MALDI-ToF Mass Spectrometry Biomarker Profiling via Multivariate Data Analysis
Application in the Biopharmaceutical Bioprocessing Industry

by

Remi A. Momo

October 2012

A Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

School of Chemical Engineering and Advanced Materials
Newcastle University
United Kingdom

Abstract

Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-ToF MS) is a technique by which protein profiles can be rapidly produced from biological samples. Proteomic profiling and biomarker identification using MALDI-ToF MS have been utilised widely in microbiology for bacteria identification and in clinical proteomics for disease-related biomarker discovery. To date, the benefits of MALDI-ToF MS have not been realised in the area of mammalian cell culture during bioprocessing.

This thesis explores the approach of 'intact-cell' MALDI-ToF MS (ICM-MS) combined with projection to latent structures – discriminant analysis (PLS-DA), to discriminate between mammalian cell lines during bioprocessing. Specifically, the industrial collaborator, Lonza Biologics is interested in adopting this approach to discriminate between IgG monoclonal antibody producing Chinese hamster ovaries (CHO) cell lines based on their productivities and identify protein biomarkers which are associated with the cell line productivities. After classifying cell lines into two categories (high/low producers; Hs/Ls), it is hypothesised that Hs and Ls CHO cells exhibit different metabolic profiles and hence differences in phenotypic expression patterns will be observed. The protein expression patterns correlate to the productivities of the cell lines, and introduce between-class variability. The chemometric method of PLS-DA can use this variability to classify the cell lines as Hs or Ls.

A number of differentially expressed proteins were matched and identified as biomarkers after a SwissProt/TrEMBL protein database search. The identified proteins revealed that proteins involved in biological processes such as protein biosynthesis, protein folding, glycolysis and cytoskeleton architecture were upregulated in Hs. This study demonstrates that ICM-MS combined with PLS-DA and a protein database search can be a rapid and valuable tool for biomarker discovery in the bioprocessing industry. It may help in providing clues to potential cell genetic engineering targets as well as a tool in process development in the bioprocessing industry. With the completion of the sequencing of the CHO genome, this study provides a foundation for rapid biomarker profiling of CHO cell lines in culture during recombinant protein manufacturing.

Acknowledgements

For the LORD gives wisdom,
and from his mouth come knowledge and understanding — *Proverbs 2:6*

Fear of the LORD is the foundation of wisdom. Knowledge of the Holy One results in good judgment — *Proverbs 9:10*

This thesis was carried out in the School of Chemical Engineering and Advanced Materials. I wish to thank everybody in this school for their moral support throughout my PhD programme.

Special thanks goes to my supervisors, Professor Elaine Martin and Professor Gary Montague for their inspiring guidance, technical support, prompt reviewing and constructive comments during the course of these studies and write-up. Further, I wish to enormously thank Prof. Mark Smales and Dr. Jane Povey of the University of Kent for the mass spectrometry data as well as their help in providing the reagents and equipments for carrying out my *E. coli* MALDI experiments.

I also wish to thank the members of the SysMA group in the School of Chemical Engineering and Advanced Materials: Richard, Chris, Bothinah, Nor, Aruna, Catherine, Ronan O'kennedy and the EngD students of BBTC, for their encouragement and help in making my stay in Newcastle enjoyable. Not forgetting the Newcastle gang of Rob, Ming and Simon for the wonderful moments we usually spend together in the pub.

I am sincerely grateful to the SAGE graduate school and Lonza Biologics for their financial support for this research.

I would also like to thank my family back home in Cameroon, my mother, elder brother, elder sister and all other extended family members as well as friends for their endless drive and support that has kept me going.

Contents

List of Abbreviations

2-D PAGE	Two dimensional polyacrylamide gel electrophoresis
96-DWP	96 deep well plate
AIChE	American Institute for Chemical Engineers
AUC	Area under the curve
BHK-12	Baby hamster kidney
BP- ANN	Back-propagation artificial neural networks
CCA	Canonical correlation analysis
CD-18	Cluster of differentiation-18
cDNA	Recombinant DNA
CDR	Complementarity determining region
CHCA	α-Cyno-4-hydroxycinnamic acid
CHO	Chinese hamster ovary
CP-ANN	Counter-propagation artificial neural networks
CSP-C	Cold chock protein-C
CWT	Continuous wavelet transform
DHB	2,5-dihydroxybenzoic acid
DHFR	Dihydrofolate reductase enzyme
DIGE	2-D gel electrophoresis
DNA	Deoxyribonucleic acid
DR	Dimensionality reduction
DWT	Discrete wavelet transforms
ELISA	Enzyme linked immunosorbent assay
EM	Expectation-maximisation
ERp72	Endoplasmic reticulum based protein-72
ESI MS	Electrospray ionisation mass spectrometry
EVD	Eigenvalue decomposition
FDA	Food and drug administration
FIR	Finite impulse response
FS	Feature selection
FT	Feature transformation
FT-IR	Fourier transform infrared
GRP78/BiP	Glucose regulated protein-78
GRP94/BiP	Glucose regulated protein-94
GS	Glutamine synthetase
H	High producer
HC	Heavy chain
hCMV-MIE	Human cytomegalovirus major immediate early
HDSS	High dimensionality and small size
HEK-293	Human embryonic kidney
hGH	Human growth hormone
HPLC	High performance liquid chromatography
ICM	Intact cell mass spectrometry
ICM-MS	'Intact-cell' MALDI-TOF MS

Ig	Immunoglobulin
IgG	Immunoglobulin G
IVC	Integral viable cell concentration
KNN	K-nearest neighbours
L	Low producer
LC	Light chain
LC-ESI MS/MS	Liquid chromatography-electrospray ionisation and tandem mass spectrometry
LDA	Linear discriminant analysis
LDI	Laser desorption/ionisation
LOOCV	Leave-one-out cross-validation
LV	Latent variables
m/z	Mass-to-charge
MAbs	Monoclonal antibodies
MALDI FT MS	MALDI Fourier transform mass spectrometry
MALDI-ToF MS or MALDI	Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry
MCC	Mammalian cell culture
MGCM	Multidimensional gauss class modeling
MS	Mass spectrometry
MSX	Methionine sulphoximine
MTX	Methotrexate
MW	Molecular weight
NIPALS	Non-iterative partial least squares
NS0, SP2/0	Mouse myeloma cell lines
PBS	Phosphate buffer saline
PCA	Principal component analysis
PCs	Principal components
PER-C6	Human-retina-derived cell line
PLS	Partial least squares
PLS-DA	PLS-discriminant analysis
PSD	Post source decay
PTMs	Post-translational modifications
QDA	Quadratic discriminant analysis
*q*P	Specific productivity
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross-validation.
SA	Sinapinic acid
SD	Standard deviation
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SEAP	Secreted alkaline phosphatase
SELDI-ToF MS	Surface enhanced laser desorption/ionisation time-of-flight mass spectrometry
SIMCA	Soft independent modelling of class analogies

SVD	Singular value decomposition
TSB	Trypticase soy broth
UNEQ	UNEQual dispersed classes
UPLC	Ultra performance liquid chromatography
WCM-MS	'Whole-cell' MALDI-TOF MS

List of Notations

E	Energy of charged ion
v	Velocity of charged ion
t, t_o	Time of flight (ToF) of charged ion/ Initial ToF
m	Mass of charged ion
z	Charge of ion
d	Length of accelerating region in the electric field
L	Length of non-accelerating region
V_o	Potential of the electric field.
C_o	Internal delay (MALDI instrument)
$C_1({}^{m}/_{z})^{1/2}$	ToF of an ion with zero initial velocity
$C_1({}^{m}/_{z})^{3/2}$	Flight time correction for an ion velocity
$f(i,j)$	Observed value for *i*th sample at *j*th *m*/*z* ratio
$b(i,j)$	Baseline value for *i*th sample at *j*th *m*/*z* ratio
$s(i,j)$	True signal for *i*th sample at *j*th *m*/*z* ratio
$\varepsilon(i,j)$	Noise for *i*th sample at *j*th *m*/*z* ratio
X_i	Signal at *i*th *m/z* ratio
n	Number of *m/z* ratio values
$\boldsymbol{X}, \boldsymbol{x}$	Matrix / Vector input (independent variables)
$\boldsymbol{Y}, \boldsymbol{y}$	Matrix / Vector input (dependent variables)
$\boldsymbol{P}, \boldsymbol{p}$	Loadings Matrix / Vector (independent variables)
M, m, M_o, m_o	Number of Samples
N, n	Number of Variables
$\bar{y}$	Sample Mean (dependent variable)
$\hat{y}$	Predicted Value (dependent variable)
$\hat{\boldsymbol{b}}$	Regression Coefficient Vector
k, A	Number of Components to retain (PCA and PLS-DA)
σ	Standard Deviation
λ_c	Classification Threshold (PLS-DA)
$\boldsymbol{T}, \boldsymbol{t}$	Scores Matrix / Vector (independent variables)
λ_i	Eigenvalue
$\boldsymbol{E}, \boldsymbol{f}$	Residual Matrix / Vector (independent and dependent variables
$\boldsymbol{Q}, \boldsymbol{q}$	Scores Matrix / Vector (dependent variables)
SEP_i	Standard Error of Prediction
ω_1, ω_o	Class of Interest / Otherwise (PLS-DA)

Publications

Journal Papers

1) Momo, R.A., Martin, E.B., Montague, G.A., Smales, C.M., Povey, J.F., O'Malley, C.J. Rapid Biomarker Profiling Of *Escherichia coli* Utilising MALDI-ToF Mass Spectrometry and a Chemometric Approach; submitted to *Analytica Chimica Acta.* (To be submitted for publication)

2) Momo, R.A., Martin, E.B., Montague, G.A., Smales, C.M., Povey, J.F., O'Malley, C.J. Biomarker Profiling Of CHO Cell Lines based on productivity Utilising MALDI-ToF Mass Spectrometry and a Chemometric Approach; submitted to *Analytica Chimica Acta*. (To be submitted for publication)

Index of Figures

Index of Tables

Chapter 1

1. Introduction of Thesis

Mammalian cell lines have the potential to synthesize, perform complex folding and post-translational modifications (e.g. glycosylation) necessary for *in vivo* biological activity; and secrete complex proteins in large-scale suspension cell culture. These characteristics give them a significant advantage over their prokaryotic counterparts such as bacteria cells (Andersen and Krummen, 2002). The increasing demand for biotherapeutic products such as monoclonal antibodies necessitates improved large-scale production of these complex heterologous proteins from mammalian cell lines. Consequently, the biopharmaceutical industry has endeavoured to enhance product titres and/or yield by investing in the engineering of cell culture processes resulting in high-producing cell cultures. Examples of mammalian cell culture cell lines which have been extensively used in the biotechnology industry for the production of recombinant proteins for biotherapeutic applications include Chinese hamster ovary (CHO), baby hamster kidney (BHK), and mouse hybridoma (NS0). Biotherapeutic protein producing cell lines generated from the same parent cell line usually display a wide range of growth, productivity, and metabolic characteristics (Browne and Al-Rubeai, 2007). This behavior is advantageous with respect to mammalian cell engineering. Producing cell lines displaying favourable characteristics have resulted in the potential to identify gene targets for genetic engineering that will improve product yield (Pascoe *et al.,* 2007).

A number of genetic engineering approaches have been applied with the aim of improving product yield. Examples include delaying programmed cell death (apoptosis), enhancing cellular metabolism and protein processing, and manipulating the cell cycle (Dietmair *et al.*, 2011). Engineering mammalian cells to reduce high concentrations of metabolic by-products such as lactate and NH_4 is an example of changing a cell’s metabolic behaviour. Hybridoma cells transfected with glutamine synthetase (GS – enzyme that converts glutamate to glutamine) were able to grow in a medium void of glutamine (NH_4 is the by-product of nutrient consumption – glutamine). This eliminated NH_4 as a by-product, otherwise whose presence will inhibit cell growth and viability, reducing maximum product yield. However this product titre gain came at a price as the stability of the transformed cells were compromised (Paredes *et al.*, 1999). Engineering approaches that target the cellular machinery responsible for protein processing can be directed towards improving protein folding. For example, overexpression of an isoform of the foldase protein disulfide isomerase (PDI) (ERp57), and other proteins that aid protein folding (calnexin and calreticulin), increased the specific productivity (*q*P) of thrombopoietin -producing CHO cells by approximately two-fold (Hwang *et al.,* 2003; Chung *et al.,* 2004). In another study, overexpression of PDI did not change the productivity of the CHO cells (Mohan *et al.,* 2007). These contrasting results indicate that the outcomes of these strategies may be cell line dependent (Dietmair *et al.,* 2012). Consequently, steps towards enhancing cellular productivity can potentially be achieved through greater understanding of cellular protein biology and how this is affected by cellular engineering.

With their track record in industry, high productivity and safety, CHO cells have gradually risen to prominence and become the most widely used platform for the production of biotherapeutics (Chu and Robinson, 2001). Despite their importance in biotechnology, limited information is available with respect to changes throughout the culture of CHO cells, due to insufficient genomic information. This is not the case with other organisms where genomic approaches (such as proteome and microarray analyses) have been applied. The sequencing of the *E. coli* (Blattner *et al.,* 1997), mouse (Waterston *et al.,* 2002), rat (Gibbs *et al.,* 2004) and human (Lander *et al.,* 2001; Venter *et al.,* 2001) genomes has provided the potential for the application of genomic tools in the study of disease, metabolism, growth, apoptosis investigation in these organisms. The lack of genomic information has materialised into proteomic profiling of CHO cells to understand their biology. This has been carried out using two-dimensional Polyacrylamide Gel Electrophoresis (2D-PAGE) (López, 2007). Proteomic profiling studies using 2D-PAGE have previously been applied to mammalian cells under various conditions including CHO cell lines undergoing temperature shifts (Kaufmann *et al.,* 1999), modified cell lines (Krawitz *et al.,* 2006), cells under hyperosmolality conditions (Lee *et al.,* 2003), cells with butyrate, zinc, and tunicamycin addition during growth (Van Dyk *et al.,* 2003), and NS0 cells displaying different productivity rates (Smales *et al.,* 2004). These studies have contributed to the understanding of the biology of mammalian cells under production, hence giving indications of potential genetic engineering targets that may be exploited to engineer improved cell lines. However, there are a number of drawbacks associated with 2D-PAGE. For example it is a lower throughput and time-consuming technique requiring about 3-4 days per run. The runs involve many steps and require a high level of laboratory skill to obtain good results.

Another drawback is the difficulties in separating high/low molecular weight proteins, low abundant proteins, hydrophobic proteins (e.g. membrane proteins). Membrane proteins are relatively insoluble in non-ionic detergents at low ionic strength and even when solubilised may precipitate at pH values close to their isoelectric points (Bunai and Yamane, 2005; Meleady, 2007).

Recent initiatives to sequence the CHO cell genome (Hammond *et al.*, 2011; Xu *et al.*, 2011), indicate that genomic approaches combined with bioinformatics could materialise in the proteomic profiling of CHO cells. The completion of the sequencing of the CHO genome has materialised into sequence-derived theoretical molecular weights (MWs) of CHO proteins to become increasingly available in compiled internet accessible protein databases (http://expasy.org/proteomics, 2012). This has provided the possibility of identifying potential protein biomarkers of CHO cells by matching the sequence-derived theoretical MWs of the database proteins with experimental MWs derived from high throughput proteomic technologies such as mass spectrometry.

‘Intact-cell’ Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-ToF MS or MALDI) has been exploited extensively in the field of microbiology for the investigation of bacterial species (Demirev *et al.*, 1999; Ryzhov and Fenselau, 2001; Warscheid *et al.*, 2004; Parisi *et al.*, 2008; Dieckmann *et al*., 2008; Ilina *et al.*, 2009; Christner *et al.* 2010; Hotta *et al*., 2011; Wang *et al*., 2012). ‘Intact cell’ or ‘whole cell’ MALDI-TOF MS (ICM-MS or WCM-MS) was developed for the rapid identification of bacteria through protein profiling. The term (whole or intact cell) indicates that the cells to be analysed are not treated or processed to specifically remove or isolate any of the cellular components. In “intact cell” analysis, the cells are only manipulated to transfer them into the mass spectrometer for analysis and no additional steps are included in the procedure to deliberately disrupt cellular membranes or separate/recover analytes from the cellular material (Wilkins and Lay, 2006).

Recently, the investigation of microorganisms by ICM-MS, primarily for bacterial identification, has been undertaken through two approaches. The first approach is by direct comparison of whole cell spectra, deemed to be a bacterial protein fingerprint, to reference spectra. This has been made possible by the creation of a bacterial fingerprint library of mass spectra from a range of known bacterial species (Mazzeo *et al.*, 2006).

An alternative approach for the identification of bacteria which does not involve the use of a fingerprinting library has been developed. This is a bioinformatics approach and is based on matching a set of protein biomarker MWs in the spectrum against those of sequence-derived theoretical protein MWs. The latter approach has seen wide applications in the field of microbiology (Demirev *et al.,* 1999; Ryzhov and Fenselau, 2001). It has been made possible by the availability of protein biomarkers of bacteria in compiled internet-accessible protein databases of microorganisms with completely sequenced genomes.

ICM-MS also raises many possibilities for the analysis of complex cellular systems such as mammalian cells. Biomarker profiles have been obtained from whole mammalian cells of neuronal origin (Li *et al.,* 2000; van Veelen *et al.,* 1993), additionally tissue sections have been profiled (Chaurand *et al.,*, 2006; Chaurand *et al.,* 2007; Crossman *et al.,* 2006; Khatib-Shahidi *et al.,* 2006; Reyzer *et al.,* 2007). Differentiation between human (K562 and GM15226) and rodent (BHK21) mammalian cell types through their protein profile 'fingerprints' (Zhang *et al.,* 2006), as well as between monocytes, T lymphocytes and polymorphonuclear leukocyte immune cells (Ouedraogo *et al.,* 2010) has also been carried out using ICM-MS. These applications of ICM-MS in bacterial and mammalian cell protein profiling, demonstrate an outcome of the approach that could be useful for mammalian cell culture (MCC) in bioprocessing for protein profiling to study their biology.

Despite the application of ICM-MS to bacteria and mammalian cells, there have been relatively few studies that have applied this approach to MCC in bioprocessing. ICM-MS has been used in the profiling of insulin/glucagon-producing pancreatic islet *α*- and *β*-cells (Buchanan *et al.,* 2007); detection of apoptosis in mammalian cells (Dong *et al.,* 2011); and characterisation of batches of IgG producing monoclonal antibody CHO cell lines (Feng *et al.,* 2010; Feng *et al.,* 2011). The application of MALDI analysis on MCC during biotherapeutic protein production leads to the issue of how to reveal the information in the mass spectra profiles relating to the state and behaviour of the cell lines in the culture. Thus the use of proper data analysis techniques for such spectra profiles is crucial in obtaining reliable information concerning potential biomarkers that may provide information on the state and behaviour of the cell lines. Multivariate data analysis (chemometric) methods can be used to simplify complex proteomic data profiles, making visualisation and classification easier, and hence the possibility of detecting biomarker patterns. Methods such as principal component analysis (PCA) (Wise *et al.,* 2005), and projection to latent structures – discriminant analysis (or partial least squares-discriminant analysis; PLS-DA) (Barker and Rayens, 2003) have previously been applied in proteomic profiling (Lee *et al.,* 2003; Eriksson *et al.,* 2004).

PLS-DA (which is the focus of this thesis) is a variant of standard PLS regression. It has previously been used in biomarker profiling of MALDI generated data sets in terms of discriminating normal from diseased blood specimens, normal from diseased urine samples, and between microorganisms (Lee *et al.,* 2003; Norden *et al.,* 2005; Pierce *et al.,* 2006). Having the dimensionality reduction advantage of PLS, PLS-DA is suited to extracting small, systematic variations from large, noisy data sets by identifying a lower dimensional latent variable space within which most of the information lies. It can potentially separate classes of samples with respect to the experimental hypothesis, forcing them to cluster together if they share a common characteristic. Separation is usually achieved on the basis of one or more peaks at certain mass-to-charge ratio (*m/z*) ratio values, offering the opportunity for the identification of specific *m/z* ratio areas in a mass spectra profile which represent potential protein biomarkers.

1.1. Aim of the thesis

The main aim of this research is to investigate the combined approach of ICM-MS and PLS-DA to discriminate between monoclonal antibody-producing mammalian cell lines based on their productivities and subsequently identify protein biomarkers (through protein database searches) that are differentially expressed in these cell lines.

The rationale behind this approach is that by applying ICM-MS to the CHO cell lines would generate mass spectra data where most of the *m/z* ratio peaks represent molecular protein/peptide ions which are singly charged. Since MALDI works through a ‘soft ionisation’ principle (that is, little or no fragmentation) each protein/peptide fragment usually produces only a single ion (MH^+) type as they are ionised by acquiring a single proton. Hence the mass of each protein/peptide fragment is equal to its *m/z* ratio value, meaning that the latter can be directly inferred as the MW of the protein/peptide fragment. Consequently, data analysis with PLS-DA would provide the possibility to rapidly identify the protein/peptide fragment ions by matching their experimental MWs (their *m/z* ratio values) to sequence-derived theoretical MWs of CHO cell proteins in internet accessible protein databases. These identified protein/peptide fragment ions can serve as potential productivity-associated protein biomarkers.

The industrial collaborators, Lonza Biologics, are interested in utilising the ICM-MS and PLS-DA approach to separate IgG producing CHO cell lines during bioprocessing based on their productivities (antibody titre in mg mL^{-1}) and identify potential protein biomarkers which are associated with the cell line productivities. Cell lines were classified into two categories (high/low producers; Hs/Ls), based on a threshold antibody titre of 1000 mg mL^{-1} above which cell lines were classed as Hs, otherwise Ls. It is hypothesised that Hs and Ls CHO cells exhibit different metabolic profiles and hence differences in protein expression patterns. These expression patterns correlate to the productivities (titre of IgG antibody produced) of the cell lines, and introduce a between-class variability. A multivariate data analysis method (PLS-DA) can be used to capture this variability and classify the cell lines based on these differences. This will be possible by training the PLS-DA model to capture the between class variability based on categories of cell lines, that is, Hs or Ls. Figure 1.1 shows an overview of the various steps involved in biomarker profiling using ICM-MS, PLS-DA and protein database search used in this thesis. It consists of the following;

- Preparation of the biological samples and analysis by MALDI mass spectrometry.
- Preprocessing (signal resampling, baseline correction, alignment, normalsation, smoothing and peak identification) of the spectra data generated to remove unwanted variation due to data acquisition issues whilst preserving biological information.
- After sampling the preprocessed samples are separated into training and test sets, the training set is overviewed using PCA to study initial trends and later analysed using PLS-DA, whilst the test sets are retained for external validation of the PCA and PLS-DA models built.
- PLS-DA scores and loadings plot are interpreted and information from the loadings plot is used for database search to identify protein biomarkers. This is followed by the biological interpretations of the results.

As mentioned above, it is hypothesised that identified protein expression patterns (or biomarkers) correlate to the productivities of the cell lines. Thus an important goal of this research would be the ability to infer the likelihood of a CHO cell line being a high or low producer based on these identified biomarkers. This could also provide insight into the biology of the mammalian cell lines during biopharmaceutical bioprocessing, and may give indications of potential genetic engineering targets that could be exploited to engineer enhanced cell lines. Predicting the likelihood of cell lines being Hs or Ls earlier in biotechnological process development may lead to early screening of Hs cell lines which will have the desired high productivity during manufacturing (bioreactor stage). The nature of other CHO cell lines can be inferred by subjecting their MALDI spectra data into the calibrated PLS-DA predictive models built with mass spectra data generated from known IgG monoclonal antibody-producing CHO cell lines.

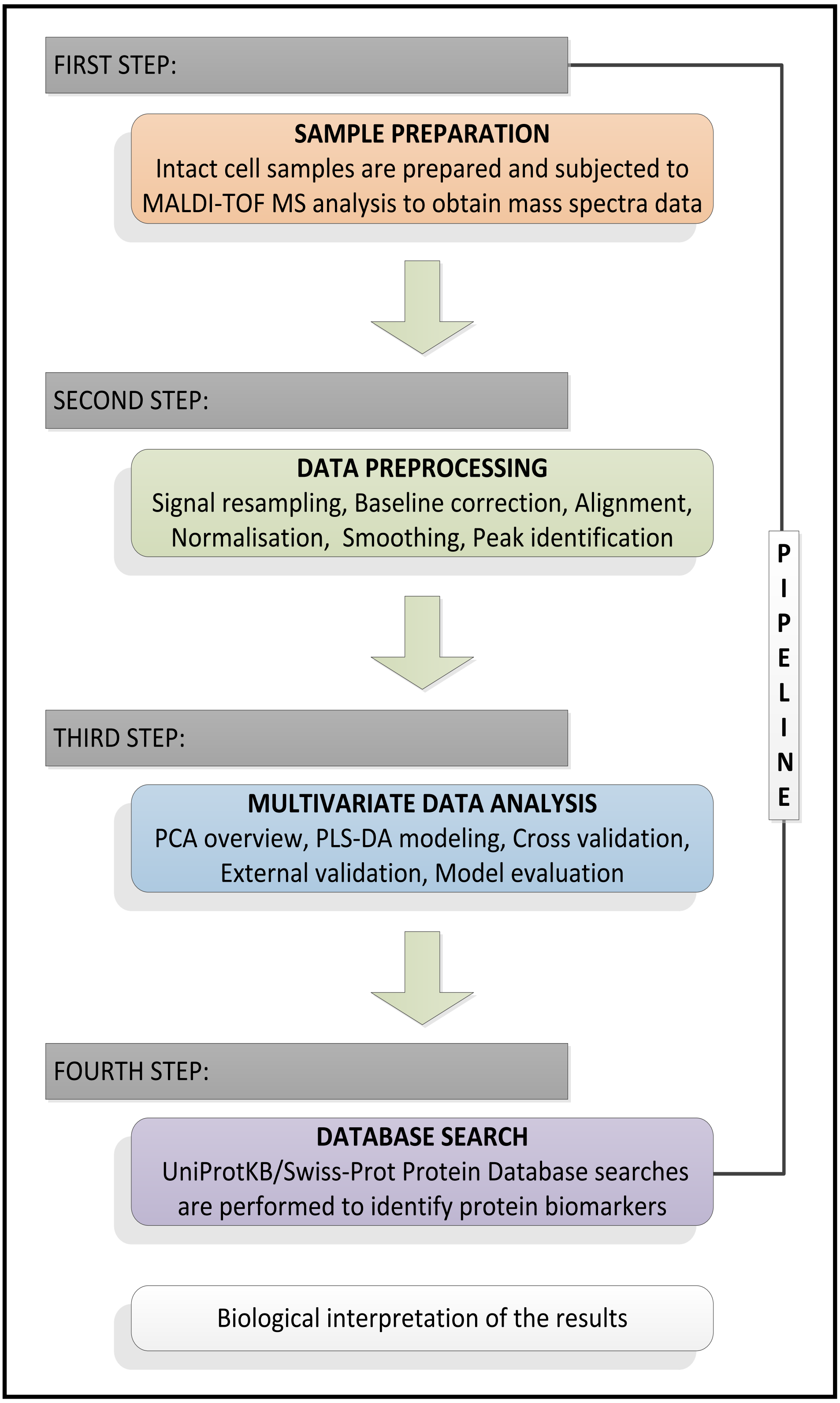

FIRST STEP:
SAMPLE PREPARATION
Intact cell samples are prepared and subjected to MALDI-TOF MS analysis to obtain mass spectra data
SECOND STEP:
DATA PREPROCESSING
Signal resampling, Baseline correction, Alignment, Normalisation, Smoothing, Peak identification
THIRD STEP:
MULTIVARIATE DATA ANALYSIS
PCA overview, PLS-DA modeling, Cross validation, External validation, Model evaluation
FOURTH STEP:
DATABASE SEARCH
UniProtKB/Swiss-Prot Protein Database searches are performed to identify protein biomarkers
Biological interpretation of the results
PIPELINE

Figure 1.1: An overview of the pipeline for biomarker discovery using 'intact-cell' MALDI-ToF MS, projection to latent structure - discriminant analysis and protein database search

1.2. Contributions of the Thesis

The main contribution of this thesis is the application of ICM-MS, PLS-DA and protein database search approach to identify protein biomarkers from monoclonal antibody producing mammalian cell cultures during biopharmaceutical bioprocessing. More specifically, the key contributions of this thesis are as follows;

1. Application of the ICM-MS, PLS-DA and protein database search approach to IgG monoclonal antibody-producing CHO cell lines during bioprocessing. While conceptually straightforward, to the knowledge of the author, no existing studies adopt this approach; consequently this is the first study where this methodology is used to identify potential mammalian cell biomarkers in the area of mammalian cell culture during bioprocessing (chapter 7).

2. An *Escherichia coli* growth-phase-associated protein biomarker model, where ICM-MS, PLS-DA and a protein database search approach is applied to *E. coli* K-12 culture to identify potential protein biomarkers associated with the different growth phases of the culture. Mammalian cells are complex systems, hence this standard *E. coli* growth-phase-associated biomarker model is used as a proof-of-concept study for the study described above. Samples were collected and analysed from the cultures at points in the three specific growth phases (exponential, stationary and decline phase). After establishing the growth curve of the bacterial cultures, the predictable timing of growth along the growth curve ensured that samples were collected at specific time points during the three growth phases. This increased the likelihood that differences between identified proteins will be related to the progression from one growth phase to another.

   Biologically, it is expected that *E. coli* cultures at the three different phases of growth exhibit different metabolic profiles and hence different protein expression patterns so that unique proteins can be induced and differentially expressed by these cultures. It is anticipated that the latter protein expression patterns correlate with the growth phase of the culture, and hence between-class (exponential, stationary and decline phase classes) variability will be evident. PLS-DA is used to capture these differences and hence classify the *E. coli* cells.

A database search by matching experimental MWs to sequence-derived theoretical MWs of *E. coli* cells using internet accessible protein databases was then carried out to identify potential growth-phase-associated protein biomarkers.

3. One challenge involved in biomarker discovery using mass spectrometry is to provide an appropriate preprocessing method for the generated mass spectra data. Preprocessing is data dependent and one of the goals is to eliminate differences between spectra profiles as a consequence of experimental and instrumental procedures, while preserving the inherent biological information within the spectra profiles. A number of combinations of data preprocessing techniques/parameters were considered and empirically investigated to enable the most appropriate selection of the combination of methods/parameters. The parameters of the preprocessing methods were modified systematically and applied to the spectra data. The preprocessed data was used to calibrate PLS-DA models. The performances (predictive ability) of the constructed models using different preprocessing techniques/parameters were assessed. The combination of preprocessing techniques/parameters that gives an improved and optimum model performance was selected as the appropriate preprocessing method (section 3.6).

