From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Statistical Methods for Biomarker Discovery in Proteomics

Chuen Seng Tan

Stockholm 2008

Published by Karolinska Institutet. Printed by

ISBN 978-91-7409-134-2

To My Family

In memory of Aunt Gik Yak Tan and Uncle Joseph Wong,
who both passed away from cancer.

Abstract

Surface-Enhanced Laser Desorption and Ionization (SELDI) is a promising proteomic technique for discovering biomarkers. However, the pre-processing of the raw data is still problematic. Integrating transcriptomic and proteomic data may enhance the search for biomarkers, but the current data integration approach results in the loss of large amounts of data.

In this thesis, we made improvements to the peak detection step in SELDI by developing the Annotated Regions of Significance (ARS) method. It uses a multi-spectral signal detection method, 'Region of Significance' (RS), to identify regions with potential biomarkers. RS had better operating characteristics (OC) than existing methods in identifying peaks. Using lung cell line data, at 80% sensitivity, the False Discovery Rates (FDRs) of existing methods were around 25% to 50%, compared to around 8% for RS. ARS extracts a peak template from all spectra in the peak region via Principal Component Analysis (PCA) and fits the template to the spectra. A refinement was made to the estimation of the amplitude via a mixture model. Using patient samples from a clinical study, we showed that ARS detected more peaks and gave more accurate peak quantifications than the standard method. We implemented ARS as an R package, *ProSpect*, and also developed a graphical user interface, *ProSpectGUI*.

Motivated by the performance of ARS in SELDI, we extended ARS to MALDI data with isotopic resolution. The extended ARS utilizes the isotopic pattern to filter out peaks which do not adhere to the expected isotopic pattern. Using the spike-in data, we validated the use of the log-transformed intensities for ARS in MALDI. Compared to the standard method, extended ARS generally had better specificity and was better in quantifying the peaks. At low FDR, extended ARS had higher sensitivity than the standard method.

We also contributed to the integration of proteomic and transcriptomic information from the same samples by investigating the use of Maximum Covariance Analysis (MCA). The estimates of the gene and protein pattern-pairs from MCA were consistent and biologically congruent, compared to Generalized Singular Value Decomposition (gSVD). Therefore MCA has the potential to enhance biomarker discovery and our understanding of the interplay between genes and proteins.

Keywords: Proteomics, mass spectrometry, SELDI, MALDI, peak detection, signal detection, peak annotation, transcriptomics, data integration, maximum covariance analysis, generalized singular value decomposition

List of publications

This thesis is based on the following papers, which are referred to in the text by their Roman numerals:

I. Tan, C.S., Ploner, A., Quandt, A., Lehtiö J., Pawitan, Y. (2006). Finding regions of significance in SELDI measurements for identifying protein biomarkers. *Bioinformatics*, 22:1515-1523.

II. Tan, C.S., Ploner, A., Quandt, A., Lehtiö, J., Pernemalm, M., Lewensohn, R., Pawitan, Y. (2006). Annotated regions of significance of SELDI-TOF-MS spectra for detecting protein biomarkers. *Proteomics*, 6:6124-6133.

III. Quandt, A., Ploner, A., Tan, C.S., Lehtiö, J., Pawitan, Y. (2005). ProSpect: An R package for analyzing SELDI measurements identifying protein biomarkers. *Lecture Notes in Computer Science*, 3695:140-150.

IV. Tan, C.S., Salim, A., Ploner, A., Lehtiö, J., Chia, K.S., Pawitan, Y. Correlating gene and protein expression data using Maximum Covariance Analysis. *Submitted*.

V. Tan, C.S., Lehtiö, J., Forshed, J., Pernemalm, M., Ploner, A., Chia, K.S., Pawitan, Y. Identifying peaks and isotopic patterns in MALDI data. *Submitted*.

Contents

1 Introduction **1**

2 Background **3**

2.1 Proteins 3

2.2 Biomarker discovery 5

2.3 Proteomic technologies 7

2.4 SELDI and MALDI 9

2.5 Existing methods for peak detection 11

2.6 Integrating transcriptomic and proteomic data 13

3 Aims **15**

4 Methods **16**

4.1 Finding peaks in SELDI (Paper I-III) 16

4.1.1 Modifications to the F-statistics 17

4.1.2 Using PCA to extract a peak template 18

4.1.3 Fitting of the template to the other spectra 20

4.1.4 *ProSpect*: An R package for ARS 20

4.2 Finding peaks and isotopic patterns in MALDI (Paper V) . . 23

4.3 Correlating gene and protein expression data (Paper IV) . . . 25

5 Results **28**

5.1 Performance of ARS (Paper I-III) 28

5.1.1 Lung cell line data (H69) 29

5.1.2 Spike-in data 30

5.1.3 Lung cancer serum data 30

5.2 Performance of extended ARS (Paper V) 32

5.2.1 Validation of ARS in MALDI 33

5.2.2 Near the spike-in regions 33

5.2.3 Across the entire mass range 36

5.3 Performance of MCA (Paper IV) 36

5.3.1 Simulated data 37

5.3.2 NCI data 38

6 Discussion **41**

6.1 ARS (Paper I-III) 41

6.2 Extended ARS (Paper V) 42

6.3 MCA (Paper IV) 43

6.4 Future research 43

7 Conclusions **46**

Acknowledgements **48**

References **50**

List of abbreviations

ANOVA	Analysis of Variance
ARS	Annotated Regions of Significance
CM10	Weak Cation Exchange Array with Carboxylate Functionality, with a Hydrophobic Barrier Coating
Da	Dalton
DIGE	Two-Dimensional Difference Gel Electrophoresis
DNA	Deoxyribonucleic Acid
ESI	Electrospray Ionization
GC	Gas Chromatography
GO	Gene Ontology
gSVD	Generalized Singular Value Decomposition
MALDI	Matrix-Assisted Laser Desorption and Ionization
MCA	Maximum Covariance Analysis
mRNA	Messenger Ribonucleic Acid
MS	Mass Spectrometry
MSE	Mean Square Error
m/z	Mass-per-charge
NMR	Nuclear Magnetic Resonance
OC	Operating Characteristics
PCA	Principal Component Analysis
Q-TOF	Quadrupole Time-of-Flight
RPLA	Reverse-Phase Protein Lysate Array
RS	Regions of Significance
SAX2	Strong Anion Exchange Array with Quaternary Amine Functionality
SELDI	Surface-Enhanced Laser Desorption and Ionization
S/N	Signal-to-Noise Ratio
TOF	Time-of-Flight
WCX2	Weak Cation Exchange Arrays with Carboxylate Functionality
wMSE	Weighted Mean Square Error
2DGE	Two-Dimensional Gel Electrophoresis

Chapter 1

Introduction

A biomarker is a molecule that indicates the physiological state of a cell (Srinivas et al., 2001). Therefore biomarkers can serve as early warning indicators for disease, help to monitor disease progression, and predict receptivity to treatment. Since proteins are effectors driving cell behavior, they are potential candidates for biomarkers (Weston and Hood, 2004).

One promising approach in the identification of biomarkers is the use of proteomic technology. The advantage of proteomic technology is that it allows us to study proteins in a high-throughput fashion. This greatly increases the chances of identifying single or even combinations of protein biomarkers. Surface-Enhanced Laser Desorption and Ionization (SELDI) is a proteomic technique that has been used for biomarker discovery (Srinivas et al., 2002). However the current peak detection method used in SELDI (Fung and Enderwick, 2002) is known to have low specificity (Coombes et al., 2003). If this limitation could be overcome, researchers would be able to identify combinations of protein biomarkers with greater ease. This has potential applications, for example, on mass cancer screening, which drives current research towards identifying combinations of biomarkers, since single biomarkers have been found to be ineffective (Etzioni et al., 2003).

Different proteomic techniques may detect different protein biomarkers (Anderson et al., 2004). Using multiple proteomic techniques to analyze the same sample could increase the number of biomarker candidates discovered. We therefore extend our method to Matrix-Assisted Laser Desorption and Ionization (MALDI) (Karas et al., 1985; Karas and Hillenkamp, 1988), another commonly used proteomic technique that has potential for biomarker discovery. Similar to SELDI, MALDI requires peak detection to be applied

to its output.

Diseases affect common protein regulatory networks as well as common gene regulatory networks. Researchers have recognized the potential of jointly analyzing gene and protein expressions in assessing the physiological state of a diseased cell (Weston and Hood, 2004). Present efforts use bioinformatic tools to integrate transcriptomic and proteomic data using deoxyribonucleic acid (DNA) and protein sequence databases (Cox et al., 2005; Waters et al., 2006). However, there are difficulties in data integration as large amounts of data could be lost due to the exclusion of the genes and proteins that are not matched (Waters et al., 2006).

In this thesis we aim to: (i) develop an improved method that performs peak detection and quantification in SELDI for biomarker discovery studies, and extend the method to MALDI, and (ii) integrate proteomic and transcriptomic information from the same samples by characterizing the patterns of correlation between the large number of gene and protein expressions, thereby detecting proteins that are jointly involved in regulating gene expressions.

Chapter 2

Background

The key focus of our work is the use of proteomics for biomarker discovery. In this chapter we will introduce some basic concepts in proteomics, focusing on topics related to protein expressions.

We begin with an overview of what a protein is and its role in biology. This is followed by a review of biomarker discovery studies in proteomics and a brief look at related proteomic technologies. Two proteomic techniques - SELDI and MALDI - are examined, with particular reference to their peak detection methods. We conclude the chapter by discussing the prospects of integrating transcriptomic and proteomic data for elucidating complex biological processes.

2.1 Proteins

Proteins play a vital role in various biological processes (Cooper and Hausman, 2004; Alterovitz et al., 2006):

- They are involved in the reading, copying and organizing of genetic code in the DNA.
- They are key agents in processes such as digesting nutrients, defending against pathogens and directing growth.
- Cells communicate with other cells via protein-based signals.

- Structural proteins are also responsible for holding an organism together.

A protein is made up of a chain of amino acids. There are 20 different amino acids and each amino acid is made up of:

- a central carbon atom (called the α carbon); and
- a hydrogen atom; and
- an amino group (NH_3^+); and
- a carboxyl group (COO^-); and
- a side chain (called the R group).

There are 20 side chains which differentiate each of the 20 amino acids. A linear sequence of amino acids is linked up by forming amide linkages through condensation polymerization of amino and carboxyl groups of adjacent amino acids. The sequence of the amino acids determines the structure and function of the protein. The size of a protein is measured by its total molecular mass, with the unit of measurement being the dalton (Da). Due to the occurrence of natural isotopes (chemical elements with the same number of protons, but different numbers of neutrons), a protein could consist of molecules with masses that are consecutively 1 Da apart. In the case of proteins, the monoisotopic mass (defined as the mass when a molecule is made up of the most common isotope of each element) happens to be the minimum mass.

Proteins are synthesized (translated) from messenger ribonucleic acids (mRNAs) that are first transcribed from sequences of DNA coding for the proteins. This process is called the central dogma of molecular biology; see Figure 2.1. The DNA consists of a double helix of nucleotides. There are a total of four different nucleotides: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Each nucleotide has a complementary pair: C is paired with G, and A is paired with T. The mRNA is a single strand of nucleotides with the nucleotide Uracil (U) instead of the Thymine. Each nucleotide triplet (i.e. codon) of the mRNA codes for either an amino acid or a stop signal. For example, the first codon of the mRNA in Figure 2.1, GUG, codes for the amino acid Valine (V). After the translation of mRNA into protein, the protein may be altered by post-translational modifications.

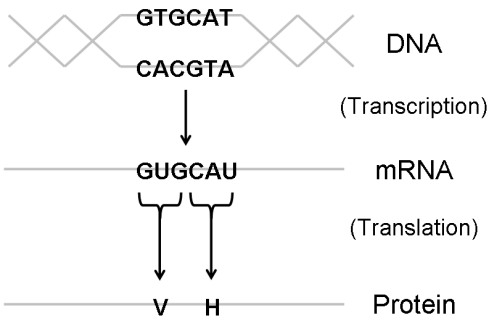

Figure 2.1: The central dogma of molecular biology.

The human genome has around 20,000 to 25,000 genes (International Human Genome Sequencing Consortium, 2004) while the genome of the worm *Caenorhabditis elegans* has around 19,000 genes (C. elegans Sequencing Consortium, 1998). Such close similarity in the number of genes between the two organisms suggests that the genome itself is not sufficient in explaining their differences. Perhaps proteins, estimated to be close to a million for humans (Bairoch and Apweiler, 2000), may provide the additional information to explain the differences.

2.2 Biomarker discovery

Biomarkers are molecules that can serve as early warning indicators for disease, help to monitor disease progression, and predict receptivity to treatment. They can be classified into three major groups: diagnostic, prognostic and predictive markers (Alaiya et al., 2005). Diagnostic markers are needed for early, accurate diagnosis of diseases to enable optimal treatment choices, while prognostic markers provide information about the future course of a disease, which would influence treatment decisions. Predictive markers offer insight on the potential responses of an individual to the various treatment options. Therefore, biomarker discovery could potentially advance the development of predictive, preventive and personalized medicine (Weston and

Hood, 2004).

