    for(i in 1:nc) {
        data[, i] <- data[, i] * nor_cor/medi[i]
    }
    Data
}
```

```
### PROcess
library(PROcess)

### shapiro-will_test
swt <- function(data) {
    npv <- 0
    nr <- nrow(data)
    pvalue <- matrix(nrow=nr, ncol=1)
    for(i in 1:nr) {
        ss <- shapiro.test(as.numeric(data[i, ]))
        if(ss$p.value < 0.05) {
            npv <- npv +1
        }
    pvalue[I,1] <- ss$p.value
    }
    rownames(pvalue) <- rownames(data)
    results <- list(num_rej = npv, pvalues = pvalue)
    Results
}

library(nortest)
### Cramér-von-miss_test
cvmt <- function(data) {
    npv <- 0
    nr <- nrow(data)
    pvalue <- matrix(nrow=nr, ncol=1)
    for(i in 1:nr) {
        ss <- cvm.test(as.numeric(data[i, ]))
        if(ss$p.value < 0.05) {
            npv <- npv +1
        }
    pvalue[I,1] <- ss$p.value
    }
    rownames(pvalue) <- rownames(data)
    results <- list(num_rej = npv, pvalues = pvalue)
    Results
}

### kolmogorov_test
klg <- function(data) {
    npv <- 0
    nr <- nrow(data)
    pvalue <- matrix(nrow=nr, ncol=1)
    for(i in 1:nr) {
        ss <- lillie.test(as.numeric(data[i, ]))
        if(ss$p.value < 0.05) {
            npv <- npv +1
        }
    pvalue[I,1] <- ss$p.value
    }
    rownames(pvalue) <- rownames(data)
    results <- list(num_rej = npv, pvalues = pvalue)
    Results
}

### pearson_test
pers <- function(data) {
    npv <- 0
    nr <- nrow(data)
    pvalue <- matrix(nrow=nr, ncol=1)
    for(i in 1:nr) {
        ss <- pearson.test(as.numeric(data[i, ]))
        if(ss$p.value < 0.05) {
            npv <- npv +1
        }
    pvalue[I,1] <- ss$p.value
    }
    rownames(pvalue) <- rownames(data)
    results <- list(num_rej = npv, pvalues = pvalue)
    Results
}

### skewness&kurtosis
ske_kur <- function(data, id, num) {
    library(e1071)
    ske <- c()
    kur <- c()
    for (i in 1:num) {
        ske <- cbind(ske, skewness(as.numeric(data[id[i],])))
        kur <- cbind(kur, kurtosis(as.numeric(data[id[i],])))
    }
    results <- list(skew = ske, kurto = kur)
    Results
}
```

```
### PCA analysis

rm(list=ls())
library(MASS)

### read in data
```

```
f <- dir("../data/peaks/pca", pattern = "peaks_imp10.csv", full.names=TRUE)
data <- read.table(f, header=TRUE, sep=",", row.name = 1)

ins <- subset(data, select=-class)
group <- as.numeric(data[,1])
group.uniq <- sort(unique(group))
control <- row.names(ins)[group==group.uniq[1]]
cancer <- row.names(ins)[group==group.uniq[2]]

###### log transformation ######
small <- min(ins)
small <- 1-small
### replace values with "smallest" to make all datas >= 1
ins <- ins + small
ins <- log2(ins)


###### scores ######
pc <- princomp(ins)
plot(pc$scores[,1:2],col="red", xlab="PC1", ylab="PC2")
plot(pc$scores[,1:2],col="red", , type="n", xlab="PC1", ylab="PC2")
points(pc$scores[, 1:2][row.names(pc$scores[, 1:2]) %in% control,], col="blue",
pch=19,lwd=0.5)
points(pc$scores[, 1:2][row.names(pc$scores[, 1:2]) %in% cancer,], col="red",
pch=19,lwd=0.5)

text(3,3.5,"control", col="blue")
text(3,3.2,"cancer", col="red")

###### loadings ######
plot(pc$loadings[,1:2],col="red", xlab="PC1", ylab="PC2")
plot(pc$loadings[,1:2],col="red", , type="n", xlab="PC1", ylab="PC2")

num_var <- nrow(pc$loadings)
num_com <- ncol(pc$loadings)
for(i in 1:num_var) {
text(pc$loadings[i,1], pc$loadings[i,2],row.names(pc$loadings)[i] , col="blue")
}
```

```
### To find peaks using PROcess package through R

rm(list=ls())
library(PROcess)
f <- dir("../data/normalised", pattern="nor_process_ovarian.csv", full.names=TRUE)
```

```
ins <- read.table(f, header = TRUE, row.names = 1, sep = ",")
inss <- as.matrix(ins)
peakfile <- paste(tempdir(), "testpeakinfo.csv", sep = "/")
getPeaks(inss, peakfile, SoN = 2,span = 81,sm.span=11, zerothrsh=1, area.w = 0.003,
ratio = 0.01)
bmkfile <- paste(tempdir(), "testbiomarker.csv", sep = "/")
testBio <- pk2bmkr(peakfile, inss, bmkfile)
write.table(testBio, file = "peak_process.csv", sep = ",", append = FALSE, col.names
= TRUE, row.names = TRUE)
```

Appendix 8-5 Programs for feature selection and cancer classification

```
### Wilcoxon test

rm (list=ls())
### To calculate p value of Wilcoxon test
pValueOfTW <- function (data){
        cNames <- colnames(data)
        nRow <- nrow(data)
        nCol <- ncol(data)
        cancer <- subset(data, data[,2] >= 1)
        control <- subset(data, data[,2] < 1)
        TWP <- matrix(nrow = 2, ncol = nCol-2)
        for (i in 3:nCol) {
                 wtestF <- wilcox.test(control[,i], cancer[,i] , paired=F)
                 wtestT <- wilcox.test(control[,i], cancer[,i], paired=T)
                 j <- i-2
                 TWP[1,j] <- wtestF$p.value
                TWP[2,j] <- wtestT$p.value
        }
        colnames(TWP) <- colnames(data)[3:nCol]
        rownames(TWP) <- c("unpaired", "paired")
        TWP
}
```

```
### To find important variables based on RF modelling

rm (list=ls())
library(randomForest)
set.seed(166)

### read in data
f <- dir("../data/", pattern = "prostate_peaks_rf.csv",full.names=TRUE)
intens <- read.table(f, header = TRUE, sep = ",", row.name = 1)

### to find the importance variables based on "mean descreas accuracy" measurment
varImp <- randomForest(class~., data=intens, ntree = 500, importance = TRUE)
write.table(varImp$importance[,3], file = "rf_err_prostate.csv", sep = ",", append =
FALSE, col.names = TRUE, row.names = TRUE)
```

```
### SVM modelling
```

```
rm (list=ls())
library(e1071)

### read in
f <- dir("../results/classification", pattern="^prostate_common_peaks.csv",
full.names=TRUE)
ins <- read.table(f, header=TRUE, sep=",", row.names =1)