   It was reasoned that improved classification and predictive model performance indicates that predictive models capture as much information as possible relating to the experimental hypothesis. This will also imply that these models are valid, that is, providing accurate biological information pertaining from the data. Thus biomarkers derived from such models will have a high probability of being related to the experimental hypothesis as well as being valid. In addition to model performance, a result-driven approach further helped to validate the appropriate preprocessing methods/parameters. That is, a combination of preprocessing methods/parameters was considered as being valid if the preprocessed spectra data (after modeling the spectra data with PLS-DA), provided accurate information (from the PLS-DA loadings plot) important for the identification of protein biomarkers previously described in the literature.

1.3. Organisation of this thesis

The following summarises the main components of the chapters of this thesis.

Chapter 2 presents an introduction to biopharmaceutical therapeutics and gives an overview of different aspects pertaining to the production of biotherapeutic proteins in cultivated mammalian cells. It also describes expression systems, factors affecting intracellular expression and highlights methods involved in enhancing productivities of recombinant proteins in mammalian cell cultures. Additionally, the chapter explores the Chinese hamster ovary (CHO) cell line highlighting the advantages it has over other mammalian cell lines. The chapter concludes with a literature survey of the methods that are currently being used for improving large-scale protein production in mammalian cell lines.

Mass spectrometry is introduced in chapter 3 with particular focus on Matrix Assisted Laser Desorption/Ionisation Time-of-Flight Mass Spectrometry (MALDI-ToF MS). It describes how the MALDI-ToF MS instrument is used to collect data from E. coli cultures at different growth phases as well as IgG monoclonal antibody-producing CHO cell lines during culturing. Methods involved with the preprocessing of mass spectra data profiles are also reviewed. An empirical evaluation of preprocessing methods on the two sets of spectra data (*E. coli* growth profile model and IgG producing CHO cell lines) is presented.

An introduction to proteomics and biomarker discovery relevant to this work is presented in chapter 4. It provides a summary of approaches used in biomarker discovery and highlights the top-down proteomics based approach of ‘intact-cell’ MALDI-ToF MS (ICM-MS), for the identification of microorganisms that may be relevant to biomarker discovery in the biopharmaceutical bioprocessing industry. The chapter also explores the potential advantage and usefulness of ICM-MS and internet-accessible protein databases for rapid biomarker profiling in the area of mammalian cell culture in biopharmaceutical bioprocessing. It concludes with a literature review on the applications of ICM-MS on both bacterial and mammalian cells.

Chapter 5 provides an introduction to multivariate data analysis (chemometrics), and an overview of chemometric methodologies underpinning this research – principal component analysis (PCA) and projection to latent structures – discriminant analysis (PLS-DA). It explores the PLS-DA algorithm involved in modeling the two mass spectra data sets discussed in chapter 4. A review on the application of PLS-DA to proteomics mass spectra data is given. The chapter ends with a discussion of results of an example where PCA and PLS-DA were used to analyse MALDI-ToF mass spectra data generated from cell lysate samples of *E. coli* K-12 cells at different growth phases.

A case study of an *E. coli* growth-phase-associated protein biomarker discovery model is presented in chapter 6, where ICM-MS, PLS-DA and the protein database search approach is applied to identify growth phase-associated protein biomarkers. The wet lab procedures for mass spectra data collection, PLS-DA modeling of the spectra data, and biomarker identification through database searches are presented. The detailed interpretations of the results as well as subsequent discussions are presented.

Chapter 7 describes the second case study where WCM-MS, PLS-DA and protein database search approach are applied to IgG monoclonal antibody-producing CHO cell lines during bioprocessing, to identify productivity-associated protein biomarkers. Mammalian cells such as CHO are complex systems, hence the standard *E. coli* growth-phase-associated biomarker model is used as a proof-of-concept or benchmark study for this case study. The wet lab procedure for mass spectra data collection, PLS-DA modelling of the data, and biomarker identification through a database searches are presented. The detailed interpretation of the results as well as discussions is presented.

A summary of the work, along with a discussion regarding the strengths and weaknesses of the proposed ICM-MS, PLS-DA and protein database search approach to biomarker discovery is presented in chapter 8. The chapter also highlights some areas of the research that could benefit from further investigation.

Chapter 2

2. Biopharmaceutical Therapeutics

2.1. Overview

The main purpose of this chapter is to provide an introduction to biopharmaceutical therapeutic proteins as well as the science and technology of *in vitro* mammalian cell culture which is used in their production. The chapter also provides an appreciation of the impact that mammalian cell culture technology has had on the health and well-being of mankind. The chapter will begin by giving an overview of the various expression systems used for recombinant protein production as well as their advantages and disadvantages which influence their ability as potential hosts. It also explores the Chinese hamster ovary (CHO) cell line highlighting the advantages it has over other cell lines, which has made it the most widely utilised mammalian cell culture expression system. Additionally, the chapter will provide an outline of the general workflow of the steps involved in the production of a recombinant protein in stirred, serum-free cultures, when beginning from a recombinant vector and a mammalian host cell line. The chapter will end with a literature survey of the methods that are currently being used for improving large-scale production of heterologous proteins from mammalian cell lines.

2.2. Introduction

Biopharmaceutical proteins can be defined as pharmaceutical substances originating from biological sources and are the basis of approximately one-third of the drugs currently in development. Biopharmaceutical drugs or biotherapeutics are large, complex protein molecules derived from living cells, used clinically for therapeutic or *in vivo* diagnostic purposes (Sekhon, 2010). Biotherapeutics include monoclonal antibodies (MAbs), recombinant proteins and viral vaccines.

Prior to the 1970’s only limited amounts of proteins for clinical use were available as they could only be sourced naturally from humans and animals. However, non-natural sources are now available with the advent of technologies such as recombinant DNA and hybridoma technology. Advances in these technologies have revolutionised the production of biotherapeutics. Recombinant DNA technology made possible the large scale production of biotherapeutics in the form of recombinant proteins. For its part hybridoma technology enabled the production of a new category of biotherapeutic proteins known as MAbs that have provided alternative treatment regimens for a number of ailments including cancers infectious, and autoimmune diseases (Birch and Onakunle, 2005).

A number of reasons resulted in the move to the production of recombinant proteins from recombinant DNA technology. Firstly recombinant proteins have been used mainly as a replacement for naturally sourced biotherapeutics such as human growth hormone (hGH) extracted from dried pituitary glands of dead humans (Birch and Onakunle, 2005). Human insulin became the first biotherapeutic to be manufactured and approved by the food and drug administration (FDA) via recombinant DNA technology in 1982. Other recombinant products including hGH, tissue plasminogen activator, erythropoietin, and blood-clotting Factor VIII have also been produced using recombinant DNA technology (Sekhon, 2010). Secondly, safety issues surrounding natural sources led to the switch to recombinant proteins. For example production of hGH was changed to *Escherichia coli* after it emerged that hGH from the pituitary glands of dead humans was a source of prion protein, the causative agent of Creutzfeldt–Jakob disease (Sekhon, 2010). Recombinant DNA technology also replaced blood serum from which the potentially dangerous hepatitis B vaccine was extracted; this viral vaccine is now produced in baker's yeast (*Saccharomyces cerevisiae*) (Birch and Onakunle, 2005).

Substantial progress has been made with respect to new approaches to treat various diseases using MAbs and recombinant proteins. These biotherapeutics have been widely used for treatment mainly in the field of cancer and other important areas like infectious diseases, autoimmune and cardiovascular disorders (Birch and Onakunle, 2005). Recent therapeutic advances in 2011 include the first new treatments for the following diseases and disorders: Benlysta is the first drug to be produced in over 50 years, used for treating lupus; Adcetris is the first drug to treat Hodgkin's lymphoma since 1977; Anascorp is a new drug effective against scorpion stings; Corifact is used for treating Factor XIII deficiency; Nulojix is the first drug used against the rejection of organ transplants in more than 10 years; and Yervoy is a new drug for the treatment of advanced melanoma (Biopharma, 2012). Globally, more than 150 biopharmaceutical drugs are currently in the market and as more drugs are gaining approval, the challenge has been to provide and maintain suitable expression systems to meet the production of biotherapeutics.

2.3. Recombinant Protein Expression Systems

The growing need for therapeutic, diagnostic and functional activity bioassay applications of recombinant proteins has enabled the advancement of bioprocessing technology for the production of recombinant proteins. However, the expression of a spectrum of these recombinant biotechnology products to meet market demands has been a major challenge (Andersen and Krummen, 2002). These demands can only be met by the heterologous synthesis of these recombinant proteins. Heterologous protein production involves the introduction of a foreign DNA into host cells for expression, a move which involves a number of considerations (Fig 2.1): isolation of the DNA to be introduced; construction of a recombinant vector (cDNA) having the DNA; and identification of a suitable expression system to accommodate the cDNA (Rai and Padh, 2001).

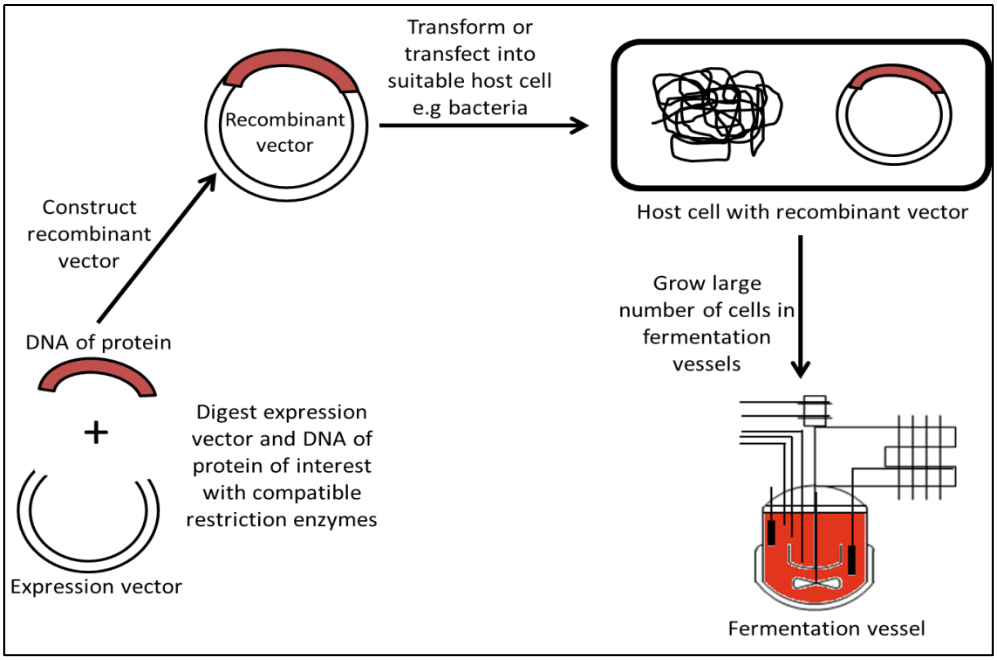

Transform or transfect into suitable host cell e.g bacteria
Recombinant vector
Host cell with recombinant vector
Construct recombinant vector
DNA of protein
Grow large number of cells in fermentation vessels
+
Digest expression vector and DNA of protein of interest with compatible restriction enzymes
Expression vector
Fermentation vessel

Figure 2.1: Schematic diagram showing the steps involved in heterologous protein production

There are a wide range of expression systems available for large-scale recombinant protein production. The most commonly used systems include bacteria, insect cells, yeast, and mammalian based systems (section 2.3.4). Each has its own advantages in terms of cost, ease of use, and post-translational modifications (PTMs) of the protein products. Selection of the most appropriate expression system for recombinant protein production must consider these elements. Most importantly, the choice of a suitable expression system is based on the system's ability to produce the protein in a form identical to that found in the cell type from which the recombinant DNA was obtained. These expression systems are briefly reviewed in the subsequent sections.

2.3.1. Bacteria

The bacteria expression system using *Escherichia coli* is perhaps the first choice for the heterologous production. This is primarily due to the organism being relatively easy to manipulate, the low cost of the culture media, the shorter time to obtain acceptable production yields, and the potential to use a great variety of strains and hence expression vectors. The limitations of this system are its lack of PTMs (for example glycosylation and disulfide bond formation), leading to incorrect modification of heterologously expressed eukaryotic or mammalian proteins. These proteins require proper PTM in order to function properly otherwise the mammalian immune system may recognise such proteins as foreign. Incorrect PTMs can also hamper secretion of large amounts of expressed eukaryotic proteins from the bacteria expression system. Another limitation of this system is the precipitation of large amounts of expressed proteins into inclusion bodies creating difficulties in terms of the purification of the final product (Hunt, 2005). Although *Escherichia coli* remains the most widely used bacteria expression system, other bacteria which have been used include *Streptococcus, Lactococcus, , Leuconostoc, Pediococcus and Lactobacillus* species.

2.3.2. Yeast

Yeasts are the favoured alternative expression systems for eukaryotic proteins. They have advantages in that they are relatively simple to manipulate and inexpensive to culture and also offer the possibility of PTM. Yeasts fall in the list of the second most commonly used expression systems and has been used as a replacement for bacteria. Expression in yeast can be both intracellular and extracellular through the application of short signaling sequences. The most common species of yeast used for intracellular heterologous protein expression is *Saccharomyces cerevisiae* since it can secrete the proteins it expresses. However *S. cerevisiae* has hyperglycosylation problems, where high mannose glycans are incorporated to the expressed proteins, a phenomenon which is not good for human therapeutics. Other yeast strains with better secretion properties are *Pichia pastoris*, *Klyveromyces lactis*, *Schizosaccharomyces pompe*, *Yarrowia* *lipolytica*, and *Hansenula polymorpha*. Well developed yeast expression systems for large scale heterologous protein expression are based on *P. pastoris*. With the exception of *S. cerevisiae*, the other yeasts are capable of producing 10 - 100-fold more secreted proteins without hyperglycosylation. The main drawback in using a yeast expression system is that yeast cells have cell walls so recovering the protein from the interior of the cell after intracellular expression is a major challenge (Reyes-Ruiz and Barrera-Saldana, 2006).

2.3.3. Insect Cells

An expression system with insect cells is based on the Baculovirus Expression Vector System. It has emerged as a popular eukaryotic expression system for expression of recombinant proteins with the ability to produce proteins having the proper PTM: folding, O-linked and N-linked glycosylation, amidation, carboxymethylation, oligomerisation, phosphorylation, disulfide bond formation, proteolytic cleavage and acylation (Reyes-Ruiz and Barrera-Saldana, 2006). Baculoviruses are present in invertebrates, primarily insect species, and can be propagated to high titres for growth in suspension cultures, creating the potential to obtain large amounts of recombinant proteins in a relatively short time (Rai and Padh, 2001). S*f*9 and S*f*21 are the most commonly used insect cell lines for expression as they can be grown in suspension and can thus be used in a bioreactor. They are derived from the *Spodoptera frugiperda* larvae and are susceptible to baculovirus infections. Other commonly used insect cell lines include High Five cells (derived from *Trichoplusia ni* egg cell homogenates) and the *Drosophilia* system, which relies on stable cell lines (Reyes-Ruiz and Barrera-Saldana, 2006).

2.3.4. Mammalian Culture Cells

Mammalian cell lines are considered the ideal eukaryotic expression system for proteins intended for human therapeutics. Proteins requiring mammalian PTMs are typically expressed in this system. The drawbacks of using this expression system include high costs of maintenance when compared to bacteria or yeast cultures due to complexity of the facilities, long culture times and nutrient requirements. In addition, handling of such facilities requires qualified and trained personnel; safety risks are involved as the required growth factors are added by calf serum which can be contaminated with viruses or prions; and there is no guarantee of always obtaining high product yields.

Despite these limitations, mammalian cell lines offer the greatest degree of product fidelity having appropriate post-translational modifications. For example mammalian cell lines express proteins with glycosylation identical to native endogenous human proteins. They are also the expression system of choice especially if clinical efficacy of the biotherapeutics is determined by its authenticity. That is, proteins not properly glycosylated may be recognised as "foreign" by the immune system of higher mammals, resulting in an immediate response against such proteins preventing them from fulfilling their therapeutic purpose (Reyes-Ruiz and Barrera-Saldana, 2006). Some commonly used mammalian cell lines for the construction of mammalian expression systems for large-scale commercial recombinant protein production are Chinese hamster ovary (CHO), mouse myeloma (NS0), and baby hamster kidney (BHK-12).

2.3.5. Cell-free Systems

Cell free systems are *in vitro* expression systems which contain cellular extracts obtained from either prokaryotic or eukaryotic cells, and provide the necessary molecular machinery and biochemical constituents required for transcription and translation. This system has the advantage of avoiding limitations associated with the *in vivo* systems of bacteria and eukaryotes, where over-expressed protein is toxic to the host cell, where the protein has insolubility problems or forms inclusion bodies, or where the protein is susceptible to rapid degradation by intracellular proteolytic enzymes. Common cell-free expression systems contain prokaryotic and eukaryotic cellular extracts (initiation, elongation and termination factors, 70S or 80S ribosomes, aminoacyltRNA synthetases, and tRNAs) from rabbit reticulocytes, wheat germ and *E. coli* (Reyes-Ruiz and Barrera-Saldana, 2006).

2.3.6. Plants

Plants are an interesting alternative to the aforementioned expression systems and are capable of providing low cost *in vitro* expression systems. The advantage with plant expression systems is the absence of contamination from endotoxins and animal viruses which are associated with bacteria and eukaryotic expression systems. Plant expression systems have drawbacks mainly from an economic perspective making them unattractive as hosts for expression. It is time consuming to express recombinant proteins in plant expression systems typically taking about two years from the initial transformation event to small-scale evaluation and production. This is a consequence of the relatively slow growth rate of terrestrial plants. Furthermore, with plant systems recombinant proteins are produced and deposited in specific organs such as leaves, fruits, and seeds rendering the protein purification complex and costly (Reyes-Ruiz and Barrera-Saldana, 2006).

2.4. Monoclonal Antibodies

2.4.1. Overview of Monoclonal Antibodies

Antibodies are protein molecules, known as immunoglobulins, synthesised by the immune cell of an animal in response to a foreign macromolecule known as an antigen (Farid, 2006). Monoclonal antibodies (MAbs) are pure populations of antibodies that recognise and attack the same molecular target. They are produced by a population of immune cells (B lymphocytes) derived from the same parent cell. MAbs began to be widely applied in research and development following the development of hybridoma technology in mice by Kohler and Milstein in 1975 for their large-scale production (Gombotz and Shire, 2010).

Recently, MAbs have become one of the fastest growing classes of all biopharmaceuticals, with a total of 26 therapeutic MAbs approved in 2007 in the US by the Food and Drug Administration (FDA). These had a market value of more than $12,612 million (Gombotz and Shire provide a list of all the commercial MAb products approved in the U.S) (Gombotz and Shire 2010). The global market for therapeutic and diagnostic MAbs increased from $26 billion in 2006 to an estimated $31 billion in 2007. This upward trend continued in 2009 and 2010. For example Fig. 2.2 shows a pie chart of US sales of nine biotherapeutics drug categories, and table 2.1 shows data for the sales and the growth rates of the nine drug categories between 2009 and 2010 (Aggarwal, 2011).

Fig. 2.2 indicates that MAbs remained the best-selling class of biologics and in 2010, US sales of MAb products reached ~$18.5 billion, 9.7% higher than 2009 sales. The table indicates that MAbs and enzymes are the drug categories that showed the fastest growth rate during that period. The global market for MAbs is expected to top $56 billion by the end of 2012 which represents a compound annual growth rate of 13% (Bccresearch, 2008). Projections for the next five years (2012 to 2017) suggest that MAbs are expected to be the biggest drivers in the global biopharmaceutical market.

This is mainly a consequence of the rich late stage pipeline and a strong uptake from both developed and emerging markets (Imarcgroup, 2012). MAbs play a major role in treating a wide variety of diseases including cancer, infectious disease, allergy, autoimmune disorders and inflammation. MAbs are produced in the following main forms: murine (100% mouse protein); chimeric (approximately 65% human and 35% mouse protein); humanised (95% human and 5% mouse protein); and fully human (100% human protein) (Gombotz and Shire, 2010).

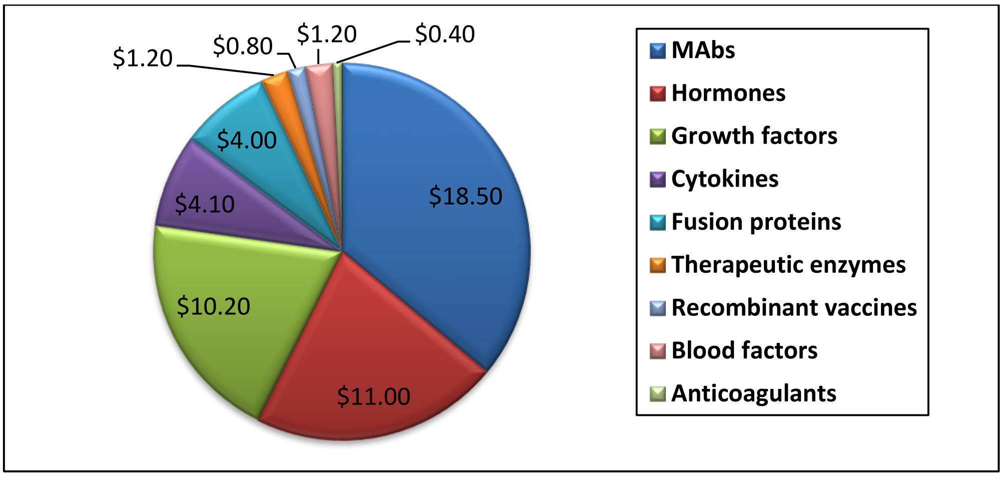


Figure 2.2: A pie chart showing US sales of nine biotherapeutics drug categories

	2009 sales ($)	**2010 sales ($)**	**2009 growth rate (%)**	**2010 growth rate (%)**
MAbs	16.9	18.5	**8.3**	**9.7**
Hormones	9.8	11.0	**14.7**	**12.3**
Growth factors	10.4	10.2	-9.1	-2.0
Cytokines	3.9	4.1	6.8	4.6
Fusion proteins	3.9	4.0	1.1	4.5
Therapeutic enzymes	1.1	1.2	-4.2	4.9
Recombinant vaccines	0.7	0.8	-37.0	13.0
Blood factors	1.3	1.2	5.0	-2.6
Anticoagulants	0.3	0.4	-1.2	7.9

Table 2.1: The table shows the growth rates of the categories between 2009 and 2010 (data obtained from Aggarwal, 2011)

2.4.2. Challenges with Monoclonal Antibody Development

The first therapeutic MAbs to be produced, Orthoclone (an immunosuppressant drug against kidney transplant rejection) had little success as far as commercialisation was concerned. Clinical trials failed as patients who received infusions of the early therapeutic MAbs developed immune responses against the administered product. Such products were also rapidly destroyed by the liver even before they could reach their therapeutic target (Ezzell *et al*., 2001). This behaviour could be attributed to the fact that MAbs were made with hybridoma technology and were of mouse origin. The human immune system recognises these as being foreign and generates human anti-mouse antibodies (HAMA response) against the administered products.

The therapeutic properties of MAbs strongly depend on their glycosylation pattern. Most of the currently approved MAbs are produced in mammalian cell lines, as they have the propensity to express proteins with almost human-like glycosylation. Moreover, the advancement of humanisation technology has facilitated the incorporation of murine (mouse) residues complimentarity determining regions (responsible for antigen binding) into a human antibody framework giving rise to antibody sequences with up to 90–95% human content (Gombotz and Shire, 2010). For example, a humanised MAb drug, mogamulizumab (Poteligeo®), with engineered glycosylation to enhance the pharmacological properties - designed to treat cancer - has recently been approved globally. This is the first glyco-engineered antibody to reach the market in the field of therapeutic antibodies (Subramaniam *et al*., 2012).

2.5. Mammalian Cell Cultures for production of MAbs

2.5.1. Introduction

The cultivation of mammalian cells *in vitro* (for example as suspension culture in bioreactors) is known as mammalian cell culture. It has evolved from an experimental science in the 20th century to a modern quantitative science in recent times. It has seen wide application in terms of research and development in the industrial, academic and medical fields (Kretzmer, 2002). Mammalian cell culture has been applied in the areas of cell and molecular biology, providing reproducible model systems in which the physiology and biochemistry of the cell can be studied; the area of medical sciences for efficacy and toxicological assessment of potential new drugs; and for the large scale production of biotherapeutics including vaccines, recombinant proteins and monoclonal antibodies.

2.5.2. Chinese Hamster (CHO) Ovary Cell

There are some key issues that affect the choice of a mammalian cell line for use in a large scale manufacturing process during therapeutic MAb production. These include the capability of attaining high product yield; the ability to produce post-translationally modified products which impact on the pharmacokinetic and pharmacodynamic properties of the product (solubility, stability and therapeutic efficiency, time taken to be cleared from the body); the ability to be amenable to genetic manipulation to easily accommodate a foreign DNA; the ability to rapidly and consistently produce safe products; and the ability for the cells to adapt and grow in suspension cultures is an important characteristic which enables volumetric scalability and use of large stirred-tank bioreactors (Jayapal *et al.*, 2007).

CHO cell lines meet these criteria and have become the most widely applied mammalian cell line industrially, for the large-scale production of therapeutic MAbs, from a range of alternative cell lines including NS0, BHK-12, mouse myoloma cell line (SP2/0), human embryonic kidney (HEK-293) and human-retina-derived cell line (PER-C6). Fig. 2.3 shows that the majority of therapeutic Mabs approved globally in 2012 were produced in CHO cell lines. Of the 30 MAbs that were approved and licenced globally as of March 2012, 12 (40%) were produced in CHO cells, whilst 7 (23%) were produced in SP2/0 cells, 5 (17%) were produced in NS0 cells, 2 (7%) were produced using hybridoma technology and 2 (7%) in *E. coli* (Reichert, 2012). Moreover, there are derivatives of CHO cell lines, (for example *dhfr*$^-$ CHO) deficient in their ability to produce the dihydrofolate reductase enzyme (DHFR), an essential enzyme which catalyses the conversion of folate to dihyfrofolate (Simonsen and McGrogan, 1994). Hence a well-established expression vector (the dihydrofolate reductase (DHFR) expression vector system) can be used as a basis to develop platform technologies which allows for the transfection, amplification, selection and expansion of high producing and stable CHO cell clones based on this property (DHFR deficiency).

Another vector system based on an amplifiable gene, glutamine synthetase (GS) is an alternative to DHFR system that provides a well characterised platform technology usable in CHO cell lines, providing them with an added advantage over other cell lines. GS is a dominant selectable marker that can be used with GS-negative and positive NS0 cells and provide a possibility of gene amplification (Birch and Onakunle, 2005). The next section provides a brief description of the production of therapeutic MAbs in a typical CHO cell suspension culture system.

2.6. Production of Recombinant Therapeutic Monoclonal Antibodies by Mammalian Cells in Suspension Culture

2.6.1. Cell Line Development

There is a general format that is typically used in producing a recombinant protein in a stirred, serum-free culture beginning from a recombinant vector (plasmid) and a mammalian host cell line. This format was used at Lonza Biologics in producing the MAb, Immunoglobulin G (IgG), using a CHO cell line generated in-house (Lonza Biologics, Slough, UK). The development of a manufacturing process for a recombinant therapeutic MAb follows a well-established procedure. In this project, the manufacturing process was carried out in accordance with established procedures explained in detail in Chu *et al*. (2005), in Smales and James, (2005). Details of the materials, reagents and protocols involved are provided in Chu *et al.*, (2005). Cell line development was carried out in this project and samples were collected for MALDI analysis as described in section 3.3.6.2.

2.7. Enhancing Productivities of Recombinant proteins in Mammalian Cell Cultures

2.7.1. Introduction

As previously mentioned mammalian cells are the host of choice for production of biotherapeutics with human-like post-translational modification. However, they are slow, expensive and productivities are low when compared to bacteria and yeast expression systems. The well-established scheme for cell line development during the manufacturing process used at Lonza Biologics in producing the MAb, IgG has described in section 2.6. Although this has provided the opportunity for optimisation of medium composition and high-producing cell line screening which has delivered significant increase in volumetric product titres over the past decades, there has been little increase in spectfic productivity (*q*P) since 1990 (Dietmair *et al*., 2012).

Mammalian cell line engineering with superior growth characteristics, and optimised for the production of high concentrations of proteins, has been noted to significantly reduce the development time for a high productivity cell line. Consequently, significant effort has been invested in using various genetic engineering strategies to improve those aspects of the cell line related to product titres. For example aspects of the cell line such as delaying apoptosis, enhancing the cell's processing capacity, increasing the rate of cell proliferation, and increasing the cell's metabolic efficiency have been carried out (Mohan *et al.*, 2008; Schroeder, 2008; Lim *et al.*, 2010).

In this respect, an approach to identifying protein biomarkers whose genes may act as potential genetic engineering targets as well as serving as a basis to identifying high producing cell lines earlier in process development is presented in this thesis (chapter 7). Genetic engineering and the selection of high producing cell lines would enable the optimisation of protein production at high concentrations. New cell lines with optimised growth and productivity properties (potentially capable of producing high yields of MAb products) can be created by manipulating the genes in the cells that are responsible for controlling growth and productivity. Selecting only high producing cell lines will also have the benefits of potentially providing high yields of MAb products. Before this approach is presented, it is important to review some of the strategies for improving mammalian cell line productivities that have already been proposed in the literature.

2.7.2. Review on Enhancing Productivities of Recombinant proteins in Mammalian Cell Cultures

Methods currently applied for improving the large-scale production of heterologous proteins from mammalian cell lines include optimisation of the cell culture environment, bioprocess design and improving expression vectors through genetic engineering (Bebbington *et al.,* 1992; Zhou *et al.,* 1997; Fussenegger *et al.,* 1998; Ibarra *et al.,* 2003). Bebbington *et al.,* (1992) reported a method in which transfectants were selected in terms of growth in a glutamine-free medium using the glutamine synthetase (GS) selectable marker. cDNA amplification as well as selection for transfectants was subsequently carried out using the specific inhibitor of GS, methionine sulfoximine. DNA sequences encoding a chimeric IgG4 antibody were expressed in NS0 cell line transfected with cDNA vectors controlled by human cytomegalovirus major immediate early (hCMV-MIE) promoters. High levels of productivities of up to 560 mg/L antibody were observed.

In another study Fussenegger *et al*., (1998) saw an increase in *q*P by decreasing the cell specific growth. In this study they reprogrammed the regulatory complex involved in the CHO cell cycle by blocking proliferation at high cell densities leading to an extended period of high production. This led to a 10–15 times increase in the production of the heterologous protein, secreted alkaline phosphatase. Alternatively, rather than causing the cessation of cell proliferation, Ibarra *et al.*, (2003) instead increased productivity by inducing the overexpression of anti-apoptotic genes, which overexpresses proteins against apoptosis (programmed cell death). An NS0 cell line, which expressed a chimeric IgG4 antibody, was further engineered to constitutively overexpress the anti-apoptotic proteins (Bcl-2 and p21CIP1). Effects of overexpression of these anti-apoptotic proteins on cell proliferation, cell viability, and antibody production were investigated in batch and continuous perfusion cultures, and mixed results were obtained. Mutant Bcl-2 expression did not show any significant improvement in cell viability of arrested cells. In contrast, p21CIP1 protein arrested cell proliferation, and gave a 4-fold increase in antibody production, the mutant Bcl-2 had observed change in cell viability.