The emergence of technologies from the field of transcriptomics and proteomics provided another avenue for seeking out potential biomarkers in biomarker discovery studies. This will improve, for example, our chances for developing diagnostic markers for early detection of cancer. Currently, single biomarkers are used to detect cancer but they are not effective for mass cancer screening. One reason for that is their inadequate specificity and sensitivity (Etzioni et al., 2003). Proteomics may provide the solution. There has been promising research showing combinations of biomarkers identified by SELDI having potentially better specificity and sensitivity than established single biomarkers in detecting cancer (Cho, 2007).

Biological samples, such as tissues or various biological fluids, are good sources to obtain biomarkers. For example, cells from a diseased tissue or its proximal biological fluid could potentially give us the biomarkers for the disease. However, a suitable biological sample, especially for early detection of diseases, should be obtained in a non-invasive and easy way. Blood, in the form of serum or plasma, is especially promising because the circulatory system of our bodies allows blood to be in constant contact with our tissues. Therefore, blood should contain proteins secreted by the diseased tissue. Unfortunately, 97% of the protein content in blood is dominated by about seven proteins such as albumin, immunoglobins and fibrinogen (Schulte et al., 2005). Methodologies developed to deplete the highly abundant proteins could improve our chances of detecting the proteins secreted by the diseased tissue (Villar-Garea et al., 2007). It is very likely that tumor secreted proteins are present in very low amounts, especially in the early stages of cancer. There is, however, growing evidence of immune response to cancer, which provides us with alternative sources of biomarkers, such as antigenic tumor proteins and the antibodies they elicit (Hanash et al., 2008).

Most biomarkers are identified through case-control study designs (Zhang and Chan, 2005). The biological samples of cases (people with the disease) and controls (people without the disease) are collected, and their protein profiles are then used to identify proteins that are differentially expressed between the two groups (i.e. protein biomarkers). These identified biomarkers are then used in a classification algorithm that assigns samples into either the diseased group or the control group. An ideal study design would be a prospective cohort study, where biological samples of individuals are collected periodically before and after their diagnosis of disease (Weston and Hood, 2004). The establishment of large and long-term cohorts such as LifeGene (`lifegene.ki.se`) and Singapore Consortium of Cohort Studies

(SCCS) (`www.nus-cme.org.sg`), will contribute immensely to the identification of biomarkers for diseases.

2.3 Proteomic technologies

Proteomic technologies can be divided according to the two approaches toward biomarker discovery: target and non-target driven approaches (Dhamoon et al., 2007). In the target-driven approach, there is a pre-selected list of proteins of interest, such as proteins involved in a particular biological pathway, which are checked for their associations with the disease. Protein microarrays, which include forward-phase and reverse-phase arrays are used. Forward-phase arrays have a variety of antibodies immobilized on them to detect the proteins present in the sample, similar to a gene expression microarray. In reverse-phase arrays, the samples are immobilized on the array and then probed with an antibody. For this technology, the antibodies need to be of high sensitivity and specificity to their target protein.

The non-target driven approach does not require a pre-selected list of proteins. The proteomic technologies used are two-dimensional gel electrophoresis (2DGE) and mass spectrometry (MS). 2DGE separates proteins based on their charge (the first dimension) and size (the second dimension), and the proteins are presented as spots on a biaxial plane; see Figure 2.2. Although 2DGE is a useful tool for biomarker discovery, it is slow and can only detect heavy protein molecules. The two-dimensional difference gel electrophoresis (DIGE) is a modification of 2DGE, with test and reference samples labeled before separating their proteins on the same gel, allowing for relative quantification of each spot.

The most promising technology in proteomics research is the mass spectrometry (Dhamoon et al., 2007), which consists of an ion source, a mass analyzer and a detector. The ion source produces ions from the biological sample. The three commonly used ionization methods are MALDI, SELDI and Electrospray Ionization (ESI). In MALDI and SELDI, each protein tends to pick up a single proton, which means that the mass-per-charge (m/z) ratio of the protein is its mass. In ESI, each protein could pick up different numbers of protons. Therefore, proteins with the same mass can have different m/z. Mass analyzers resolve the ions into their respective m/z values. Some basic types of mass analyzers for mass spectrometry are time–of–flight (TOF) and quadrupole time–of–flight (Q–TOF). Finally, the detector counts the number of ions that the mass analyzer resolves.

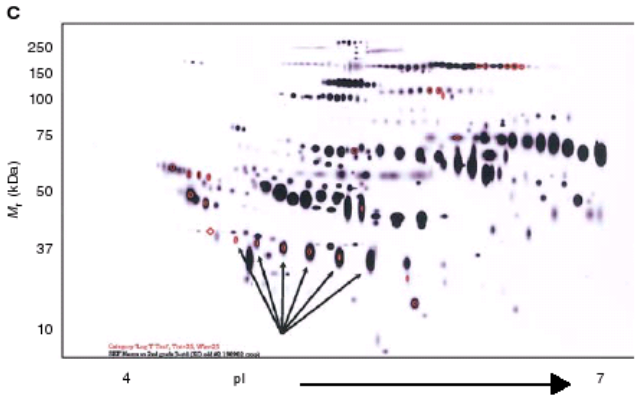

Figure 2.2: An example of a 2DGE. Haptoglobin-I identified as a potential biomarker for ovarian cancer. Reprinted by permission from Macmillan Publishers Ltd: British Journal of Cancer (Ahmed et al., 2004), copyright 2004, `http://www.nature.com/bjc/index.html`.

A protein profile of a sample, obtained by mass spectrometry, is shown graphically as a line plot with m/z and intensity as the horizontal and vertical axis respectively in Figure 2.3. A peak in the spectral profile suggests the presence of a protein in the sample. We can then identify the protein by the m/z of the peak and determine its amount from the intensity value of the peak. This is called the 'top-down' proteomics approach. However, the measurement accuracy in the mass spectrometry decreases as the mass of the protein increases, making identification of large proteins difficult. Post-translational modification further complicates the identification of proteins, because the sequence of amino acids remains unchanged, but the mass is changed.

The other proteomics approach, 'bottom-up' or 'shotgun' proteomics, breaks proteins into smaller units, called peptides. The protein is digested by adding protease, which cleaves the proteins at predictable amino acid locations. This results in better measurement accuracy with peptides that are of lower masses. If enough peptides remain unmodified, it could be possible to identify proteins with post-translational modifications. To identify the proteins, bioinformatic tools are used to compare the detected masses from the mass spectrometry with the theoretical masses of proteins from the genome of the organism.

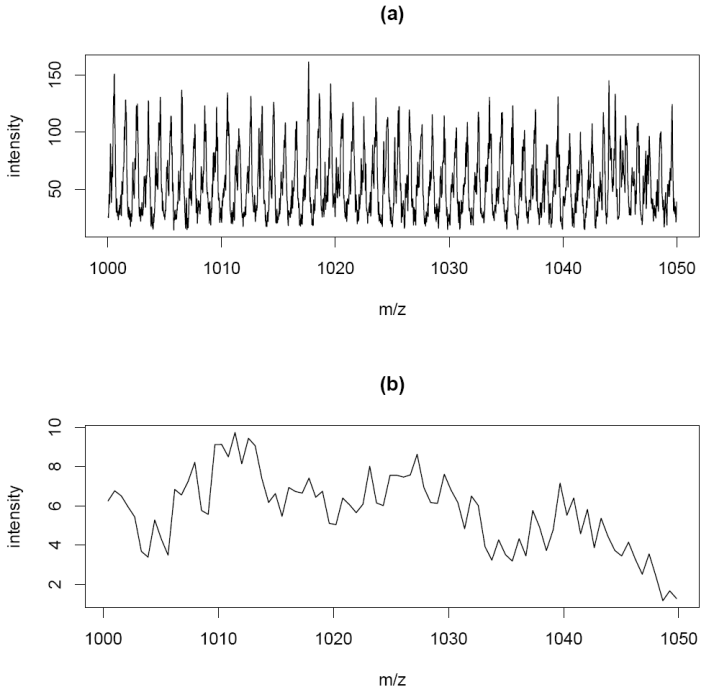

Figure 2.3: The blank spectra profile from (a) MALDI and (b) SELDI in the 1000 to 1050 Da range.

2.4 SELDI and MALDI

Surface-Enhanced Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (SELDI-TOF-MS), developed by Ciphergen Biosystems, and Matrix-Assisted Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF-MS) are two well recognized mass spectrometry techniques for obtaining protein profiles from biological samples (Hutchens and Yip, 1993; Karas et al., 1985; Karas and Hillenkamp, 1988). Both techniques have been used to explore large clinical cohort materials, such as plasma, because of their high-throughput capability and their ability to analyze a large number of samples within a short span of time.

Both techniques co-crystallize the samples and matrix on a surface; see Figure 2.4. The matrix enables the transfer of laser energy for the desorption and ionization of the proteins in the samples. The ionized proteins are then accelerated into the vacuum time-of-flight (TOF) tube. In the TOF tube, similar protein molecules gather together and hit the detector at the same time. The detector records the TOF and the number of ions that hit the detector.

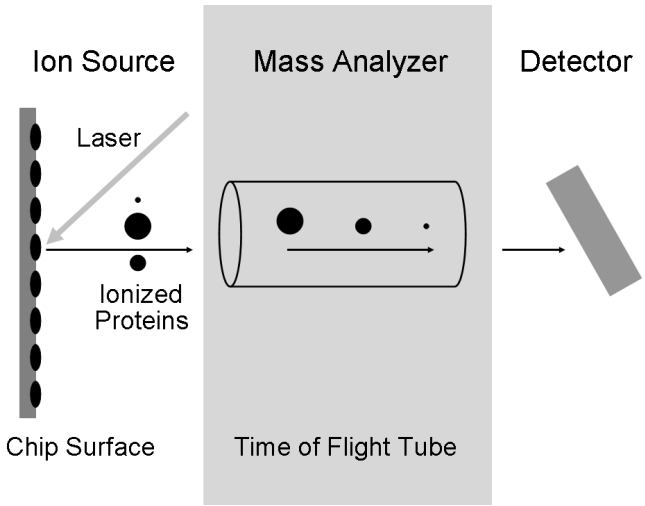

Figure 2.4: A schematic overview of either MALDI-TOF-MS or SELDI-TOF-MS.

The pairs of (TOF, intensity) data undergo three pre-processing stages before any data is analyzed: mass calibration, baseline subtraction and normalization. Firstly, mass calibration converts raw TOF values into m/z. Normally, proteins with known molecular weights are used to calibrate the quadratic relation between m/z and TOF. Baseline subtraction is then carried out to eliminate baseline signal caused by chemical noise from the matrix molecules. Finally, normalization is performed to eliminate any variation between samples that is not due to biological differences.

One major difference between SELDI-TOF-MS and MALDI-TOF-MS lies in the ionization surface. Unlike MALDI, SELDI ionization surfaces are coated with various activated and patented chemistries (i.e. pre-coated chromatographic chips) that select a subset of proteins based on their physiochemical

property, enabling an integrated fractionation of the protein sample. Because of the ease in sample preparation, there is greater potential in the use of SELDI in clinical applications.

The resolution of current MALDI-MS/MS instruments (i.e. tandem mass spectrometry) is higher than most of the SELDI-MS instruments used for protein profiling. This is illustrated in Figure 2.3, where the spectral profile of blanks (i.e. only the matrix is applied to the chip) from a MALDI-MS/MS instrument (AB4800) and SELDI-MS instrument (PBSIIc) are plotted using 2635 and 85 (m/z, intensity)-pairs respectively in the 1000 to 1050 Da region. Proteins signals from MALDI-MS/MS instruments can be resolved into their isotopic patterns.

One common limitation of SELDI and MALDI is that both techniques are unsuitable for detecting proteins of high molecular weight (>100 kDa). A review paper by Kiehntopf et al. (2007) and Poon (2007) discussed in further details the limitations of SELDI for biomarker discovery, such as competitive binding and competitive ionization, while Engwegen et al. (2006) summarized the advantages and disadvantages of selected proteomic technologies.

2.5 Existing methods for peak detection

The presence of proteins in the sample is indicated by peaks in the spectral profile from SELDI and MALDI. Therefore developing an effective peak detection method is critical. There are currently two general approaches in peak detection: intensity threshold-based methods and matching spectral intensities to a reference peak shape. In intensity threshold-based approaches, an instrument noise level is first established. This could be the standard deviation of the intensity values in the absence of peaks. This noise level is then used to define a critical threshold that flags intensities exceeding the threshold as peaks.

An example of a threshold-based method in SELDI is by Coombes et al. (2003). They propose using the first differences of successive intensities to identify local maxima and local minima, and the median of the absolute values of first differences as the noise level. Maximum points, whose distances to their nearest local minimum points are greater than the noise level, are marked as potential peaks. In a more recent method by Coombes et al. (2005), the authors estimate the noise from the residuals in a wavelet denoising method that does baseline subtraction on each spectrum. Yasui et al.