### building
acc <- c()
for (i in 1:100) {
      m <- svm(class~., data = ins, cross =10, gamma=0.02, cost=64)
      acc <- c(acc, m$tot.accuracy)
}
write.table(acc, file = "common_svm_model.csv", sep = ",", append = FALSE,
col.names = TRUE, row.names = FALSE)
```

```
### RF modelling

rm (list=ls())
library(randomForest)
set.seed(366)

### read in
f <- dir("../results/classification", pattern="prostate_common_peaks.csv",
full.names=TRUE)
ins <- read.table(f, header=TRUE, sep=",", row.names =1)

### modeling
acc <- c()
for (i in 1:100) {
    rf <- randomForest(class~., data=ins, ntree = 500, importance = TRUE)
    acc <- c(acc, mean(rf$err.rate[,1]))
}
write.table(acc, file = "common_rf_model.csv", sep = ",", append = FALSE,
col.names = TRUE, row.names = FALSE)
```

```
### kNN modelling

rm (list=ls())
library(class)

### read in
f <- dir("../results/classification", pattern="prostate_common_peaks.csv",
full.names=TRUE)
ins <- read.table(f, header=TRUE, sep=",", row.names =1)
```

```
train <- subset(ins, select=-class)
group <- factor(c(rep("n",63), rep("c",69)))

### buinding
acc <- c()
for (i in 1:100) {
   kn <- knn.cv(train, group, k = 3, l = 0, prob = TRUE)
   k <- 0
   for(j in 1:132) {
     if(kn[j] != group[j]) {
        k <- k+1
     }
   }
   acc <- c(acc, k/132)
}
write.table(acc, file = "common_knn_model.csv", sep = ",", append = FALSE,
col.names = TRUE, row.names = FALSE)
```

To draw matrix of scatter plot and Venn diagram

```
rm (list=ls())
library(limma)

### read in data ###

f<- dir("../results/venn", pattern = "Pvalue_ovarian_data.csv",full.names=TRUE)
Wilcoxon_test <- read.table(f, header = FALSE, sep = ",")
f<- dir("../results/venn", pattern = "auc_ovarian_data.csv",full.names=TRUE)
ROC_curve <- read.table(f, header = FALSE, sep = ",")
f<- dir("../results/venn", pattern = "rf_err_ovarian.csv",full.names=TRUE)
Random_forest <- read.table(f, header = FALSE, sep = ",")

### scatter plot ###
imp <- t(rbind(Wilcoxon_test, ROC_curve, Random_forest))
pairs(imp, pch=20, labels=c("Wilcoxon test", "ROC curve", "Random forest"),
col="red", font.labels=1.5, cex.labels=1.5)