2.8. Summary

As the demand for biotherapeutics continues to increase, there is the need for this to be matched by increased global large-scale recombinant protein production. There are a wide range of expression systems available for large-scale recombinant protein production, with the most commonly used ones being bacteria, insect cells, yeast, and mammalian cell based systems including CHO, NS0, BHK-12. The choice of expression system for the production of a recombinant protein is influenced by a number of factors. All the expression systems have some advantages as well as limitations which must be considered before selecting the most appropriate one. The factors include cost, yield, propensity for post-translational modification and the economics of scale-up. Mammalian cell lines are considered to be the ideal eukaryotic expression system for the production of biotherapeutics as they are capable of producing proteins with posttranslational modification patterns closest to those in humans. Proteins with glycosylation patterns not resembling those found in higher mammals will not work in humans therapeutically, hence failing to fulfil the very purpose for which they were manufactured.

Therapeutic MAbs have become one of the fastest growing classes of all biotherapeutics, expecting to top $56 billion by the end of 2012 representing a compound annual growth rate of 13% (Bccresearch, 2008). Advances in technology as well as the increased global capacity of mammalian cell culture have made the latter the most utilised system for large-scale commercial manufacture of therapeutic MAbs. Although there are numerous mammalian cell lines that can serve as suspension cultures in bioreactors (for example NS0, BHK, SP2/0, CHO, and PER-C6) CHO cell lines have emerged as the cell line of choice industrially, for the large-scale production of therapeutic MAbs. Compared to others, CHO cell lines have the advantage of well-established expression vector systems (the DHFR and GS systems) with platform technologies that allow for the transfection, amplification, selection and expansion of high producing and stable CHO cell clones.

During the production of a recombinant protein in a stirred and serum-free culture, cell line development (involving transfection, amplification, selection and expansion of cell lines) is very important for generating high producing cell lines with the potential to increase productivity. However, cell line development times are usually long (about 14 weeks) and there has not always been an increase in *q*P after successful screening and selection of high producing cell lines.

Methods such as optimisation of the cell culture environment, bioprocess design and expression vector genetic engineering are currently being applied to improve the large-scale production of recombinant proteins therapeutics from mammalian cell cultures. Specifically, emphasis has been placed on using various genetic engineering strategies to improve cell line product titres. Cellular aspects including delaying apoptosis, increasing the rate of cell proliferation, and increasing the cell's metabolic efficiency have been assessed and manipulated to engineer useful variants of cell lines. However these have had mixed results where an increase in productivity is not guaranteed for all the cell lines. This suggests that the understanding of molecular details of mammalian cells is limited, and the ability to isolate cell lines and eventually identify potential genetic engineering targets will come from a better understanding of the cell's phenotypic biology.

Chapter 3

3. Mass Spectrometry

3.1. Overview

In chapter 2, the production of therapeutic proteins in mammalian cell cultures was discussed along with a literature review of some of the genetic engineering strategies that have been used to modify cell lines with the aim of enhancing cellular productivity. Mixed results were obtained with only a few studies reporting significant improvements of productivity. This suggests that more genes that can be manipulated by genetic engineering may be required. Furthermore, the ability to isolate cell lines and eventually identify potential genetic engineering targets will require a better understanding of the cell's phenotypic biology. The area of highthroughput data-rich biology is termed "-omics" and include proteomics, transcriptomics, genomics and metabolomics. These approaches can be used to study the overall biology of a cell and they have played a major role in providing knowledge in the biological sciences. Examples of such techniques include two-dimensional gel electrophoresis (2-D-PAGE) and mass spectrometry used for proteomic profiling.

In this chapter, the mass spectrometry technique of matrix assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-ToF MS) is introduced. The chapter begins with an overview of mass spectrometry with specific emphasis on describing the MALDI-ToF mass spectrometer. A description of how the MALDI-ToF MS instrument was used to collect data from *E. coli* cultures at different growth phases as well as from the IgG monoclonal antibody- producing CHO cell lines during culturing. Aspects relating to mass spectrometry data preprocessing are also discussed and preprocessing results from the aforementioned spectral data sets are reported. Finally, the chapter concludes with a review of applications of mass spectra data preprocessing.

3.2. Introduction

Over the years, advancements in the field of proteomics have led to the development of a number of analytical techniques based primarily on chromatography and electrophoresis, including 2-D PAGE, two-hybrid analysis, high performance liquid chromatography (HPLC), and protein microarrays. This toolkit of techniques, have been developed to a high technical standard, but are inappropriate with respect to selectivity, sensitivity, cost-efficiency ratio, accuracy and speed (McGuire *et al.*, 2008). The ability of mass spectrometry (MS) to meet these demands and handle the complexities associated with the proteome has led to its increased popularity (Han *et al.*, 2008).

MS is a microanalytical technique used for the identification of unknown compounds, quantification of known compounds, and for helping understand the structure and chemical properties of a given analyte. The basic principle of MS is the experimental measurement of the mass (in terms of mass-to-charge (*m/z*) ratio) of gas-phase molecular ions produced from the molecules of a sample. Unique features of MS include its ability to directly determine the nominal mass of a sample, and to produce and detect fragments of the molecule that correspond to discrete groups of atoms of different elements that reveal structural features (Watson and Sparksman, 2007).

The fundamental basis of MS is a mass spectrometer with the data generated termed mass spectrum. The components of a mass spectrometer are: (i) a sample inlet (ii) an ion source, (iii) a mass analyser, (iv) a detector and (v) a software system to store and analyse data (Fig. 3.1). From Fig. 3.1, the full meanings of the abbreviations are APCI (atmospheric pressure chemical ionisation), AP-MALDI (atmospheric pressure MALDI), CI (chemical ionisation), ESI (electrospray ionisation), FI (field ionisation), and LSIMS (liquid secondary ion mass spectrometry).

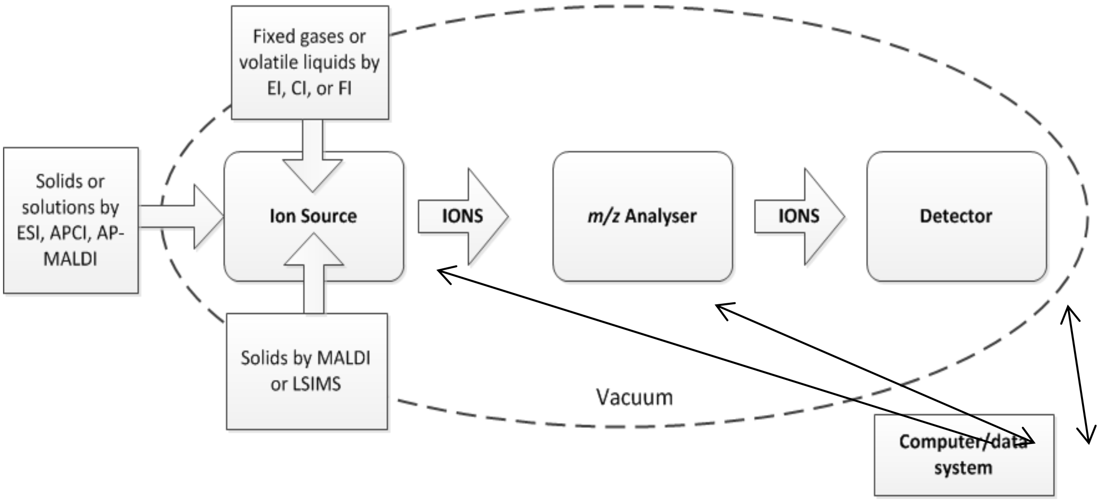

Fixed gases or volatile liquids by EI, CI, or FI
Solids or solutions by ESI, APCI, AP-MALDI
Ion Source
IONS
m/z Analyser
IONS
Detector
Solids by MALDI or LSIMS
Vacuum
Computer/data system

Figure 3.1: A conceptual illustration of the mass spectrometer showing the various components

The sample to be analysed is introduced into the ionisation source of the instrument. Sample molecules are then ionised into gaseous ions through the application of electric and magnetic fields. These ions are accelerated and transferred into the analyser region of the mass spectrometer where they are separated according to their individual *m/z* ratios. The separated ions are detected and this signal is sent to the data system where the *m/z* ratios of the ions are stored together with their relative abundance (Bergquist *et al.*, 2007). The vacuum system removes molecules thereby providing a collision-free path for the ions from the ion source to the detector. The software system coordinates the functions of the individual components and records and stores the data in the format of a mass spectrum (Watson and Sparksman, 2007). A mass spectrum is a pattern representing the distribution of molecular ions by *m/z* ratio in a sample, with the length of each *m/z* ratio peak representing the relative abundance (intensity) of the molecular ion. Fig. 3.2 is an example of a MALDI-ToF mass spectrum from a CHO cell line (a) and an *E. coli* cell sample (b).

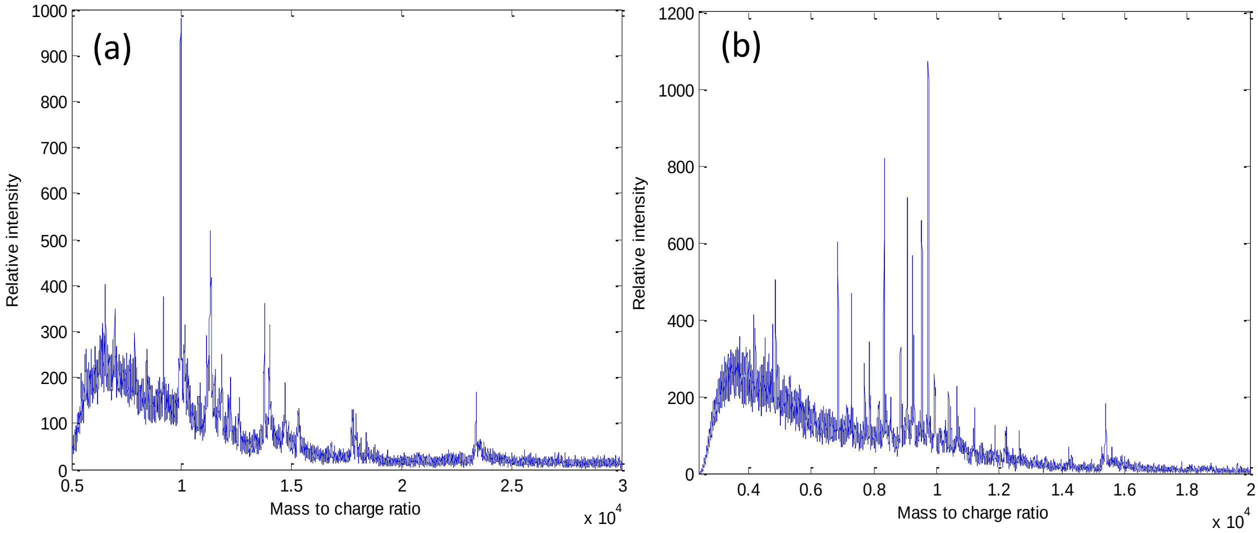


Figure 3.2: Matrix-assisted laser desorption/ionisation time-of-flight (MALDI-ToF) mass spectra: (a) CHO cell line and (b) E. coli cell

3.3. Matrix-assisted Laser Desorption/Ionisation Time-of-Flight (MALDI-ToF) Mass Spectrometry

3.3.1. Introduction to MALDI

When first introduced to mass spectrometry, lasers were directly applied to the analyte without the use of a matrix, an organic compound within which the analyte is embedded prior to analysis. In the 1960s, a technique called laser desorption (LD) used ultra violet (UV) lasers to transfer energy and ionise the analyte through electron excitation (or vibrational excitation when infrared (IR) lasers were used) (Watson and Sparksman, 2007). The limitation of this was that as the lasers were applied with intense pulses for short durations, compounds with large molecular weights and those that were thermally labile could not be analysed. Applying the lasers directly led to destruction of the analyte. Previous experiments in resonant and non-resonant laser desorption had demonstrated that large molecules may be thermally dissociated upon energy transfer (Gantt *et al.*, 1999). Furthermore, ions produced from molecules with masses greater than 500 Daltons (Da) were likely to undergo fragmentation. These limitations were overcome through the development of MALDI in the late 1980s (De Hoffmann and Stroobant, 2008).

The MALDI technique was introduced principally by Karas and Tanaka in 1987, for which they were awarded part of the 2002 Nobel Prize in Chemistry (Watson and Sparksman, 2007). MALDI has since become a widespread and powerful source for the production of intact gas-phase ions from a wide variety of large thermally labile compounds including proteins, oligonucleotides, synthetic polymers and large inorganic compounds. The routine technique for MALDI where the analyte is embedded in an organic matrix, as used today was developed by Karas and Tanaka as described in the following sections.

3.3.2. Principles of MALDI

The MALDI technique is a two-step process. In the first step, the sample is pre-mixed with a UV-light absorbing matrix solution, usually weak organic acids. The matrix-analyte solution mixture is then dried to remove any liquid solvent. The result is a co-crystal of matrix-analyte where the analyte molecules become incorporated into the matrix crystals so that they are completely isolated from one another. The second step, which occurs under vacuum conditions inside the ion source, involves irradiation of the matrix-analyte mixture with intense UV laser (337 nm) pulses. Fig. 3.3 is an illustration of the MALDI desorption ionisation process. The exact mechanism of the process is still not fully understood. However, as shown in Fig. 3.3, one thought is that irradiation by the laser induces rapid heating of the co-crystals through the accumulation of a large amount of energy resulting in the vaporisation of the matrix. The analyte molecules within the co-crystal vaporise as well but without having to directly absorb energy (De Hoffmann and Stroobant, 2008).

A number of chemical and physical processes for the formation of ions have been proposed including gas-phase photoionisation, excited state proton transfer, ion–molecule reactions, and desorption of preformed ions. The most widely accepted ion formation mechanism involves proton transfer to the analyte molecules in the solid phase before desorption. Alternatively, it can be a gas-phase proton transfer to analyte molecules in the vaporised co-crystal complex from photoionised matrix molecules. Protein molecules are usually ionised by adding a proton (H^+) to the molecule (M) to create a singly charged protein molecular ion $[M+H]^+$, but there may also be some doubly charged proteins $[M+2H]^+$ (Fig. 3.3) (Watson and Sparkman, 2007). The ions in the gas phase are then accelerated by an electrostatic field towards the analyser.

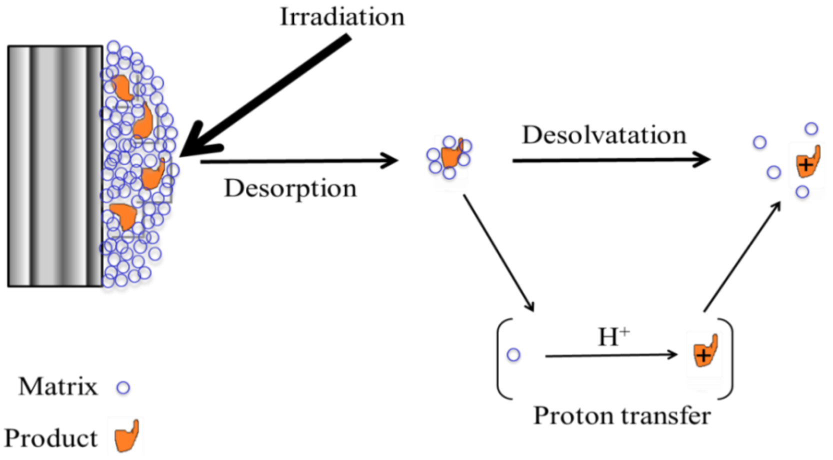


Figure 3.3: Diagram illustrating the ionisation principle of MALDI

3.3.3. Matrix

The matrix which consists of an organic solid or liquid species, performs two important functions: (1) it absorbs photon energy from the laser beam and transfers it into excitation energy, and (2) it serves as a solvent for the analyte, so that the intermolecular forces are reduced and aggregation of the analyte molecules is kept to a minimum. Desirable attributes of a typical MALDI matrix are:

- Strong light absorption at the wavelength of the laser flux.
- The ability to form micro-crystals with the analyte.
- A low sublimation temperature, which facilitates the formation of an instantaneous high-pressure plume of matrix-analyte material during the laser pulse duration.
- The participation in a photochemical reaction so that the analysed molecules can be ionised to produce large amounts of ions.

The matrix is usually a solution of organic molecules. However, when the matrix is in a solid form, the analyte and the matrix are mixed together in a mutually soluble organic solvent, and allowed to co-crystallise. The co-crystallisation of the analyte and the matrix is critical to the success of the MALDI experiment. Studies of protein ground up in a dry crystalline matrix failed to produce any spectra (Horneffer *et al.*, 2006). The ratio of matrix molecule to analyte molecules are typically between 500:1 and 5000:1. This ratio ensures that the analyte molecules are diluted in the matrix hence separating the analyte molecules to prevent analyte-analyte molecular interaction during the ionisation process. Typically examples of MALDI matrices are α-cyno-4-

hydroxycinnamic acid (CHCA), 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid), 3-amino-4 hydroxybenzoic acid and 2,5-dihydroxy-benzoic acid (DHB).

3.3.4. The MALDI -Time-of-Flight (ToF) Mass Spectrometer

The linear time-of-flight (ToF) analyser is the simplest analyser compared to others such as the reflectron and orthogonal acceleration analysers. It is widely used alongside the MALDI (the MALDI–ToF mass spectrometry platform). It has recently seen wide applications to electrospray as well as gas chromatography electron ionisation mass spectrometry (GC/MS). A MALDI instrument can be used in either linear or reflectron mode (Fig. 3.4). In linear mode (which is the focus of this thesis) the ions travel down a linear flight path and their *m/z* ratios are determined based on the time taken for the ion to reach the linear detector. A reflectron MALDI has an ion mirror at its end which reflects the ions back (at a slight angle) to the detector, into a different flight path. These deflected ions are detected at the level of the microchannel detector. Hence, this instrument is called a time-of-flight (ToF) instrument.

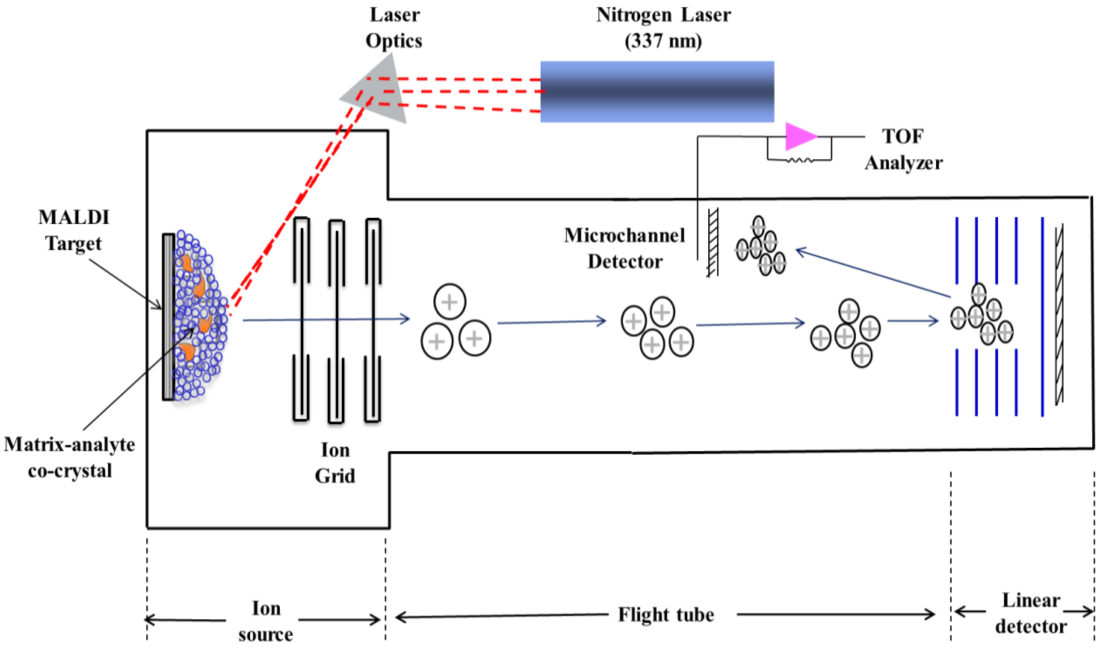


Figure 3.4: Schematic representation of a linear or reflectron mode MALDI-ToF Mass Spectrometer

The relationship that allows the *m/z* ratio to be determined is:

$$E = \frac{1}{2}\left(\frac{m}{z}\right)v^2 \qquad (3.1)$$

where E is the energy imparted to the charged ions as a result of the voltage that is applied by the instrument and, v, is the velocity of the ions along the flight path. Since all the ions are exposed to the same electric field, all similarly charged ions will have similar energies. Therefore, based on equation 3.1, heavier ions will have lower velocities and hence ions take longer to reach the detector because of their lower velocity, whilst the lighter ions will reach the detector first because of their greater velocity (Watson and Sparkman, 2007).

The time for an ion to reach a detector from the source is given by equation 3.2, in which (t - t_o) is the time-of-flight for an ion from the source to the detector, m is the mass of the ion, z is the charge of the ion, d is the length of accelerating region in the electric field, L is the length of non-accelerating region without the electric field, and V_o is the potential of the electric field.

$$t - t_o = \left[\frac{2md}{zE}\right]^{\frac{1}{2}} + L\left[\frac{m}{2zV_o}\right]^{\frac{1}{2}} \qquad (3.2)$$

After rearranging equation 3.2 for *m/z* ratio, the quadratic relationship between *m/z* ratio and ToF is apparent in equation 3.3. The constants a and b depends on the instrument, potential applied at the source, electric field, and length of the flight tube:

$$\frac{m}{z} = a(t - t_o)^2 + b \qquad (3.3)$$

The quadratic relationship between *m/z* ratio and ToF (equation 3.3) means that ions having the same *m*/*z* ratio will also have the same ToF and thus impact the detector at the same time. A cascade of secondary electrons is released when the ions strike the detector. This electron current is captured by an anode and converted to a voltage using a preamplifier. The resulting voltage is recorded by a digital storage oscilloscope or by a digitizer card in a computer, and the amplitude of the signal corresponds to the number of ions that struck the detector in each bin of ion flight time. Other sources of noise from physical and electrical components of the mass spectrometer are also recorded (for instance, high frequency noise) (Shin and Markey, 2006). After the instrument is calibrated with compounds of known mass, the constants in the quadratic equation 3.2

relating time to *m/z* ratio are determined, and the *m/z* ratio of the detected ions are calculated (section 3.3.6.1) (Watson and Sparkman, 2007).

3.3.5. Advantages of MALDI-ToF MS

MALDI-ToF was used in this thesis for biomarker studies because of its various advantages. Firstly, it is more sensitive than other laser ionisation techniques and the matrix separates single analyte molecules from each other and minimises the aggregation due to a large excess of matrix molecules. Furthermore, energy is transferred from the matrix molecules to the analyte molecules to help them to ionise. In this way, the energy imparted by the laser is absorbed and retained by the matrix molecules. This prevents the analytes from dissociation and increases the efficiency of energy transfer from the laser to the analyte leading to an increase in sensitivity (Harrington *et al*., 2006).

Generally, MALDI-ToF is considered a soft ionisation technique, that is, there is little or no fragmentation of the protonated ions formed during the MALDI process. This absence of fragmentation is a key advantage of MALDI-ToF MS compared to other classical mass spectrometry proteomic approaches such as electrospray ionisation (ESI) MS (Bergquist *et al*., 2007). Since MALDI has the inherent advantage that most ions are singly charged, the mass of the molecular ion usually is equal to the *m/z* ratio and each analyte typically only produces a single ion type. By recording the ToF, the mass of the molecular ion (MH^+) and hence peptide/protein can be directly determined. This makes it particularly appealing for protein biomarker discovery and identification. The experimental molecular weights (MWs) of MALDI protein molecular ions can be matched against sequence MWs of organisms with sequenced genomes, directly identifying the protein (chapters 6 and 7).

MALDI is also more widely applicable than the other laser ionisation techniques. Adjusting the wavelength to match the absorption frequency of each analyte is not required since it is the matrix that absorbs the laser pulse. Furthermore, because the process is independent of the absorption properties and size of the compound to be analysed, MALDI allows the desorption and ionisation of analytes with very high molecular mass including those in excess of 100 000 Da. For example, MALDI allows the detection of femtomoles of proteins with MWs of up to 300 000 Da (De Hoffmann and Stroobant, 2008).

3.3.6. MALDI-ToF Data Analysis

In this thesis two case studies were carried out involving the use of MALDI-ToF mass spectrometry. In the first case study *Escherichia coli* (*E. coli*) K-12 cells were grown in culture and samples were collected when the cultures reached exponential, stationary and decline phases. Cell pellets prepared from the samples collected were subjected to MALDI-ToF mass spectrometry analysis. In the second case study, IgG monoclonal antibody Chinese hamster ovaries (CHO) cell lines (Lonza Biologics, Slough, UK) were harvested during cell line development. Samples were then prepared for MALDI-ToF analysis. The sample preparation procedures as well as the MALDI-ToF analysis are described in the following sections.

3.3.6.1. MALDI-ToF Analysis *E. coli* K-12 cells in different growth phases

Intact Cell Pellet Sample Preparation: Cell pellets to be analysed that had been stored under -70^{o}C were re-suspended in 300µl distilled water, followed by the addition of 900µL of pure ethanol to give a final ethanol concentration of 75%. The suspension was mixed thoroughly by vortexing. A two-layer method was used for matrix/analyte sample preparation. In this two-layer method, 1µL of suspension from the previous step was deposited onto a sample spot of a MALDI target plate (MSP 96 target, ground steel; Bruker Daltonics, Germany) and was allowed to air-dry at room temperature. Finally, 1µL of the saturated sinapinic acid (SA) matrix solution was deposited onto the dried sample and allowed to dry at room temperature for co-crystallisation to take place. At this stage, the plate was ready for MALDI analysis. For the matrix preparation, SA (20 mg/mL), was prepared fresh by weighing 10mg of the SA powder and dissolving in 500µL of 40:60 ACN/TFA (0.1%, v/v). This was mixed thoroughly and sonicated for 15 minutes, then vortexed to mix and completely dissolve all traces of powder. The saturated solution was then ready for use.

Cell Lysate Sample Preparation: The ethanol suspension was made, as described for the intact cell pellet sample preparation procedure. Fifty microlitres of 70% formic acid was added to the pellets and mixed well. This was followed by the addition of 50µL pure acetonitrile and centrifuged at 17949 × *g* for 2 minutes. One-microlitre of the bacterial lysate was spotted onto a sample spot of a MALDI target plate and overlaid with 1μL of the saturated SA matrix solution. This was allowed to dry at room temperature and was then ready for MALDI-ToF analysis. All the bacterial suspensions were prepared on the same day.

Instrumentation and Data Analysis

The mass spectra were acquired with an Ultra Flex MALDI-ToF mass spectrometer (Bruker Daltonics, GmbH, Germany) (Fig. 3.5), based in the School of Biosciences, University of Kent, Canterbury, Kent, UK. Before the samples could be analysed, a series of cell pellet samples were first spotted onto the MALDI plate and used as test samples for instrument optimisation studies. During optimisation, the parameter settings of the instrument were modified to identify the combination of parameters that would give visible and intense spectra signals with less noise. Factorial design was used with different MALDI-ToF parameters as design variables. Table A.1 of Appendix A shows the optimisation results. Based on the results of instrument optimisation, the following parameters were set in the mass spectrometer for MALDI spectra acquisition: accelerating voltage 24.24 kV, a pulse ion voltage 88% (21.23 kV) of the total accelerating voltage, a laser firing rate of 20 Hz, a delayed extraction time of 300 ns, a lens voltage of 5.5kV, matrix suppression 2 kDa, and a linear detector voltage of 1.681 kV. The instrument was controlled by FlexControl v2.4 (Bruker Daltonics), operated in positive polarity and linear mode targeting a mass range of 2,000 – 30,000 *m/z* ratio.

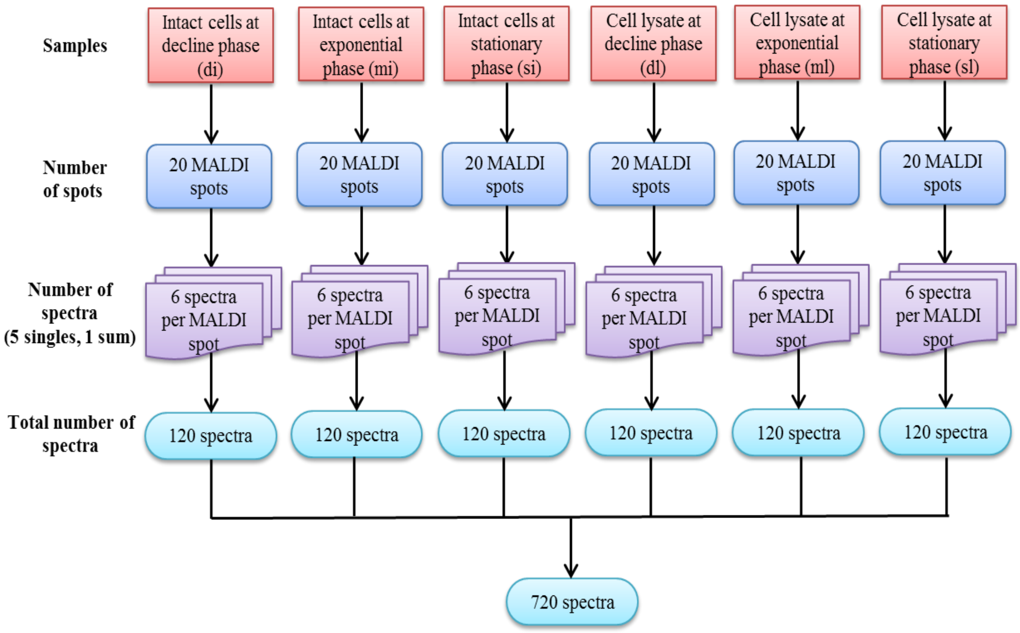

Samples
Intact cells at decline phase (di)
Intact cells at exponential phase (mi)
Intact cells at stationary phase (si)
Cell lysate at decline phase (dl)
Cell lysate at exponential phase (ml)
Cell lysate at stationary phase (sl)
Number of spots
20 MALDI spots
Number of spectra (5 singles, 1 sum)
6 spectra per MALDI spot
Total number of spectra
120 spectra
720 spectra

Figure 3.5: Schematic diagram to illustrate the MALDI spectra data collection

MALDI data acquisition was performed manually. Five single composite spectral scans were acquired from each sample spot, which were summed to give a 6th spectrum. Each composite spectra scan was the average of 20 single laser shots fired from the same location. Once the signal was depleted (that is, areas in sample consumed due to the laser irradiation), a new scan position was selected manually. As shown in Fig. 3.6, each of the 6 groups of sample was spotted 20 times to obtain sufficient spectra for data modelling. The samples were named as di (intact cell pellets for cultures at decline phase), mi (intact cell pellets for cultures at mid-log or exponential phase), si (intact cell pellets for cultures at stationary phase), dl (cell lysate for cultures at decline phase), ml (cell lysate for cultures at exponential phase), and sl (cell lysate for cultures at stationary phase). So for each of the 20 spots per sample, 6 spectra per MALDI spot (5 single and 1 sum), 120 spectra were obtained to give a total of 720 spectra for the six samples. After data acquisition, the raw mass spectra data were saved and exported as text files.