(2003) identify points as peaks if their intensities are the maximum in a neighborhood containing a prespecified number of points. A smoother is subsequently used to estimate the local noise.

Spectral matching approaches compare the intensity values within a window to a reference peak for detection and subsequent characterization. In general, this requires the specification of a predefined reference peak and a suitable distance measure to determine the similarity of a window of intensity values with the reference peak. For MALDI, Kempka et al. (2004) suggest the sum of two Gaussian functions as the reference peak and the least-squares as the distance measure. By describing the spectral profile with a mixture model, which consists of components such as peak signals (reference peaks) and background noise, Dijkstra et al. (2006) and Wang et al. (2008) have recently developed approaches that perform peak detection and baseline correction simultaneously on SELDI data. Dijkstra et al. (2006) use the EM-algorithm to solve for the parameters that contain peak information, while Wang et al. (2008) use the reversible jump Markov Chain Monte Carlo approach. Jarman et al. (2003) develop an approach for MALDI data, where the reference peak does not need to be specified explicitly. They use a histogram-based model for spectral intensity and detect peaks by comparing the estimated variance of the observations to the expected variance when no peak is present in a window.

However, most of the peak detection methods do not utilize information across spectra. This seems unnatural, especially in protein profiling studies for biomarker discovery, because we expect a potential biomarker to be consistently detected across spectra of samples with similar conditions. Pooling spectra together can potentially improve the characterization of the background noise and reduce the number of falsely declared peaks (i.e. false positives). Yu et al. (2008) suggest borrowing peak information across spectra by aligning multiple peaks across the spectra and keeping those peaks that are consistent across spectra, while Morris et al. (2005) suggest doing peak detection on the mean spectrum because it is less affected by noise than individual spectra.

In order to incorporate the isotopic pattern in MALDI data, additional steps have been proposed. After peak detection, Senko et al. (1995) and Breen et al. (2000) propose using averagine (an average amino acid) to model isotopic distributions. Instead of averagine, Wehofsky et al. (2001) enumerate all possible amino acid formulae, and use the relative intensity of the second and third peak to the first peak to model the isotopic pattern. A commonly used method for MALDI data analysis, PeakExplorer, picks a 'typical' peak

model and employs a non-linear iterative algorithm to fit detected peaks to the peak model (Applied Biosystems, 2008).

2.6 Integrating transcriptomic and proteomic data

Advanced technology enables us to measure thousands of gene and protein expressions simultaneously. Joint analysis of these gene and protein expressions from the same sample has the potential to discover complex biological processes. Present efforts use bioinformatic tools to integrate transcriptomic and proteomic data through DNA and protein sequence databases (Cox et al., 2005; Waters et al., 2006). Briefly, the current approach matches genes and proteins through a common identifier from the databases, before computing the pairwise correlation.

One disadvantage of the current approach is that large amounts of data can be lost in the matching process. This is well illustrated by Waters et al. (2006), who found 60% of the proteins from liquid chromatography-mass spectrometry analysis did not match the sequence identifiers from two microarray platforms, Affymetrix and Nimblegen. At least 29% and 46% of the genes from Affymetrix and Nimblegen, respectively, did not overlap with the proteins.

Although the central dogma of molecular biology suggests a strong correlation between gene and protein expressions, past studies suggest only a modest correlation (Nie et al., 2007). Factors that potentially mask the correlation are: analytical variability of the measurement technologies, post-transcriptional mechanisms affecting mRNA stability and protein degradation, and timing differences between gene and protein expressions. Since genes and proteins are connected in pathways or processes, a global correlation between genes and proteins will be more informative, as a result of pooling signals across genes and proteins.

Furthermore, proteomic technology is still not as comprehensive in its coverage as compared to transcriptomic technology. Therefore, protein expressions corresponding to some genes might not be measured and their expression values are set to zero. In order to account for the excess number of proteins with expression values at zero, the zero-inflated Poisson regression model was proposed, with the mean protein expression value defined as a function of the gene expression value (Nie et al., 2006).

In addition, diseases affect both common protein and gene regulatory networks. Integrating mRNA and protein level expressions may be a way to improve our ability of finding biomarkers and reduce the number of falsely identified protein or gene biomarkers. From data integration, we can also potentially gain understanding of the interplay of genes and proteins in diseases.

Chapter 3

Aims

The aims of this thesis were motivated by the two issues mentioned in Chapter 1. The first is the low specificity problem of the peak detection step in the analysis of SELDI data (Coombes et al., 2003). To overcome this problem, scientists visually inspect multiple spectra in parallel, a time consuming task which slows down the biomarker discovery process. The second is the loss of large amounts of data when integrating transcriptomic and proteomic data. Given the insufficiency of using one 'omic' technology to gain a comprehensive understanding of the biological processes (Hegde et al., 2003), improvements to data integration could pave the way to better biomarker discovery techniques.

The overall objective of this thesis was to address the above issues and the specific aims were:

1. To develop an improved method that performs peak detection and quantification in SELDI for biomarker discovery studies (Paper I and II).
    - Develop an R package for our method (Paper III).
    - Extend our method to MALDI data that have isotopic resolution (Paper V).
2. To integrate transcriptomic and proteomic data by characterizing their correlations through Maximum Covariance Analysis (Paper IV).

Chapter 4

Methods

This chapter starts with a description of the method we have developed for detecting and quantifying peaks in SELDI (Paper I and II), and its accompanying R package, called *ProSpect* (Paper III). Next, we describe the extension of our method to MALDI data with isotopic resolution (Paper V). We conclude the chapter with a brief presentation of the selected approaches for integrating transcriptomic and proteomic data, such as the Maximum Covariance Analysis (Paper IV).

4.1 Finding peaks in SELDI (Paper I-III)

Our method for peak detection, called Annotated Regions of Significance (ARS), consists of two steps (Paper II). In the first step, our aim is to detect a signal, which is defined as a spectral region containing potential biomarkers. The algorithm called Regions of Significance (RS), uses a modified F-statistic (F^*) to pick out regions with significant intensity variability between spectra (Paper I). Since all the spectra are analyzed simultaneously in our method, its characterization of the background noise is likely to be better than methods that detect peaks for each spectrum individually.

The second step is a peak quantification procedure, which focuses on the potential biomarker regions detected by RS. It extracts peak templates through Principal Component Analysis (PCA) across all spectra (Anderson, 1984; Stoyanova et al., 1995). With the templates, ARS estimates the amplitude and location of the peak in each spectrum, using the weighted least-squares method, and refines the estimation of the amplitude via a mixture model.

R is a widely used statistical programming environment (R Development Core Team, 2008) and we have implemented ARS as an R package called *ProSpect* (Paper III). Users are able to call up key functions to run different stages of ARS and have control over the tuning parameters. A graphical user interface version of *ProSpect*, called $ProSpectGUI$, has also been developed to make our method accessible to users not familiar with the R command line.

As mentioned earlier, our method identifies *potential* biomarker peaks. Apart from detecting peaks, it also filters out those which do not have the potential to be biomarkers, because they have similar intensities across the spectra. This should be seen as an advantage. To verify whether the peak is a biomarker, additional analysis is required to test for association between the peaks and the clinical or experimental outcome of interest.

In this section, we describe the modifications we made to the F-statistic. This is followed by a description on how PCA is used to identify a peak template, and our approach to fitting the template to a spectrum. We conclude this section with a description of *ProSpect*.

4.1.1 Modifications to the F-statistics

We obtained four blanks from SAX2 chips to understand the null distribution of the F-statistics from the one-way analysis of variance (ANOVA). Blanks are spectra generated from chips that carry no biological tissue. Therefore, theoretically, their intensities do not contain any biological signal for differentiating one spectrum from another, making them good null data candidates.

From our investigation of the null distribution of the F-statistics by using blanks, the F-statistic was observed to be inflated. Under the null hypothesis, where the intensities of the spectra are the same, the standard F-statistic with normal but dependent intensities can be inflated by a multiplicative factor c, i.e. distributed as cF (Scariano and Davenport, 1987). This dependency is not surprising, given that the raw spectra are in fact time series data. In addition, fluctuation in the mean square error (MSE) was observed. To deal with the fluctuation in the MSE and the dependency between intensities of a spectrum, we proposed two modifications to the standard F-statistic.

The first modification deals with the fluctuation in MSE by smoothing the MSE with the mean or median of its neighboring MSEs, denoted by MSE$'$. MSE$'$ estimates the variance of the error term better than MSE, thereby in-

creasing the sensitivity of the F-statistic. The smoothing parameter, M_{mse}, is expressed as a percentage of the total number of measurement points that form the spectral trace. Using the local median provides some robustness, whereas using the local mean allows us to apply the Satterthwaite's approximation to estimate the degrees of freedom of MSE$'$ (Satterthwaite, 1946). The F-statistic corresponding to MSE$'$ is denoted by F'.

The second modification aims to remove the effect of the dependency between the intensities of a spectrum. We expect a local correction factor to remove c and this is achieved by dividing F' by the mean or median of its neighboring F's and adjusting F' to its expected mean or median. The smoothing parameter, M_F, is also expressed as a percentage of the total number of measurement points. The final modified F is denoted by F^*.

When the local mean is used in the first modification, we expect F^* to follow an F distribution with estimated degrees of freedom from Satterthwaite's approximation. When a local running median is used, there is no simple way to compute the degrees of freedom of MSE$'$. Since MSE$'$ is based on a large number of points, we use the fact that F with degrees of freedom (df_1, df_2) is approximately $\chi^2_{df_1}/df_1$ for large df_2, where 1 and 2 denotes the χ^2-distribution statistic in the numerator and denominator of F, respectively. In practice, df_2 will be in the safe order of several hundreds.

4.1.2 Using PCA to extract a peak template

The first step of our peak quantification approach is to identify a peak template in each peak region. By performing PCA across all spectra in the peak region, we obtain a template that best captures the peak shape. However, we often encounter a misalignment problem along the m/z axis for real data. This is illustrated in Figure 4.1 (a) where a common peak shape can be observed across most spectra, but the unknown shifts in the peaks along the m/z axis create difficulties in having a clear view of the common peak shape. Hence, in order to depict a clearer view, we have aligned the peaks along a straight vertical line, as shown by the vertical dotted line in Figure 4.1 (b).

Alignment of the peaks and extraction of the peak template are performed simultaneously by linearizing the misalignment problem due to a shift in location. This is done through applying a first-order Taylor expansion to the spectrum, $S(t)$,

$$S(t) = \beta_0 + Af(t-\delta) \approx \beta_0 + Af(t) - A\delta f'(t), \tag{4.1}$$

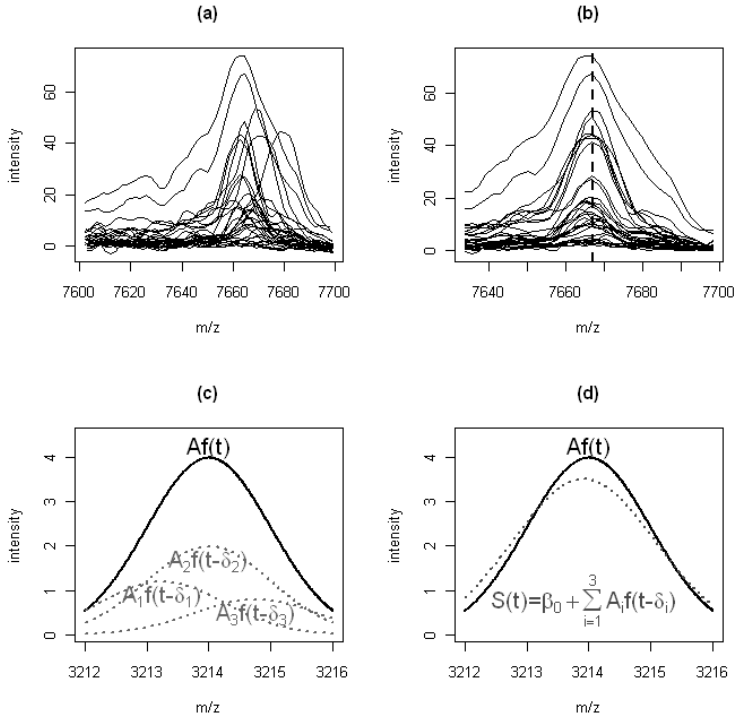

Figure 4.1: (a) Illustrates the misalignment problem. (b) Shows a common peak shape after aligning the peaks along the vertical dotted line. (c) The expected protein peak shape is $Af(t)$ (black solid line), but the protein was perturbed at three different locations, represented by shifts δ_1-δ_3, and with the corresponding amplitudes A_1-A_3. This results in three individual peaks, $A_1f(t-\delta_1)$-$A_3f(t-\delta_3)$ (gray dotted lines). (d) Instead of observing $Af(t)$, we observe the aggregation of the three individual peaks $\beta_0+\sum_{i=1}^{3} A_if(t-\delta_i)$.

where β_0 is an additive parameter to make the intensity values non-negative, $f(t)$ is the common template, A is the amplitude and δ is the shift.