### venn diagram ###
set1 <- sort(as.matrix(Wilcoxon_test))
set2 <- sort(as.matrix(ROC_curve))
set3 <- sort(as.matrix(Random_forest))
names <- c("Wilcoxon", "ROC", "RF")
Venn3(set1, set2, set3, names)
```

9 REFERENCES

Adam, B. L., Qu, Y., Davis, J. W. et al (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. **Cancer Res.,** 62 (13): 3609-3614

Aebersold, R. and Goodlett, D. R. (2001) Mass spectrometry in proteomics. **Chem Rev,** 101 (2): 269-295

Agresti, A. (1990) **Categorical data analysis.** New York: Wiley.

Alaoui-Jamali, M. A. and Xu, Y. J. (2006) Proteomic technology for biomarker profiling in cancer: an update. **J Zhejiang.Univ Sci.B,** 7 (6): 411-420

Alterovitz, G., Aivado, M., Spentzos, D. et al (2004) Analysis and robot pipelined automation for SELDI-TOF mass spectrometry. **Conf.Proc.IEEE Eng Med.Biol.Soc.,** 4 3068-3071

Amundadottir, L. T., Sulem, P., Gudmundsson, J. et al (2006) A common variant associated with prostate cancer in European and African populations. **Nat.Genet.,** 38 (6): 652-658

Assikis, V. J., Do, K. A., Wen, S. et al (2004) Clinical and biomarker correlates of androgen-independent, locally aggressive prostate cancer with limited metastatic potential. **Clin Cancer Res.,** 10 (20): 6770-6778

Austen, B. M., Frears, E. R. and Davies, H. (2000) The use of seldi proteinchip arrays to monitor production of Alzheimer's betaamyloid in transfected cells. **J.Pept.Sci.,** 6 (9): 459-469

Baggerly, K. A., Morris, J. S. and Coombes, K. R. (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. **Bioinformatics.,** 20 (5): 777-785

Baggerly, K. A., Morris, J. S., Wang, J. et al (2003) A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. **Proteomics.,** 3 (9): 1667-1672

Baggio, L. and Prodi, G. A. (2005) False discovery rate: setting the probability of false claim of detection. **Class.Quantum Grav.,** 22 p. s1373-s1379

Bai, Z. D. and Cheng, L. (2003) Weighted W test for normality and asymptotics a revisit of Chen-Shapiro test for normality. **Journal of Statistical Planning and Inference,** 113 (2003): 485-503

Ball, G., Mian, S., Holding, F. et al (2002) An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. **Bioinformatics.,** 18 (3): 395-404

Balmain, A., Gray, J. and Ponder, B. (2003) The genetics and genomics of cancer. **Nat.Genet.,** 33 Suppl 238-244

Banez, L. L., Prasanna, P., Sun, L. et al (2003) Diagnostic potential of serum proteomic patterns in prostate cancer. **J.Urol.,** 170 (2 Pt 1): 442-446

Banks, R. E., Stanley, A. J., Cairns, D. A. et al (2005) Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. **Clin.Chem.,** 51 (9): 1637-1649

Bender, R. and Lange, S. (2001) Adjusting for multiple testing--when and how? **J Clin Epidemiol.,** 54 (4): 343-349

Benjamini, Y. and Hochberg, Y. (1995a) Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society B,** 57 p. 289

Benjamini, Y. and Hochberg, Y. (1995b) Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society,** 57 (1): 289-300

Benjamini, Y. and Yekutieli, D. (2005) Quantitative trait Loci analysis using the false discovery rate. **Genetics,** 171 (2): 783-790

Bensmail, H. and Haoudi, A. (2003) Postgenomics: Proteomics and Bioinformatics in Cancer Research. **J.Biomed.Biotechnol.,** 2003 (4): 217-230

Beyer, S., Walter, Y., Hellmann, J. et al (2006) Comparison of software tools to improve the detection of carcinogen induced changes in the rat liver proteome by analyzing SELDI-TOF-MS spectra. **J.Proteome.Res.,** 5 (2): 254-261

Bhanot, G., Alexe, G., Venkataraghavan, B. et al (2006) A robust meta-classification strategy for cancer detection from MS data. **Proteomics.,** 6 (2): 592-604

Bioconductor (2007a) Bioconductor Packages **[online]**. www.bioconductor.org. Accessed in Jan. 2007

Bioconductor (2007b) Bioconductor Vignette 'PROcess' **[online]**. http://www.bioconductor.org/. Accessed in May 2007

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. **Clin Pharmacol Ther.,** 69 (3): 89-95

Birkner, M. D., Hubbard, A. E., van der Laan, M. J. et al (2006) Issues of processing and multiple testing of SELDI-TOF MS proteomic data. **Stat.Appl.Genet.Mol.Biol.,** 5 p. Article11

Birnbaum, Z. W. (1952) Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size. **Journal of the American Statistical Association,** 47 425-441

Bolstad, B. M., Irizarry, R. A., Astrand, M. et al (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. **Bioinformatics.,** 19 (2): 185-193

Bradley, A. P. (1997) The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. **Pattern Recognition,** 30 (7): 1145-1159

Breiman, L. and Spector, P. (1992) Submodel selection and evaluation in regression. The X-random case. **Int statist Rev.,** 60 291-319

Breiman, L. (1992) The Little Bootstrap and Other Methods for Dimenionality Selection in Regression: X-Fixed Prediction Error. **Journal of the American Statistical Association,** 87 (419): 738-754

Breiman, L. (2001) Random Forest. **Machine Learning,** 45 5-32

Bureau, A., Dupuis, J., Falls, K. et al (2005) Identifying SNPs predictive of phenotype using random forests. **Genet.Epidemiol.,** 28 (2): 171-182

caBIG (2007) Biomarkers **[online]**. http://cabig.cancer.gov/resources/glossary.asp. Accessed in June 2007

Callister, S. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. **Journal of proteome research,** 5 277-286

Cancer Research UK (2007) UK Cancer mortality statistics **[online]**. http://info.cancerresearchuk.org/cancerstats/mortality/. Accessed in Jan. 2007

Cancerbackup (2007) There is no effective treatment for cancer **[online]**. http://kiosks.cancerbackup.org.uk/Resourcessupport/Practicalissues/Cancerandolderpeople/Commonmyths/?section:int=2. Accessed in Jan. 2007

Cardone, M. H., Roy, N., Stennicke, H. R. et al (1998) Regulation of cell death protease caspase-9 by phosphorylation. **Science,** 282 (5392): 1318-1321

Cavalierie, E. and Rogan, E. (2006) Catechol Quinones of Estrogens in the Initiation of Breast, Prostate, and Other Human Cancers. Keynote Lecture. **Annals of the New York Academy of Sciences,** 1089 p. 286

Cazares, L. H., Adam, B. L., Ward, M. D. et al (2002) Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ionization mass spectrometry. **Clin.Cancer Res.,** 8 (8): 2541-2552

Chaurand, P. and Caprioli, R. M. (2002) Direct profiling and imaging of peptides and proteins from mammalian cells and tissue sections by mass spectrometry. **Electrophoresis,** 23 (18): 3125-3135

Chen, G., Gharib, T. G., Huang, C. C. et al (2002) Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors. **Clin.Cancer Res.,** 8 (7): 2298-2305

Chen, L., Klebanov, L. and Yakovlev, A. (2005) Normality of gene expression revisited. **BMC.Bioinformatics.,** 5 p. 77

Cheng, Y., Nyangoma, S. O., & Johnson, P. J. (2008) "Identifying Biomarkers from SELDI-TOF Protein Profiles: An Integrative Approach". In: **XXIVth International Biometric Conference,13-18th July, 2008, Ireland**.