Instrument Calibration

Before MALDI analysis, external mass calibration was used to calibrate the instrument. The calibrant used as a standard was a protein mixture containing insulin (MW 5,735), ubiquitin (MW 8,566), cytochrome C (MW 12,361), myoglobin (MW 16,952) and myoglobin (MW 8,477) (Bruker Daltoniks, Germany), all covering the mass range 4,000 - 20,000Da. About 50µL of SA was pre-mixed with calibrant by pipetting up and down until the solution became cloudy. The solution was kept ready for use. About 1µl of calibrant was spotted onto of the MALDI target plate onto six specific sample spots surrounded by 20 spots containing samples to be analysed (Table A.2, Appendix A). The instrument used the conventional three-term calibration equation as follows:

$$ToF = C_o + C_1(m/z)^{1/2} + C_2(m/z)^{3/2} \tag{3.4}$$

where, C_o represents any internal delay in the acquisition system, $C_1(m/z)^{1/2}$ is the time-of-flight (ToF) of an ion with zero initial velocity from the target surface to the detector, and $C_2(m/z)^{3/2}$ is a small flight time correction for the ion velocities at the onset of the extraction pulse (Moskovets & Karger, 2003). As can be seen from Tables 3.1 and 3.2 the ToF of the molecular ions (with their accurately known reference masses) recorded by the instrument were used to calculate the constants C_0, C_1 and C_2 from equation 3.4.

Fit result			
Equation terms	**Values**		
C_0	425.47		
C_1	1256359.66		
C_2	0.00		
Initial ppm	1500.00		
Result ppm	266.81		
Calibration result			
Ions	**Reference mass**	**Current mass**	**Error (100 ppm)**
Insulin $[M+H]^+$ avg	5734.52	5736.51	347.28
Ubiquitin $[M+H]^+$ avg	8565.83	/	/
Cytochrome C $[M+H]^+$ avg	12360.97	12357.95	-244.03
Myoglobin $[M+H]^+$ avg	16952.31	16955.55	191.00
Cytochrome C $[M+2H]^{2+}$ avg	6181.05	/	/
Myoglobin $[M+2H]^{2+}$ avg	8476.66	8474.45	-261.00

Table 3.1: Linear calibration information for E. coli cell lysate samples

Fit result			
Equation terms	**Values**		
C_0	427.10		
C_1	1259863.64		
C_2	0.00		
Initial ppm	1500.00		
Result ppm	250.20		
Calibration result			
Ions	**Reference mass**	**Current mass**	**Error (100 ppm)**
Insulin $[M+H]^+$ average (avg)	5734.52	5735.97	252.43
Ubiquitin $[M+H]^+$ avg	8565.83	/	/
Cytochrome C $[M+H]^+$ avg	12360.97	12356.47	-364.45
Myoglobin $[M+H]^+$ avg	16952.31	16956.02	218.91
Cytochrome C $[M+2H]^{2+}$ avg	6181.05	/	/
Myoglobin $[M+2H]^{2+}$ avg	8476.66	8476.01	-77.03

Table 3.2: Linear calibration information for intact E. coli cell samples

The calibration molecular ions were a singly charged protonated molecules of insulin, ubiquitin, cytochrome C and myoglobin, and doubly charged protonated molecules cytochrome C and myoglobin. The constants were then used to calculate the current masses of the molecular ion to generate a calibration. The ToF of molecular ions from samples analysed by the instrument are subsequently converted to *m/z* ratio via the calibration (equation 3.3). The error given in parts per million (ppm) is the difference between the reference and current masses of the calibration molecular ions. The mass accuracy in the instrument was 100 ppm, i.e., for the proteins, experimental MW masses should be within 0.01% of their theoretical MWs (+/- 1 mass unit for 10,000 MW protein. After performing the calibration, the spectra were preprocessed in order to remove systemic errors.

3.3.6.2. MALDI-ToF Analysis of IgG Monoclonal Antibody Producing CHO Cell Lines

Sample Preparation for MALDI-ToF Analysis

Samples were prepared in accordance with protocols determined by collaborators of this project at the School of Biosciences, University of Kent, UK. Aliquots of required volumes of intact cells (0.5×10^6 cells) from exponential cultures in 96 deep well plate (DWP) were pelleted in microfuge tubes using a centrifuge at 956 × *g* for 5 minutes. The supernatant was removed and 1mL of ice-cold PBS buffer was added to the cell pellet, with gentle pipetting up and down to resuspend cells. Cells were centrifuged at 3000 rpm for 5 minutes again and the PBS buffer was removed. Cell pellets were then resuspended in 0.35M of sucrose (previously prepared and stored at -20^{o}C) with gentle pipetting up and down. The cell samples were kept in an ice bath in-between the washing steps and centrifugation. Finally, after another round of centrifugation, at956 × *g* for 5 minutes, the cell pellets were transferred to -70^{o}C.

Sinapinic acid (SA, Sigma; 20 mg/mL), was prepared fresh for 20 samples, by dissolving in 1.2mL of buffer and made up with buffer. This was mixed thoroughly and sonicated for 15 minutes in a water bath. The matrix solution was spun at 956 × *g* for 5 minutes after which 50µL of the solution was added to each cell pellet sample which had been removed from -70^{o}C, and allowed to reach room temperature. Cells were then re-suspended in the sinapinic acid solution by pipetting up and down in order to dislodge clumps of cells. All samples were then transferred to 4^{o}C and left for 3hours prior to spotting (1µL of sample each) on the MALDI plate.

A two-layer method was used for matrix/analyte sample preparation to spot the sample for MALDI analysis. In this two-layer method, 1µL of sample suspension was deposited onto a sample spot of a MALDI target plate (MSP 96 target, ground steel; Bruker Daltonics, Germany) and was allowed to air-dry at room temperature. Finally, 1µl of the saturated SA matrix solution was deposited onto the dried sample and allowed to dry at room temperature for co-crystallisation to take place. The plate was ready for MALDI analysis. MALDI data analysis was carried out as explained in section 3.3.6.1.

3.4. Mass Spectra Data Preprocessing

3.4.1. Introduction

A MALDI-ToF mass spectrometry (MS) instrument generates a mass spectrum with each being the result of two measurements, *m/z* ratio and intensity, that is corrupted by noise due to data acquisition issues. The first step in the analysis is to preprocess the spectra to remove systematic noise and bias while preserving the information content inherent within the profiles. The goal of preprocessing is to take the MS proteomic data set, and generate a data set whereby statistical techniques can be applied.

Issues with the data acquisition can be divided into two areas:

- Flawed experimental technique. This includes samples prepared utilising different procedures, spectra data sets not acquired randomly to minimise systematic errors; and comparing spectra acquired with different MALDI instruments. In this situation the experimental processes need to be addressed (Baggerly *et al.*, 2004).
- Instrument miscalibration, noise, and variation in sensitivity. Problems associated with these issues can be minimised by applying preprocessing methods described in this section (Monchamp *et al*., 2007).

Preprocessing aims to consider the issues of (i) reducing noise (ii) reducing the amount of data, and (iii) ensuring the spectra are comparable.

Preprocessing of mass spectra includes a number of techniques including signal resampling, baseline correction, *m/z* ratio alignment, intensity normalisation smoothing/filtering, and peak identification (Hilario and Kalousis, 2008; Monchamp *et al*., 2007). However, these tasks are inter-related and different combinations may have to be tested to identify an acceptable procedure. A good preprocessing method should be able to eliminate differences between spectra profiles as a consequence of experimental and instrumental procedures, while preserving the inherent biological information within the spectra profiles (section 3.6). A number of techniques for MALDI-ToF MS data preprocessing have been described in the literature and the sequence has been proposed by Monchamp *et al*., (2007) (Fig. 3.6). In this thesis, the proprocessing techniques are applied using functions from the Bioinformatics Toolbox of MathWorks.

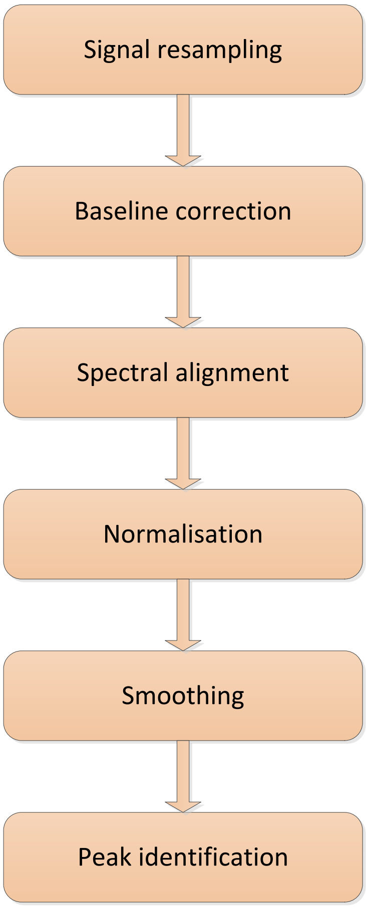


Figure 3.6: Typical preprocessing task workflow for mass spectra data (Monchamp et al., (2007))

3.4.2. Signal Resampling

As proposed by (Monchamp *et al*., 2007), resampling is the first preprocessing technique to be performed. Resampling is the process of calculating a new signal with intensity values at selected *m/z* ratio points. By selecting *m/z* ratio points, the signal can be down-sampled (fewer points than the original signal), up-sampled (more points) or synchronised (approximately the same amount of points).

Resampling can be used to create a constant scale for the *m/z* ratio values, allowing for the comparison of different spectra utilising the same reference *m/z* ratio vector and the same resolution. Unequal spacing of the *m/z* ratio values can occur (due to random and systematic errors) with the same instrument or when different instruments are used to generate the spectra. Resampling has several advantages:

- During down-sampling, the *m/z* ratio vector can be converted into a vector with fewer data points, whilst preserving the information content of the spectra. This is important when working high-resolution data sets, as the quantity of data can be impractical to work with using computationally intensive algorithms. If the sampling rate is higher than the resolution of the instrument redundant values may become immersed in noise. These values may be removed by down-sampling (Monchamp *et al*., 2007).
- Resampling can be used to fill "missing" values. Abundances may be "missing" for certain *m/z* ratio values so resampling can provide a value. A mass spectrometer may have trouble detecting the weak signals of low-abundance peptides during MS experiments. Even when the signal is detected by the instrument, the peak intensities may be too low to be distinguished from background noise during data processing. Therefore, the lower the ion abundance, the more likely the peptide will be "missing" in the mass spectra data. Bias maybe introduced during subsequent analyses if these "missing" values are ignored. Filling the "missing" values helps when data needs to be visualised (Wang *et al*., 2006).

Care must be taken when resampling the mass spectra profiles not to set the number of resampling units too low since information may be masked or removed due to the signals losing resolution. The '*msresample*' function from the bioinformatics toolbox of MathWorks is used for resampling (http://www.mathworks.com/products/bioinfo/demos.html). With this function, the selection of the *m/z* ratio vector is carried out by down-sampling a raw mass spectrum to give an output spectrum with the spacing between the points increasing linearly within the specified range.

Prior to down-sampling, the function prefilters the spectrum to prevent aliasing by using an antialias filter, a linear-phase finite impulse response (FIR) filter with least-squares error minimisation. Aliasing is a phenomenon whereby it is difficult to distinguish between high frequency and low frequency signals with the latter being mistaken for the former in the down-sampled spectra. The high frequency signals in MS data are mostly noise. The '*msresample*' function automatically sets the cut-off frequency to a value equal to the minimum distance between two contiguous data points within the *m/z* ratio vector range (MATLAB, 2008). This filter allows those frequencies below the cut-off frequency to remain unchanged while it suppresses those above the cut-off frequency.

As an example, the effects of resampling on raw spectra data are shown in Fig. 3.7. Fig. 3.7(a) and (b) shows a CHO cell line spectrum before and after down-sampling from 24000 to 10000 bins. Note the change in the thickness of the plots as a result of a reduction in the number of *m/z* ratio data points by eliminating redundant ones, but the relative intensity pattern of the down-sampled spectrum (Fig. 3.7(b)) looks very much like the raw spectrum.

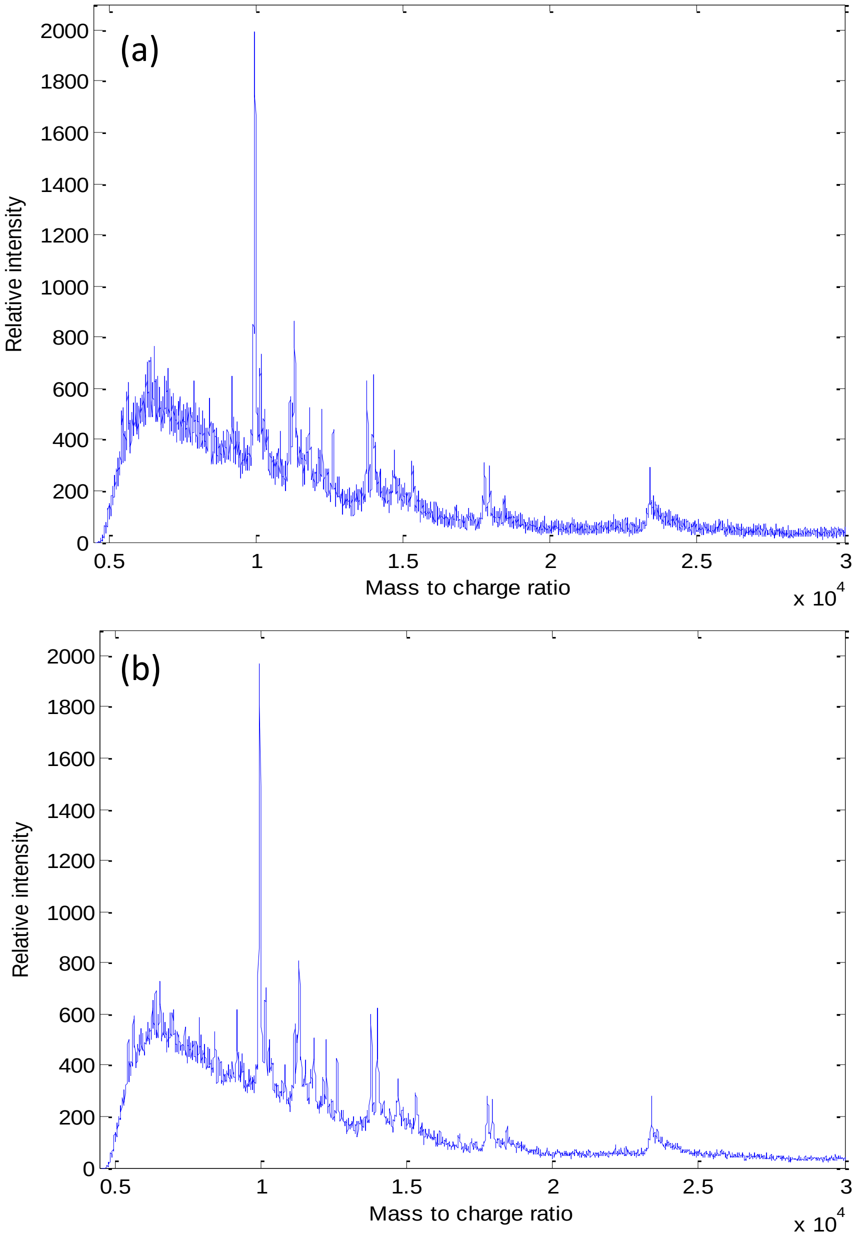


Figure 3.7: Mass spectra of CHO cell line samples illustrating the effects of resampling

Fig. 3.8(a) shows the process by which the *'msresample'* function reduces original *m/z* ratio data points in a raw spectrum of an *E. coli* cell sample to a subset which is much easier to handle. The red spots and line in the graph represent the newly calculated *m/z* ratio data points that would best fit the original raw spectra data. The blue spots represent the original *m/z* points. Fig. 3.8(b) indicates the *E. coli* cell sample spectrum with the resampled *m/z* ratio data points. Note how the resampled spectra (Fig. 3.8(b) and Fig. 3.8(b)) still show variation in baseline after the number of data points has been reduced suggesting that baseline correction is required.

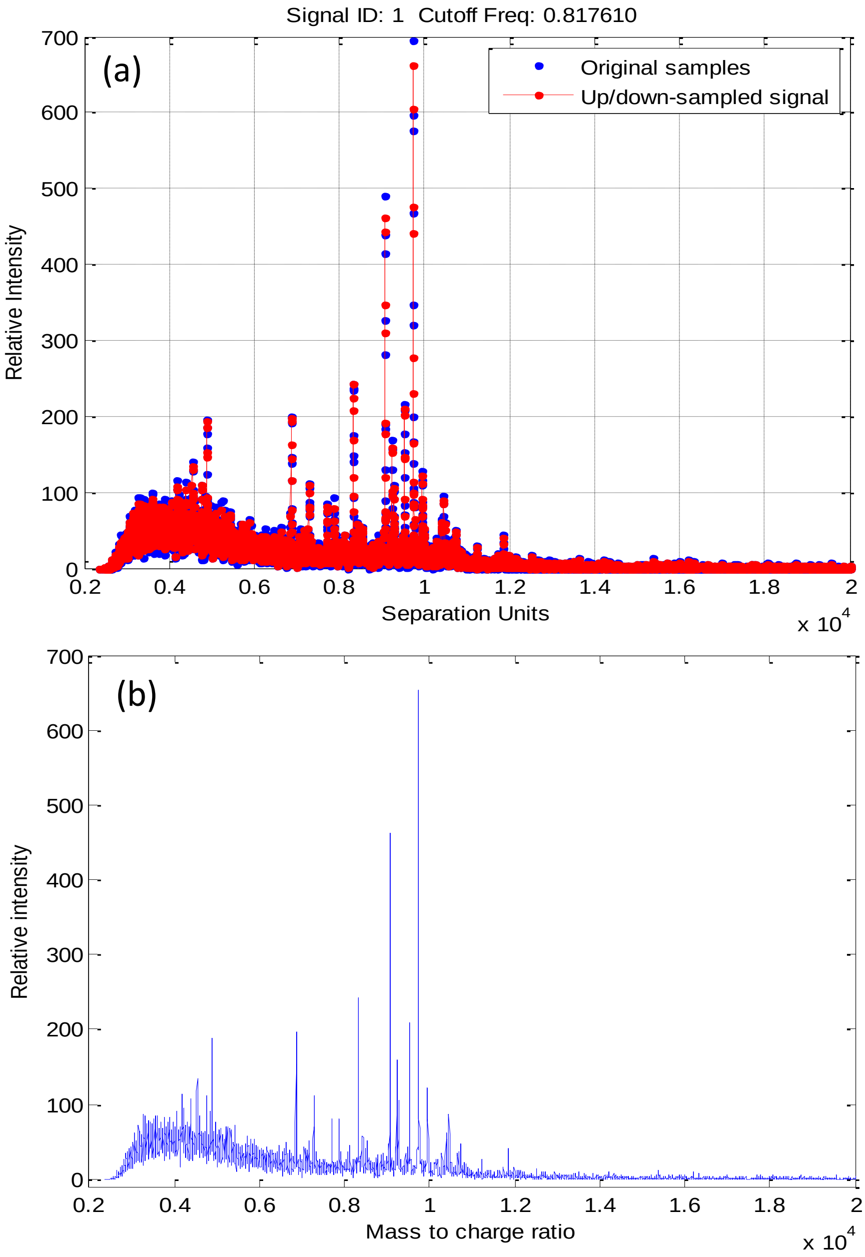


Figure 3.8: Mass spectra of E. coli cell samples illustrating the effects of resampling

3.4.3. Baseline Correction

After signal resampling, baseline correction is applied on the data. Varying baseline is caused by chemical noise from sources such as the sample ion dispensing, matrix chemical contamination, and data collection. Baseline correction is applied after resampling as it uses spectra with aligned *m/z* ratio vectors.

Observed mass spectra can theoretically be decomposed into three components (Guangtao and Wong, 2008);

$$f(i,j) = b(i,j) + s(i,j) + \varepsilon(i,j) \qquad (3.4)$$

where $f(i, j)$ is the observed value, $b(i, j)$ is the baseline value, $s(i, j)$ is the true signal and $\varepsilon(i, j)$ is the noise for ith sample at jth *m*/*z* ratio. The baseline is considered to be the low frequency component of the observed signal. Baseline variation is especially significant at low peak intensities because the signal to noise ratio is larger. The consistent decrease in baseline exhibited by MS data may be caused by the interaction of the matrix material with itself as well as with the sample proteins, during the MALDI analysis. More specifically, the baseline originates from small clusters of the matrix material and since the likelihood of cluster formation decreases with cluster size, the baseline diminishes consistently with an increase in *m/z* ratio (Shin and Markey, 2006).

Baseline correction estimates a low-frequency baseline, which is latent within the high-frequency noise and signal peaks and then subtracts this baseline to give a baseline corrected spectrum. This is done in three steps: (1) the baseline in a small window of width 200 *m/z* ratio is first estimated; (2) spline interpolation is then performed to regress the varying baseline to the estimated window baseline; and (3) the estimated and regressed baseline is subtracted from the spectrum (MATLAB, 2008). The bioinformatics toolbox of MathWorks uses the '*msbackadj*' function to perform baseline correction (http://www.mathworks.com).

In baseline correction, an iterative algorithm uses quantile value (i.e. equal proportions of spectral points are taken at regular intervals), for the observed value within a window to remove the varying baseline from a spectrum by iteratively calculating the best fit straight line through a set of estimated baseline points (Fig. 3.9). As demonstrated in

Fig. 3.9, the number of points above and below the best fit straight line (line in red) is then counted. If there are fewer points above the line than below, they are considered peaks and discarded. A new line is then fitted through the remaining data points. After repeating the process until the number of points above the line is greater than or equal to those below the line, this final line is subtracted from the spectrum to get the baseline corrected spectrum (Veltri *et al*., 2008).

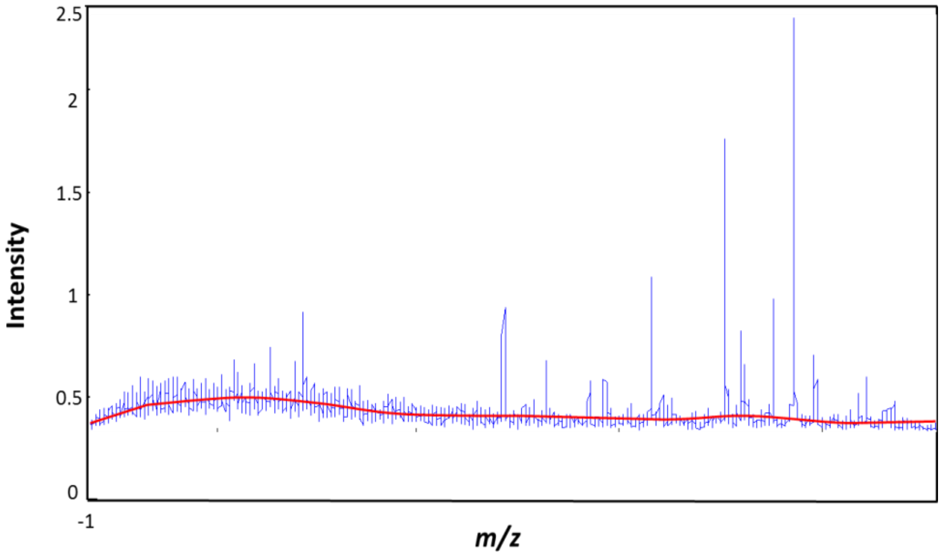


Figure 3.9: An example baseline correction within a window

Fig. 3.10 shows examples of raw mass spectrum of a CHO cell line (a), and the baseline corrected spectrum (b). Fig 3.11(a) shows how the '*msbackadj'* function calculates the new baseline of the resampled spectrum of *E. coli* cell samples. Note the red line indicates the newly estimated baseline. Fig. 3.11(b) is the baseline corrected spectrum of the *E. coli* cell clearly showing a uniform baseline.

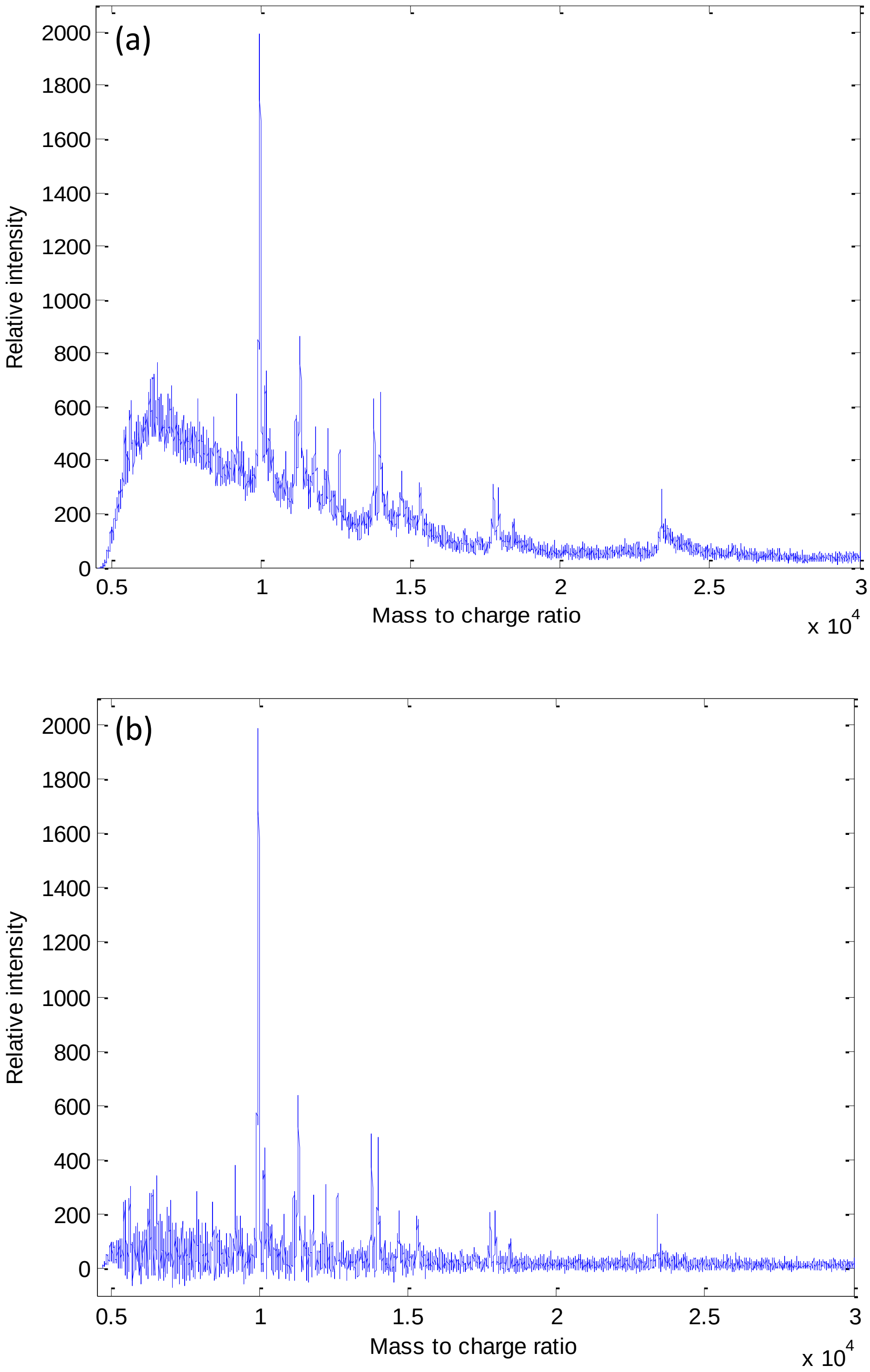


Figure 3.10: Mass spectra of CHO cell line samples demonstrating the effects of baseline correction

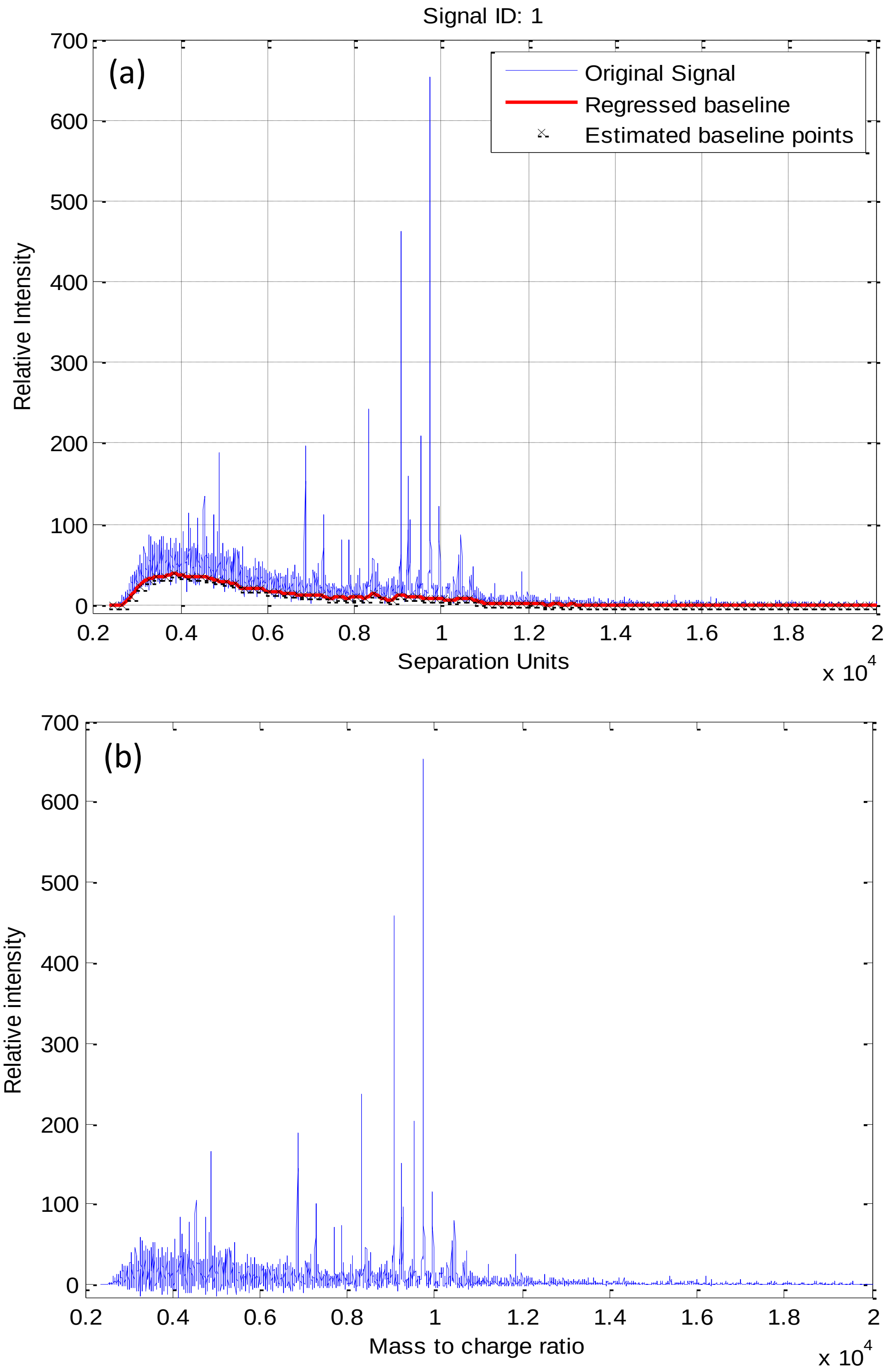


Figure 3.11: Mass spectra of E. coli cell samples demonstrating the effects of baseline correction

3.4.4. Spectra Alignment

Instrument measurement error (0.03% - 0.06%) can cause miscalibration leading to variations in the relationship between the observed *m/z* ratio vector and the true ToF of the ions. Therefore, systematic shifts can appear in repeated experiments and the spectra for two identical proteins acquired can have different *m/z* ratio values. Alignment consists of aligning corresponding peaks across samples to address this problem (Veltri, 2008).