Equation 4.1 suggests that the observed intensities can be decomposed into two-components, where the first Principal Component (PC), PC_1, provides a template for the peak shape, the second PC, PC_2, captures the remaining

signal-associated PC due to the shift and the remaining PCs are associated with noise (Stoyanova et al., 1995). We estimate β_0 to be the smallest non-positive intensity in the spectrum and estimate the other parameters by performing a least-squares regression with constrained parameters.

4.1.3 Fitting of the template to the other spectra

The fitting of the template $f(t)$ is done by minimizing the weighted mean squared error (wMSE) between the template $f(t)$ and the measurements $S(t)$:

$$\text{wMSE} = w \sum_t \{S(t) - \beta_0 - Af(t-\delta)\}^2/n, \tag{4.2}$$

where the weight w is defined as the reciprocal of the median intensity of the spectrum for the region and n is the number of points in the template. The wMSE is also used to compare the quality of the fit.

From Figure 4.1 (b), it is obvious that some spectra have slightly different peak shapes from the common peak shape. This might be expected because the peak shapes are an aggregation of the template perturbed around the m/z of the protein, as illustrated in Figures 4.1 (c)-(d). If the protein with amplitude A is perturbed at three different locations, represented by shifts δ_1-δ_3 with the corresponding amplitudes A_1-A_3, this will result in three individual peaks, $A_1 f(t-\delta_1)$-$A_3 f(t-\delta_3)$, represented as gray dotted lines in Figure 4.1 (c). Instead of observing the peak shown as black solid line, we observe the peak shown as gray dotted line in Figure 4.1 (d), which is the aggregation of the three individual peaks.

By re-formulating $S(t) = \beta_0 + \sum_{k=1}^{d} A_k f(t-\delta_k)$ as a mixture model problem, where the amplitude of the peak is naturally defined as the sum of the individual amplitudes, it turns out that

$$A = \sum_{k=1}^{d} A_k = \sum_t [S(t) - \beta_0] / \sum_t f(t). \tag{4.3}$$

Therefore no re-estimation is needed for the mixture model and we can use Equation 4.3 to refine the estimate of the amplitude.

4.1.4 *ProSpect*: An R package for ARS

We have implemented our method, ARS (Paper I and II), in an R package called *ProSpect*. R is a widely used statistical programming environment

that provides a base system and a large repository of modules called R packages (R Development Core Team, 2008). Since R mainly works on a command line interface to allow for the rapid creation of analysis workflow, it may not be a very friendly environment for beginners. Therefore, we developed a package *ProSpectGUI* which allows access to the functionality of *ProSpect* via a graphical user interface.

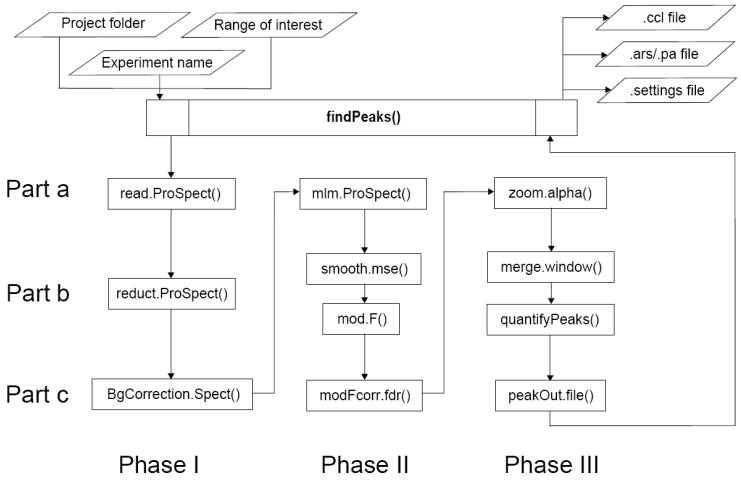

Figure 4.2: Basic workflow of the key function $findPeaks()$.

In *ProSpect*, we use key functions to facilitate (i) user control of the algorithms' parameters and (ii) maintenance of the codes. A key function controls separate and independent blocks of codes which usually have specific roles. There are three key functions in *ProSpect*:

- $findPeaks()$: Identifies and quantifies the peaks in the spectra, and exports the results.
- $summaryPeaks()$: Summarizes the basic statistics of the intensity and m/z of the spectra.
- $plotPeaks()$: Plots the spectra at different steps of the algorithm.

The three key functions have two kinds of inputs: required parameters and optional parameters. The required parameters are a minimum set of arguments that has to be specified before the key function can be executed. For the optional parameters, default values exist which can be changed by advanced R-users. In the following, we focus on the key function $findPeaks()$ that performs the ARS algorithm.

Table 4.1: The ARS algorithm implemented in $findPeaks()$.

Step	**Task**	**Description**
I	Data preparation	Preparation of the data to calculate the F-statistics
Ia	Importation	Read the .csv files exported by Ciphergen's software
Ib	Reduction	Reduce data to the region of interest
Ic	Baseline correction	Correct the intensity for baseline noise
II	Calculation of different F-statistics	Spectra are reduced to one spectrum of F^*
IIa	F	Calculate the F-statistic
IIb	F'	Smooth the MSE to get MSE'
IIc	F^*	Scale F' to its null distribution
III	Peak detection and exportation of data	Export information of peaks in potential biomarker regions
IIIa	Identification of potential biomarker regions	Flag regions that are significant from the user specified criterion and cut-off level
IIIb	Peak quantification	Quantify the peaks in the potential biomarker regions
IIIc	Exportation of data	Export peak information via .ccl, .ars and .pa files

Figure 4.2 and Table 4.1 briefly describe the workflow of $findPeaks()$. In Phase I, $findPeaks()$ reads in comma separated value (csv) files containing intensity and m/z information of the spectra, and pre-processes the data for identifying and quantifying potential biomarker peaks. Functions in Phase II flag out potential biomarker regions via the RS algorithm, and those in Phase III quantify the peaks via the ARS algorithm. The estimates of the m/z and intensity of peaks are then stored for further analysis. Recently, we updated *ProSpect* by adding a parametric peak template (i.e. a mixture of log-normal distribution) option to quantify the peaks in a cluster (Dijkstra et al., 2006).

After each run of $findPeaks()$, it generates .settings, .ars, .pa and .ccl files. The full name of the output files depends on the specified project name and gives the user the possibility of distinguishing between different calculations done for one dataset. The .settings file records all the options specified to run $findPeaks()$; .ars and .pa files contain the estimated m/z and intensity of peaks from ARS and the parametric peak template approach respectively; and the .ccl file contains information for importing peaks detected by ARS into Ciphergen software.

Users who are unfamiliar with R may have difficulties manipulating *ProSpect* because of the command line interface environment in R. To increase the usability of *ProSpect*, we used an R package, called *tcltk* (Dalgaard, 2001), to develop a graphical user interface version, *ProSpectGUI*, which is also available as an R package.

4.2 Finding peaks and isotopic patterns in MALDI (Paper V)

In this section, we briefly describe the extension of ARS for MALDI data. For SELDI, a protein biomarker is represented by a peak, while MALDI - which can resolve a protein signal into its isotopic pattern - a protein biomarker is represented by a series of peaks that are consecutively 1 Da apart. To pinpoint a biomarker using MALDI, we need to consider the m/z and intensity for each peak in the isotopic pattern.

From ARS, we have obtained potential biomarker peaks, like the blue vertical lines in Figure 4.3 that represent the m/z and intensity estimates of the peaks for a particular spectrum. Given that an isotopic pattern of a protein is made up of a series of peaks that are consecutively 1 Da apart, and that we expect a protein biomarker to be made up of a series of biomarker peaks, we first filter out stand-alone potential biomarker peaks from ARS that cannot be formed as part of a series of peaks which are more or less consecutively 1 Da apart. The series of blue vertical lines in Figure 4.3 is made up of biomarker peaks that are approximately 1 Da apart from their immediate neighbors.

Another aspect of an isotopic pattern is its intensity. The expected isotopic pattern of the intensity approximately follows a Poisson distribution. We have used the approach by Breen et al. (2000) to get the expected isotopic pattern at a given m/z value. The red vertical lines in Figure 4.3 represent the expected isotopic pattern in the region. Capitalizing on this information,

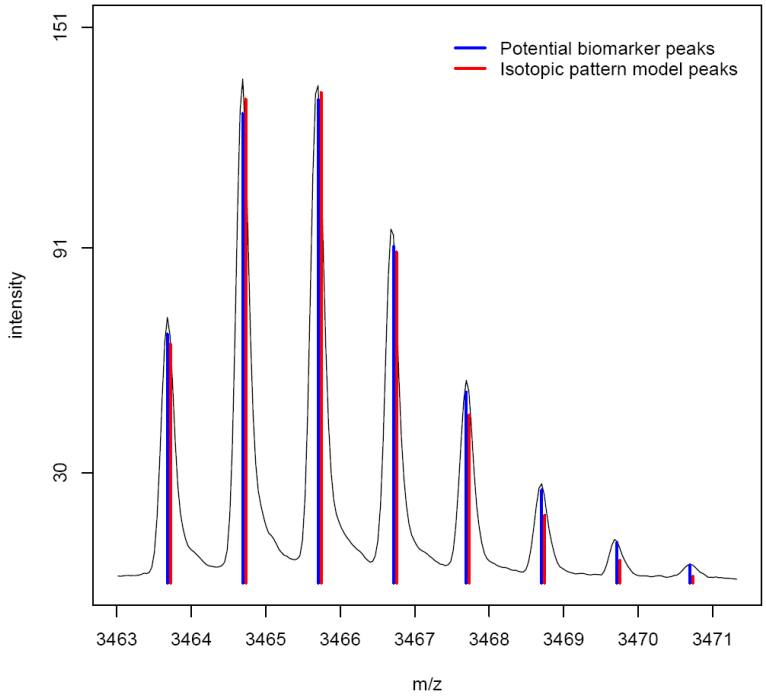

Figure 4.3: An illustration of the extended ARS for MALDI data.

we define a goodness-of-fit measure for the isotopic peak, r, as a ratio of the sum of squared residuals between isotopic pattern and constant intensity models. If $r < 1$, a potential protein biomarker is detected. If $r > 1$, the series of peaks is probably noise. The series of peaks in Figure 4.3 has $r < 1$, suggesting a detection of a potential protein biomarker. This process of eliminating the series of peaks is done sequentially with the series of peaks ordered according to its minimum m/z in an ascending manner.

4.3 Correlating gene and protein expression data (Paper IV)

In this section, we describe the multivariate statistical method called Maximum Covariance Analysis (MCA). To the best of our knowledge, this is the first time MCA is applied to the problem of integrating transcriptomic ($\mathbf{X}_{p\times n}$) and proteomic ($\mathbf{Y}_{q\times n}$) data with p genes and q proteins from the same n samples. This is followed by a brief description of Gene Ontology (GO) enrichment analysis which is used to gain biological insight into the results obtained from MCA. We conclude this section with a description of a closely related technique called the Generalized Singular Value Decomposition (gSVD), which has been previously used to jointly analyze gene expression and copy number variation information from the same samples (Berger et al., 2006). A comparison of the two methods, MCA and gSVD, will be presented in Chapter 5

By considering an extension of the factor analysis model (Salim and Pawitan, 2007), where the factors are allowed to correlated, MCA can be used to estimate gene ($\mathbf{a_j}$s) and protein ($\mathbf{b_j}$s) patterns. Let $\boldsymbol{x}_j$ be a p-vector of gene expression, and $\boldsymbol{y}_j$ a q-vector of protein expression data from sample j, for $j = 1, \ldots, n$. Assuming co-expression in r pathways are reflected in r gene patterns and protein patterns:

$$\boldsymbol{x}_j = \sum_{k=1}^{r} g_{jk}\mathbf{a_k} + \boldsymbol{\epsilon}_j^x, \text{ and } \boldsymbol{y}_j = \sum_{k=1}^{r} h_{jk}\mathbf{b_k} + \boldsymbol{\epsilon}_j^y, \tag{4.4}$$

where g_{jk}s and h_{jk}s are random scalars associated with the unobserved r factors. Let $\mathbf{A}_{p\times r} \equiv [\mathbf{a}_1 \ldots \mathbf{a}_r]$ and $\mathbf{B}_{q\times r} \equiv [\mathbf{b}_1 \ldots \mathbf{b}_r]$, $\boldsymbol{g}_j \equiv (g_{j1}, \ldots, g_{jr})'$ and $\boldsymbol{h}_j \equiv (h_{j1}, \ldots, h_{jr})'$. To avoid non-identifiability, we assume orthogonality: $\mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B} = I_r$, and $\text{cov}(\boldsymbol{g}_j, \boldsymbol{h}_j) \equiv \Lambda$ is a diagonal matrix with decreasing values. The cross-covariance between $\boldsymbol{x}_j$ and $\boldsymbol{y}_j$ is given by:

$$\mathbf{A}\Lambda\mathbf{B}', \tag{4.5}$$

so the correlation is captured by the r values on the diagonal of Λ and the corresponding pattern-pairs given by $\mathbf{A}$ and $\mathbf{B}$.