Cheng, Y., Ward, D. G., Wen, W., Billingham, L. J., & Johnson, P. I. (2006) "Plasma proteome profiling of liver tumours in flatfish". In: **NCRI conference, 8-11th October, 2006, Birmingham**.

Chernoff, H. and Lehmann, E. L. (1954) The Use of Maximum Likelihood Estimates in Chi2 Tests for Goodness of Fit. **The Annals of Mathematical Statistics,** 25 (3): 579-586

Chernyak, A., Karavanov, A., Ogawa, Y. et al (2001) Conjugating oligosaccharides to proteins by squaric acid diester chemistry: rapid monitoring of the progress of conjugation, and recovery of the unused ligand. **Carbohydr.Res.,** 330 (4): 479-486

Cho, W. C. S. (2007) Contribution of oncoproteomics to cancer biomarker discovery. **Molecular Cancer,** 6 (25):

Chu, W., Ghahramani, Z., Falciani, F. et al (2005) Biomarker discovery in microarray gene expression data with Gaussian processes. **Bioinformatics.,** 21 (16): 3385-3393

Coombes, K. R., Fritsche, H. A., Jr., Clarke, C. et al (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. **Clin.Chem.,** 49 (10): 1615-1623

Coombes, K. R., Tsavachidis, S., Morris, J. S. et al (2005) Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. **Proteomics,** 5 (16): 4107-4117

Corder, E. H., Guess, H. A., Hulka, B. S. et al (1993) Vitamin D and prostate cancer: a prediagnostic study with stored sera. **Cancer Epidemiol.Biomarkers Prev.,** 2 (5): 467-472

Cordingley, H. C., Roberts, S. L., Tooke, P. et al (2003) Multifactorial screening design and analysis of SELDI-TOF ProteinChip array optimization experiments. **Biotechniques,** 34 (2): 364-373

CRAN (2007) The Comprehensive R Archive Network **[online]**. http://cran.r-project.org/ Accessed Jan. 2007

Cristianin, N. & Shawe-Taylor, J. (2000) **An Introduction to Support Vector Machines and other kernel-based learning methods.** Cambridge University Press.

Curtis, D., North, B. V. and Sham, P. C. (2001) Use of an artificial neural network to detect association between a disease and multiple marker genotypes. **Ann Hum.Genet.,** 65 (Pt 1): 95-107

Dallal, G. E. and Wilkison, L. (1986) An analytic approximation to the distribution of Lilliefors' test for normality. **The American Statistician,** 40 (11): 294-296

Dalmasso, E. A. (2008) Planning for Success in Biomarker Discovery. **Genetic Engineering & Biotechnology News,** 28 (12): 28-30

Datta, S. and Lara, M. D. (2006) Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. **Statistical Methodology,** 3 (1): 79-92

Datta, S., Satten, G. A., Benos, D. J. et al (2004) An empirical bayes adjustment to increase the sensitivity of detecting differentially expressed genes in microarray experiments. **Bioinformatics.,** 20 (2): 235-242

De Boor, C. (2001) **A Practical Guide to Splines.** Applied Mathematical Science, 27 Springer.

De Torre, C., Ying, S. X., Munson, P. J. et al (2006) Proteomic analysis of inflammatory biomarkers in bronchoalveolar lavage. **Proteomics,** 6 (13): 3949-3957

Dekking, F., Kraaikamp, C., Lopuhaä, H. et al (2007) **A Modern Introduction to Probability and Statistics.** Springer-Verlag.

Diamandis, E. P. (2003b) Re: Serum proteomic patterns for detection of prostate cancer. **J.Natl.Cancer Inst.,** 95 (6): 489-490

Diamandis, E. P. (2003a) Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? **Clin.Chem.,** 49 (8): 1272-1275

Diamandis, E. P. (2002) Cancer Diagnostics: Discovery and Clinical Applications - Introduction. **Clinical Chemistry,** 48 (8): 1145-1146

Diamandis, E. P. (2004a) Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. **J.Natl.Cancer Inst.,** 96 (5): 353-356

Diamandis, E. P. (2004b) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. **Mol.Cell Proteomics.,** 3 (4): 367-378

Duda, R. O., Hart, P. E., & Stork, D. G. (2000) **Pattern Classification (2nd ed).** Wiley.

Easton, D. F., Pooley, K. A., Dunning, A. M. et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. **Nature,** 447 (7148): 1087-1093

Ebert, M. P., Meuer, J., Wiemer, J. C. et al (2004) Identification of gastric cancer patients by serum protein profiling. **J.Proteome.Res.,** 3 (6): 1261-1266

Editorial (2004) Recent Advances in Cancer Biomarkers. **Clinical Biochemistry,** 37 503-504

Etzioni, R., Urban, N., Ramsey, S. et al (2003) The case for early detection. **Nat.Rev.Cancer,** 3 (4): 243-252

Fayers, P. M., Ashby, D. and Parmar, M. K. (1997) Tutorial in biostatistics Bayesian data monitoring in clinical trials. **Stat.Med.,** 16 (12): 1413-1430

Finne, P., Finne, R. and Stenman, U. H. (2001) Neural network analysis of clinicopathological factors in urological disease: a critical evaluation of available techniques. **BJU Int,** 88 (8): 825-831

Forde, C. E., Gonzales, A. D., Smessaert, J. M. et al (2002) A rapid method to capture and screen for transcription factors by SELDI mass spectrometry. **Biochem.Biophys.Res.Commun.,** 290 (4): 1328-1335

Forrest, M. S., Lan, Q., Hubbard, A. E. et al (2005) Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. **Environ.Health Perspect.,** 113 (6): 801-807

Forshed, J., Schuppe-Koistinen, I. and Jacobsson, S. P. (2003) Peak alignment of NMR signals by means of a genetic algorithm. **Analytica Chimica Acta,** 487 (2): 189-199

Forshed, J., Torgrip, R. J., Aberg, K. M. et al (2005) A comparison of methods for alignment of NMR peaks in the context of cluster analysis. **J Pharm.Biomed.Anal.,** 38 (5): 824-832

Fung, E. T. and Enderwick, C. (2002) ProteinChip clinical proteomics: computational challenges and solutions. **Biotechniques,** Suppl 34-1

Gao, C. L., Rawal, S. K., Sun, L. et al (2003) Diagnostic potential of prostate-specific antigen expressing epithelial cells in blood of prostate cancer patients. **Clin.Cancer Res.,** 9 (7): 2545-2550

Gavin, A. (2007) Northern Ireland Cancer Registry **[online]**. http://www.qub.ac.uk/research-centres/nicr. Accessed in Jan. 2007

General Register Office for Scotland (2007) GRO for Scotland Registrar General's Annual Report, 2006 **[online]**. http://www.gro-scotland.gov.uk/statistics/annrep/index.html. Accessed in Jan. 2007

Gentleman, R., Carey, V. J., Huber, W. et al (2005) **Statistics for Biology and Health.** Springer Science+Business Media,Inc.

Gentleman, R. (2005) **Bioinformatics and Computational Biology solutions using R and Bioconductor.** Springer Verlag.

Geurts, P., Fillet, M., de, S. D. et al (2005) Proteomic mass spectra classification using decision tree based ensemble methods. **Bioinformatics,** 21 (14): 3138-3145

Giles, P. J. and Kipling, D. (2003) Normality of oligonucleotide microarray data and implications for parametric statistical analyses. **Bioinformatics.,** 19 (17): 2254-2262

Gloeckler Ries, L. A., Reichman, M. E., Lewis, D. R. et al (2003) Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. **Oncologist.,** 8 (6): 541-552