Alignment is usually applied when:

- Known profiles of peaks are expected in the spectrogram of a biological sample which can be used to standardise the *m/z* ratio values. This can allow an easy and effective comparison of different spectra;
- Internal and external calibration standards of known spectral profiles to aid calibration show variation in *m/z* ratio peak positions. Calibration standards are usually a set of known proteins which are expected to appear with reference peaks at specific *m/z* ratio points. However, there may be a slight shift across spectra with respect to the *m/z* ratio points of the reference peaks. So alignment needs to be performed to adjust the known peaks from the celebrants to their correct location.

During alignment, a smooth function (which can be any higher-order polynomial) twists the spectral signals by resampling them. It builds a new signal with two or more peaks represented by a normal distribution. The *m/z* ratios of the new signal are shifted and adjusted until a cross-correlation between the mass spectrum and the new signal reaches a maximum value closest to the true peak location. After determination of a new *m/z* ratio vector, a new spectrum is calculated by piecewise cubic interpolation and shifted from the original *m/z* ratio vector (Monchamp *et al*., 2007).

Alignment can be carried out by the '*msalign'* function from the bioinformatics toolbox of MathWorks (http://www.mathworks.com). Fig. 3.12 and Fig. 3.13 shows an overall spectra heat map (Fig. 3.12(a) and (b)) and two overlaid spectra (Fig. 3.13(a) and (b)) of an *E. coli* cell sample demonstrating the alignment of peaks in the original spectra. Baseline-corrected spectra were aligned to the reference peaks of 6411, 6855, 7273, 7333, 7869, 9061, 9218, 9532, and 9736. A close look at the heat maps show the spectra peaks are more aligned along the *m/z* ratio reference peaks after alignment (Fig. 3.12(b)) than before alignment (Fig. 3.12(a)). After alignment, Figs. 3.13(a) and (b) show observed improvements in peak alignment between spectra based on peak height.

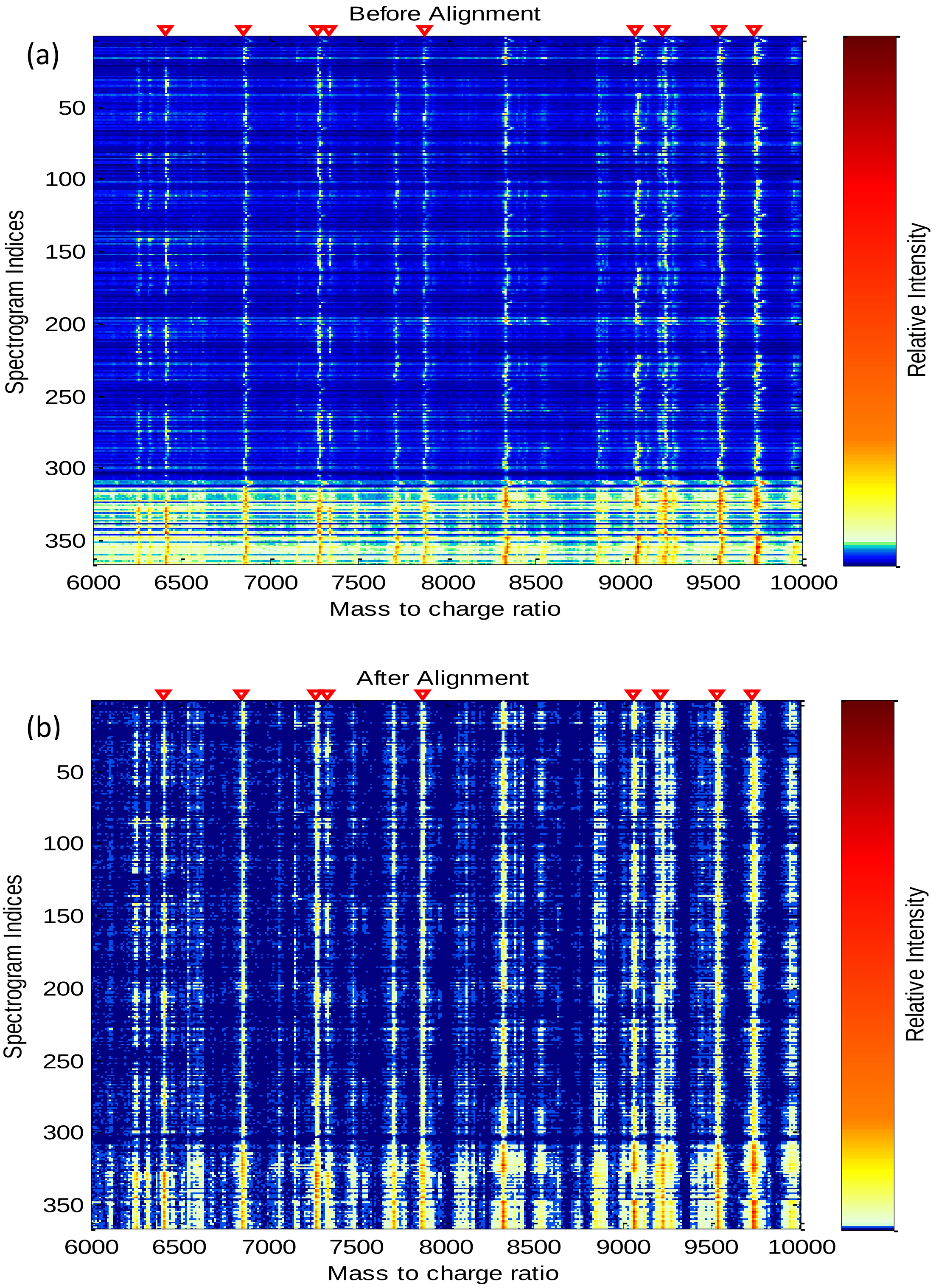


Figure 3.12: Heat maps of mass spectra of E. coli cell samples illustrating the effects of alignment

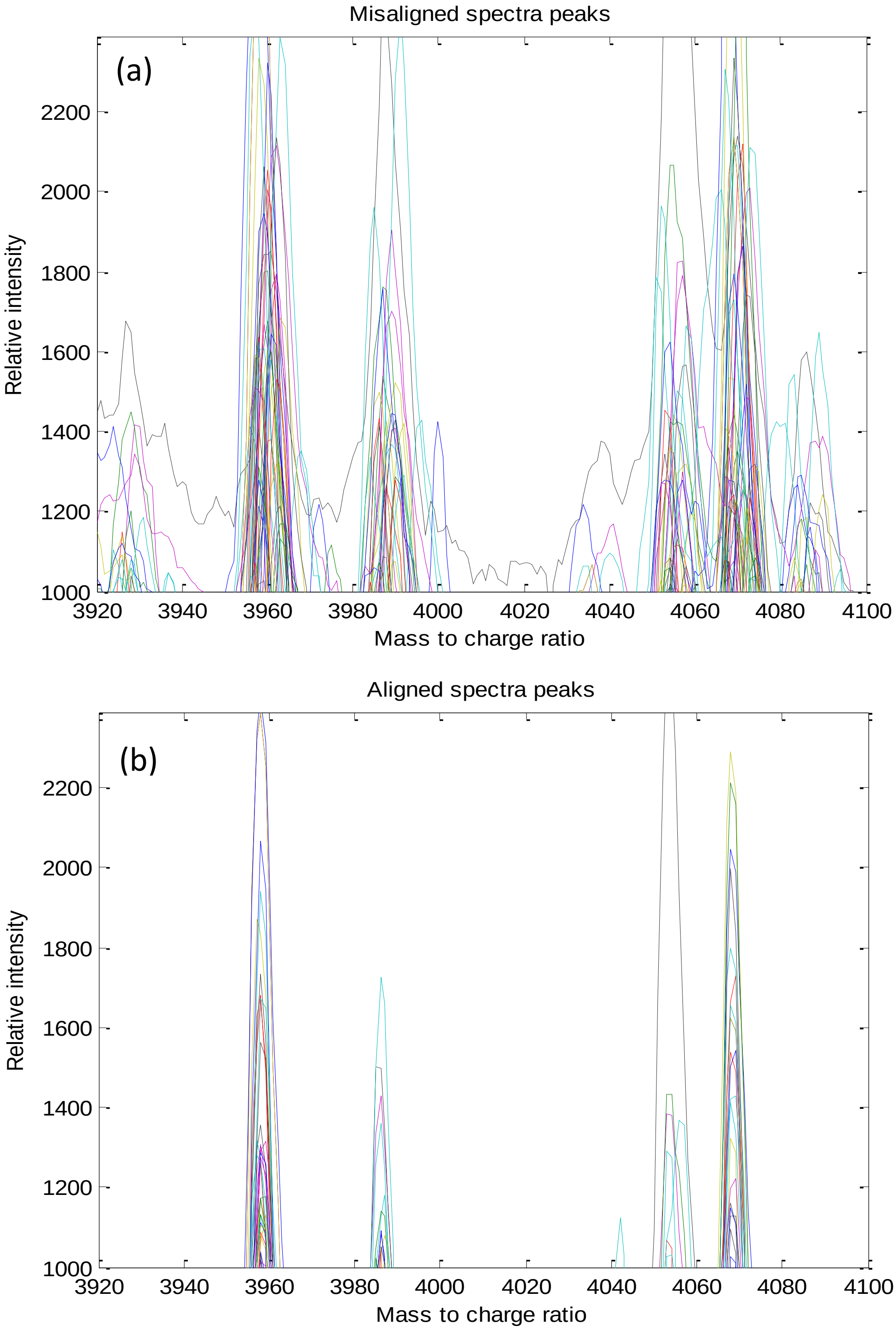


Figure 3.13: Mass spectra of E. coli cell samples illustrating the effects of alignment

3.4.5. Normalising the Relative Intensity

In repeated experiments, it is also common to find systematic differences in the total amount of desorbed and ionised proteins. Sample size may differ, sample preparation may not be consistent between different technicians, there may be ion saturation or changes in the sensitivity of the instrument. This may result in variations in the amplitude of the ion intensities. Normalisation is a row-oriented transformation used to force the intensities of the data to the same scale thereby enabling the comparison of different samples (Veltri, 2008).

There are a number of normalisation methods (Fung and Enderwick, 2002; Wang *et al.,* 2003). In the bioinformatics toolbox of MathWorks, the normalisation function, '*msnorm*', provide a number of options (http://www.mathworks.com):

Area normalisation: To compensate for systemic differences, the relative intensities of the spectra are normalised to the average area under the spectra curves or the height of a selected peak. The area under the spectra curve (AUC) can be defined as,

$$AUC = \sum_{i=1}^{n} X_i \tag{3.5}$$

where X_i is the signal at *i*th *m/z* ratio, and *n* is number of *m/z* ratio values. Normalisation can be carried out by dividing the spectra signals by a constant,

$$X_{i,1:n}^{normalised} = \frac{X_{i,1:n}}{A_i} \tag{3.6}$$

where X_i is the signal at *i*th *m/z* ratio, *n* is number of *m/z* ratio values, and A_i is a constant.

Area or height of an internal standard: A second normalisation method uses the area or height of an internal standard. The internal standard can be a compound with a known mass and the same amount of the substance is added to each analyte. Differences in the area of the internal standard should be proportional to the differences in the area for proteins in the analyte. For example, the maximum intensity of every signal can be rescaled to a specific value, for instance 100, with respect to the highest peak in the signal. It is also possible to ignore problematic regions; for example, the low-mass region (*m/z* ratio < 4000 Da) which may be due to matrix molecules may be ignored.

This is done by choosing a threshold value that eliminates the large amount of noise at the lower *m/z* ratio values but does not remove any important proteins (MATLAB, 2008). The example shown in Fig. 3.14 and Fig. 3.15 shows the effects of normalisation on spectra. It is a mass spectrum before and after normalisation of a CHO cell line sample (Fig. 3.14(a) and (b)) and an *E. coli* cell sample (Fig. 3.15(a) and (b)). Notice the change in scale of the relative intensity axis of the spectra before and after normalisation because the base peak (largest peak in the spectrum) was rescaled to 100%. All other peaks were then normalised based on the respective base peaks.

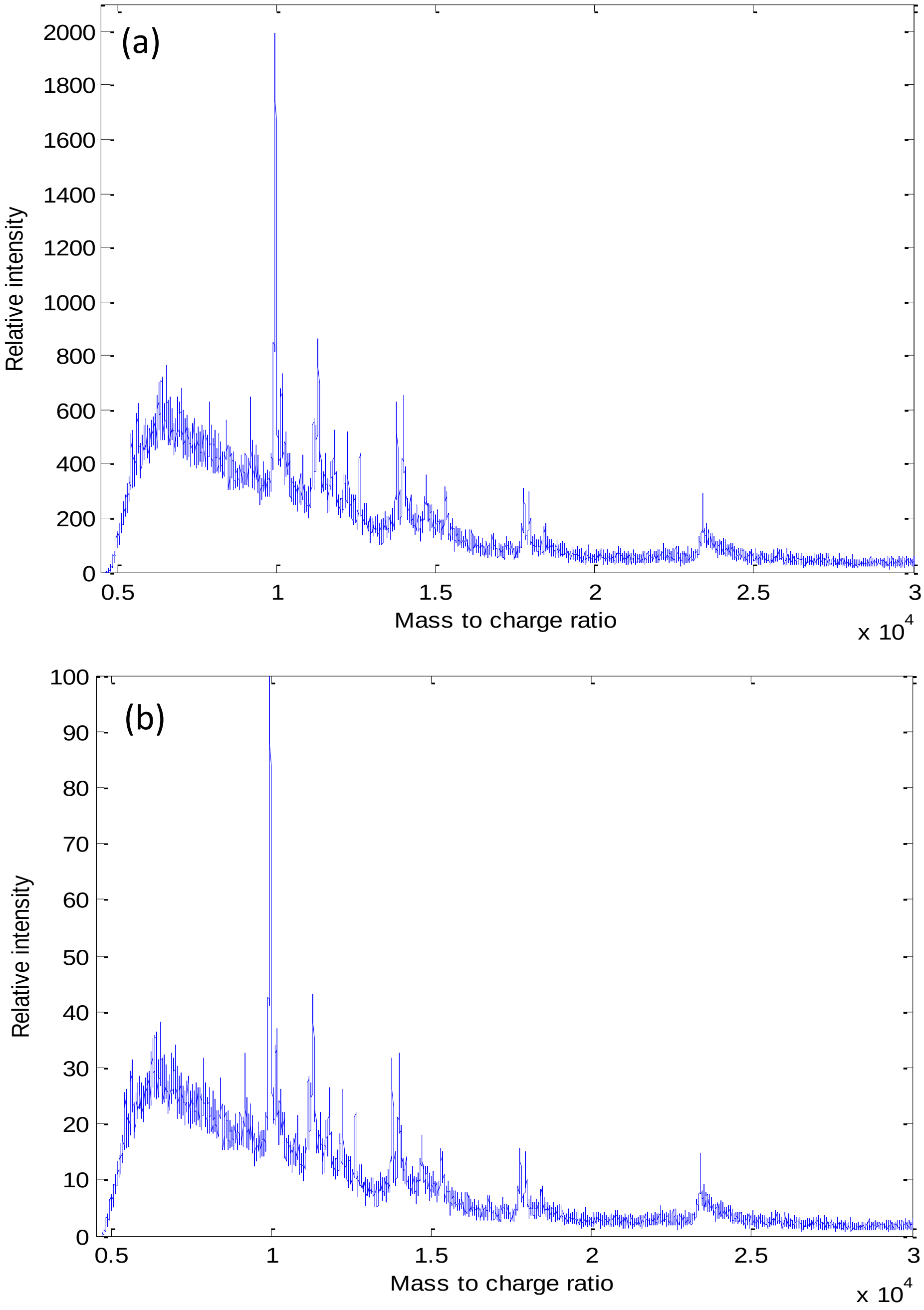


Figure 3.14: Mass spectra of a CHO cell lines sample illustrating the effects of normalisation

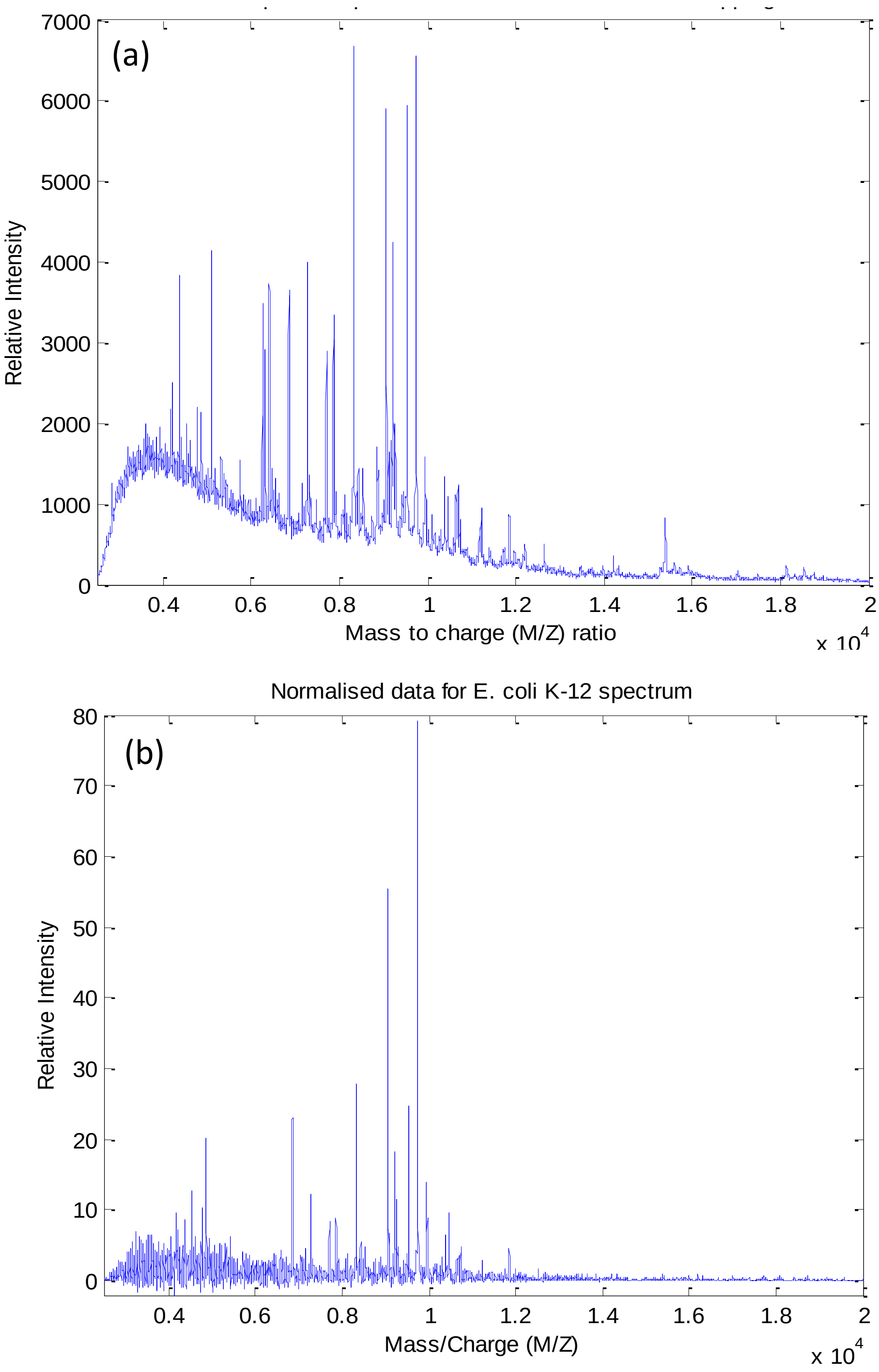


Figure 3.15: Mass spectra of E. coli cell samples illustrating the effects of normalisation

3.4.6. Smoothing (Noise Filtering)

Standardised spectra usually contain a mixture of noise and signal. Noise can be defined as unwanted signal interfering with the clarity of desired signals. Some applications require the denoising of the spectrograms to improve the validity and precision of the observed *m/z* ratio values of the peaks in the spectra. Additionally, denoising facilitates the application of peak detection algorithms to select significant features as noise which may be confused for peaks are removed. Filtering usually involves removing high frequency noise (Monchamp *et al*., 2007). Filtering is usually carried out after resampling, baseline correction, alignment and normalisation. There are two main smoothing techniques reported in the literature;

Lowess filter smoothing: Lowess filters smoothes a mass spectrum by using a locally weighted linear regression method. In summary, Lowess smoothing finds a data value by averaging the values within a span of data points. Smoothing is directly proportional to the span size (i.e. the segment size containing a specific number of *m/z* ratio data points), so care must be taken when choosing the span size as a large span size may lead to information loss. For example a span size of 10 means performing a locally weighted regression smoothing algorithm by applying a full least-square fit to the 10 *m/z* ratio data points within the span. The step is repeated for every point in the signal (Monchamp *et al*., 2007).

Savitzky-Golay filters: This technique (Savitzky-Golay, 1964), smooths mass spectra using a least-squares digital polynomial filter. Digital polynomials have points with coordinates often referred to as pixels. This method of smoothing is basically a generalisation of the Lowess method. The filter coefficients can be derived by performing a linear least squares fit using a polynomial of a given degree. As a result, the algorithm preserves signal features such as the resolution between ion peaks and the height of the peaks. A higher degree of polynomial will fit the data better. Smoothing is controlled by the span size and the polynomial order. The data at both ends are truncated and the larger the segment size, the more the smoothing. For instance for low-resolution mass spectra data, the span sizes commonly used are 15 - 20 (Monchamp *et al*., 2007).

The '*mssgolay*' function of the bioinformatics toolbox of MathWorks uses the Savitzky-Golay filter to carry out smoothing (http://www.mathworks.com). Fig. 3.16 shows an example of mass spectra before and after smoothing. Fig. 3.16(a) and (b) shows the spectrum of a CHO cell line sample before and after smoothing respectively. A portion of the spectrum has been enlarged so that the high frequency noise is apparent. As expected the scale of the noise within the raw spectrum is decreased after filtering. Fig. 3.17(a) demonstrates how the *'mssgolay'* function filters a resampled, baseline corrected and normalised spectra of an *E. coli* cell sample. In the figure, a raw spectrum (blue) can be seen with overlays of Savitzky-Golay smooths (green). Fig. 3.17(b) shows the completely smoothed spectrum.

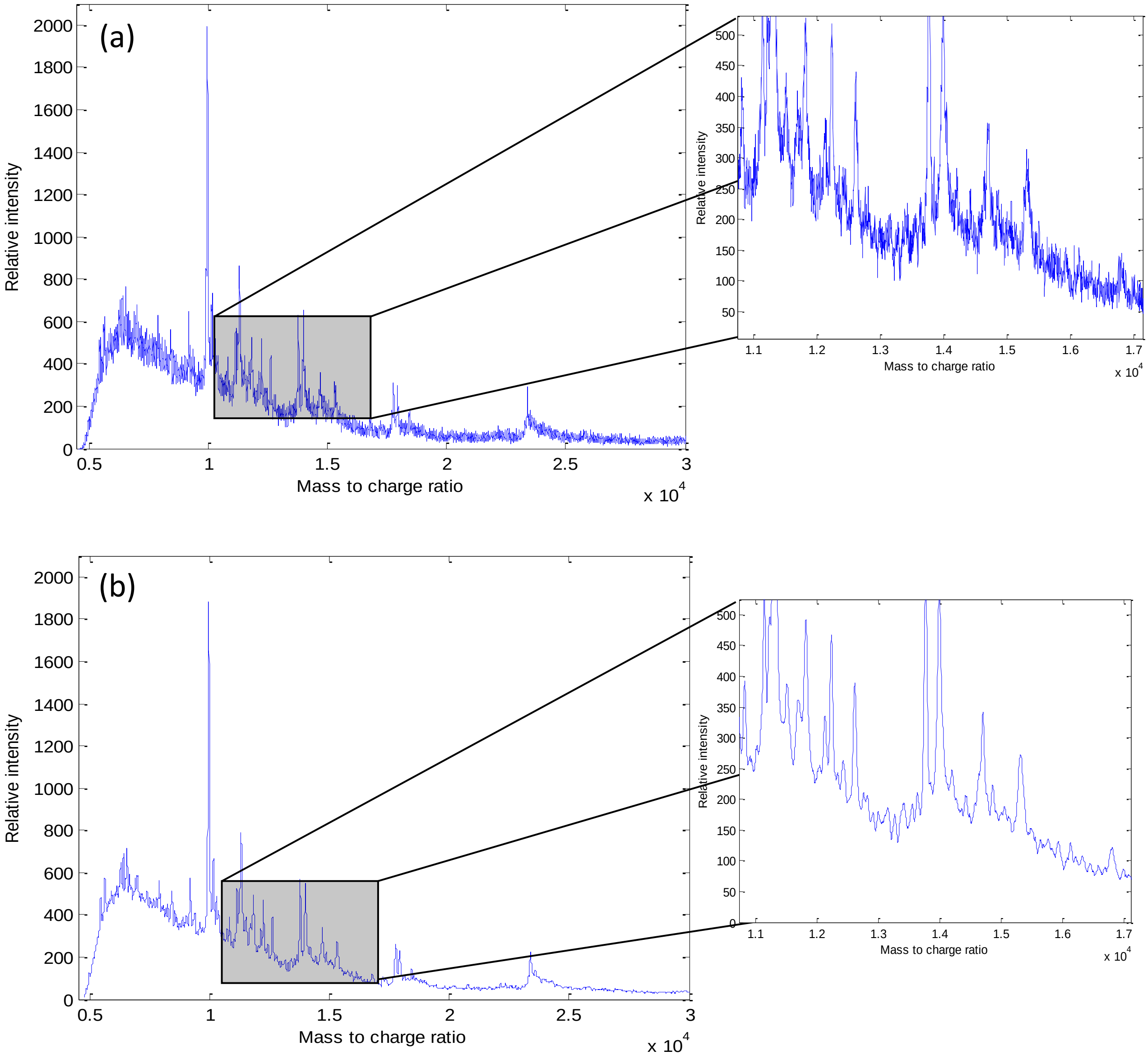


Figure 3.16: Mass spectra of CHO cell line samples illustrating the effects of noise filtering

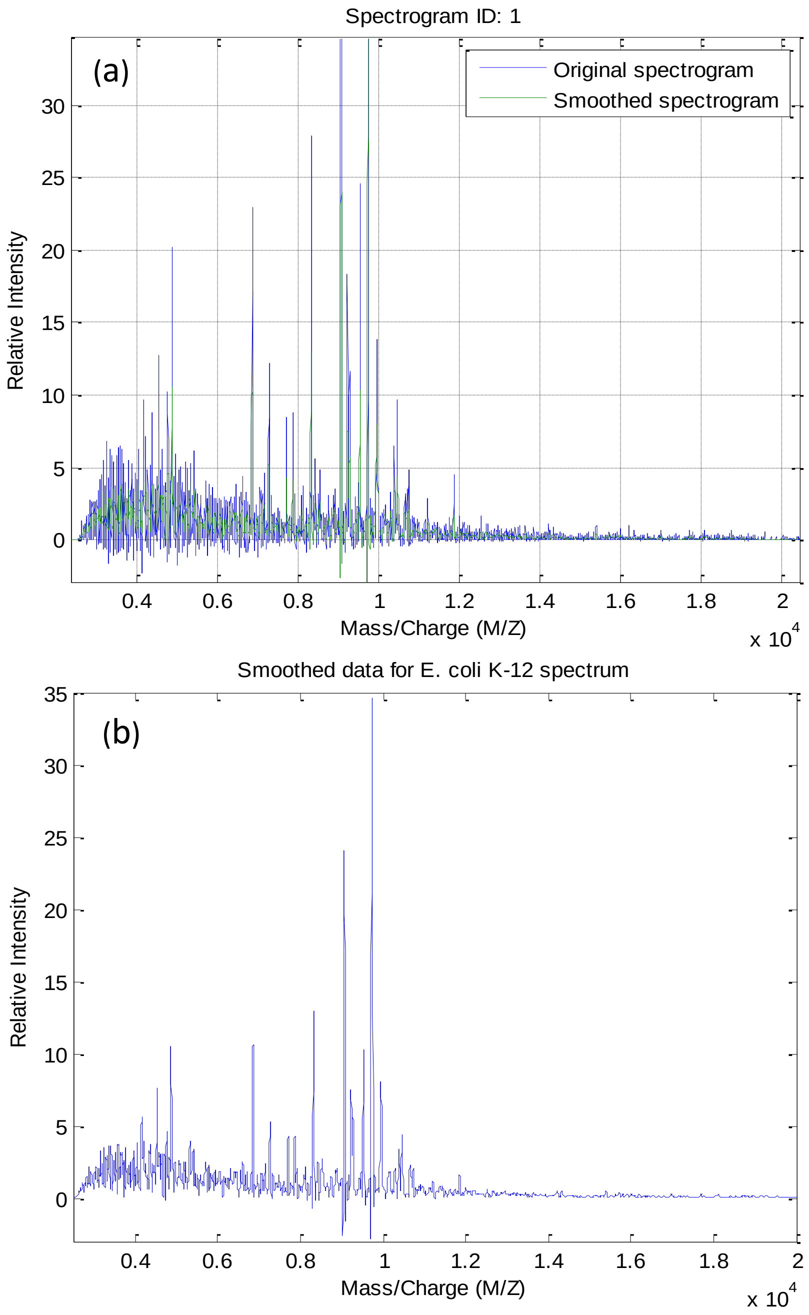


Figure 3.17: Mass spectra of E. coli cell samples illustrating the effects of noise filtering

3.4.7. Peak Identification

After baseline adjustment, alignment, normalisation of the intensities and smoothing of spectra, peak identification can be considered. One approach of doing this is by looking at the first derivative of the smoothed spectra. Alternative peak detection methods uses descrete wavelet transforms (DWTs). The '*mspeaks*' function of the bioinformatics toolbox of MathWorks uses this approach to perform peak identification (http://www.mathworks.com).