MCA can be used to obtain estimates of the pattern matrices $\mathbf{A}$ and $\mathbf{B}$ in model (4.4). The MCA maximizes the *sample* covariance of linear combinations from two datasets. The objective is to find $\mathbf{a_1}$ and $\mathbf{b_1}$ such that

$$\lambda_1 \equiv \text{cov}(\mathbf{X}'\mathbf{a_1}, \mathbf{Y}'\mathbf{b_1}),$$

is maximized over all choices of $\mathbf{a_1}$ and $\mathbf{b_1}$, with $\mathbf{a_1}'\mathbf{a_1} = \mathbf{b_1}'\mathbf{b_1} = 1$. The constraints on the pattern-pair are needed because λ_1 can be made as large as possible by multiplying $\mathbf{a_1}$ and $\mathbf{b_1}$ with a constant scalar. The second pair of gene and protein patterns are found by maximizing:

$$\lambda_2 \equiv \text{cov}(\mathbf{X}'\mathbf{a_2}, \mathbf{Y}'\mathbf{b_2}),$$

over all unit vectors $\mathbf{a_2}$ and $\mathbf{b_2}$ that are orthogonal to $\mathbf{a_1}$ and $\mathbf{b_1}$ respectively. In summary, the k-th pattern-pair is found by maximizing:

$$\lambda_k \equiv \text{cov}(\mathbf{X}'\mathbf{a_k}, \mathbf{Y}'\mathbf{b_k}),$$

with constraints that $\mathbf{a_k}'\mathbf{a_k} = \mathbf{b_k}'\mathbf{b_k} = 1$, $\mathbf{a_i}'\mathbf{a_j} = 0$ for $i \neq j$, and $\mathbf{b_i}'\mathbf{b_j} = 0$ for $i \neq j$, where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k$.

The MCA estimates are obtained by performing a Singular Value Decomposition (SVD) on the cross-covariance matrix $\mathbf{\Sigma}_{p\times q} = \mathbf{X}\mathbf{Y}'/n$ with $\mathbf{X}$ and $\mathbf{Y}$ previously centered across the rows. The relative amount of covariances explained by the j-th component is $\lambda_j^2 / \sum_{k=1}^{r} \lambda_k^2$, where r is the rank of $\mathbf{\Sigma}$. To determine the number of pattern-pairs, we used a permutation approach.

The genes and proteins with large absolute pattern values in their respective pattern pair have strong influence on the cross-covariance matrix. Therefore we propose performing a GO enrichment analysis on the set of genes which have the top 5% absolute gene pattern values in its pattern-pair. In a GO analysis, we test whether a subset of genes is enriched with a particular GO term when compared to all genes on the microarray. Therefore a GO analysis reduces the test to a 2×2 table test of association between gene membership in a GO term and the set of genes having the top 5% gene pattern values. This allows us to make biological inferences about each pattern pair from MCA and hence identify associations between proteins and biological processes.

The gSVD, which has been used to integrate two datasets from the same samples, simultaneously reduces $\mathbf{X}$ and $\mathbf{Y}$ to a $s \times s$ metagene-array space:

$$\begin{aligned} \mathbf{X}_{p\times n} &= \mathbf{A}_{p\times p}[\mathbf{D}_{\mathbf{X}\,p\times s}, \mathbf{0}_{p\times(n-s)}]\mathbf{G}^{-1}_{n\times n} \\ \mathbf{Y}_{q\times n} &= \mathbf{B}_{q\times q}[\mathbf{D}_{\mathbf{Y}\,q\times s}, \mathbf{0}_{q\times(n-s)}]\mathbf{G}^{-1}_{n\times n}, \end{aligned}$$

where s is the rank of $[\mathbf{X}', \mathbf{Y}']'$, $\mathbf{A}$ and $\mathbf{B}$ are orthogonal matrices, and $\mathbf{D_X}$ and $\mathbf{D_Y}$ are matrices, such that their (i, j)-entries are zero when $i \neq j$, and non-negative when $i = j$, where $\mathbf{D_X}'\mathbf{D_X} + \mathbf{D_Y}'\mathbf{D_Y} = I_s$ (Paige and Saunders, 1981).

Similar to MCA, the significance of the i-th metagene and its corresponding meta-array for dataset $j = \mathbf{X}, \mathbf{Y}$ is quantified by:

$$P_{ij} = d_{ji}^2 / \sum_{t=1}^{s} d_{jt}^2, \tag{4.6}$$

where $d_{\mathbf{X}i}$ and $d_{\mathbf{Y}i}$ are the (i,i)-entries of $\mathbf{D_X}$ and $\mathbf{D_Y}$ respectively, which carry the expression information of the i-th metagene and its corresponding meta-array in $\mathbf{X}$ and $\mathbf{Y}$ respectively.

The relative significance of the i-th metagene is assessed through the ratio of the expression information from the datasets (Alter et al., 2003):

$$\theta_i = \arctan(d_{\mathbf{X}i}/d_{\mathbf{Y}i}) - \pi/4,$$

where $-\pi/4 \leq \theta_i \leq \pi/4$. When the angular distance is 0, the i-th metagene may be equally significant in both datasets. However, when the angular distance is $\pi/4$, the i-th metagene may have no significance in $\mathbf{Y}$ relative to $\mathbf{X}$. And when the angular distance is $-\pi/4$, the i-th metagene may have no significance in $\mathbf{X}$ relative to $\mathbf{Y}$.

Chapter 5

Results

In this chapter we present the results for (i) Paper I-III: Performance of ARS, (ii) Paper V: Performance of extended ARS for MALDI and (iii) Paper IV: Performance of MCA in integrating gene and protein expression data.

5.1 Performance of ARS (Paper I-III)

We used the following datasets to assess the performance of our method, ARS:

- *Lung cell line data (H69).* Two types of lung cell lines were studied, resistant versus sensitive to chemotherapy, with four spectra for each type of cell lines. A low intensity laser setting was used on SAX2 chips, and the analysis was restricted to the 3-10 kDa range. The scientist who was familiar with the data manually identified 51 regions in the spectra with biologically plausible peaks. These regions were taken to be the gold standard in determining true and false positives.
- *Spike-in data.* Bovine insulin at approximately 5733 Da was spiked into human blood serum at seven levels of dilution. Each dilution was performed in independent duplicates and applied to WCX2 chips. The chips were scanned at low laser intensity, resulting in 14 spectra. The analysis was restricted to the 5-6.5 kDa range.
- *Lung cancer serum data.* The data consisted of eight serum samples from patients diagnosed with adenocarcinoma and squamous-cell car-

cinoma, respectively. Duplicates of the sample were applied to CM10 chips with optimized mass between 2-10 kDa. The settings were obtained by an experienced SELDI analyst.

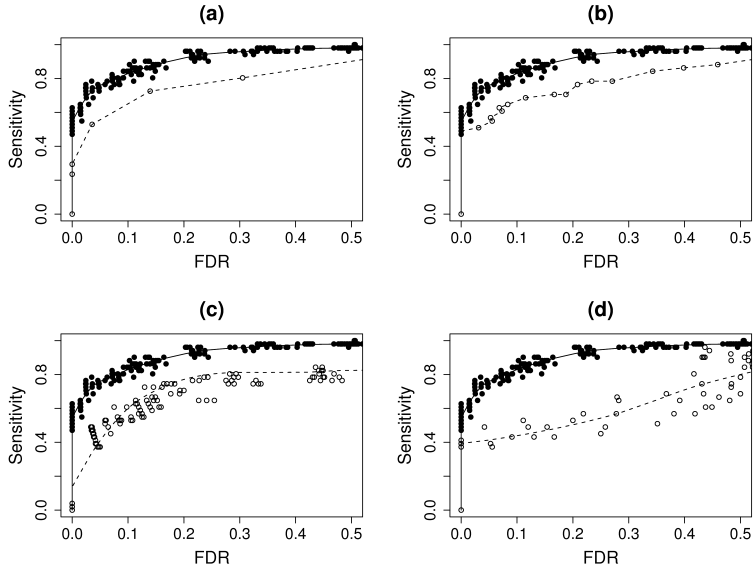

Figure 5.1: Comparing RS with the (a) standard method (Fung and Enderwick, 2002), (b) Coombes method (Coombes et al., 2003), (c) Yasui method (Yasui et al., 2003) and (d) Cromwell method (Coombes et al., 2005) respectively by using the lung cell line data in the 3 kDa to 10 kDa region. The scattered solid and open circles are the (sensitivity, empirical FDR)-pairs for RS and the other methods, respectively. The OC curve of RS is a solid line and the other methods are dashed lines.

5.1.1 Lung cell line data (H69)

Using the lung cell lines data (H69), we studied the operating characteristics (OC) curves of RS and four other methods, namely the standard method (Fung and Enderwick, 2002), 'Coombes' method (Coombes et al., 2003),

'Yasui' method (Yasui et al., 2003) and 'Cromwell' method (Coombes et al., 2005). We modified the traditional OC curves by replacing the false positive rate on the horizontal axis with the FDR (Choe et al., 2005). While the empirical FDR is available for all methods, the theoretical FDR is available for RS by converting the p-values from F^* to FDR.

Figures 5.1 (a)-(d) compare the OC curves of RS to the four methods mentioned above by using the empirical FDR. From the plots we observed that RS (solid curve) has better OC than the other four methods (dashed curves). At 80% sensitivity, the FDRs of the four methods are around 25% to 50%, compared to around 8% for RS.

We briefly summarize the other observations made from the same data:

- The local running median for smoothing the MSE and F' performed better than the other options, such as local running mean for smoothing either the MSE or F'. Therefore robust smoothing was necessary.
- The similarity of the OC curves for the theoretical and empirical FDRs corroborated our distribution theory of F^*.

5.1.2 Spike-in data

By using the standard method (Fung and Enderwick, 2002), the insulin peak was detected together with 30 peak regions across the whole mass range from 5.5-6 kDa; see bottom row of Figure 5.2. In comparison, RS identified nine significant windows in the 5.5-6 kDa range at FDR cut-off of 5%, out of which eight corresponded to the insulin peak; see Figure 5.2. Thus, this showed that RS had a higher specificity than the standard method.

5.1.3 Lung cancer serum data

We compared the performance of ARS with the standard method (Fung and Enderwick, 2002) on the SELDI data of serum samples obtained from lung cancer patients. While ARS detected 151 peak regions, the standard method only detected 89 peak regions. We investigated all peak regions which are not detected by both methods and visually verified that (i) 60 out of 68 ARS peak regions and (ii) all 11 standard peak regions were plausible. Using the McNemar test (Lachenbruch, 1998), ARS classified significantly more

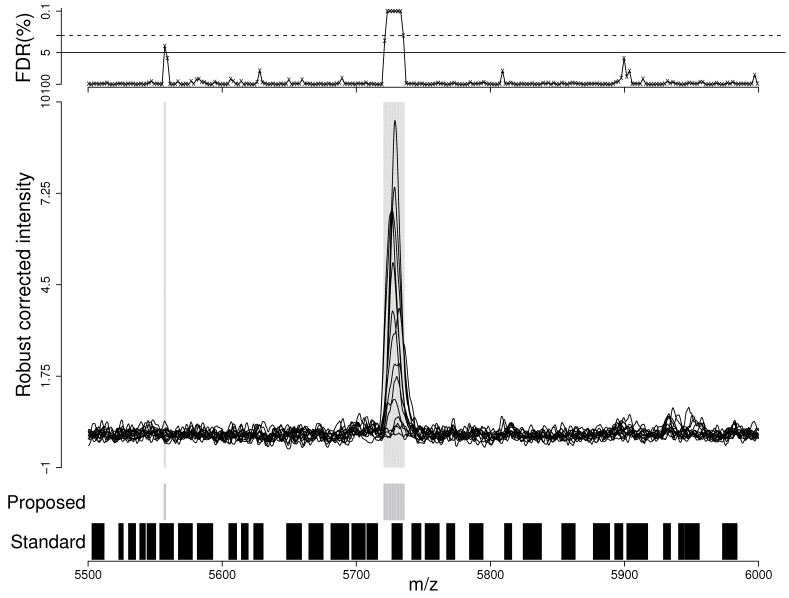

Figure 5.2: Analysis of the spike-in data in the 5.5 kDa to 6 kDa region; the insulin peak is at approximately 5733 Da. The top plot shows the FDR as a function of the m/z-value. The main plot shows intensity vs. m/z-value for all fourteen spectra. The two rows of vertical bars in the bottom correspond to regions flagged by RS (proposed, gray) and standard method (black) respectively.

regions correctly into peak and non-peak regions than the standard method (p-value=4.0×10^{-6}).