Gonzalez, J. R., Armengol, L., Sole, X. et al (2007) SNPassoc: an R package to perform whole genome association studies. **Bioinformatics.,** 23 (5): 644-645

Grizzle, W. E., Adam, B. L., Bigbee, W. L. et al (2003) Serum protein expression profiling for cancer detection: validation of a SELDI-based approach for prostate cancer. **Dis.Markers,** 19 (4-5): 185-195

Gudmundsson, J., Sulem, P., Manolescu, A. et al (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. **Nat.Genet.,** 39 (5): 631-637

Hampel, D. J., Sansome, C., Sha, M. et al (2001) Toward proteomics in uroscopy: urinary protein profiles after radiocontrast medium administration. **J.Am.Soc.Nephrol.,** 12 (5): 1026-1035

Hanley, J. A. and McNeil, B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology,** 143 (1): 29-36

HapMap (2007b) The Origins of Haplotypes **[online]**. http://www.hapmap.org/originhaplotype.html. Accessed in March 2007

HapMap (2007a) The international HapMap project **[online]**. http://www.hapmap.org. Accessed in March 2007

Hartwell, L., Mankoff, D., Paulovich, A. et al (2006) Cancer biomarkers: a systems approach. **Nat.Biotechnol.,** 24 (8): 905-908

Henderson, B. E., Ross, R. K. and Pike, M. C. (1991) Toward the primary prevention of cancer. **Science,** 254 (5035): 1131-1138

Henderson, N. A. and Steele, R. J. (2005) SELDI-TOF proteomic analysis and cancer detection. **Surgeon,** 3 (6): 383-90, 422

Hinshelwood, J., Spencer, D. I., Edwards, Y. J. et al (1999) Identification of the C3b binding site in a recombinant vWF-A domain of complement factor B by surface-enhanced laser desorption-ionisation affinity mass spectrometry and homology modelling: implications for the activity of factor B. **J.Mol.Biol.,** 294 (2): 587-599

Hoheisel, J. D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. **Nat.Rev Genet.,** 7 (3): 200-210

Hong, H., Dragan, Y., Epstein, J. et al (2005) Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of flight (TOF) mass spectrometry (MS). **BMC.Bioinformatics,** 6 Suppl 2 p. S5

Huber, W., von, H. A., Sultmann, H. et al (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. **Bioinformatics.,** 18 Suppl 1 S96-104

Human Genome Project (2007) From the Genome to the Proteome **[online]**. http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml. Accessed in June 2007

Hunter, D. J., Thomas, G., Hoover, R. et al (2007) Scanning the horizon: what is the future of genome-wide association studies in accelerating discoveries in cancer etiology and prevention? **Cancer Causes Control,** 18 (5): 479-484

Hutchens, T. W. and Yip, T. T. (1993) New desorption strategies for the mass spectrometric analysis of macromolecules. **Rapid.Commun.Mass Spectrom,** 7 576-580

Isaaq, H. J., Veenstra, T. D., Conrads, T. P. et al (2002) The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. Biochem. Biophys. **Res.Commun,** 292 587-592

IUPAC (2007) Reproducibility **[online]**. http://www.iupac.org/goldbook/R05305.pdf. Accessed in March 2007

Izmirlian, G. (2004) Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. **Ann.N.Y.Acad.Sci,** 1020 154-174

Jeffries, N. (2005) Algorithms for alignment of mass spectrometry proteomic data. **Bioinformatics.,** 21 (14): 3066-3073

Joanes, D. and Gill, C. (1998) Comparing measures of sample skewness and kurtosis. **Journal of the Royal Statistical Society (Series D): The Statistician,** 47 (1): 183-189

Jock, C. A., Paulauskis, J. D., Baker, D. et al (2004) Influence of matrix application timing on spectral reproducibility and quality in SELDI-TOF-MS. **Biotechniques,** 37 (1): 30-2, 34

Juergen, G. (2007a) Lilliefors (Kolmogorov-Smirnov) test for normality **[online]**. http://pbil.univ-lyon1.fr/library/nortest/html/lillie.test.html. Accessed in June 2007

Juergen, G. (2007b) Pearson chi-square test for normality **[online]**. http://pbil.univ-lyon1.fr/library/nortest/html/pearson.test.html. Accessed in June 2007

Kachman, M. T., Wang, H., Schwartz, D. R. et al (2002) A 2-D liquid separations/mass mapping method for interlysate comparison of ovarian cancers. **Anal.Chem,** 74 (8): 1779-1791

Karas, M., Bachmann, D. and Hillenkamp, F. (1985) Influence of the Wavelength in High-Irradiance Ultraviolet Laser Desorption Mass Spectrometry of Organic Molecules. **Anal.Chem,** 57 2935-2939

Karas, M. and Bahr, U. (1990) Laser Desorption Ionization Mass Spectrometry of Large Biomolecules. **Trends Anal.Chem.,** 9 321-325

Karsan, A., Eigl, B. J., Flibotte, S. et al (2005) Analytical and preanalytical biases in serum proteomic pattern analysis for breast cancer diagnosis. **Clin.Chem.,** 51 (8): 1525-1528

Kasthuri, R. S., Verneris, M. R., Ibrahim, H. N. et al (2006) Studying multiple protein profiles over time to assess biomarker validity. **Expert Rev Proteomics.,** 3 (4): 455-464

Kelly, D. J. and Ghosh, S. (2005) RNA profiling for biomarker discovery: practical considerations for limiting sample sizes. **Dis Markers,** 21 (1): 43-48

Kerr, M. K., Martin, M. and Churchill, G. A. (2000) Analysis of variance for gene expression microarray data. **J.Comput.Biol.,** 7 (6): 819-837

Khodavirdi, A. C., Song, Z., Yang, S. et al (2006) Increased expression of osteopontin contributes to the progression of prostate cancer. **Cancer Res.,** 66 (2): 883-888

Klein, R. J., Zeiss, C., Chew, E. Y. et al (2005) Complement factor H polymorphism in age-related macular degeneration. **Science,** 308 (5720): 385-389

Kohavi, G. R. J. (1997) Wrappers for feature subset selection. **Artificial Intelligence,** 97 (97): 273-324

Kozak, K. R., Amneus, M. W., Pusey, S. M. et al (2003) Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis. **Proc.Natl.Acad.Sci.U.S.A,** 100 (21): 12343-12348

Kozak, K. R., Su, F., Whitelegge, J. P. et al (2005) Characterization of serum biomarkers for detection of early stage ovarian cancer. **Proteomics,** 5 (17): 4589-4596

Kroczak, T. J., Baran, J., Pryjma, J. et al (2006) The emerging importance of DNA mapping and other comprehensive screening techniques, as tools to identify new drug targets and as a means of (cancer) therapy personalisation. **Expert Opin.Ther Targets.,** 10 (2): 289-302

Kuwata, H., Yip, T. T., Yip, C. L. et al (1998) Bactericidal domain of lactoferrin: detection, quantitation, and characterization of lactoferricin in serum by SELDI affinity mass spectrometry. **Biochem.Biophys.Res.Commun,** 245 (3): 764-773

Laiko, V. V., Moyer, S. C. and Cotter, R. J. (2000) Atmospheric pressure MALDI/ion trap mass spectrometry. **Anal.Chem.,** 72 (21): 5239-5243

Lee, G. C. and Woodruff, D. L. (2004) Beam search for peak alignment of NMR signals. **Anal.Chim.Acta,** 513 413-416

Levner, I. (2005) Feature selection and nearest centroid classification for protein mass spectrometry. **BMC Bioinformatics.,** 6 p. 68

Li, J., Zhang, Z., Rosenzweig, J. et al (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. **Clin.Chem.,** 48 (8): 1296-1304