As previously mentioned, a mass spectrum can contain tens of thousands (10,000 up to 1,000,000) of *m/z* ratios, each with a corresponding signal intensities. However, mass spectra can contain regions that do not contain useful information. Extracting the relevant signals from a mass spectrum is therefore a means to reduce its dimensionality. Fig. 3.18 shows two mass spectra of *E. coli* cell samples demonstrating peak identification. A relative intensity of 10 was specified for peak identification in the algorithm, that is, only peaks with a relative intensity of 10 and above were identified.

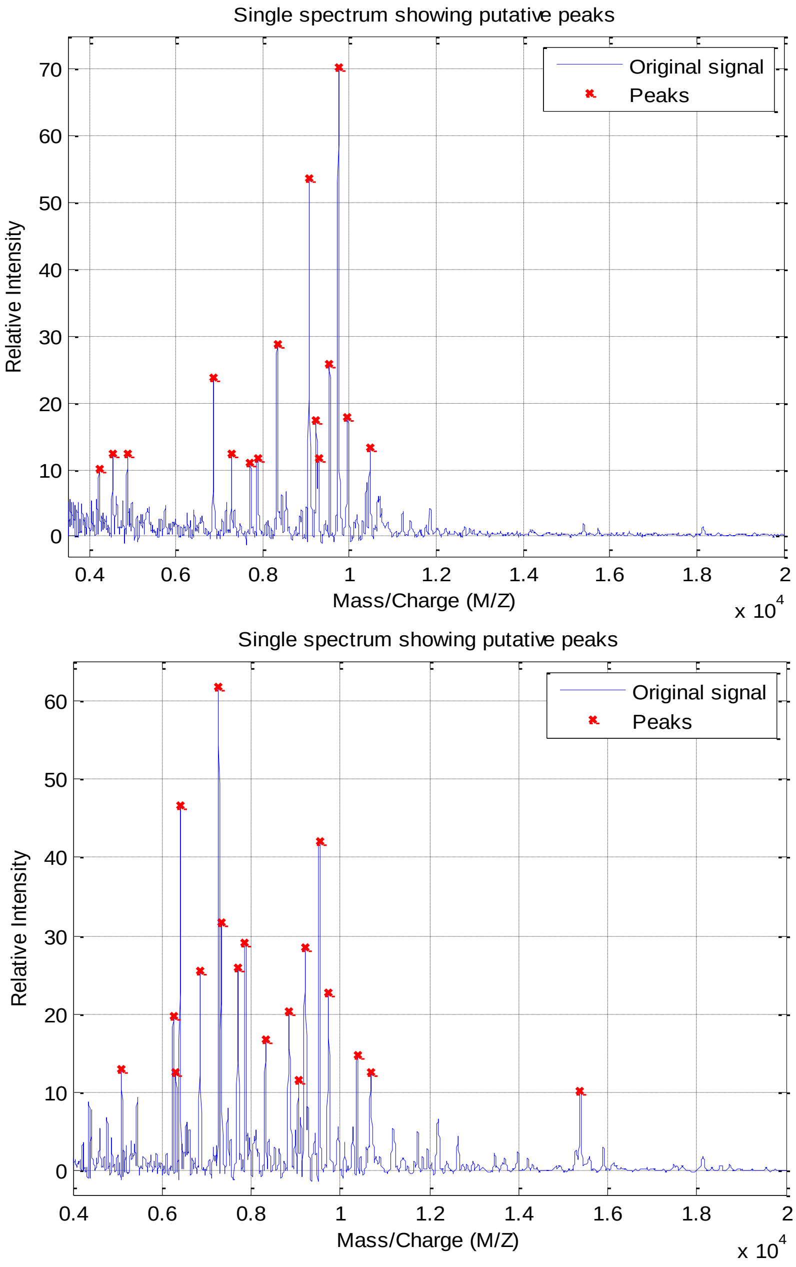


Figure 3.18: Two mass spectra of E. coli cell samples demonstrating peak identification

3.5. Review of the Applications of Mass Spectra Data Preprocessing

Several studies have been carried out to demonstrate the importance of preprocessing techniques on mass spectra data. A number of baseline correction methods have been reported in previous mass spectra proteomic studies. Tibshirani *et al.* (2004) used a logarithmic transformation on MALDI-ToF data, to reduce the dependence of signal width on mass values. By log-transforming the data, the peak widths were approximately constant across the *m/z* ratio range, correcting the baseline and facilitating peak detection, and multivariate data analysis. Filters with fast Fourier transform technique have also been used on MALDI-ToF mass spectra data to estimate the baseline. In this approach, the filters first to estimate the baseline level of the spectra and remove unwanted maxima and minima (Breen *et al.* 2000). Using a two-step algorithm, Coombes *et al.* (2003) combined baseline correction and peak detection. The algorithm first detected peaks as well as the base of the peaks. It then interpolated across the bases linearly after removing the peaks. The baseline was computed subsequently as the local minimum in a window of specified width. Finally a revised spectrum is constructed by subtracting the baseline from the original spectrum to give a final baseline-corrected spectrum.

Other approaches to alignment such as cubic splines have been proposed for mass spectra data sets (Jeffries, 2005). Du *et al.* (2006) provided a different approach where a continuous wavelet transform (CWT)-based peak detection algorithm was applied to identify peaks with different scales and amplitudes. By transforming the spectrum into wavelet space, the pattern-matching problem was simplified and in addition provided a powerful technique for identifying and separating the signal from the spike noise and coloured noise. The algorithm evaluated with surface-enhanced laser desorption/ionisation time-of-flight (SELDI-ToF) spectra data showed that no baseline removal or peak smoothing preprocessing steps were required before peak detection; and it improved the peak detection across different scales of the CWT algorithm and on spectra showing varying peak intensities.

Some studies have been undertaken where various normalisation techniques were used (Satten *et al.,* 2004; Wagner *et al.,* 2003). In one such study, Satten *et al.* (2004) proposed a standardisation procedure where the spectra is centred using a local estimate of the median spectral intensity, and divided by a local estimate of the interquartile range. The interquartile range is chosen over the standard deviation as a measure of scale because it is less likely to be sensitive to peak intensity values. Wagner *et al.* (2003) choose to normalise with respect to the sum of the intensities. Each peak intensity was divided by the sum of all peak intensities and multiplied by 1000, so that the processed intensities could be interpreted over a uniform range across fractions and samples.

Other smoothing techniques have been proposed in various studies. Barak (1995) suggested an extension to the Savitzky-Golay algorithm, in which the window width and the degree of the polynomial can be defined for adaptively every *m/z* ratio data point window size. The digital filter used varies the degree of the fitting polynomial as it slides down the *m/z* ratio data point window, leading to an improvement in noise reduction compared to Savitzky-Golay filters with optimally-chosen polynomial degrees.

3.6. Preprocessing Studies for MALDI-ToF data generated from *E. coli* K-12 cells in different growth phases

As previously explained preprocessing is important especially if biomarker identification is the final goal of mass spectra data modeling. Incorrect or inadequate preprocessing may lead to the incorrect identification of biomarkers and make it difficult to reach meaningful biological conclusions, which could have serious implications if such biomarkers are clinically relevant. In this thesis, before the mass spectra generated were subjected to multivariate data analysis for biomarker identification, preprocessing was carried out.

Preprocessing was carefully considered and empirically investigated to select an appropriate combination of preprocessing techniques and associated parameters. The parameters of the preprocessing methods were modified systematically and applied to the spectra data. The preprocessed spectra data was used to calibrate PLS-DA models (chapter 5). The qualities and performances of the constructed models using different preprocessing techniques/parameters were assessed. Factorial designs (and random designs) was used with different parameters of the preprocessing techniques defining the design variables and the PLS-DA root mean square error of prediction (RMSEP) (model performance), RMSE cross validation (RMSECV) (model quality), and R^2 of calibration and prediction (model quality and performance respectively) as the response variables to determine the optimal combination. In random design, the parameters were modified systematically (but not through factorial designs) and tested in order to increase the range of the factorial design and test parameter combinations that were not covered in the factorial design.

The preprocessing was carried out using MATLAB® v.7.6.0.324 (R2008a the MathWorks, Inc.) and functions from the Bioinformatics toolbox of MathWorks (v 3.1, R2008a, Eigenvector Research, Inc.). Sixty *E. coli* cell sample mass spectra data sets were preprocessed using the sequence in accordance with the recommendation by Monchamp *et al.,* (2007) (Fig. 3.7). The data sets were first cropped and down sampled from 12722 to 8423 data points prior to the analyses. Cropping involved removing extremes of the intensity vector where missing *m/z* ratio values were found. For example the *m/z* ratio region from 0-4000 is usually considered as noise since signals in this area are from matrix ion molecules. This region was removed during cropping.

3.6.1. Selecting the Appropriate Combination of Spectral Preprocessing Technique

3.6.1.1. Cropping and Signal Resampling

Initially, cropping and signal resampling were investigated the model quality (RMSECV and R^2) of the constructed models using different cropping and signal resampling parameters were used as output to verify the suitability of the two preprocessing methods on these mass spectra data sets. Table 3.3 and 3.4 show the percentage R^2 and RMSECV values across a number of latent variables (2 to 11) demonstrating the effect of cropping and signal resampling on the quality of PLS-DA models. As see from the tables, the percentage RMSECV of the three growth phases representing three classes were recorded as RMSECV1 (decline phase), RMSECV2 (exponential phase) and RMSECV3 (stationary phase) respectively. The average value of the three latter values was calculated and recorded. This procedure was repeated for the R^2 values.

Number of latent variables	RMSECV 1 (%)	RMSECV 2 (%)	RMSECV 3 (%)	Average RMSECV (%)	$R^2$1 (%)	$R^2$2 (%)	$R^2$3 (%)	Average R^2 (%)
2	3.3	60.7	34.1	32.7	3.3	60.7	34.1	32.7
3	35.4	63.4	45.1	48	35.4	63.4	45.1	48
4	54.2	65.4	71.3	63.6	54.2	65.4	71.3	63.6
5	60.8	64.2	72.5	65.8	60.8	64.2	72.5	65.8
6	59.6	70.9	73.9	68.1	59.6	70.9	73.9	68.1
7	64.6	71.5	74.6	70.2	64.6	71.5	74.6	70.2
8	65.4	73.2	70.7	69.8	65.4	73.2	70.7	69.8
9	67.3	72.4	73	70.9	67.3	72.4	73	70.9
10	66.9	72.5	73.9	71.1	66.9	72.5	73.9	71.1
11	66.7	68.2	74.4	69.8	66.7	68.2	74.4	69.8

Number of latent variables	RMSECV 1 (%)	RMSECV 2 (%)	RMSECV 3 (%)	Average RMSECV (%)	$R^2$1 (%)	$R^2$2 (%)	$R^2$3 (%)	Average R^2 (%)
2	49.4	29.9	38.8	39.4	2.9	61.8	32.8	32.5
3	40.1	28.9	35.2	34.7	34.8	64.9	44.8	48.2
4	33.1	28.5	24.6	28.7	55	66.1	71.7	64.3
5	30.2	29	25.9	28.4	62.3	65.3	73.2	66.9
6	31.7	27.1	24	27.6	60.3	70.1	74.9	68.5
7	31.5	26.4	24.1	27.4	64	70.7	75.4	70
8	29.9	25.8	25.8	27.2	65.7	72	72.4	70
9	29.6	26.5	27.1	27.7	67.4	71.2	71.1	69.9
10	29.3	26.6	26.9	27.6	68.8	72	71	70.6
11	28	28.9	26.7	27.9	70.5	68.1	71.7	70.1

Table 3.3: The effect of cropping (left table), no cropping (right table) and number of latent variables on quality of PLS-DA models

Number of latent variables	Resampling data points (*m/z*)	Resampling range (*m/z*)	RMSECV 1 (%)	RMSECV 2 (%)	RMSECV 3 (%)	Average RMSECV (%)	$R^2 1$ (%)	$R^2 2$ (%)	$R^2 3$ (%)	Average R^2 (%)
2	7000	2350 to 20400	49.3	29.8	38.9	39.3	3.1	62	32.6	32.6
3	7000	2351 to 20400	40.1	28.9	35.2	34.7	34.8	64.9	44.8	48.2
4	7000	2352 to 20400	32.8	28.5	25.2	28.8	55.6	66.1	71.9	64.6
5	7000	2353 to 20400	30.1	28.9	24.5	27.9	62.6	65.5	73.4	67.1
6	7000	2354 to 20400	31.4	27.2	23.9	27.5	60.9	69.9	75	68.6
7	7000	2355 to 20400	31.6	26.5	24.1	27.4	64	70.6	75.5	70
8	7000	2356 to 20400	29.9	25.8	25.7	27.2	65.8	72	72.5	70.1
9	7000	2357 to 20400	29.7	26.4	27.1	27.7	67.3	71.2	71.1	69.9
10	7000	2358 to 20400	29.5	26.6	26.9	27.7	68.6	72	70.9	70.5
11	7000	2359 to 20400	28.1	28.8	26.8	27.9	70.6	68.1	71.8	70.2

Table 3.4: The influence of resampling and number of latent variables on quality of PLS-DA model

Resampling was carried out after cropping the data. The data sets were resampled by down sampling to 7000 bins within the range 2350 to 20400 *m/z* ratio values. Initially, attempts to resample to other data points such as 8000, 9000, 10000 or 11000 *m/z* ratio values were unsuccessful, rejected by the software suggesting that discriminatory information may be lost from data. The RMSECV and R^2 values were plotted against the latent variables (Table A.3 in Appendix A and Fig. 3.19). As seen in Fig. 3.19, results suggest that LV2 had the worst model qualities with low average R^2 (32.5%) with a correspondingly high RMSECV (39.3%), which increases until LV4 is reached, from where the graphs start levelling out. This suggests that to obtain models with good qualities, LV4 to LV11 should be considered for the models. Furthermore, there was no change on R^2 or RMSECV to raw spectra, and spectra that was cropped and/or resampled. This suggest that cropping and resampling may not be necessary or suitable for this data set, at least on their own, as far as the qualities of predictive models are concerned. Probably they may be more effective if applied alongside the other preprocessing techniques of baseline correction, normalisation, alignment and smoothing.

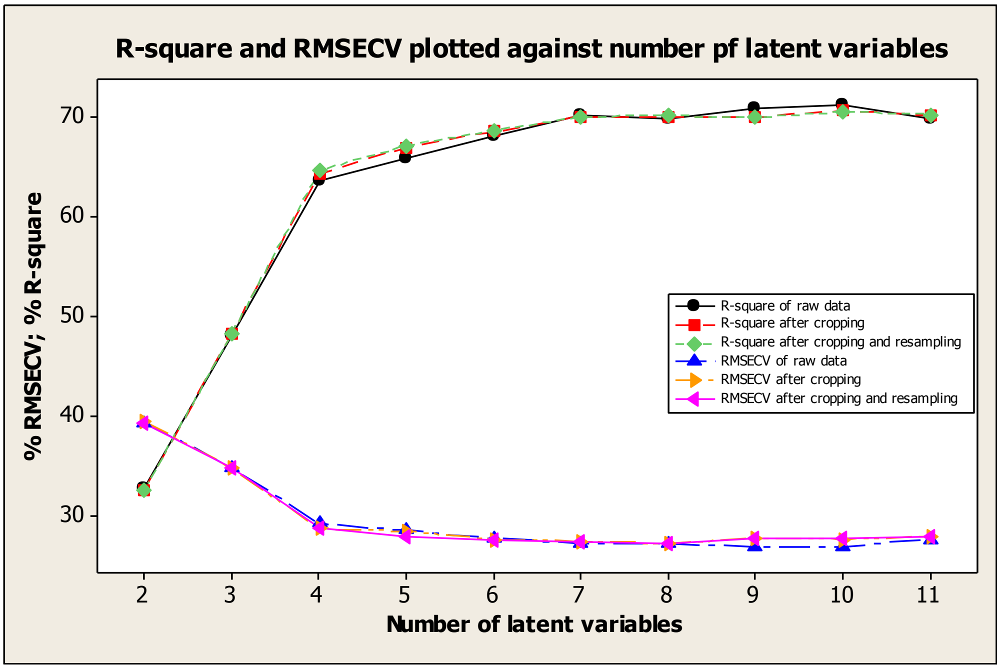


Figure 3.19: Graphs showing the effect of cropping and resampling on quality of PLS-DA models across latent variables 2 to 11

3.6.1.2. Baseline Correction

Table 3.5 and 3.6 show the recorded values of RMSECV and R^2 for the three classes and their averages respectively for different latent variables and parameter settings after baseline correction. Since the raw data had a vertical shift of in the mass spectra profile at the lower *m/z* ratio region, baseline correction on these mass spectra data set was a necessity. As summarised in Table 3.5, the window size was fixed at 200 and the quantile value at 0.2; LV4 gave the optimal model with the highest R^2 value (62.5%) and the lowest RMSECV (3.01%) This is demonstrated in Fig. 3.20. As seen in Fig. 3.20, R^2 increased sharply until LV4 is reached while RMSECV fell sharply to the same LV; from where they level out up to LV11. Thus the optimal parameter settings (window size and quantile vale) for baseline correction were tested on LV4 (Table 3.6).

Number of latent variables	Window size	Quantile value	RMSECV 1 (%)	RMSECV 2 (%)	RMSECV 3 (%)	Average RMSECV (%)	$R^2$1 (%)	$R^2$2 (%)	$R^2$3 (%)	Average R^2 (%)
2	200	0.2	48.9	31	40.3	40.1	3.4	60	27.9	30.5
3	200	0.2	39.6	29	37.8	35.5	36.1	65.3	37.9	46.4
4	**200**	**0.2**	**33.9**	**28.9**	**27.5**	**30.1**	**54.6**	**65.9**	**66.9**	**62.5**
5	200	0.2	32.6	27	28.3	29.3	56.8	69.7	65.4	64
6	200	0.2	32	26.7	29	29.2	60.3	70.6	65.1	65.3
7	200	0.2	29.7	26.2	26	27.3	67.4	71.6	71.9	70.3
8	200	0.2	27.8	25.3	26.3	26.2	69.2	73.2	71	71.1
9	200	0.2	27.4	25.6	25.8	26.3	71	72.7	72.4	72
10	200	0.2	25.9	27.4	27.3	26.9	73.2	69.8	70.1	71.1
11	200	0.2	22.5	25.2	25.1	24.3	78.5	75.8	74.2	76.2

Table 3.5: Influence of number of latent variables after baseline correction on quality of PLS-DA models

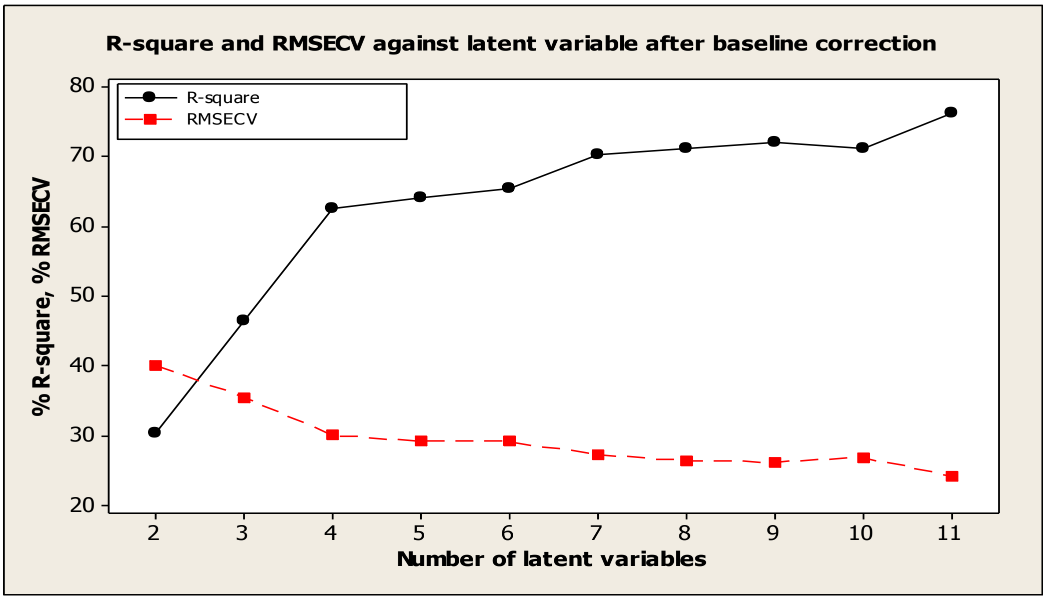


Figure 3.20: Graph showing the influence of latent variable on the quality of PLS-DA models

As shown in Table 3.6, a 2^2 full factorial design matrix (with 2 center points) was set up with two variables, window size and quantile value set at 100-500 and 0.1-1 as lower and higher levels. A combination of window size, 500 and quantile value 0.1 gave the optimal model with average R^2 of 72.4% and RMSECV of 23.6%. Thus two different random designs were set up to investigate the window of 500 and 0.1 quantile value more closely.

In the first random design (random design 1) the quantile value was fixed at 0.1 whilst the window size was increased progressively from 100 to 1000 at intervals of 100. The optimal model (R^2 of 72.3% and RMSECV of 24%) was found when window size was 400 and quantile value 0.1. In the second random design (random design 2) the quantile value was increased progressively from 0.1 to 1 at intervals of 0.1 whilst the window size was fixed at 500. No window size and quantile value combination gave a model with R^2 better that 72% (Table 3.6). These R^2 values however were better than that obtained (69.8%) when no preprocessing was applied to the data. These suggest that baseline correction, at least on its own, will improve model quality of the models built with the spectra data. The optimal parameter setting for baseline correction was also found to be a window size of 500 and a quantile value of 0.1 (Table 3.6 in bold).

Standard order	Latent variable	Run order	Center point	Window size	Quantile value	RMSECV 1 (%)	RMSECV 2 (%)	RMSECV 3 (%)	Average RMSECV (%)	$R^2$1 (%)	$R^2$2 (%)	$R^2$3 (%)	Average R^2 (%)
None	**None**	**None**	**None**	**None**	**None**	**29.5**	**25.2**	**26.8**	**27.2**	**65.4**	**73.2**	**70.7**	**69.8**
Factorial Design													
2	**4**	**1**	**1**	**500**	**0.1**	**28.5**	**25.2**	**23.6**	**25.8**	**68.1**	**73.1**	**76.2**	**72.4**
6	4	2	0	300	0.55	28.3	25.3	26.4	26.7	68.4	73	71.1	70.8
5	4	3	0	300	0.55	28.3	25.3	26.4	26.7	68.4	73	71.1	70.8
3	4	4	1	100	1	63.9	34.2	66.2	54.8	9.4	52.8	4.5	22.2
4	4	5	1	500	1	62.2	40.9	47.1	50.1	0.1	38.2	18.6	19
1	4	6	1	100	0.1	28.2	25.1	26.1	26.5	68.5	73.1	71.3	71
Random design 1													
None	4	7	None	100	0.1	28.2	25.2	26.4	26.6	68.5	73	70.9	70.8
None	4	8	None	200	0.1	28.4	25.1	25.2	26.2	68.3	73	73	71.4
None	4	9	None	300	0.1	28.2	25.1	24.4	25.9	68.6	73.1	74.6	72.1
None	4	10	None	400	0.1	28.3	25.2	24	25.8	68.3	73	75.4	72.3
None	4	11	None	600	0.1	29.5	25.2	24.5	26.4	65.8	73.1	74.5	71.1
None	4	12	None	700	0.1	29.3	25.4	23.8	26.2	67.1	72.6	75.7	71.8
None	4	13	None	800	0.1	31.2	25.6	23	26.6	64.7	72.2	77.2	71.4
None	4	14	None	900	0.1	30.7	26	22.7	26.4	64.9	71.5	77.8	71.4
None	4	15	None	1000	0.1	31.6	25.8	23.6	27	64	72	76.3	70.8
Random design 2													
None	4	16	None	500	0.2	28.5	25.3	24.2	26	68	72.9	75.1	72
None	4	17	None	500	0.3	28.4	25.3	25.2	26.3	68.1	72.9	73.2	71.4
None	4	18	None	500	0.4	28.3	25.3	26.4	26.7	68.3	72.9	70.9	70.7
None	4	19	None	500	0.6	28.5	25.2	26.8	26.8	68	73.2	70.2	70.5
None	4	20	None	500	0.7	34.9	29.1	27.9	30.6	52.6	65.7	66	61.4
None	4	21	None	500	0.8	28.6	25.3	24.9	26.3	67.8	72.9	73.7	71.5
None	4	22	None	500	0.9	31.7	26.1	25.1	27.6	63.1	71.6	73.3	69.3

Table 3.6: Design matrix showing the influence of different baseline correction parameter settings on the quality of PLS-DA models

3.6.1.3. Alignment, Normalisation and Smoothing

Table 3.7 shows the effect of alignment across a number of latent variables. Alignment of the spectra data was carried out along five *m/z* ratio peaks 6411, 6855, 7273, 7333, 7869, 9061, 9218, 9532, 9736 which were found to be common among most of the spectra profiles. Results in Table 3.7 suggest that applying alignment on the data slightly improved the qualities of the PLS-DA models built using the mass spectra data, on average. This can be clearly seen from Fig. 3.21 (Table A.4, Appendix A). This suggests that using alignment may be necessary for this spectra data sets, at least on its own.

Number of latent variables	RMSECV 1 (%)	RMSECV 2 (%)	RMSECV 3 (%)	Average RMSECV (%)	$R^2$1 (%)	$R^2$2 (%)	$R^2$3 (%)	Average R^2 (%)
2	47.2	28.1	39.5	38.2	9.1	67.2	30.9	35.7
3	38.9	28.4	30.6	32.6	38	66.3	58.2	54.2
4	31.5	29.3	24.7	28.5	59.1	64.8	72.9	65.6
5	32.1	26.9	25	28	59.4	69.6	73.5	67.5
6	31.1	25.6	25.4	27.4	62.8	72.3	72.7	69.3
7	27.6	25.4	25.6	26.2	69.6	70.7	73.6	71.3
8	26.1	25.8	25.8	25.9	72.9	73.1	73.2	73
9	26.9	25.6	26.4	26.3	72.2	73.4	72.5	72.7
10	26.7	25.8	26.1	26.2	72.2	73.6	72.5	72.8
11	26.2	26	26.5	26.2	73.1	73.4	73.1	73.2

Table 3.7: Influence of alignment and number of latent variables on the quality of PLS-DA models

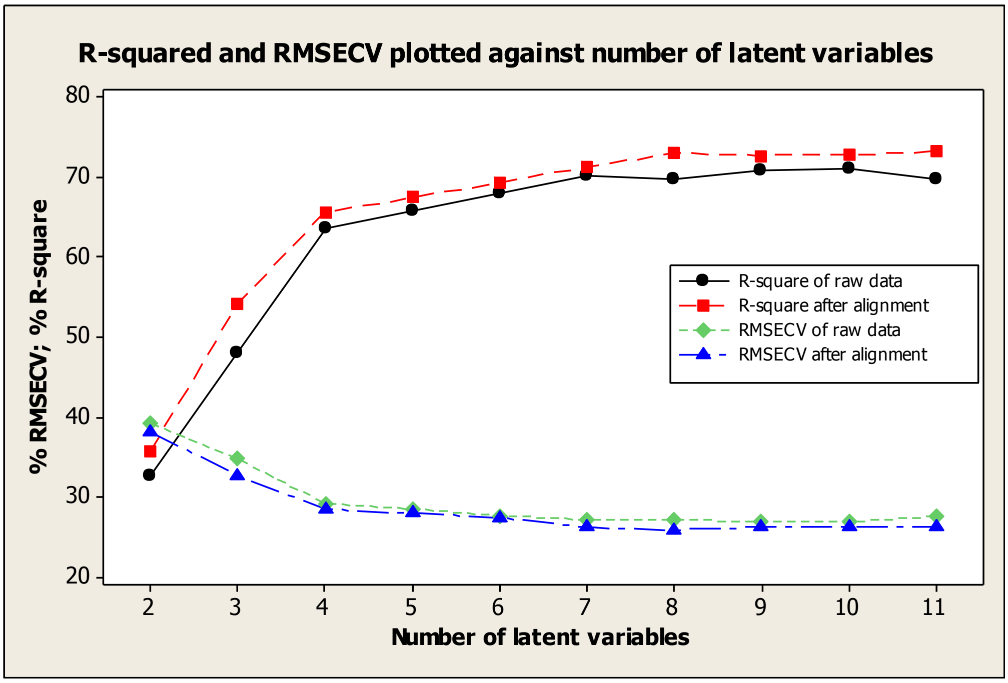


Figure 3.21: Graphs showing the effect of alignment on quality of PLS-DA models across latent variables

Table 3.8 shows results of the quality of PLS-DA classification models obtained after the data used to build the models was normalised at with varied parameter settings. The model evaluations were carried out at a fixed LV (LV 4). A 2^3 full factorial design matrix (with no center points) was set up with two variables, rescaling and quantile value. Eight runs were carried out with the rescaling fixed at 100 and the quantile value was increased progressively from fixed at 0.1 to 0.6 at intervals of 0.1. A random design was subsequently set up and quantile values from 0.7 to 1 were evaluated against a fixed rescaling value of 100; whilst a fixed quantile value of 0.1 was evaluated against rescaling values of 20, 50 and 80. Results suggested that better models (R^2 of 86% and RMSECV of 17.4%) were produced after normalisation, than when no preprocessing was performed on the data (R^2 of 69. 8% and RMSECV of 65.4%). This suggests that data normalisation improves the quality of the PLS-DA models built with the mass spectra data.