We noticed that 78 peak regions detected by the standard method overlapped with 83 peak regions detected by ARS. This discordance in the number of peak regions was due to the fact that five (single) peak regions detected by the standard method overlapped with two peaks from multiple peaks clusters of ARS. Further investigation revealed that the standard method was unable to distinguish multiple peaks in close vicinity of each other. In contrast, ARS was more robust in distinguishing them.

The other results obtained in the comparison of ARS with the standard

method showed the following:

- The standard method missed an obvious peak which ARS could identify.
- ARS performed better than the standard method under severe misalignment.

5.2 Performance of extended ARS (Paper V)

We used the spike-in data to assess the performance of extended ARS for MALDI. All spike-in samples contained Bovine Serum Albumin (BSA) tryptic digested. The following peptides were added into the samples at various quantities: (A) Angiotensin, (B) [Glu1]-Fibrinopeptide B, (C) Dynorphin A, (D) Adrenocorticotropic hormone (ACTH) and (E) β-Endorphin. Table 5.1 gives the detailed composition for each sample. Each sample was spotted once on the plate, except for Sample 12 which was spotted on five different spots. Five CHCA blanks were also spotted in parallel with the samples. The samples were then analyzed in an AB4800 MALDI-TOF/TOF Mass Spectrometer (Applied Biosystems) with optimized mass between 0.7-4kDa.

Table 5.1: Description of peptide composition for each sample (μl).

	Peptides						
Sample	A	B	C	D	E	BSA	MilliQ
1	10	3.33	5	0	8.33	0.38	72.96
2	0	5	6.67	5	10	0.38	72.95
3	3.33	10	0	6.67	3.33	0.38	76.29
4	5	0	10	8.33	1.67	0.38	74.62
5	1.67	8.33	8.33	10	0	0.38	71.29
6	6.67	1.67	1.67	3.33	5	0.38	81.28
7	8.33	6.67	3.33	1.67	6.67	0.38	72.95
8-11	5	5	5	5	5	0.38	74.62
12	0	0	0	0	0	0.38	99.62

5.2.1 Validation of ARS in MALDI

We applied ARS separately to the following groups of spectra:

- Blank spectra generated from spots that contained no sample.
- BSA only spectra generated from spots that contained Sample 12.
- Spike-in(8-11) spectra generated from spots of Sample 8 to Sample 11, which had the same quantity of spike-ins.

We varied the nominal p-value cut-offs for flagging significant windows within the region 0.7-4.01 kDa at the following values: 0.1, 0.05 and 0.01. At each nominal p-value, we computed the empirical false positive rates (total number of significant windows/total number of windows): (i) as a whole (overall empirical false positive rates) and (ii) for each of the 30 sub-regions of equal length (local empirical false positive rates). We also fitted a smoothed curve (loess fit) to the 30 (m/z, local empirical false positive)-pairs.

The first row of plots in Figure 5.3 shows a lack of agreement between the nominal p-values and the empirical false positive rates. However, there was strong agreement when the intensities of the blanks are log-transformed; see second row of Figure 5.3. This was also observed in: (i) log-transformed spectra from BSA only (log-BSA) and (ii) log-transformed spectra from Spike-in(8-11) (log-Spike-in(8-11)); see the third and fourth rows of Figure 5.3 respectively.

The above validates the use of the log-transformed intensities of MALDI for ARS. The log-blanks have the closest agreement between the nominal p-value and empirical false positive rates, followed by log-BSA and log-Spike-in(8-11).

5.2.2 Near the spike-in regions

Using Sample 1 to Sample 7, we computed the Pearson correlation of the quantity spiked and the estimated peak intensity from ARS at approximately 0, 1, 2, 3 and 4 Da to the right of the observed monoisotopic peak of Spike-in A to Spike-in E. In general, correlations were greater than 0.8 for 0-4 Da. The monotonic increasing relationship between the estimated peak intensity and the quantity spiked was linear in Spike-in A, B, C and E, but curvilinear

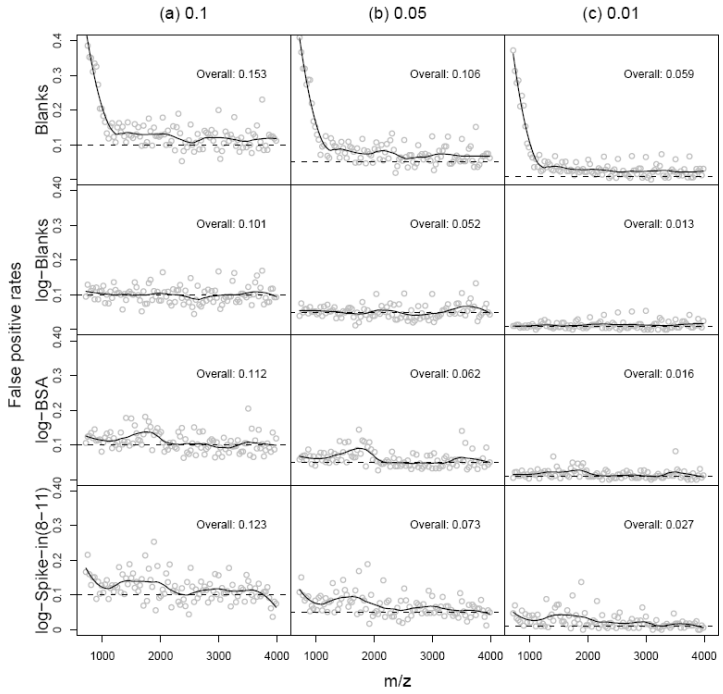

Figure 5.3: The empirical false positive rates at various nominal p-value cut-offs of: (a) 0.1, (b) 0.05 and (c) 0.01. The first, second, third and fourth rows correspond to blanks (Blanks), log-transformed blanks (log-Blanks), log-transformed BSA (log-BSA) and log-transformed spike-ins which have the same quantity spiked (log-Spike-in(8-11)). The gray circles are the 30 (m/z, local empirical false positive rate)-pairs; the black solid lines are the smoothed curve of the 30 local points and the horizontal black broken lines correspond to the nominal p-value. The overall empirical false positive rate is presented at the top right hand corner of each plot.

in Spike-in D for 0-4 Da. Therefore, peaks in an isotopic pattern of a protein contained information on the protein's quantity.

Using the same spike-in samples as above, we compared the extended ARS and the standard method, PeakExplorer, by investigating their abilities in capturing the quantity of proteins spiked. The ratio of residuals cut-offs for

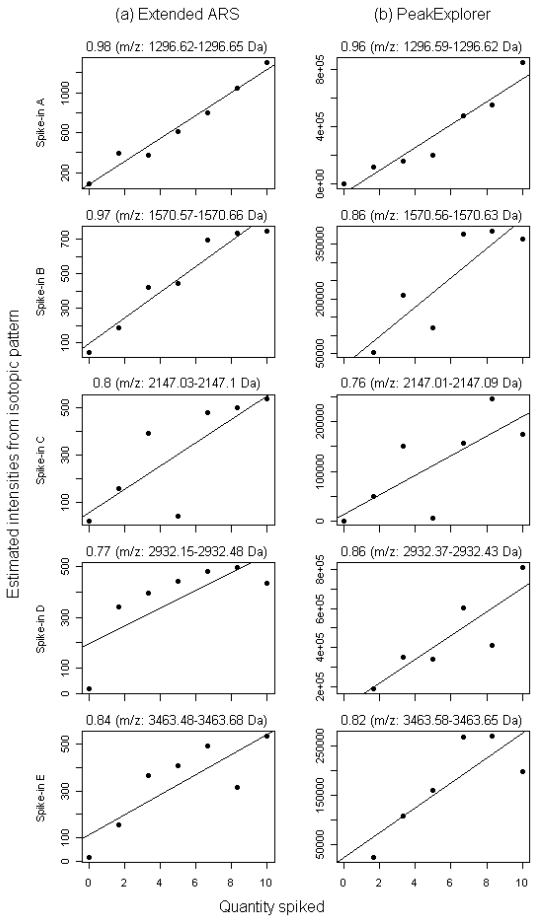

Figure 5.4: Scatter plot of the quantity spiked for Spike-in A to E in Sample 1 to 7 and the estimated protein intensity obtained from (a) extended ARS using $r = 0.1$, and (b) PeakExplorer using S/N=3, with the fitted linear regression line (black line). Above each plot, we report the Pearson correlation between the estimated protein intensity and quantity spiked, with the m/z range of the monoisotopic peak in parentheses.

the extended ARS were: 0.1, 0.2, 0.4, 0.6, 0.8, 1 and 1.2. For PeakExplorer, the signal-to-noise ratio (S/N) cut-offs were: 3, 5, 10, 15, 20, 50, 100 and 150.

In general, the Pearson correlation between the quantity spiked and the estimated protein intensity was higher in extended ARS than PeakExplorer; see Figure 5.4. The estimation of the location of the monoisotopic peaks was similar between the two methods with extended ARS having a wider monoisotopic peak m/z range. Therefore the two methods performed well in detecting the spike-ins, but extended ARS quantified the spike-ins better than PeakExplorer.

We investigated the detection of monoisotopic peaks in the neighborhood of $\pm$ 5 Da of the observed monoisotopic peak of Spike-in A to Spike-in E. For extended ARS, it only detected a monoisotopic peak that was 3 Da to the left of Spike-in B's observed monoisotopic peak. However, PeakExplorer detected more monoisotopic peaks in the neighborhood. Therefore, ARS is potentially more specific than PeakExplorer.

5.2.3 Across the entire mass range

Using the same spike-in samples in Section 5.2.2, we compared the performance of extended ARS and PeakExplorer across the m/z range by studying their OC curves. Similar to Section 5.1.1, we modified the traditional OC curves by replacing the false positive rate on the horizontal axis with the FDR. All the monoisotopic peaks detected by the two methods were visually checked to confirm if they were real monoisotopic peaks. The FDR is the proportion of unique monoisotopic peaks detected by the method that were verified real. Sensitivity is the proportion of all unique and verified real monoisotopic peaks that were detected by the method. The OC curves for the two methods were obtained by varying their cut-offs. When we compared their OC curves, at low FDR, extended ARS had higher sensitivity than PeakExplorer.

5.3 Performance of MCA (Paper IV)

We used the following datasets to investigate the integration of transcriptomic and proteomic data:

- *NCI data.* Microarray and proteomic datasets from the same human cell line of a variety of cancers were downloaded from the CellMiner program package, National Cancer Institute (`http://discover.nci.nih.gov/cellminer/`). One of the 60 human cancer cell lines was excluded from the analysis, because it had missing microarray information on the Affymetrix HG-U133A chip. The gene expression data were normalized using the GCRMA method (Shankavaram et al., 2007). For the proteomic data, reverse-phase protein lysate arrays (RPLA) were used to obtain 89 protein expressions. RPLA used an antibody to measure the amount of protein presented across samples by spotting many samples on one slide. For the microarray dataset, genes with expression variances lower than the 25th percentile were filtered out, leaving 15918 genes. For the proteomic dataset, no filtering was performed.

- *Simulated data.* An extended standard factor analysis model (Salim and Pawitan, 2007) was used to simulate correlated gene and protein expression data. Sample size was set to be $n = 59$, or 500, with $p = 1000$ genes, $q = 89$ proteins and $r = 2$ pairs of patterns. Sub-sampling of 1000 genes from the NCI data without replacement was performed to obtain realistic parameters and pattern-pairs, which were used to generate 250 simulated sets of correlated gene and protein expression data.

5.3.1 Simulated data

The simulated data were used to investigate the consistency of the MCA approach in estimating the pairs of patterns ($\mathbf{a_j}$s and $\mathbf{b_j}$s), and to compare MCA against gSVD. For MCA, estimates were close to the true patterns values when the sample size was large ($n = 500$). This suggested that MCA produced consistent estimates of the gene and protein patterns. However, a small bias was observed in the small sample ($n = 59$).

For gSVD, we considered gene and protein patterns that had the highest absolute correlation with the corresponding true patterns values. The result for large sample size ($n = 500$) suggested that gSVD captured some correlation patterns in the data, but it did not estimate them consistently, especially the gene patterns. Only 11% of these gene and protein patterns were from the same pair. In small samples ($n = 59$), we observed a smaller bias, but higher variability.

We also investigated if the use of angular distances would improve the strength of correlation between gSVD and the true patterns. There was no evidence of improvement.

5.3.2 NCI data

Using MCA

For MCA, we determined through permutation that the number of significant pattern-pairs was three; the three pattern-pairs explained 74.8% of the covariation. For subsequent pattern-pairs, the cumulative profile of the covariation started to plateau off to 100%. Therefore, we concluded the first three pattern-pairs were adequate in capturing the structure of the cross-covariance matrix between genes and proteins.