Lilliefors, H. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. **Journal of the American Statistical Association,** 62 399-402

Listgarten, J., Damaraju, S., Poulin, B. et al (2004) Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. **Clin.Cancer Res.,** 10 (8): 2725-2737

Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. **Mol.Cell Proteomics,** 4 (4): 419-434

Ludwig, J. A. and Weinstein, J. N. (2005) Biomarkers in cancer staging, prognosis and treatment selection. **Nat.Rev.Cancer,** 5 (11): 845-856

Lunetta, K. L., Hayward, L. B., Segal, J. et al (2004) Screening large-scale association study data: exploiting interactions using random forests. **BMC Genet.,** 5 (1): p. 32

Malyarenko, D. I., Cooke, W. E., Adam, B. L. et al (2005) Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. **Clin.Chem.,** 51 (1): 65-74

Maraganore, D. M., de, A. M., Lesnick, T. G. et al (2005) High-resolution whole-genome association study of Parkinson disease. **Am J Hum.Genet.,** 77 (5): 685-693

Marchiori, E., Heegaard, N. H. H., West-Nielsen, M. et al (2005) Feature Selection for Classification with Proteomic Data of Mixed Quality. **Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology,** 385-391

Marshall, J., Kupchak, P., Zhu, W. et al (2003) Processing of serum proteins underlies the mass spectral fingerprinting of myocardial infarction. **J.Proteome.Res.,** 2 (4): 361-372

Massey, F. J. (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. **Journal of the American Statistical Association,** 46 68-78

McCormack, A. L., Schieltz, D. M., Goode, B. et al (1997) Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. **Anal.Chem,** 69 (4): 767-776

Menon, U. and Jacobs, I. (2002) Screening for ovarian cancer. **Best Pract.Res.Clin Obstet.Gynaecol.,** 16 (4): 469-482

Merchant, M. and Weinberger, S. R. (2000) Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. **Electrophoresis,** 21 (6): 1164-1177

Miyaki, K., Omae, K., Murata, M. et al (2004) High throughput multiple combination extraction from large scale polymorphism data by exact tree method. **J.Hum.Genet.,** 49 (9): 455-462

Miyamae, T., Malehorn, D. E., Lemster, B. et al (2005) Serum protein profile in systemic-onset juvenile idiopathic arthritis differentiates response versus nonresponse to therapy. **Arthritis Res.Ther.,** 7 (4): p. R746-R755

Moriss, J. S., Coombes, K. R., Koomen, J. et al (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. **Bioinformatics,** 21 (9): 1764-1775

Morra, R., Munteanu, M., Bedossa, P. et al (2007) Diagnostic value of serum protein profiling by SELDI-TOF ProteinChip compared with a biochemical marker, FibroTest, for the diagnosis of advanced fibrosis in patients with chronic hepatitis C. **Aliment.Pharmacol.Ther.,** 26 (6): 847-858

Moshkovskii, S. A., Serebryakova, M. V., Kuteykin-Teplyakov, K. B. et al (2005) Ovarian cancer marker of 11.7 kDa detected by proteomics is a serum amyloid A1. **Proteomics,** 5 (14): 3790-3797

National Cancer Institute (2007b) Ovarian Data set 4_3_2 and Prostate Data set 7_3_2 **[online]**. http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. Accessed in Jan. 2007

National Cancer Institute (2007a) National Cancer Institute Announces Awards to Accelerate Cancer Biomarker Discovery **[online]**. http://www.nci.nih.gov/newscenter/pressreleases/ProteomicBiomarkerAwards. Accessed in June 2007

Negm, R. S., Verma, M. and Srivastava, S. (2002) The promise of biomarkers in cancer screening and detection. **Trends Mol.Med.,** 8 (6): 288-293

Notterman, D. A., Alon, U., Sierk, A. J. et al (2001) Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. **Cancer Res.,** 61 (7): 3124-3130

Nyangoma, S. O., van Kampen, A. A., Reijmers, T. H. et al (2007) Multiple testing issues in discriminating compound-related peaks and Chromatograms from high frequency noise, spikes and solvent-based noise in LC-MS data sets. **Statistical Applications in Genetics and Molecular Biology,** 6 (1): 2-23

Office for National Statistics (2007) Mortality Statistics:Cause England & Wales, 2005 **[online]**. http://www.statistics.gov.uk/downloads/theme_health/Dh2_32/DH2_No32_2005.pdf . Accessed in Jan. 2007

Orchekowski, R., Hamelinck, D., Li, L. et al (2005) Antibody microarray profiling reveals individual and combined serum proteins associated with pancreatic cancer. **Cancer Res.,** 65 (23): 11193-11202

Osborne, J. W. (2002) Notes on the use of data transformations. **Practical Assessment, Research & Evaluation,** 8 (6):

Ozaki, K., Ohnishi, Y., Iida, A. et al (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. **Nat.Genet.,** 32 (4): 650-654

Ozier, O., Amin, N. and Ideker, T. (2003) Global architecture of genetic interactions on the protein network. **Nat.Biotechnol.,** 21 (5): 490-491

PacificLife (2007) Cancer Biomarkers: Still Controversial **[online]**. http://news.creaders.net/headline/newsViewer.php?nid=343827&id=799678&dcid=4 . Accessed in June 2007

Pang, R. T., Poon, T. C., Chan, K. C. et al (2006) Serum proteomic fingerprints of adult patients with severe acute respiratory syndrome. **Clin.Chem.,** 52 (3): 421-429

Pang, R. T. K., Johnson, P. J., Chan, C. M. L. et al (2004) Technical Evaluation of MALDI-TOF Mass Spectrometry for Quantitative Proteomic Profiling: Matrix Formulation and Application. **Clinical Proteomics,** 1 259-270

Pasinetti, G. (2006) Identification of potential CSF biomarkers in ALS. **Neurology,** 66 1218-1222

Pavlidis, P. (2003) Using ANOVA for gene selection from microarray studies of the nervous system. **Methods,** 31 (4): 282-289

Paweletz, C. P., Trock, B., Pennanen, M. et al (2001) Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. **Dis.Markers,** 17 (4): 301-307

Pawlik, T. M., Hawke, D. H., Liu, Y. et al (2006) Proteomic analysis of nipple aspirate fluid from women with early-stage breast cancer using isotope-coded affinity tags and tandem mass spectrometry reveals differential expression of vitamin D binding protein. **BMC.Cancer,** 6 p. 68

Penfield (2007) Human Proteome Organisation **[online]**. www.hupo.org. Accessed in June 2007

Perneger, T. V. (1998) What's wrong with Bonferroni adjustments. **BMJ,** 316 (7139): 1236-1238

Petricoin, E., III and Liotta, L. A. (2003) Counterpoint: The vision for a new diagnostic paradigm. **Clin.Chem.,** 49 (8): 1276-1278

Petricoin, E. F., Ardekani, A. M., Hitt, B. A. et al (2002a) Use of proteomic patterns in serum to identify ovarian cancer. **Lancet,** 359 (9306): 572-577

Petricoin, E. F., Belluco, C., Araujo, R. P. et al (2006) The blood peptidome: a higher dimension of information content for cancer biomarker discovery. **Nat.Rev.Cancer,** 6 (12): 961-967

Petricoin, E. F. and Liotta, L. A. (2004) SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. **Curr.Opin.Biotechnol.,** 15 (1): 24-30