Standard order	Latent variable	Run order	Center point	Rescaling	Quantile value	RMSECV 1 (%)	RMSECV 2 (%)	RMSECV 3 (%)	Average RMSECV (%)	$R^2$1 (%)	$R^2$2 (%)	$R^2$3 (%)	Average R^2 (%)
No preprocessing	**None**	**None**	**None**	**None**	**None**	**29.5**	**25.2**	**26.8**	**27.2**	**65.4**	**73.2**	**70.7**	**69.8**
Factorial Design													
2	4	1	1	100	0.1	19.1	12.2	21	17.4	84.1	93.3	80.7	86
6	4	2	0	100	0.2	19.1	12.2	21	17.4	84.1	93.3	80.7	86
5	4	3	0	100	0.3	19.1	12.2	21	17.4	84.1	93.3	80.7	86
3	4	4	1	100	0.4	19.1	12.2	21	17.4	84.1	93.3	80.7	86
4	4	5	1	100	0.5	19.1	12.2	21	17.4	84.1	93.3	80.7	86
1	4	6	1	100	0.6	19.1	12.2	21	17.4	84.1	93.3	80.7	86
Random design													
None	4	7	None	100	0.7	19.1	12.2	21	17.4	84.1	93.3	80.7	86
None	4	8	None	100	0.8	19.1	12.2	21	17.4	84.1	93.3	80.7	86
None	4	9	None	100	0.9	19.1	12.2	21	17.4	84.1	93.3	80.7	86
None	4	10	None	50	0.1	19.1	12.2	21	17.4	84.1	93.3	80.7	86
None	4	11	None	80	0.1	19.1	12.2	21	17.4	84.1	93.3	80.7	86
None	4	12	None	20	0.1	19.1	12.2	21	17.4	84.1	93.3	80.7	86
None	4	13	None	100	0.7	19.1	12.2	21	17.4	84.1	93.3	80.7	86
None	4	14	None	100	None	19.1	12.2	21	17.4	84.1	93.3	80.7	86
None	4	15	None	100	1	19.1	12.2	21	17.4	84.1	93.3	80.7	86

Table 3.8: Design matrix showing the influence of different normalisation parameter settings on the quality of PLS-DA models

Table 3.9 shows results of the quality of PLS-DA classification models obtained built with smoothed mass spectra data. Smoothing was performed with varied parameter settings. The parameter involved during smoothing was the smoothing span value (i.e. the segment size containing a specific number of *m/z* ratio data points). This was systematically changed from 10 to 100 at intervals of 5 units. The model evaluations were carried out at a fixed LV (LV4).

As can be seen from the Table 3.9, span values of 20 and 25 produced better models, with an R^2 of 72.2% and RMSECV of approximately 27% (indicated in bold in Table 3.9). This also suggests that smoothing improves the quality of classification models. To conclude, results thus far suggest that evaluating the preprocessing methods in isolation remove undesired effects, increasing between class variations and hence the classification models as seen through the R^2 and RMSECV. In the next section, the preprocessing methods were combined and evaluated together to verify if this will be of value.

Run order	Latent variable	Span value	RMSECV 1 (%)	RMSECV 2 (%)	RMSECV 3 (%)	Average RMSECV (%)	$R^2 1$ (%)	$R^2 2$ (%)	$R^2 3$ (%)	Average R^2 (%)
1	4	10	29.9	25.8	25.8	27.2	65.7	72	72.4	70
2	4	15	29.2	26	25.6	26.9	66.9	72	72.7	70.5
3	**4**	**30**	**26.4**	**26.7**	**25.7**	**27**	**72.4**	**71**	**73.3**	**72.2**
4	**4**	**35**	**26.5**	**26.7**	**25.8**	**26.8**	**72.3**	**71.1**	**73.2**	**72.2**
5	4	20	28.6	26.5	26.4	27.2	68.1	71.5	71.4	70.3
6	4	25	27.7	26.6	25.8	26.7	69.4	71.1	72.5	71
7	4	40	27.6	26.8	25.6	26.7	69.6	70.9	72.9	71.2
8	4	45	28	26.9	27	27.3	69.9	70.9	71.1	70.6
9	4	50	27.7	26.9	26.8	27.1	70.7	71	71.3	71
10	4	55	27.4	26.8	26.6	27	71.2	71.1	71.6	71.3
11	4	60	27.4	26.8	26.6	27	71.2	71.1	71.6	71.3
12	4	65	27.1	26.8	26.4	26.8	71.8	71	71.9	71.6
13	4	70	26.9	26.8	26.3	26.7	72	71	72.1	71.7
14	4	75	26.7	26.8	26.2	26.6	72.2	71.1	72.3	71.8
15	4	80	26.6	26.8	26.1	26.5	72.3	71.1	72.4	72
16	4	85	26.5	26.7	26	26.4	72.4	71.1	72.6	72
17	4	90	26.4	26.7	25.9	26.3	72.4	71.1	72.9	72.1
18	4	95	27.7	26.9	25.8	26.3	69.4	70.8	72.9	71
19	4	100	27.9	26.9	26.1	26.3	69.3	70.8	72.1	70

Table 3.9: Influence of different smoothing values on PLS-DA model optimisation

3.6.1.4. Using all Preprocessing Techniques for the *E. coli* Spectra Profiles

Tables 3.10 and 3.11 shows the effect of the spectra preprocessing techniques on the quality and performance of PLS-DA classification techniques obtained after the data used to build the models was preprocessed with all the techniques, with varied parameter settings. The model evaluations were carried out at a fixed LV (LV 4). A 2^4 full factorial design matrix (with 2 center points per block) was set up with four factors, i.e. baseline correction quantile value, baseline correction window size, normalisation quantile value and smoothing span value. All these factors were held at 2 levels (upper and lower levels).

Eighteen runs were performed with baseline correction quantile value at 0.1-0.2, baseline correction window size at 300-500, normalisation quantile value at 0.1-1 and smoothing span value at 20-25. These intervals were those that gave optimal results (high R^2 and low RMSECV) when the corresponding preprocessing techniques were used in isolation. A random design was subsequently set up with baseline correction quantile value at 0.1-0.2, baseline correction window size at 100-200, and normalisation quantile value at 0.1-0.2, evaluated against a fixed smoothing span value of 25. The qualities of the models were evaluated through the R^2 and the RMSECV whilst the performance was evaluated against the root mean square error of prediction (RMSEP) and the R^2 of prediction.

As can be seen from the results (Tables 3.10 and 3.11, run order 21 and 24) the best models were obtained (with R^2 of calibration 33.46%; RMSECV of 82.01%; R^2 of prediction 89.19%; and RMSEP of 15.85%), when the baseline correction quantile value was at 0.1, baseline correction window size was set to 200, normalisation quantile value, 0.1 or 0.2, and smoothing span value was at 25. These parameter settings were used as standard settings for the preprocessing approach applied to the *E. coli* mass spectra data. The standard preprocessing algorithm used for all spectra data sets can be found in Appendix A, Fig. A.1.

Standard order	Run order	Center point	Blocks	Latent variable	Baseline correction window size	Baseline correction quantile value	Normalisation quantile value	Smoothing span value	Average RMSECV	Average R^2
No preprocessing	None	None	None	None	None	None	None	None	85.96	16.57
Factorial design										
13	1	1	1	4	300	0.1	1	25	81.79	33.89
4	2	1	1	4	500	0.2	0.1	20	81.84	33.74
9	3	1	1	4	300	0.1	0.1	25	81.79	33.89
18	4	0	1	4	400	0.15	0.55	22.5	81.91	33.59
3	5	1	1	4	300	0.2	0.1	20	81.71	33.74
6	6	1	1	4	500	0.1	1	20	81.75	33.54
16	7	1	1	4	500	0.2	1	25	82.04	33.48
11	8	1	1	4	300	0.2	0.1	25	81.91	33.59
2	9	1	1	4	500	0.1	0.1	20	81.75	33.54
1	10	1	1	4	300	0.1	0.1	20	81.58	34.06
10	11	1	1	4	500	0.1	0.1	25	82.08	33.08
14	12	1	1	4	500	0.1	1	25	82.08	33.08
12	13	1	1	4	500	0.2	0.1	25	82.04	33.48
7	14	1	1	4	300	0.2	1	20	81.71	33.74
5	15	1	1	4	300	0.1	1	20	81.58	34.06
15	16	1	1	4	300	0.2	1	25	81.91	33.59
17	17	0	1	4	400	0.15	0.55	22.5	81.91	33.59
8	18	1	1	4	500	0.2	1	20	81.84	33.74
Random design										
None	20	None	None	4	100	0.2	0.1	25	82.13	33.23
None	**21**	**None**	**None**	**4**	**200**	**0.1**	**0.2**	**25**	**82.01**	**33.46**
None	22	None	None	4	100	0.1	0.2	25	82.13	33.23
None	23	None	None	4	100	0.1	0.1	25	82.13	33.23
None	**24**	**None**	**None**	**4**	**200**	**0.1**	**0.1**	**25**	**82.01**	**33.46**

Table 3.10: Design matrix showing the influence of different preprocessing techniques of various parameter settings on the quality of PLS-DA mode

Standard order	Run order	Center point	Blocks	Latent variable	Baseline correction window size	Baseline correction quantile value	Normalisation quantile value	Smoothing span value	Average RMSEP	Average R^2 of prediction
No preprocessing	None	None	None	None	None	None	None	None	35.36	48.83
Factorial design										
13	1	1	1	4	300	0.1	1	25	16.16	89.96
4	2	1	1	4	500	0.2	0.1	20	20.42	84.39
9	3	1	1	4	300	0.1	0.1	25	16.16	89.96
18	4	0	1	4	400	0.15	0.55	22.5	17.43	86.87
3	5	1	1	4	300	0.2	0.1	20	19.35	83.87
6	6	1	1	4	500	0.1	1	20	17.75	86.22
16	7	1	1	4	500	0.2	1	25	18.01	86.1
11	8	1	1	4	300	0.2	0.1	25	18.24	85.71
2	9	1	1	4	500	0.1	0.1	20	17.75	86.22
1	10	1	1	4	300	0.1	0.1	20	17.81	86.25
10	11	1	1	4	500	0.1	0.1	25	16.36	88.31
14	12	1	1	4	500	0.1	1	25	16.36	88.31
12	13	1	1	4	500	0.2	0.1	25	18.01	86.1
7	14	1	1	4	300	0.2	1	20	19.35	83.87
5	15	1	1	4	300	0.1	1	20	17.81	86.25
15	16	1	1	4	300	0.2	1	25	18.24	85.71
17	17	0	1	4	400	0.15	0.55	22.5	17.43	86.87
8	18	1	1	4	500	0.2	1	20	19.1	84.29
Random design										
None	20	None	None	4	100	0.2	0.1	25	19.07	84.57
None	**21**	**None**	**None**	**4**	**200**	**0.1**	**0.2**	**25**	**15.85**	**89.19**
None	22	None	None	4	100	0.1	0.2	25	19.07	84.57
None	23	None	None	4	100	0.1	0.1	25	19.07	84.57
None	**24**	**None**	**None**	**4**	**200**	**0.1**	**0.1**	**25**	**15.85**	**89.19**

Table 3.11: Design matrix showing the influence of different preprocessing techniques/parameter settings on the performance or quality of PLS-DA models

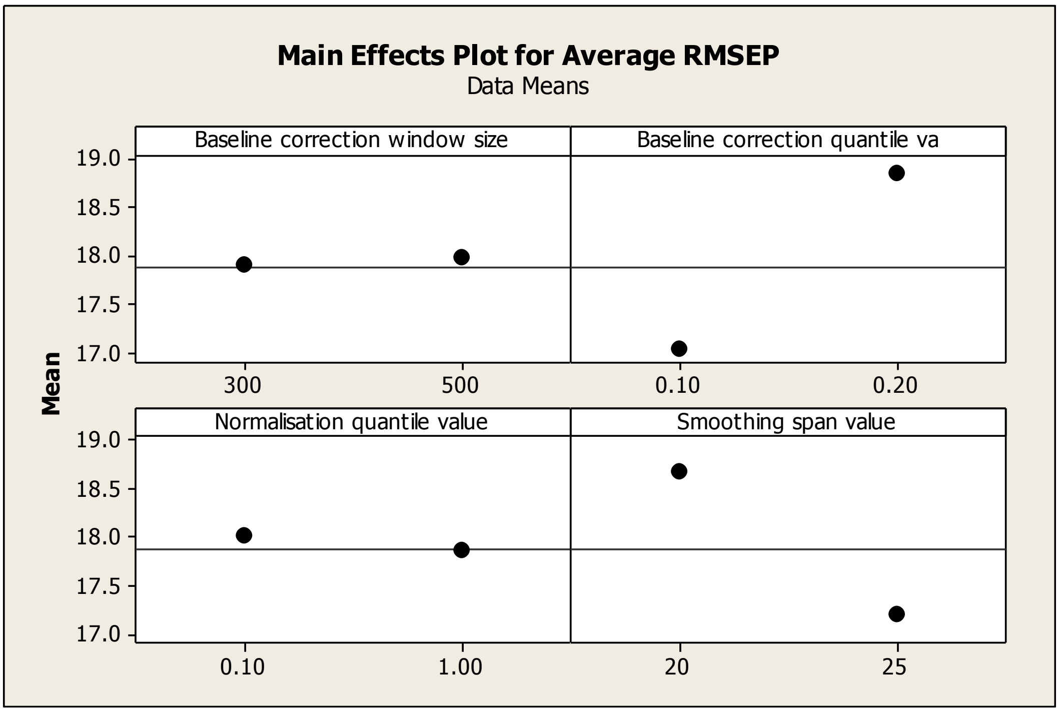
Fig. 3.22 shows the main effects and interaction plots for the preprocessing techniques with output being the RMSEP. The latter was used because prediction error is an absolute measure. Furthermore, using a parameter which test for model performance will help increase confidence on conclusions drawn with respect to model quality, and hence the preprocessing method.

A seen in Fig. 3.22, the main effects plot suggests that all the preprocessing techniques are significant, with baseline correction quantile value and smoothing span value being the most significant preprocessing techniques affecting the model performance. This goes to support the earlier view that all these techniques were essential for preprocessing the spectra data. The main effects for average RMEP are maximised when baseline correction quantile value was set at 0.2 and smoothing span value to 20.

From the interaction plots, the following interactions could be observed;

- baseline correction window size and normalisation quantile value;
- smoothing span value and normalisation quantile value;
- baseline correction window size and smoothing span value; and
- baseline correction window size and quantile value.

At baseline correction window sizes of 300 or 500, model performance are increased (with lower average RMSEP) when the baseline correction quantile value is reduced from 0.2 to 0.1, and when the smoothing span value is increased from 20 to 25. At baseline correction window sizes of 300 or 500, model performance is increased when the normalisation quantile value is increased from 0.1 to 1. At normalisation quantile values of 0.1, 0.2 or 1, model performance is improved when the smoothing span value is increased from 20 to 25. At baseline correction quantile value of 0.2, model performance is improved when the normalisation quantile value is increased from 0.1 to 1. However the gain is not as large as simply setting the baseline correction quantile value at low levels of 0.1. These interactions prove the notion that looking at one preprocessing technique in isolation from the others may be of limited value, an argument which was stressed by Baggerly *et al*, (2003). The factorial design enabled the incorporation of the possibility of interactions between the different preprocessing techniques whilst helping to evaluate the relative importance of the individual methods.



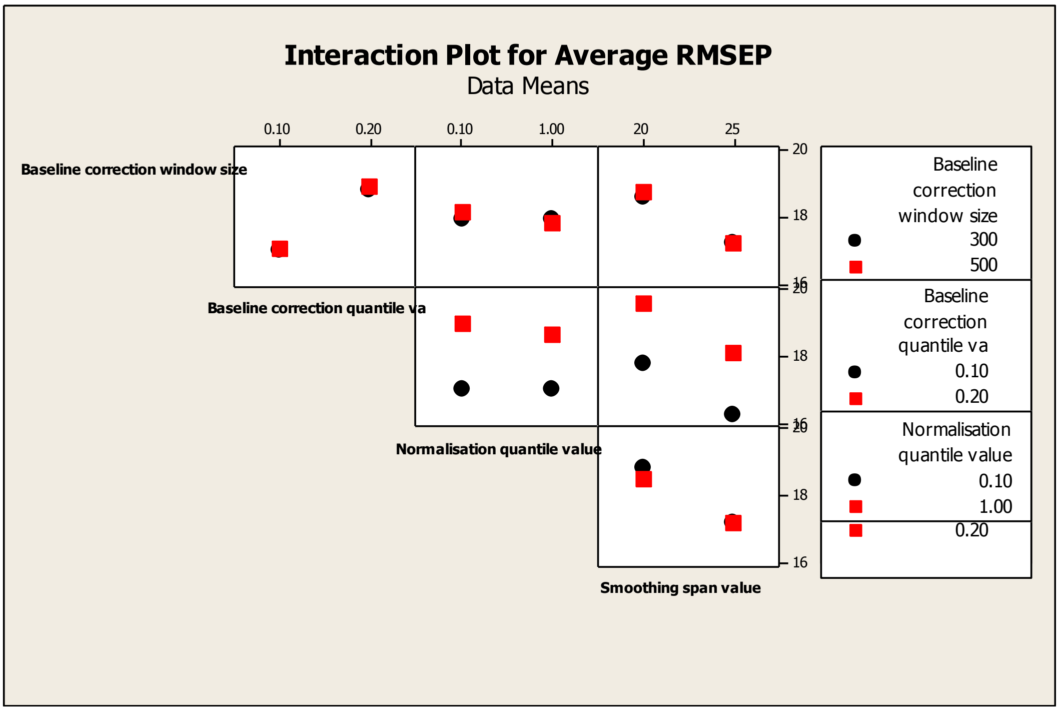


Figure 3.22: Main effects and interaction plots of preprocessing techniques for RMSEP

3.7. Preprocessing Studies for MALDI-ToF data generated from Monoclonal Antibody Producing CHO cell line Spectra Profiles

3.7.1. Using all Preprocessing Techniques for the CHO Cell Line Spectra Profiles

The procedure described in section 3.6 was repeated for the CHO cell line spectra data sets. All the preprocessing techniques were evaluated for the CHO cell line spectra data. The qualities and performances of the constructed models using different preprocessing techniques/parameters were assessed. Factorial and random designs was used with different parameters of the preprocessing techniques defining the design variables and the PLS-DA RMSEP and R^2 of prediction as the response variables to determine the optimal combination.

Tables 3.13 shows the effect of the spectra preprocessing techniques on the quality and performance of PLS-DA classification techniques obtained after the CHO cell line spectra data used to build the models was preprocessed with all the techniques varying the parameter settings. The model evaluations were carried out at a fixed LV (LV 5). A 24 full factorial design matrix (with 2 centre points per block) was set up with four factors, i.e. baseline correction quantile value, baseline correction window size, normalisation quantile value and smoothing span value. All these factors were held at 2 levels (upper and lower levels).

Eighteen runs were performed with baseline correction quantile value at 0.1-0.2, baseline correction window size at 300-500, normalisation quantile value at 0.1-1 and smoothing span value at 20-25. These intervals were those that gave optimal results (high R^2 and low RMSECV) when the corresponding preprocessing techniques were used in isolation. A random design was subsequently set up with baseline correction quantile value at 0.1-0.2, baseline correction window size at 100-200, and normalisation quantile value at 0.1-0.2, evaluated against a fixed smoothing span value of 25. The qualities of the models were evaluated through the R^2 and the RMSECV whilst the performance was evaluated against the root mean square error of prediction (RMSEP) and the R^2 of prediction.

As can be seen from the results (Tables 3.12, run order 20) the best models were obtained (with R^2 of calibration 48.5%; RMSECV of 46.2%; R^2 of prediction 33.5%; and RMSEP of 50.3%), when the baseline correction quantile value was at 0.2, baseline correction window size was set to 100, normalisation quantile value, 0.1, and smoothing span value was at 25. These parameter settings were used as standard settings for the preprocessing approach applied to the CHO cell line mass spectra data. The standard preprocessing algorithm used for all spectra data sets can be found in Appendix A, Fig. A.2.

Standard order	**Run order**	**Center point**	**Blocks**	**Latent variable**	**Baseline correction window size**	**Baseline correction quantile value**	**Normalisation quantile value**	**Smoothing span value**	**Average RMSECV**	**Average RMSEP**	**Average R^2 of cal**	**Average R^2 of pred**
Factorial design												
13	1	1	1	5	300	0.1	1	25	0.442	0.511	0.482	0.32
4	2	1	1	5	500	0.2	0.1	20	0.441	0.504	0.484	0.292
9	3	1	1	5	300	0.1	0.1	25	0.442	0.512	0.482	0.32
18	4	0	1	5	400	0.15	0.55	22.5	0.446	0.533	0.469	0.272
3	5	1	1	5	300	0.2	0.1	20	0.446	0.516	0.477	0.324
6	6	1	1	5	500	0.1	1	20	0.449	0.521	0.475	0.287
16	7	1	1	5	500	0.2	1	25	0.44	0.513	0.478	0.281
11	8	1	1	5	300	0.2	0.1	25	0.445	0.513	0.47	0.206
2	9	1	1	5	500	0.1	0.1	20	0.446	0.521	0.475	0.287
1	10	1	1	5	300	0.1	0.1	20	0.443	0.507	0.487	0.228
10	11	1	1	5	500	0.1	0.1	25	0.446	0.519	0.469	0.281
14	12	1	1	5	500	0.1	1	25	0.446	0.519	0.469	0.281
12	13	1	1	5	500	0.2	0.1	25	0.44	0.513	0.477	0.281
7	14	1	1	5	300	0.2	1	20	0.446	0.516	0.477	0.324
5	15	1	1	5	300	0.1	1	20	0.443	0.513	0.487	0.327
15	16	1	1	5	300	0.2	1	25	0.445	0.515	0.47	0.312
17	17	0	1	5	400	0.15	0.55	22.5	0.446	0.533	0.469	0.272
8	18	1	1	5	500	0.2	1	20	0.441	0.513	0.484	0.293
Random design												
None	19	None	None	5	200	0.2	0.1	25	0.449	0.516	0.469	0.306
None	**20**	**None**	**None**	**5**	**100**	**0.2**	**0.1**	**25**	**0.462**	**0.503**	**0.485**	**0.335**
None	21	None	None	5	200	0.1	0.2	25	0.457	0.495	0.482	0.239

Table 3.12: Design matrix showing the influence of different preprocessing techniques/parameter settings on the performances and qualities of PLS-DA models

3.8. Summary

In this chapter, mass spectrometry was introduced as an important ‘omics’ tool. In particular, the capabilities and advantages of the MALDI-ToF mass spectrometry platform was explored. The relative ease of operation of MALDI coupled with ToF detection and its characteristic generation of mostly singly charged peptide and protein ions makes a useful mass spectrometry technique. Ions are generated from high-mass and non-volatile protein molecules through laser irradiation. A key aspect of MALDI is the use of an energy absorbing matrix which can co-crystallise with the analyte preventing decomposition of the latter due to the laser energy. Gas phase protein molecular ions produced by the matrix-analyte co-crystal traverse a field-free flight tube and are then separated according to their *m/z* ratio. MALDI-ToF was used for data analysis in this thesis because of the advantages it has such as high sensitivity, little or no fragmentation, widespread use and predominance of singly charged molecules as MALDI uses a soft ionisation approach. The fact that mostly singly charged molecular ions (with the mass equal to *m/z* ratio) are generated from the instrument is a key advantage as it enables the direct determination of the mass of the protein molecular ion and hence protein molecules. This offers the opportunity for rapid biomarker discovery as the experimental molecular weight of MALDI molecular ions can be matched against sequence molecular weight of organisms with sequenced genomes, directly identifying the protein (chapters 6 and 7).

However, before any data mining is carried out on mass spectra data for biomarker identification, the raw spectra profiles have to be preprocessed. Preprocessing serves to reduce the spectral noise, reduce the amount of data, and ensure the spectra comparable. It includes a number of techniques including resampling, baseline correction, alignment, normalisation smoothing, and peak identification. Studies have shown that these preprocessing techniques are interrelated, and several combinations of the different techniques may have to be tested to identify an appropriate preprocessing approach (Baggerly *et al.,* 2004). A review of the applications of the application of preprocessing on mass spectra data was provided, emphasising the importance of preprocessing.

In this chapter, preprocessing studies were carried out using the MALDI-ToF mass spectra profiles generated from *E. coli* cells and CHO cell lines. A quantitative assessment of the impact of different combinations of preprocessing techniques on the mass spectral data was carried out by factorial designs with the parameter settings of the various preprocessing techniques used as factors, and root mean square error of prediction (RMSEP), RMSE cross validation (RMSECV), R^2 of prediction and calibration as the response. After combining all the preprocessing techniques, the combinations that gave favourable responses for the RMSEP and R^2 of prediction were used as the preprocessing procedures for the *E. coli* cell spectra profiles. Moreover, identification of protein biomarkers from the reprocessed spectra profiles, consistent with those already described in the literature further helped validate the preprocessing methods/parameters combination used (chapter 6 and 7). The next chapter explores proteomic profiling for biomarker discovery as well as some important mass spectrometry based techniques. It focuses on intact-cell MALDI-ToF mass spectrometry (ICM), the technique that was used for analysing the samples in this thesis.

Chapter 4

4. Protein Profiling for Biomarker Discovery

4.1. Overview

The previous chapter introduced the technique of mass spectrometry with a description of MALDI-ToF MS which is the main analytical technique used in this thesis. It also described the importance of pre-treatment of the high dimensionality mass spectra data generated from the MALDI, a number of preprocessing techniques, substantiated by examples using suitable mass spectra data from *E. coli* cells at different growth phases as well as from the IgG monoclonal antibody producing CHO cell lines during culturing.

This chapter describes the concept of proteomic biomarker discovery and introduces the top-down proteomics based approach of 'intact-cell' MALDI-ToF MS (ICM-MS). The application of ICM-MS for the analysis of microorganism and mammalian cell line is given. The chapter also explores the potential advantages and usefulness of ICM-MS and internet-accessible protein databases for rapid biomarker profiling in the area of mammalian cell culture in biopharmaceutical bioprocessing. Finally the chapter concludes with a literature review on the applications of the application of ICM-MS on both bacterial and mammalian cells.

4.2. Introduction

Recent advances in genomics with the sequencing of the human genome (Venter *et al.,*, 2001) and many other species including *E. coli* (Blattner *et al*., 1997), mouse (Waterston *et al*., 2002), and rat (Gibbs *et al*., 2004) are providing increasing knowledge in terms of the fundamental genetic code that characterises signal transduction pathways, and the control of important cellular events like growth, differentiation, and cell death. Although gene-related-information is of value, evidence within the context of clinical research suggests that analysing genome sequences alone provides insufficient indicators for the development of new therapies to fight human disease. Potentially of greater value will be to have knowledge of global patterns of protein content and activity and how these are altered during development or in response to disease. This type of information will be required to facilitate the discovery of novel drug targets and new therapies (Gygi *et al.*, 1999).

Contrary to genomic studies, the analysis of proteomes is significantly more challenging and complicated mainly because the proteome is dynamic and is in constant flux. Other factors such as alternative splicing of the respective mRNAs, and posttranslational modifications can result in important functional differences in proteins of higher eukaryotes (Wery, 2012). The term proteome refers to all proteins expressed by a cell, tissue or a body fluid and is thus a complex mixture. Proteomics can be defined as *'the systematic analysis of protein populations and the protein complement of cells, including the concurrent identification, modification, quantification, and localisation of large numbers of proteins in a functional context'* (Pothur *et al.*, 2001). Among the many vital functions they perform, proteins catalyse a variety of chemical reactions, support a range of skeletal structures, control membrane permeability, modulate the concentration of metabolites, and control gene expression. Thus information at the level of the cellular proteome is essential for determining which proteins or groups of proteins are responsible for a specific function or phenotype in cells. This could be in relation to health and diseases, as in clinical proteomics, or cell systems involved in the production of a therapeutic protein in bioprocessing.

Protein profiling, a sub-discipline of proteomics, has provided significant insight into biological events such as transcription and translation. It is the generation of extended protein expression data sets for analysing changes in global protein expression patterns in biological systems as a function of developmental, physiological, and disease processes. Protein profiling can promote understanding of aspects of a disease such as pathogenesis, improved early detection, staging, therapeutic monitoring and prognosis (Bakry *et al.*, 2011). In clinical proteomics, biomarkers are among the important tools critical to understanding these disease aspects.

4.3. Protein Biomarker Discovery

Biomarkers can be defined as biological molecules that correlate to a specific biological or pathological state, pharmacologic response or a therapeutic intervention. As advances in the fields of genomics and proteomics continue to contribute to a wide-range of scientific disciplines (e.g. industrial manufacture of therapeutic proteins) through new technologies, the field of biomarker discovery, development and application has become an area of considerable research focus and activity.

The discovery of biomarkers can be carried out through the use of a variety of approaches depending on the nature of the biomarker involved. Approaches can range from transcriptional profiling and DNA methylation studies which have shown strong potential for biomarker discovery in cancer, to metabolomic approaches as demonstrated for metabolic disease, drug and toxicity studies (Rifai *et al.,* 2006). However, with protein domains being the most affected entities during a pathological condition, protein biomarkers have become one of the most valuable classes of biomarkers over the past 100 years. As a consequence, proteomic biomarker discovery studies have become commonplace.

In the context of clinical proteomics, protein biomarkers serve as important tools in terms of acting as an early indicator of a disease, for the monitoring of disease progression, and for assisting with disease detection. Protein biomarkers are often low-molecular-weight proteins and their secretion and appearance in the bloodstream is triggered by the onset and progression of a disease process and their prominence has made them the cornerstone of medical care (Frank and Hargreaves, 2003). Blood biomarker measurements can provide indicators relating to the source of patient symptoms such as abdominal pain (transaminase biomarkers - hepatitis, alkaline phosphatase - biliary disorders, and human chorionic gonadotropin (*β*-hCG) - pregnancy) or chest pain (troponin biomarkers - heart attack). The demonstrated success of biomarkers in terms of impacting the prognosis of numerous patients and could providing insights into the appropriate patient therapy in situations where immediate treatment is necessary; the development of more and better therapeutic and diagnostic biomarkers has now become a priority area in the clinical sciences (Paulovich *et al.,* 2008).

4.3.1. Biomarker Discovery Pipeline

As reviewed in Rifai *et al.* (2006), the mass spectrometry-based biomarker development pipeline (Fig. 4.1) in clinical research usually consists of several phases. Discovery or identification is a step that involves the definition of differential protein expression between biological states. Biomarker discovery or identification is usually carried out using a series of proteomic technologies including gel-based mass spectrometry (MS) (e.g. two-dimensional polyacrylamide gel electrophoresis with matrix assisted laser desorption ionisation time of flight MS; 2D-PAGE/MALDI-ToF-MS), or gel-free MS approaches, i.e., MudPIT (multidimensional protein identification technology) or 'shot-gun proteomics'. An example of 'shot-gun proteomics' is liquid chromatography with electrospray ionisation MS; LC/ESI MS. Sample systems used in discovery include model systems (e.g. mouse models or cell lines) or materials of human biological origin, and usually consist of a binary comparison between diseased and normal tissues. The outcome of the discovery phase is a compiled list of specific proteins differentially expressed between the normal and diseased states.