For each pattern-pair, we considered the genes from the top 5% absolute gene pattern values as interesting and performed a GO analysis on the biological processes. The p-value cut-off was set at 0.01 for evaluating over-representation of biological processes (i.e. enriched GO terms). By using the top 10 most significant enriched GO terms, we inferred the biological processes of each MCA pattern-pair. The inferred biological processes were associated with cancer, such as angiogenesis and blood vessel morphogenesis. Therefore MCA suggests that there is a strong association between the inferred biological processes and the proteins with high absolute protein pattern values.

Next, we investigated whether both gene and protein patterns from MCA gave congruent signals. We made the reasonable assumption that the top 10 proteins with the largest absolute protein pattern values were likely to be involved in the biological processes of the pattern-pair, while the bottom 10 were not. Thus, the GO terms from top 10 proteins were more likely to match the 100 most significant GO terms obtained from the GO analysis of the genes compared to the bottom 10. A GO term of a gene matched a GO term of a protein when either their GO terms, or their GO terms' parents, or their GO terms' children overlapped. The p-values of the 100 most significant GO terms were ranked in descending order (i.e. the largest p-value had the lowest rank, while the smallest p-value had the highest rank). We computed the mean ranking, M, for each protein's GO term. The median of M for the top 10 proteins was significantly higher than the bottom 10 (p-value=0.005 using Wilcoxon test). Therefore the gene and protein pattern-pairs from

MCA were extracting similar biological signals.

Using gSVD

For gSVD, we determined the interesting pattern-pairs by considering their angular distances. All of the 59 angular distances were positive and ranged from 0.485 to 0.778. The generalized variance explained by the microarray data was quite uniform, while the generalized variance explained by the proteomic data was high when the angular distance was low. In view of the generalized variance explained, we further investigated the pattern-pairs with the lowest three angular distances (0.485, 0.548 and 0.556).

Similar to the MCA, we defined genes from the top 5% absolute gene pattern values as interesting and performed a GO analysis on the biological processes. The inferred biological processes from the enriched GO terms were also associated with cancer. We analyzed the concordance between the gene and protein patterns for gSVD by applying the same approach used in MCA.

The median of M for the top 10 proteins was *significantly lower* than the bottom 10 (p=0.016 using the Wilcoxon test). This indicated that gSVD gene and protein pairs were not internally congruent, with each referring to different processes.

Comparing MCA and gSVD

To compare the two methods, we tried to match the MCA and gSVD results as much as possible, by identifying pattern-pairs from gSVD that had the highest absolute correlation with the first three pattern-pairs from MCA.

Similar to the previous sub-sections, we defined a set of interesting genes from the absolute gene pattern values and performed a GO analysis on the biological processes. The inferred biological processes were associated with cancer. However, the median of M from the top 10 proteins was not significantly different from the bottom 10 (p=0.325). Again, this indicated that these gSVD gene and protein pairs were not internally congruent.

Using a similarity measure between highly significant GO terms from genes and GO terms from proteins, which were grouped into their top and bottom 10 absolute protein pattern values, we observed that all the three MCA pattern-pairs had a higher similarity value for their top 10 proteins than their bottom 10. For gSVD there was one pattern-pair where the bottom

10 proteins had a higher similarity value than the top 10. Therefore all the pattern-pairs from MCA were having similar biological signals in the genes and proteins, while gSVD had a pattern-pair with dissimilar biological signals.

Chapter 6

Discussion

6.1 ARS (Paper I-III)

Our approach, called ARS, contains a signal detection step, followed by a peak quantification step. The signal detection step effectively reduces the spectra of intensities to a spectrum of F^*, before zooming in on regions that contain potential biomarkers for peak quantification. This reduces the number of peaks to be inspected visually in parallel with multiple spectra for differences in intensities. If a peak has the same intensity across all spectra, it will not be identified as significant. Hence, RS functions as a filter for common but uninformative proteins.

The advantage of investigating the null distribution of the F-statistic through blanks is the ability to use an objective selection criterion, such as FDR, which accounts for multiple testing. Existing methods use arbitrary criteria, such as signal-to-noise ratio, which gives only a vague notion of the level of false positive rate or false discovery rate. At 80% sensitivity, the FDRs of the four methods are around 25% to 50%, compared to around 8% for RS. This observation could be explained by the fact that RS analyzes the spectra simultaneously, which is likely to improve the characterization of noise compared to the other four methods, which detect peaks for each spectrum individually.

For the peak quantification step, the appeal of ARS lies in using the data to obtain peak templates instead of specifying potentially unrealistic parametric templates. In addition, we refined the estimation of the amplitude by using a mixture model that mimics an elongated cloud of ionized molecules from

the same protein hitting the detector of the mass spectrometry.

We have also demonstrated that ARS can detect more peaks than the standard method. Attempting to reduce false positives by adjusting the settings of the standard method reduces its sensitivity substantially. Furthermore, we have shown that improvements in peak annotation in ARS can potentially benefit downstream data analysis in biomarker research.

6.2 Extended ARS (Paper V)

We validated the use of ARS on the log-transformed intensities of MALDI. This suggests that the log-transformation reduces variation in the intensities which reduces false positives. In the validation process, we noticed that the agreement between the nominal p-value and the empirical false positive rate was the closest for log-blanks, followed by log-BSA and log-Spike-in(8-11). It is suspected that variation in trypsin digestion and competitive ionization are responsible for the larger discrepancy between the nominal p-value and the empirical false positive rate.

The monotonic increasing relationship between the estimated peak intensity and the quantity spiked is linear for Spike-in A, B, C and E, but curvilinear for Spike-in D. This demonstrates the ability of MALDI in detecting the quantity of protein in the sample. At the same time, the results suggest that proteins may ionize differently from each other.

Correlation between the intensities of peaks identified by ARS and the quantity spiked is generally high (> 0.8) in the neighborhood of the isotopic region of the spike-ins. This suggests that protein biomarkers consist of biomarker peaks. Detection of potential protein biomarkers in the extended ARS requires a sustained series of potential biomarker peaks that are more or less consecutively 1 Da apart. This is a good feature, as the peaks corroborate among themselves the presence of the potential protein biomarker.

While extended ARS and PeakExplorer perform well in detecting the spike-ins, extended ARS is more specific than PeakExplorer. Extended ARS quantifies the intensities better and has higher sensitivity at low FDR than PeakExplorer, although ARS generally has a wider monoisotopic peak m/z range. The generally better performance of ARS may be a consequence of the corroborative feature in ARS when detecting a potential biomarker.

6.3 MCA (Paper IV)

The SVD has been used to study dominant patterns of variation in a single phenotype such as gene expression (Alter et al., 2003). Here we apply SVD on the cross-covariance matrix to study dominant patterns of correlation between two phenotypes.

Our simulation study indicates that MCA gives consistent estimates of the pattern-pairs, while the gSVD does not. For MCA, analysis was done through the cross-covariance matrix, while for gSVD, it was done through the appended gene and protein expression matrices. Although gSVD does capture some portion of the correlation, it is not designed to capture it completely. From the NCI data analysis, we demonstrated that the gene and protein pattern-pairs found by MCA were biologically congruent, but not those found by gSVD. This suggests that MCA could be used to gain biological insight into the interplay between genes and proteins.

From our gSVD results on the NCI datasets, we considered the three pattern-pairs with the lowest angular distance. Their total generalized variances are 4.8% and 29.6% for microarray and proteomic datasets respectively. However, for the highest three angular distances, the total generalized variance for the proteomic dataset drops dramatically (0.06%). This suggests that the selection of pattern pairs from gSVD requires both angular distance and total generalized variance.

However, the angular distance carrying information on the significance of metagenes in one dataset over the other may not be applicable in our context. Independence between the two datasets is required for a meaningful interpretation of the angular distances (Alter et al., 2003). Therefore, the use of angular distance may not be appropriate in our situation, since they are from the same samples. Furthermore, the angular distance profile changes with increasing sample size in the simulation study.

6.4 Future research

ARS can be extended to other technologies that require peak detection to be made before further downstream analysis can be performed. We have already demonstrated the feasibility of extending ARS to MALDI in this thesis.

We could extend ARS to other MS techniques, such as Gas Chromatography Mass Spectrometry (GC-MS) which consists of two major components: the gas chromatography and the mass spectrometry (Dunn et al., 2005). The gas chromatography separates out molecules according to their retention time, which is the time taken to travel through the column. At the end of the column, the molecule proceeds to the mass spectrometry. Therefore, apart from the m/z dimension from the mass spectrometry component, GC-MS also has a retention time dimension. The output can be visualized as a biaxial plane, similar to the 2DGE in Figure 2.2, where the axes correspond to retention time and m/z, and level of the intensity from the mass spectrometry corresponds to the color intensity of the spot. When extending ARS, the retention time dimension needs to be addressed.

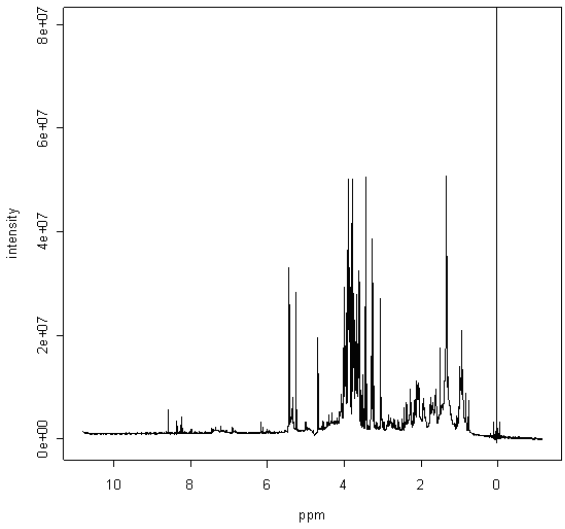

Figure 6.1: An example of an NMR data. The x-axis is the ppm and the y-axis is the intensity.

We could extend ARS to non-MS technologies, such as, Nuclear Magnetic Resonance (NMR) spectroscopy, which requires little sample preparation and is non-destructive (Dunn et al., 2005). It is used in the field of metabolomics to profile metabolites from tissues or biological fluids in a high throughput

fashion. By using the fact that nuclei absorb electromagnetic radiation in a strong magnetic field, NMR obtains information on the structure and concentration of the metabolites. The metabolite profile of the sample generated by NMR can be represented graphically where the horizontal and vertical axes are the parts per million (ppm) and the intensity respectively; see Figure 6.1. The peak intensity is proportional to the total number of nuclei, indicating the concentration of the metabolites in the sample, while the peak position indicates the molecular group and molecular environment of the metabolites. The metabolite profile could potentially be used in biomarker discovery.

Systems biology 'aims at a system-level understanding of genetic or metabolic pathways by investigating interrelationships (organisation or structure) and interactions (dynamics or behavior) of genes, proteins and metabolites' (Wolkenhauer, 2001). The integration of datasets from various biological levels, such as DNA, mRNA, proteins and metabolites, is one aspect of it. In our thesis, we have illustrated how MCA could be used to integrate two such datasets - mRNA and proteins - to gain understanding of the interplay between mRNA and proteins. To integrate more than two datasets, we could consider formulating the MCA under the duality diagram theory, a unifying mathematical tool which includes PCA or correspondence analysis (Dray et al., 2003; Dray and Dufour, 2007).

Briefly, the duality diagram is based on the statistical triplet, which is composed of three matrices: the data matrix, $\mathbf{X}'_{p\times n}$, and two positive symmetric matrices $\mathbf{Q}_{p\times p}$ and $\mathbf{D}_{n\times n}$. $\mathbf{Q}$ is a metric used as an inner product in $\mathbb{R}^p$ to measure the distances between n individuals, while $\mathbf{D}$ is a metric used as an inner product in $\mathbb{R}^n$ to measure the relationships between p variables. Different definitions of $\mathbf{X}'$, $\mathbf{Q}$ and $\mathbf{D}$ correspond to different multivariate methods. We can obtain Canonical Correlation Analysis (CCA) from Co-inertia Analysis (CIA), which uses the duality diagram theory to define two statistical triplets from two datasets and co-inertia criterion for measuring the adequacy between the two datasets. An R package, *ade4*, runs the multivariate methods under the duality diagram theory for any number of datasets (Dray and Dufour, 2007).

Chapter 7

Conclusions

We have developed an improved method that performs peak detection and quantification in SELDI for biomarker discovery studies (Paper I and II), and an accompanying R package, called *ProSpect*, which has a graphical user interface version, called *ProSpectGUI* (Paper III):

- RS uses an objective selection criterion for peak detection. RS has better OC than existing methods. At 80% sensitivity, the FDRs of comparable methods are around 25% to 50%, compared to around 8% for RS.
- ARS captures several peak regions in the spectral data that are missed by the standard method. It is more robust than the standard method, as two or more neighboring peaks are not mistaken as a single peak. It is also able to detect peaks in the presence of m/z-misalignment.
- ARS is accessible through R packages *ProSpect* and *ProSpectGUI*.

We extended ARS to MALDI data (Paper V):

- Extended ARS is generally better than the standard method in quantifying the intensities of proteins.
- Extended ARS has higher specificity than the standard method. At low FDR, extended ARS has higher sensitivity than the standard method.