Petricoin, E. F. I., Ornstein, D. K., Paweletz, C. P. et al (2002b) Serum proteomic patterns for detection of prostate cancer. **J.Natl.Cancer Inst.,** 94 (20): 1576-1578

Pinheiro, J. et al (2009) nlme:Linear and Nonlinear Mixed Effects Models **[online].** http://cran.r-project.org/web/packages/nlme/index.html/nlme_3.1-92.tar.gz Accessed May 2009

Plebani, M. (2005) Proteomics: the next revolution in laboratory medicine? **Clin Chim.Acta,** 357 (2): 113-122

Poon, T. C. (2007) Opportunities and limitations of SELDI-TOF-MS in biomedical research: practical advices. **Expert.Rev.Proteomics,** 4 (1): 51-65

Poon, T. C., Chan, K. C., Ng, P. C. et al (2004) Serial analysis of plasma proteomic signatures in pediatric patients with severe acute respiratory syndrome and correlation with viral load. **Clin.Chem.,** 50 (8): 1452-1455

Poon, T. C., Hui, A. Y., Chan, H. L. et al (2005) Prediction of liver fibrosis and cirrhosis in chronic hepatitis B infection by serum proteomic fingerprinting: a pilot study. **Clin.Chem.,** 51 (2): 328-335

Poon, T. C., Yip, T. T., Chan, A. T. et al (2003) Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. **Clin Chem,** 49 (5): 752-760

Purohit, P. V., Rocke, D. M., Viant, M. R. et al (2004) Discrimination models using variance-stabilizing transformation of metabolomic NMR data. **OMICS.,** 8 (2): 118-130

Qu, Y., Adam, B. L., Yasui, Y. et al (2002) Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. **Clin.Chem.,** 48 (10): 1835-1843

Raychaudhuri, S., Sutphin, P. D., Chang, J. T. et al (2001) Basic microarray analysis: grouping and feature reduction. **Trends Biotechnol.,** 19 (5): 189-193

Reddy, G. and Dalmasso, E. A. (2003) SELDI ProteinChip(R) Array Technology: Protein-Based Predictive Medicine and Drug Discovery Applications. **J.Biomed.Biotechnol.,** 2003 (4): 237-241

Rice, J. (1995) **Mathematical Statistics and Data Analysis (Second ed.).** Duxbury Press.

Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. **Science,** 273 (5281): 1516-1517

Risch, N. J. (2000) Searching for genetic determinants in the new millennium. **Nature,** 405 (6788): 847-856

Rise, M. L., Jones, S. R., Brown, G. D. et al (2004) Microarray analyses identify molecular biomarkers of Atlantic salmon macrophage and hematopoietic kidney response to Piscirickettsia salmonis infection. **Physiol Genomics,** 20 (1): 21-35

Robnik-Sikonja, M. (2004) Improving random forests. In J.F. Boulicaut et al. (eds): machine Learning, EMCL 2004 Proceedings. **Springer, Berlin,**

Rocke, D. M. and Durbin, B. (2003) Approximate variance-stabilizing transformations for gene-expression microarray data. **Bioinformatics.,** 19 (8): 966-972

Rodin, A. S. and Boerwinkle, E. (2005) Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). **Bioinformatics.,** 21 (15): 3273-3278

Rogers, M. A., Clarke, P., Noble, J. et al (2003) Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility. **Cancer Res.,** 63 (20): 6971-6983

Rosty, C., Christa, L., Kuzdzal, S. et al (2002) Identification of hepatocarcinoma-intestine-pancreas/pancreatitis-associated protein I as a biomarker for pancreatic ductal adenocarcinoma by protein biochip technology. **Cancer Res.,** 62 (6): 1868-1875

Royston, P. (1995) A Remark on Algorithm AS 181: The *W* Test for Normality. **Applied Statistics.,** 44 (5): 547-551

Royston, P. (1993) A toolkit for testing for non-normality in complete and censored samples. **The Statistician,** 42 37-43

Sato, K., Sasaki, K., Akiyama, Y. et al (2001) Mass spectrometric high-throughput analysis of serum-free conditioned medium from cancer cell lines. **Cancer Lett.,** 170 (2): 153-159

Sauve, A.C. and Speed, T.P. (2003) Normalization, baseline correction and alignment of high-throughput mass spectrometry data **[online]**. http://www.stat.berkeley.edu/~terry/Group/publications/Final2Gensips2004Sauve.pdf Accessed in June 2007

Scharpf, R. B., Ting, J. C., Pevsner, J. et al (2007) SNPchip: R classes and methods for SNP array data. **Bioinformatics.,** 23 (5): 627-628

Schmid, F. and Trede, M. (1996) An L_1-variant of the Cramer-von Mises test. **Statistics and Probability Letters,** 26 91-96

Schork, N. J., Cardon, L. R. and Xu, X. (1998) The future of genetic epidemiology. **Trends Genet.,** 14 (7): 266-272

Semmes, O. J., Feng, Z., Adam, B. L. et al (2005) Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. **Clin Chem,** 51 (1): 102-112

Shapiro, S. S., Wilk, M. B. and . (1965) An analysis of variance test for normality (complete samples). **Biometrika,** 52 591-611

Shier, R. (2004) The Mann-Whitney U Test **[online]**. http://mlsc.lboro.ac.uk/resources/statistics/Mannwhitney.pdf. Accessed in June 2007

Simon, R., Radmacher, M. D., Dobbin, K. et al (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. **J.Natl.Cancer Inst.,** 95 (1): 14-18

Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. **Stat.Appl.Genet.Mol.Biol.,** 3 p. Article3

SNP Fact Sheet (2009) What are SNPs? **[online]**.
http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml#snps
Accessed in March 2009

Spackman, K. A. (1989) "Signal Detection Theory: Valuable Tools for. Evaluating Inductive Learning". In: **Proc.6th International Workshop on Machine Learning**.

Spellman, P. T., Sherlock, G., Zhang, M. Q. et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. **Mol.Biol.Cell,** 9 (12): 3273-3297

Srinivas, P. R., Verma, M., Zhao, Y. et al (2002) Proteomics for cancer biomarker discovery. **Clin.Chem.,** 48 (8): 1160-1169

Staunton, J. E., Slonim, D. K., Coller, H. A. et al (2001) Chemosensitivity prediction by transcriptional profiling. **Proc.Natl.Acad.Sci.U.S A,** 98 (19): 10787-10792

Stephan, C., Vogel, B., Cammann, H. et al (2003) [An artificial neural network as a tool in risk evaluation of prostate cancer. Indication for biopsy with the PSA range of 2-20 microg/l]. **Urologe A,** 42 (9): 1221-1229

Stephens, M. A. (1986) **Goodness-of-Fit Techniques.** 26 edn. Marcel Dekker, New York.

Stoeckli, M., Chaurand, P., Hallahan, D. E. et al (2001) Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. **Nat.Med.,** 7 (4): 493-496

Stoica, G. E., Kuo, A., Aigner, A. et al (2001) Identification of anaplastic lymphoma kinase as a receptor for the growth factor pleiotrophin. **J.Biol.Chem.,** 276 (20): 16772-16779

Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. **Science,** 240 (4857): 1285-1293

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. **Nature,** 447 661-678

Thygesen, H. H. and Zwinderman, A. H. (2004) Comparing transformation methods for DNA microarray data. **BMC.Bioinformatics.,** 5 p. 77

Tibshirani, R., Hastie, T., Narasimhan, B. et al (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. **Bioinformatics.,** 20 (17): 3034-3044

Toivonen, H. T., Onkamo, P., Vasko, K. et al (2000) Data mining applied to linkage disequilibrium mapping. **Am J Hum.Genet.,** 67 (1): 133-145

Tomita, Y., Tomida, S., Hasegawa, Y. et al (2004) Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. **BMC Bioinformatics.,** 5 p. 120

Tong, W., Xie, Q., Hong, H. et al (2004) Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence. **Environ.Health Perspect.,** 112 (16): 1622-1627

Unneberg, P., Stromberg, M. and Sterky, F. (2005) SNP discovery using advanced algorithms and neural networks. **Bioinformatics.,** 21 (10): 2528-2530

van de Vijver, M. J., He, Y. D., van't Veer, L. J. et al (2002) A gene-expression signature as a predictor of survival in breast cancer. **N Engl.J Med,** 347 (25): 1999-2009

Viant, M. R. (2003) Improved methods for the acquisition and interpretation of NMR metabolomic data. **Biochem.Biophys.Res.Commun,** 310 (3): 943-948

Vlahou, A., Laronga, C., Wilson, L. et al (2003a) A novel approach toward development of a rapid blood test for breast cancer. **Clin.Breast Cancer,** 4 (3): 203-209

Vlahou, A., Schellhammer, P. F., Mendrinos, S. et al (2001) Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. **Am.J.Pathol.,** 158 (4): 1491-1502

Vlahou, A., Schellhammer, P. F. and Wright, G. L., Jr. (2003b) Application of a novel protein chip mass spectrometry technology for the identification of bladder cancer-associated biomarkers. **Adv.Exp.Med.Biol.,** 539 (Pt A): 47-60

Vlahou, A., Schorge, J. O., Gregory, B. W. et al (2003c) Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data. **J.Biomed.Biotechnol.,** 2003 (5): 308-314

Vorderwülbecke, S., Cleverley, S., Weinberger, S. R. et al (2007) Protein quantification by the SELDI-TOF-MS-based ProteinChip® System. **Nature Method,** 2 393-395

Wada-Isoe, K., Michio, K., Imamura, K. et al (2007) Serum proteomic profiling of dementia with Lewy bodies: diagnostic potential of SELDI-TOF MS analysis. **J.Neural Transm.,** 114 (12): 1579-1583

Wadsworth, J. T., Somers, K. D., Cazares, L. H. et al (2004) Serum protein profiles to identify head and neck cancer. **Clin.Cancer Res.,** 10 (5): 1625-1632

Wagner, M., Naik, D. N., Pothen, A. et al (2004) Computational protein biomarker prediction: a case study for prostate cancer. **BMC Bioinformatics,** 5 (26):

Ward, D. G., Cheng, Y., N'Kontchou, G. et al (2006a) Changes in the serum proteome associated with the development of hepatocellular carcinoma in hepatitis C-related cirrhosis. **Br.J.Cancer,** 94 (2): 287-292

Ward, D. G., Cheng, Y., N'Kontchou, G. et al (2006b) Preclinical and post-treatment changes in the HCC-associated serum proteome. **Br.J.Cancer,** 95 (10): 1379-1383

Ward, D. G., Suggett, N., Cheng, Y. et al (2006c) Identification of serum biomarkers for colon cancer by proteomic analysis. **Br.J Cancer,** 94 (12): 1898-1905

Ward, D. G., Wei, W., Cheng, Y. et al (2006d) Plasma proteome analysis reveals the geographical origin and liver tumor status of Dab (Limanda limanda) from UK marine waters. **Environ.Sci.Technol.,** 40 (12): 4031-4036

William, M.K. (2007) Theory of Reliability **[online]**. http://www.socialresearchmethods.net/kb/reliablt.php. Accessed in June 2007

Wolters, D. A., Washburn, M. P. and Yates, J. R., III (2001) An automated multidimensional protein identification technology for shotgun proteomics. **Anal.Chem,** 73 (23): 5683-5690

Won, Y., Song, H. J., Kang, T. W. et al (2003) Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons. **Proteomics.,** 3 (12): 2310-2316

Wong, J. W., Cagney, G. and Cartwright, H. M. (2005a) SpecAlign--processing and alignment of mass spectra datasets. **Bioinformatics,** 21 (9): 2088-2090

Wong, J. W., Durante, C. and Cartwright, H. M. (2005b) Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. **Anal.Chem,** 77 (17): 5655-5661

Wu, B., Abbott, T., Fishman, D. et al (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. **Bioinformatics.,** 19 (13): 1636-1643

Wu, W. (2007) Random forest R package **[online]**. http://www.r-project.org. Accessed in March 2007

Wulfkuhle, J. D., McLean, K. C., Paweletz, C. P. et al (2001) New approaches to proteomic analysis of breast cancer. **Proteomics.,** 1 (10): 1205-1215

Xiao, X., Liu, D., Tang, Y. et al (2003) Development of proteomic patterns for detecting lung cancer. **Dis.Markers,** 19 (1): 33-39

Xiao, Y., Gordon, A. and Yakovlev, A. (2006) The L(1)-Version of the Cramer-von Mises Test for Two-Sample Comparisons in Microarray Data Analysis. **EURASIP.J.Bioinform.Syst.Biol.,** p. 85769

Yasui, Y., Pepe, M., Thompson, M. L. et al (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. **Biostatistics.,** 4 (3): 449-463

Ye, B., Cramer, D. W., Skates, S. J. et al (2003) Haptoglobin-alpha subunit as potential serum biomarker in ovarian cancer: identification and characterization using proteomic profiling and mass spectrometry. **Clin.Cancer Res.,** 9 (8): 2904-2911

Yoon, Y., Song, J., Hong, S. H. et al (2003) Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. **Clin.Chem.Lab Med.,** 41 (4): 529-534

Yu, F. L. (2002) 17Beta-estradiol epoxidation as the molecular basis for breast cancer initiation and prevention. **Asia Pac.J Clin Nutr.,** 11 Suppl 7 p. S460-S466

Yu, J. K., Chen, Y. D. and Zheng, S. (2004) An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics. **World J Gastroenterol.,** 10 (21): 3127-3131

Yu, K. H., Rustgi, A. K. and Blair, I. A. (2005) Characterization of proteins in human pancreatic cancer serum using differential gel electrophoresis and tandem mass spectrometry. **J.Proteome.Res.,** 4 (5): 1742-1751

Zanke, B. W., Greenwood, C. M., Rangrej, J. et al (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. **Nat.Genet.,**

Zhang, X., Lu, X., Shi, Q. et al (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. **BMC.Bioinformatics.,** 7 p. 197

Zhang, Y. F., Wu, D. L., Guan, M. et al (2004a) Tree analysis of mass spectral urine profiles discriminates transitional cell carcinoma of the bladder from noncancer patient. **Clin Biochem.,** 37 (9): 772-779

Zhang, Z., Bast, R. C., Jr., Yu, Y. et al (2004b) Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. **Cancer Res.,** 64 (16): 5882-5890

Zhao, G., Gao, C. F., Song, G. Y. et al (2004) Identification of colorectal cancer using proteomic patterns in serum. **Ai.Zheng.,** 23 (6): 614-618

Zhukov, T. A., Johanson, R. A., Cantor, A. B. et al (2003) Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. **Lung Cancer,** 40 (3): 267-279

Ziauddin, J. and Sabatini, D. M. (2001) Microarrays of cells expressing defined cDNAs. **Nature,** 411 (6833): 107-110

Zimmermann, E. (2007) First biomarker discovered that predicts prostate cancer outcome **[online]**. http://www.eurekalert.org/pub_releases/2007-08/mc-fbd081307.php. Accessed in June 2007

Zolg, W. (2007) The Proteomics Search for Diagnostic Biomarkers: Lost in Translation? **[online]**. http://www.mcponline.org/cgi/reprint/R600001-MCP200v1.pdf Accessed in June 2007