We are able to integrate transcriptomic and proteomic data using MCA (Paper IV):

- By circumventing the step of matching genes and proteins, MCA exploits all information in the analysis. The estimates of the gene and protein pattern-pairs from MCA are consistent and biologically congruent.
- MCA allows proteins to correlate with genes throughout the genome, reflecting the biological phenomenon of proteins and genes being interconnected in various pathways. This increases the chances of uncovering novel biological relationships between genes and proteins.

Acknowledgements

During the four years of my PhD studies, I spent half of it in Sweden and the other half in Singapore. I am grateful to the many people who in different ways have contributed to this work. Specifically, I would like to thank:

Yudi Pawitan, Kee Seng Chia and Alexander Ploner, my supervisors, for their patient guidance and encouragement throughout my PhD studies. Their passion for scientific research has truly been inspirational and I can only hope that one day, through hard work and perseverance, I can attain to a fraction of the breadth and depth of the knowledge they possess. It is indeed a privilege to be their doctoral student.

Janne Lehtö, Jenny Forshed, Andreas Quandt, Maria Pernemalm, Rolf Lewensohn, my wonderful co-authors, for providing a fruitful research collaboration. Thanks for the many insightful and rich discussions on proteomics.

Stefano Calza, my mentor, for his guidance and advice. He has made my PhD experience a wonderful one. Tuttavia, è stata un'esperienza dolorosa allo stesso tempo proprio per colpa tua... i tuoi scherzi sono sempre così divertenti che ogni volta non posso fare a meno di ridere fino a quando non mi fa male lo stomaco.

All the friends at MEB, for creating an open and friendly working environment. The Biostat group, that made my time at KI all the more memorable, with activities, such as Thursday *Fika* and Movie Night. A special thanks to Therese Andersson, Rino Bellocco, Paul Dickman, Sandra Eloranta, Keith Humphreys, Marie Jansson, Anna Johansson, Paul Lambert, Cecilia Lundholm, Barbara Mascialino, Juni Palmgren, Marie Reilly, Samuli Ripatti, Sven Sandin, Davide Valentini, Fredrik Wiklund and Li Yin. The wonderful IT group, especially for the help rendered when my laptop died on me just two months prior to my thesis defense. I am grateful to them for providing the desktop, which was used to write this very thesis. Kamila Czene (Director of Postgraduate Studies), the education administrators, Camilla Ahlqvist and Marie Dokken, and Marie Jansson and Monica Rundgren for their help in my thesis defense application.

All present and former doctoral students in MEB, for the company and encouragement in the journey of learning. It would have been most lonesome and dull if not for them. A special thanks to Hatef Darabi, Annica Dominicus, Ulrika Eriksson, Fang Fang, Elinor Fondell, Arief Gusnanto, Gudrun Jonasdottir, Junmei Jonasson Miao, Kenji Kato, Monica Leu, Juhua Luo, Dariush Nesheli, Arvid Sjölander, Ben Yip and Zongli Zheng.

All the friends in Singapore at CME and COFM, for introducing me to biostatistical and epidemiological research, and giving me ample opportunities to grow as a researcher. Thanks to Sin Eng Chia, Wei Gao, David Koh, Jeannette Lee, Daniel Ng, Choon Nam Ong, Agus Salim, Seang Mei Saw, Bee Choo Tai, E-Shyong Tai and Yik Ying Teo for being such excellent and inspiring seniors. Thanks to Kwok Hang Cheung, Kar-Wai Tan and Sharon Wee for the comradeship. Thanks to all the research fellows and assistants for the company and encouragement in our journey of learning, especially Gek Hsiang Lim and Xueling Sim, my fellow colleagues in biostatistics who graciously helped to proof read my thesis. I am also grateful to the non-research staff for their support, especially Muhammad Hazrin Bin Abdul Rahi, Saadiah Binte Awek, Doris Chen, Po Jan Chen, Moira Khaw, Sock Fan Koh, Teck Ngee Lee, Eng Jee Lim, Poh Choo Lim, Ai-Leen Ng and Gim Choo Soh.

My family members, who have supported and cheered me on all the way. Their unconditional love has always been a source of strength and comfort.

Last but not least, it is only fitting that I thank God, for, 'every good thing bestowed and every perfect gift is from above' (James 1:17).

References

Ahmed, N., Barker, G., Oliva, K., Hoffmann, P., Riley, C., Reeve, S., Smith, A., Kemp, B., Quinn, M., and Rice, G. (2004). Proteomic-based identification of haptoglobin-1 precursor as a novel circulating biomarker of ovarian cancer. *Br J Cancer*, 91:129–140.

Alaiya, A., Al-Mohanna, M., and Linder, S. (2005). Clinical cancer proteomics: promises and pitfalls. *J Proteome Res*, 4:1213–1222.

Alter, O., Brown, P. O., and Botstein, D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A*, 100:3351–3356.

Alterovitz, G., Patek, D., Kohane, I. S., and M., R. (2006). *Encyclopedia of Biomedical Engineering*, chapter Proteomics. John Wiley & Sons.

Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R. S., Conrads, T. P., Veenstra, T. D., Adkins, J. N., Pounds, J. G., Fagan, R., and Lobley, A. (2004). The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol Cell Proteomics*, 3:311–326.

Anderson, T. (1984). *An introduction to multivariate statistical analysis*. Wiley, second edition.

Applied Biosystems (2008). *Data Explorer Version ®4.6 Software Online help*. Applied Biosystems.

Bairoch, A. and Apweiler, R. (2000). The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res*, 28:45–48.

Berger, J. A., Hautaniemi, S., Mitra, S. K., and Astola, J. (2006). Jointly analyzing gene expression and copy number data in breast cancer using

data reduction models. *IEEE/ACM Trans Comput Biol Bioinform*, 3:2–16.

Breen, E. J., Hopwood, F. G., Williams, K. L., and Wilkins, M. R. (2000). Automatic poisson peak harvesting for high throughput protein identification. *Electrophoresis*, 21:2243–2251.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, 282:2012–2018.

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

Choe, S., Boutros, M., Michelson, A., Church, G., and Halfon, M. (2005). Preferred analysis methods for Affymetrix genechips revealed by a wholly defined control dataset. *Genome Biol*, 6:R16.

Coombes, K., Fritsche, H. J., Clarke, C., Chen, J.-N., Baggerly, K., Morris, J., Xiao, L.-C., Hung, M.-C., and Kuerer, H. (2003). Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem*, 4:1615–1623.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M.-C., and Kuerer, H. M. (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*, 5:4107–4117.

Cooper, G. and Hausman, R. (2004). *The Cell: A Molecular Approach*. Sinauer Associates.

Cox, B., Kislinger, T., and Emili, A. (2005). Integrating gene and protein expression data: pattern analysis and profile mining. *Methods*, 35:303–314.

Dalgaard, P. (2001). The r-tcl/tk interface. In Hornik, K. and Leisch, F., editors, *DSC 2001 Proceedings of the 2nd International Workshop on Distributed Statistical Computing*.

Dhamoon, A. S., Kohn, E. C., and Azad, N. S. (2007). The ongoing evolution of proteomics in malignancy. *Drug Discov Today*, 12:700–708.

Dijkstra, M., Roelofsen, H., Vonk, R. J., and Jansen, R. C. (2006). Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics*, 6:5106–5116.

Dray, S., Chessel, D., and Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology*, 84:3078–3089.

Dray, S. and Dufour, A. (2007). The ade4 package: Implementing the duality diagram for ecologists. *J Stat Softw*, 22:Issue 4.

Dunn, W. B., Bailey, N. J. C., and Johnson, H. E. (2005). Measuring the metabolome: current analytical technologies. *Analyst*, 130:606–625.

Engwegen, J. Y. M. N., Gast, M.-C. W., Schellens, J. H. M., and Beijnen, J. H. (2006). Clinical proteomics: searching for better tumour markers with seldi-tof mass spectrometry. *Trends Pharmacol Sci*, 27:251–259.

Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., Radich, J., Anderson, G., and Hartwell, L. (2003). The case for early detection. *Nat Rev Cancer*, 3:243–252.

Fung, E. T. and Enderwick, C. (2002). Proteinchip clinical proteomics: computational challenges and solutions. *Biotechniques*, Suppl:34–8, 40–1.

Hanash, S. M., Pitteri, S. J., and Faca, V. M. (2008). Mining the plasma proteome for cancer biomarkers. *Nature*, 452:571–579.

Hegde, P. S., White, I. R., and Debouck, C. (2003). Interplay of transcriptomics and proteomics. *Curr Opin Biotechnol*, 14:647–651.

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.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945.

Jarman, K., Daly, D., Anderson, K., and Wahl, K. (2003). A new approach to automated peak detection. *Chemom Intell Lab Syst*, 69:61–76.

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 Hillenkamp, F. (1988). Laser desorption ionisation of proteins with molecular masses exceeding 10.000 daltons. *Anal Chem*, 60:2299–2301.

Kempka, M., Sjdahl, J., Bjrk, A., and Roeraade, J. (2004). Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom*, 18:1208–1212.

Kiehntopf, M., Siegmund, R., and Deufel, T. (2007). Use of seldi-tof mass spectrometry for identification of new biomarkers: potential and limitations. *Clin Chem Lab Med*, 45:1435–1449.

Lachenbruch, P. (1998). *Encyclopedia of Biostatistics*, chapter McNemar Test, 2486–2487. Wiley.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21:1764–1775.

Nie, L., Wu, G., Brockman, F. J., and Zhang, W. (2006). Integrated analysis of transcriptomic and proteomic data of desulfovibrio vulgaris: zero-inflated poisson regression models to predict abundance of undetected proteins. *Bioinformatics*, 22:1641–1647.

Nie, L., Wu, G., Culley, D. E., Scholten, J. C. M., and Zhang, W. (2007). Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit Rev Biotechnol*, 27:63–75.

Paige, C. and Saunders, M. (1981). Towards a generalized singular value decomposition. *SIAM J Numer Anal*, 18:398–405.

Poon, T. C. W. (2007). Opportunities and limitations of seldi-tof-ms in biomedical research: practical advices. *Expert Rev Proteomics*, 4:51–65.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Salim, A. and Pawitan, Y. (2007). Model-based maximum covariance analysis for irregularly observed climatological data. *J Agric Biol Environ Stat*, 12:1–24.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2:110–114.

Scariano, S. and Davenport, J. (1987). The effects of violations of independence assumptions in the one-way anova. *Am Stat*, 41:123–129.

Schulte, I., Tammen, H., Selle, H., and Schulz-Knappe, P. (2005). Peptides in body fluids and tissues as markers of disease. *Expert Rev Mol Diagn*, 5:145–157.

Senko, M., Beu, S., and F.W., M. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J Am Soc Mass Spectrom*, 6:229–233.

Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Scherf, U., Kahn, A., Dolginow, D., Cossman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K., and Weinstein, J. N. (2007). Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study. *Mol Cancer Ther*, 6:820–832.

Srinivas, P. R., Kramer, B. S., and Srivastava, S. (2001). Trends in biomarker research for cancer detection. *Lancet Oncol*, 2:698–704.

Srinivas, P. R., Verma, M., Zhao, Y., and Srivastava, S. (2002). Proteomics for cancer biomarker discovery. *Clin Chem*, 48:1160–1169.

Stoyanova, R., Kuesel, A., and Brown, T. (1995). Application of principal-component analysis for nmr spectral quantificaton. *J Magn Reson A*, 115:265–269.

Villar-Garea, A., Griese, M., and Imhof, A. (2007). Biomarker discovery from body fluids using mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*, 849:105–114.

Wang, Y., Zhou, X., Wang, H., Li, K., Yao, L., and Wong, S. T. C. (2008). Reversible jump mcmc approach for peak identification for stroke seldi mass spectrometry using mixture model. *Bioinformatics*, 24:i407–i413.

Waters, K. M., Pounds, J. G., and Thrall, B. D. (2006). Data merging for integrated microarray and proteomic analysis. *Brief Funct Genomic Proteomic*, 5:261–272.

Wehofsky, M., Hoffman, R., Hubert, M., and Spengler, B. (2001). Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substance-class specific analysis of complex samples. *Eur J Mass Spectrom*, 7:39–46.

Weston, A. D. and Hood, L. (2004). Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res*, 3:179–196.

Wolkenhauer, O. (2001). Systems biology: the reincarnation of systems theory applied in biology? *Brief Bioinform*, 2:258–270.

Yasui, Y., Pepe, M., Thompson, M., Adam, B., Wright, G. J., Qu, Y., Potter, J., Winget, M., Thornquist, M., and Feng, Z. (2003). A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4:449–463.

Yu, W., He, Z., Liu, J., and Zhao, H. (2008). Improving mass spectrometry peak detection using multiple peak alignment results. *J Proteome Res*, 7:123–129.

Zhang, Z. and Chan, D. W. (2005). Cancer proteomics: in pursuit of "true" biomarker discovery. *Cancer Epidemiol Biomarkers Prev*, 14:2283–2286.