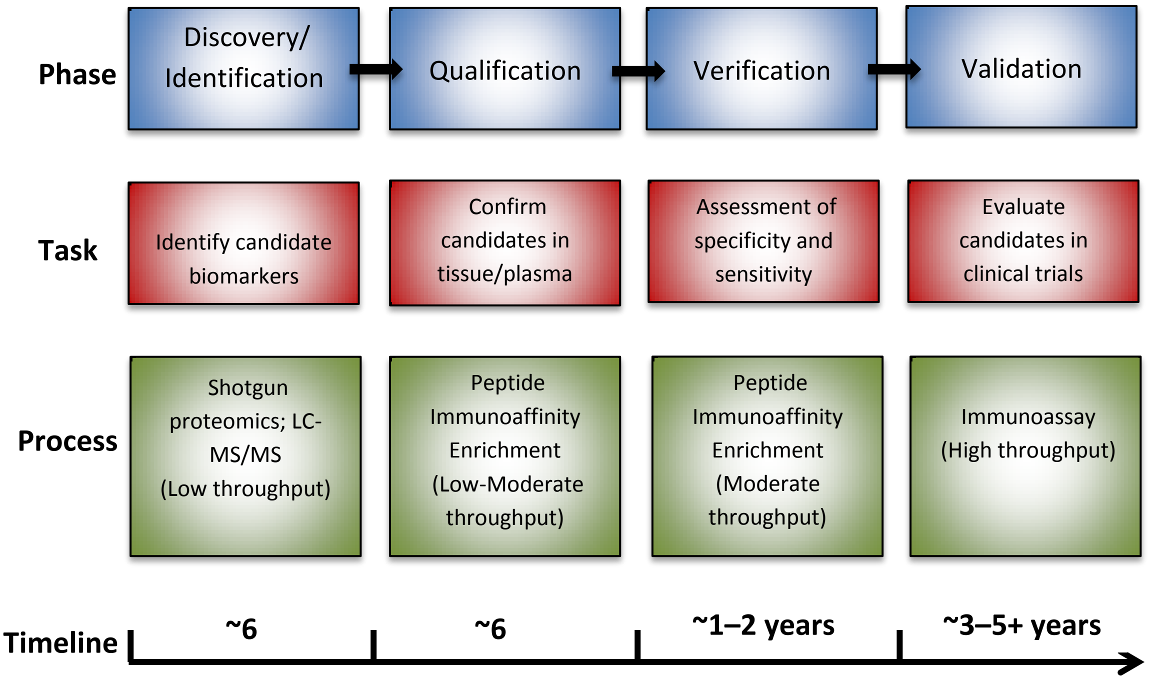

Phase
Discovery/ Identification
Qualification
Verification
Validation
Task
Identify candidate biomarkers
Confirm candidates in tissue/plasma
Assessment of specificity and sensitivity
Evaluate candidates in clinical trials
Process
Shotgun proteomics; LC-MS/MS (Low throughput)
Peptide Immunoaffinity Enrichment (Low-Moderate throughput)
Peptide Immunoaffinity Enrichment (Moderate throughput)
Immunoassay (High throughput)
Timeline
~6
~6
~1–2 years
~3–5+ years

Figure 4.1: Process flow for protein biomarker development

As shown in Fig. 4.1, the next phase in biomarker development is qualification. Qualification confirms the differentially expressed candidate biomarkers, i.e., it links the biomarkers with a biological state. Qualification may also serve to confirm the candidate biomarkers in comparisons of diseased and normal human samples, if discovery was not initially performed in such samples. Discovery and qualification are mainly involved in verifying that the candidate biomarkers are consistently associated with the disease. Principally, they both demonstrate the sensitivity of a candidate biomarker (the likelihood that a diseased sample will test positive) over specificity (the likelihood that an unaffected sample will test negative).

Qualification is followed by the verification step, which serve to determine if there is sufficient evidence for potential clinical utility of a given candidate biomarker to warrant further investment in that candidate for clinical validation studies. Hundreds of human samples are involved, where factors such as environmental, genetic, biological and variation in the population are tested, to confirm the sensitivity and specificity of the biomarkers. Biomarkers which perform well under verification are taken forward to clinical validation. This stage involves the use of quantitative methods on large populations of samples to confirm that the candidate biomarkers have clinical utility, i.e., the biomarker candidates are evaluated in clinical trials in a context most relevant to their eventual clinical application. If not performed during the verification phase, immunoassays are used to assess the diagnostic abilities of the biomarkers, followed by assessment of performance characteristics such as reproducibility and accuracy (e.g. the ability to be accurately used as a disease indicator). Successfully validated biomarkers are then selected for commercialisation.

4.3.2. Overview of Proteomic Approaches for Biomarker Discovery

Currently, protein biomarker discovery is undertaken by one of two approaches (Fig. 4.2), the analysis of intact proteins (top-down proteomics) and the analysis of peptide mixtures from digested proteins (bottom-up proteomics). Proteomic studies in the past have always been based on a trade-off between throughput (top-down proteomics) and resolution (bottom-up proteomics) (Dalmasso *et al.*, 2009). These two proteomic approaches are briefly described in the subsequent sections.

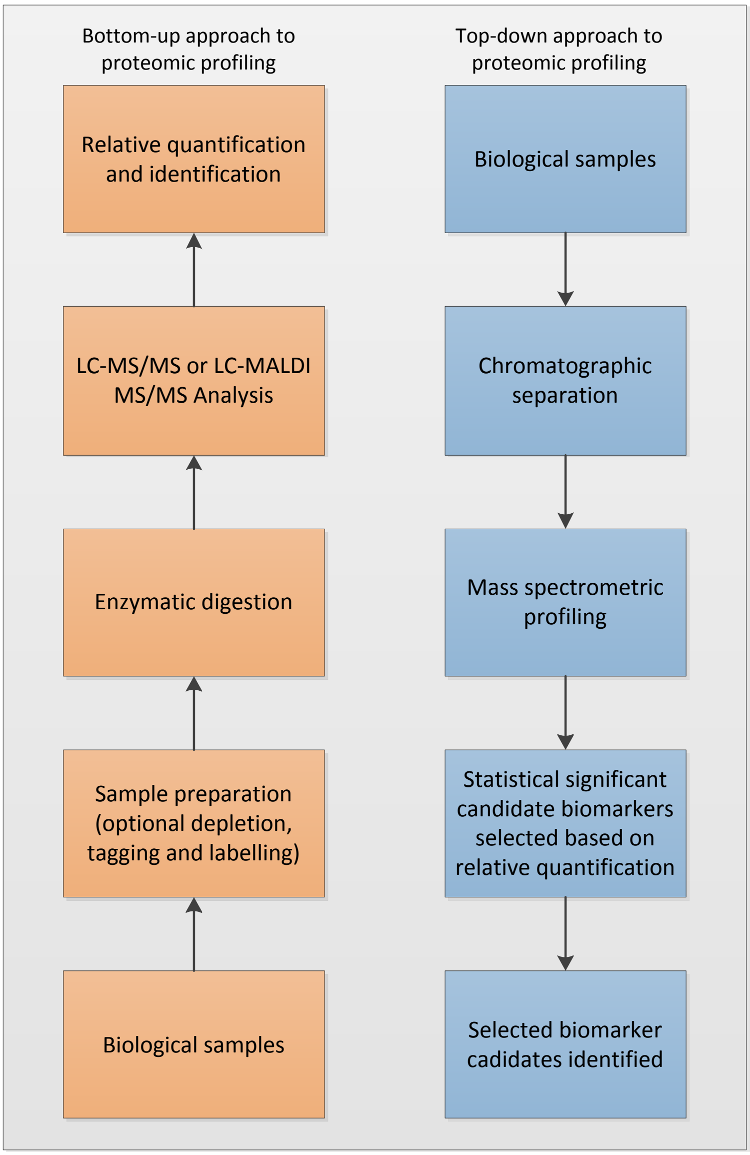

Bottom-up approach to proteomic profiling
Relative quantification and identification
LC-MS/MS or LC-MALDI MS/MS Analysis
Enzymatic digestion
Sample preparation (optional depletion, tagging and labelling)
Biological samples
Top-down approach to proteomic profiling
Biological samples
Chromatographic separation
Mass spectrometric profiling
Statistical significant candidate biomarkers selected based on relative quantification
Selected biomarker cadidates identified

Figure 4.2: General workflows for bottom-up and top-down proteomic profiling

4.3.3. Bottom-up Proteomics

Bottom-up proteomics, also known as ‘shot-gun’ proteomics, involves mass spectrometry analysis of purified proteins, or complex protein mixtures, that have initially undergone enzymatic digestion, using enzymes such as trypsin (Fig. 4.3). Protein mixtures are treated with a proteolytic enzyme (e.g., trypsin) to fragment proteins into peptides. The resulting peptides can be subjected to either mass spectrometry (MS) or MS/MS for protein identification. Proteins purified by employing gel electrophoresis or chromatography containing only one or a few proteins are subjected to MS. In MS, the samples are ionised in the ionisation chamber, analysed by *m/z* ratio in the mass analyser, and detected by the ion detector (Fig. 4.3).

Alternatively (Fig. 4.3), if the analysis involves a complex protein mixture, MS/MS (or tandem MS) is used. In MS/MS (or tandem MS), peptides are fragmented in the collision cell at their peptide bonds before entering the second MS. Better sample resolution can be achieved by processing proteins/peptides before running MS. This is achieved by 2-dimensional gel electrophoresis (2D-GE) and gel purification of protein bands before tryptic digestion. The digested product may contain hundreds of thousands of peptides, and may require separation in liquid chromatography (LC) columns or capillary electrophoresis (CE) before MS analysis (Wehr, 2006). Examples of some quantitative shotgun proteomic technologies are stable isotope labelling by amino acids in cell culture (SILAC), isotope-coded affinity tagging (ICAT) and isobaric tags for relative and absolute quantification (iTRAQ) technology.

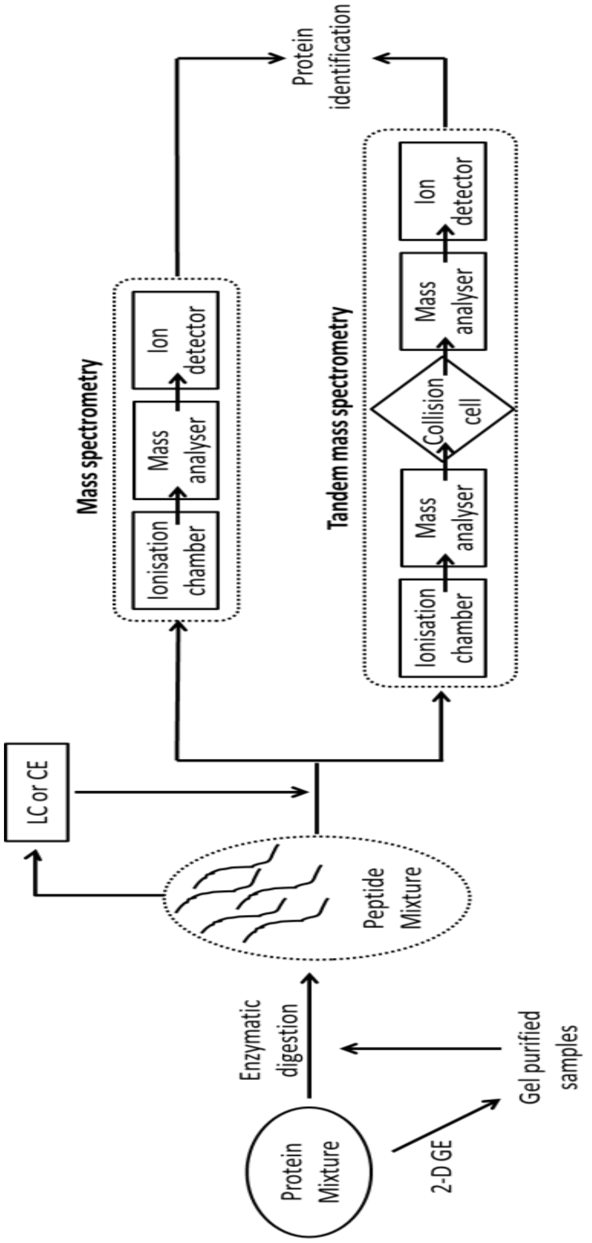


Figure 4.3: Schematic diagram showing steps involved in bottom-up proteomics

The main advantage of bottom-up proteomics is its ability to achieve high resolution proteomic separations, for example, HPLC provides high-resolution separations of proteolytic protein products. The approaches are amenable to automation, with automated on-line nano-scale reversed-phase LC–ESI–MS–MS being widely used for bottom-up proteomics (Wehr, 2006). The identification of proteins from complex proteolytic mixtures such as cell lysates has been most successfully carried out using the bottom-up strategy using on-line multidimensional capillary HPLC–MS-MS (Link *et al.,* 1999).

Bottom-up approaches have several practical limitations. They are generally time-consuming, as an on-line multidimensional LC–MS-MS proteomic analysis using ion-exchange coupled to reversed-phase columns have run times of up to 15 hours or more. Furthermore, as each protein in the samples is reduced to multiple individual peptides, there is an overall increase in complexity during the analysis. This reduction into individual peptide leads to loss of some information concerning specific proteins and so there is no thorough coverage of the protein sequence. This limited sequence coverage and fragmentation process commonly used during bottom-up approaches leads to a loss of information about posttranslational modifications (PTMs) (phosphorylation, glycosylation, and methylation), which are potential biomarkers in clinical proteomics. This is because smaller proteins (below 30 kDa) and peptides have fewer proteolytic cleavage sites and do not generate enough peptides for confident identification PTMs. Other problems may include the masking of low-abundance peptides by high-abundance species in the generated spectra, leading to a loss of information about low-abundance peptides (Wehr, 2006)

4.3.4. Top-down Proteomics

Top-down proteomics involves separating intact proteins from complex mixtures using conventional separation techniques such as liquid chromatography or 2-DE followed by differential expression analysis using spectrum analysis (such as MALDI or ESI). The generated intact molecular ions are then subjected to gas-phase fragmentation by MS/MS, for subsequent bioinformatics data analysis through database searches (Dalmasso *et al.,* 2009). Top-down proteomics is essential for the identification of small proteins with MWs below 20 kDa. Examples of top-down proteomic techniques are, fourier-transform ion cyclotron resonance (FT-ICR), and surface enhanced laser desorption ionisation time-of-flight mass spectrometry (SELDI-ToF MS).

There are two major advantages of the top-down approach. First, there is the possibility to gain access to the protein sequence and detect the native molecular mass (MW) of the protein; as well as locating and characterising PTMs. Secondly, it is less time consuming compared to the bottom-up approaches due to a simplified sample preparation procedure and the absence of time-consuming protein digestion that is required for bottom-up approaches. The absence of multiple digested peptides means there is an overall reduction in the complexity of the samples to be analysed (Dalmasso *et al.,* 2009).

Top-down proteomics is a relatively new field and is not as widespread as bottom-up proteomics. Limitations of the approach include the fact that it is restricted to isolated proteins or simple protein mixtures because its output is a complex spectra which comprises mainly multiply charged protein ions. Moreover, the fragmentation behaviour of the latter ions is not well understood. Secondly, the favoured FT-ICR instrumentation in top-down proteomics is expensive to purchase and operate. Finally, there are fewer bioinformatics tools for top-down proteomics compared with those for bottom-up proteomics making protein identification more challenging (Wehr, 2006).

4.3.5. Top-down Proteomics without digestion: ‘Intact-cell’ MS by MALDI-ToF MS

The development of a new MS-based top-down protein profiling technique has created a new wave of excitement amongst microbiologists as well as the biomarker discovery community. This approach is called intact- or whole-cell MALDI-ToF MS. The power of MALDI-ToF MS in protein profiling resides in its minimal sample preparation as well as rapid computer-assisted data handling. It should be emphasized that this approach differs from classical proteomics based-approaches. The latter may either employ 2-DE, followed by proteolytic digestion, and extraction prior to MALDI-ToF MS analysis (top-down proteomics); or the digested sample is subjected to clean up prior to tandem MS analysis (LC-ESI-MS/MS) (bottom-up proteomics).

In intact-cell MALDI-ToF MS (ICM), the word “intact” literally means that the cell samples to be analysed are not treated or processed in any way specifically for the removal or isolation of any cellular components. In “intact-cell” analysis, the cells are manipulated only as necessary to transfer them into the mass spectrometer for analysis. The cells are usually applied to the MALDI target plate along with a matrix compound and solvents in which the matrix has been dissolved. However, intimate contact between cells, matrix and matrix solvents means that osmotic pressure may result in stress on the cellular membrane and hence cell disruption. Moreover, in the case of bacteria cells, suspension of cells in solvents to minimise exposure to the organisms in the laboratory may lead to the disruption of the cells. Because none of these steps are intended to disrupt cells, and no steps are added to isolate proteins or other analytes, the technique is called “whole-cells” or “intact cell” MALDI analysis to differentiate it from procedures where additional steps are included in the procedure to deliberately disrupt cellular membranes or separate/recover analytes from the cellular material (Wilkins and Lay, 2006). Consequently, this techniques has been rapidly exploited by microbiologists for the investigation of microorganisms.

4.3.5.1. Intact-cell MALDI-ToF (ICM) MS-based Biomarker Identification in Microorganisms

MALDI-ToF MS analysis of bacterial whole cell proteins, as well as viruses, fungal vegetative cells, and spores, is now well established as reviewed by Lay, (2001); and Fenselau and Demirev (2001). For the analysis of bacteria through ICM, one proposed approach has been the rapid identification of bacteria (Holland *et al.*, 1996). Identification of bacteria is based on the direct comparison of whole cell spectra considered to be the bacterial protein fingerprint, to reference spectra. This has been made possible by the creation of a bacterial fingerprint library of mass spectra from a wide range of known bacterial species (Mazzeo *et al.*, 2006).

A bioinformatics approach for microorganism identification and characterisation which does not involve the use of a fingerprinting library has also been used. This approach is based on matching a set of protein biomarker signal ion molecular weights (MWs) in the spectrum with those of sequence-derived theoretical MWs of proteins (Demirev *et al.*, 1999; Fenselau and Ryzhov, 2001), to identify the protein biomarker signal ions. This is a consequence of the availability of protein biomarkers of bacteria in compiled internet-accessible protein databases of microorganisms with completely sequenced genomes (http://www.uniprot.org/uniprot/).

4.3.5.2. Bioinformatics and Internet Accessible Protein Databases

Bioinformatics is the storage, organisation and analysis of huge amounts of biological data using computational tools and information technology. Such data are typically generated in the form of sequences and structures of proteins and nucleic acids. Genome sequencing of organisms has generated an exponential growth in biological data compiled in databases that are structured, searchable and up-to-date. Protein databases in particular have become a crucial part of modern biology. One of the first steps in the study of a new protein usually involves searching protein databases. Information about the relationship between proteins within a genome or across different species can be obtained by comparing with different proteins or protein families, offering much more valuable information than when studying an isolated protein due to coevolution (evolution of two or more interdependent proteins, each adapting to changes in the other). Since cellular proteins are connected to each other (through pathways and interaction networks), comparing proteins can expose functional interactions between molecules in the cell, generating insights into biological processes of interactions important for cellular function (Tillier and Charlebois, 2009). An example of such a protein database is the UniProt Knowledgebase (UniProtKB).

4.3.5.3. UniProtKB/Swiss-Prot and UniProtKB/TrEMBL Protein Databases

In 2002, the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI) and the Protein Information Resource (PIR) group at the Georgetown University Medical Center and National Biomedical Research Foundation in the U.S, joined forces to create what is known as the UniProt consortium. The mission of UniProt is to provide the scientific community with a single, comprehensive, high quality and freely accessible protein sequence database, UniProtKB (www.uniprot.org). The UniProt Knowledgebase (UniProtKB) is the central access point and consists of two sections: UniProtKB/Swiss-Prot, a manually annotated and reviewed section; and the UniProtKB/TrEMBL, an in-silico annotated section which is not reviewed. Both databases provide sequences and theoretical MWs of 'complete proteome sets', which are the entire set of proteins thought to be expressed by organisms. The bulk of the UniProt complete proteome sets are derived from the translation of completely sequenced genomes of organisms, and normally includes sequences that are derived from extra-chromosomal elements such as plasmids or organellar genomes in organisms (www.uniprot.org).

The sequencing of organisms such as *E. coli* (Blattner *et al.*, 1997), mouse (Waterston *et al.*, 2002), rat (Gibbs *et al.*, 2004) and human (Lander *et al.*, 2001; Venter *et al.*, 2001) genomes, with their eventual placing in the UniProt databases, has provided the possibility of applying the novel bioinformatics approach in protein profiling studies for biomarker discovery and identification in these organisms. Initiatives to completely sequence the CHO cell genome (Hammond *et al.*, 2011; Xu *et al.*, 2011), implies that the bioinformatics approach could also be used for the proteomic profiling of CHO cells. Recently, the complete sequencing of the CHO genome has materialised into sequence-derived theoretical molecular weights (MWs) of CHO proteins becoming increasingly available in compiled internet accessible protein databases such as the UniProtKB database (http://expasy.org/proteomics, 2012). The advantage of requiring minimal sample preparation with MALDI-ToF MS, along with the production of singly charged ions, and the ease with which biomarker signal assignments can be sought from internet accessible databases raises the possibility for the rapid identification of biomarkers for industrially relevant production platforms such as CHO. With the importance of the CHO cell platform for the commercial production of biopharmaceutical therapeutic products, this approach could be an asset for the biotechnology industry.

4.3.5.4. Applications of Intact-cell MALDI-ToF MS (ICM) for the Analysis of Microorganisms

It is now well-established that MALDI-ToF can be used to identify biomolecules above 4kDa. These biomolecules are readily desorbed from unprocessed microorganisms and are intact protein ions (Ryzhov and Fenselau, 2001). Demirev *et al.*, (1999) showed that the proteins of prokaryotic microorganisms have the propensity to fall in the range 4±15kDa MALDI and this is supported in the literature using experimental evidence (Demirev *et al.*, 1999; Ryzhov and Fenselau, 2001; Dieckmann *et al*., 2008; Ilina *et al.*, 2009; Christner *et al.* 2010; Hotta *et al*., 2011; Wang *et al*., 2012).

In the last few years MALDI-ToF MS has been increasingly applied for the identification using the ‘fingerprint’ matching approach, a method that was first introduced by Holland *et al.,* (1996). More recently, MALDI-ToF MS has been applied clinically for microbiological diagnostics. Wang *et al*. (2012) applied the technique for the identification of 65 *Streptococcus pyogenes* isolates, a gram-positive pathogen, obtained from patients. 'Intact cell' measurements obtained using MALDI were matched with reference spectra found in the MALDI Biotyper software through a pattern recognition algorithm. This 'fingerprint' matching facilitated the isentification of 61 of the 65 *S. pyogenes* isolates with 93.85% accuracy. In a similar study involving cilinical proteomics, Ilina *et al.* (2009) sought to identify and categirise the genus *Neisseria* into its pathogenic and non-pathogenic subtypes. Human pathogens, *Neisseria meningitidis* and *Neisseria gonorrhoeae*, as well as several nonpathogenic *Neisseria* species was profiled via MALDI and visual inspections coupled with 'fingerprint' matching using the MALDI Biotyper software successfully distinguished the pathogenic from the the non-pathogenic *Neisseria* isolates.

Several other clinically relevant studies have used the MALDI technique for the identification, diagnosis and hence earlier treatment of bloodstream bacterial infections mainly through the MALDI Biotyper software. Vlek *et al.* (2012) directly profiled a suspension of bacteria and blood cell samples for Methicillin-resistant *Staphylococcus aureus* and vancomycin resistant enterococci to allow for earlier implementation of appropriate antimicrobial treatments. In another of such studies aimed at improving the clinical outcomes of bloodstream infections, Christner *et al.* (2010) used 'fingerprint' matching to reference spectra to identify aerobic and anaerobic bacteria in culture broth, providing identification rates as hiog as 87% with mismatching mostly resulting from insufficient bacterial numbers.

A bioinformatics-based approach for microorganism analysis and identification applies only to microorganisms with sequenced genomes. It exploits the wealth of information contained in prokaryotic genome and protein biomarker sequences in internet-accessible databases like UniProtKB or SwissPROT and uses the fact that the majority of observed biomarkers above *m/z* ratio 4000 in MALDI spectra of intact organisms are proteins. This approach for microorganism identification was first introduced by Demirev *et al.* (1999), who analysed *B. subtilis* and *E. coli*, two organisms with completely sequenced genomes. The spectra *m/z* ratio peaks or signal ions of the microorganisms were tentatively assigned to protein biomarkers, by matching experimental spectra MWs against theoretical MWs for the protein biomarkers based on genome sequences in internet-accessible protein databases. Subsequent ranking of the organisms corresponding to matched ions resulted in the identification of the microorganism.

To further explore the bioinformatics aoproach, a second study was carried out by Ryzhov and Fenselau (2000). In the latter study, various features of the proteins rapidly desorbed by MALDI from intact *E. coli* K-12 cells, mass spectra *m/z* ratio ion peaks were also matched and correlated to protein biomarkers found in internet-accessible databases. Forty *m/z* ratio peaks observed in the mass range 4-20kDa, and the matched proteins were analysed for hydrophobicity, basicity, copy number and location within the cell. It was shown that the bulk of matched proteins originating from the cytosol, were ribosomal, which are abundant within the cell, and are basic in nature with medium hydrophilicity.

More recently, another bioinformatics-based approach has been applied for the identification of the genus *Bacillus* (Hotta *et al.*, 2011). MALDI-ToF MS analysis of ribosomal proteins as biomarkers i.e. S10, S14, S19, L18, L22, L24, L29, and L30, coded in *S10* and *spc* operons successfully distinguished *Bacillus subtilis* subsp. *subtilis* from *B. subtilis* subsp. *spizizenii*. Identification was possible by matching the experimental mass spectrum of the ribosomal protein biomarkers against the in silico–predicted masses of the genome-sequenced *Bacillus* strains. In another recent study, Dieckmann *et al.* (2008) identified and classified various Salmonella subspecies, i.e. *Salmonella enterica* subsp. *enterica*, *S. enterica* subsp. *salamae*, *S. enterica* subsp. *arizonae*, *S. enterica* subsp. *diarizonae*, *S. enterica* subsp. *houtenae*, and *S. enterica* subsp. *indica*, and *Salmonella bongori,* using ‘intact cell’ MALDI analysis, based on matching 200 spectra protein biomarker peaks. These spectra biomarker peaks with masses corresponding mainly to abundant and highly basic ribosomal proteins were matched to biomarker masses of Salmonella genome sequence data in internet accessible protein databases.

4.3.5.5. Applications of Intact-cell MALDI-ToF MS (ICM) for the Analysis of Mammalian cells

‘Intact-cell’ MALDI-ToF MS (ICM-MS) raises many possibilities for the analysis of complex cellular systems like those of mammalian cells. Biomarker profiles have been obtained from whole mammalian cells of neuronal origin (Li *et al.*, 2000; van Veelen *et al.*, 1993) and tissue sections have been profiled (Chaurand *et al.*, 2006; Chaurand *et al.*, 2007; Crossman *et al.,* 2006; Khatib-Shahidi *et al.,* 2006; Reyzer *et al.,* 2007). Differentiation between human (K562 and GM15226) and rodent (BHK21) mammalian cell types through their protein profile 'fingerprints' (Zhang *et al.,* 2006), as well as between monocytes, T lymphocytes and polymorphonuclear leukocyte immune cells (Ouedraogo *et al.,* 2010) have also been carried out using ICM-MS. These applications of ICM-MS in bacterial and mammalian cell protein profiling, demonstrate an outcome of the approach that could be useful for mammalian cell culture (MCC) in bioprocessing.

Despite the numerous applications of ICM-MS to bacteria and mammalian cells, there have been relatively few studies that have applied this approach to MCCs in bioprocessing. ICM-MS has been used in profiling insulin/glucagon-producing pancreatic islet *α*- and *β*-cells (Buchanan *et al.,* 2007); detection of apoptosis in mammalian cells (Dong *et al.,* 2011); and characterisation of batches of monoclonal IgG-producing CHO cell lines (Feng *et al.,* 2010; Feng *et al.,* 2011). The 2006 work from Zhang's group is an example of a MALDI approach where minimal sample pretreatment is involved (Zhang *et al.,* 2006). The paper described a fast and simple approach to cellular protein profiling in which mammalian cells were lysed directly in the MALDI matrix 2,5-dihydroxybenzoic acid (DHB) and mass analysed using MALDI-ToF MS. Similar to the 'fingerprint' approach for microorganism identification, a unique MALDI mass spectral 'fingerprint' was generated in this analysis, to demonstrate that it was possible to differentiate between several different mammalian cell lines.

Buchanan *et al.,* (2007) applied ICM-MS in a direct analysis to cells from two cell lines representative of pancreatic islet α- and β-cells and acquired data in the 2000–20000 *m/z* ratio range. They identified the expected secretory products (i.e. insulin and glucagon) from these intact cultured endocrine cells. Moreover, mass consistent with a protein oxyntomodulin was visualised in the cultured α -cells, a finding that had not been previously reported.

A recent study conducted by Feng *et al.,* (2010) has clearly demonstrated how biomarker profiling by ICM could be exploited to screen cultured mammalian cell lines in bioprocessing. They succeeded in distinguishing viabilities of CHO cells through the different 'fingerprints' of mass spectra after rapid and simple cell pretreatments. A chemometric method (PLS) was used to discriminate between these cell lines with different productivities. As a follow-up of this work, the latter group more recently used PLS-DA to model spectra data sets obtained from another batch of monoclonal IgG-producing CHO cell lines (Feng *et al.*, 2011). In both studies mass spectra peaks from the spectra generated were identified, and hypothesised as being associated with potential protein biomarkers which can be correlated to the productivity of the cell lines.

4.4. Summary

In this chapter a brief introduction to biomarker discovery was presented with particular emphasis on protein biomarker discovery, as well as the biomarker discovery pipeline in clinical research. Moreover, an overview of MS-based has associated advantages and drawbacks.

Several key points are as follows:

- The main advantage of bottom-up proteomics approaches is the ability to achieve high resolution separation, the approaches are amenable to automation, and there is the availability of sophisticated quantitative proteomic technologies.
- The limitations of bottom-up proteomics are that they are generally time-consuming, the samples to be analysed are usually complex, and there is a lack of complete coverage of the protein sequence during analysis.
- The advantages of the top-down approach is the direct determination of the MWs of the proteins, the simplified sample preparation procedure and it is less time-consuming.
- The limitations of top-down proteomics is that it is relatively novel, generates high dimensionality data which is difficult to analyse, and there are fewer bioinformatic tools available for top-down proteomics for protein identification.
- ICM-MS can be applied for the identification of microorganisms through two main approaches: a ‘fingerprint’ approach through direct comparison of intact-cell spectra considered to be a bacterial protein fingerprint to reference spectra; and a bioinformatics approach, by matching a set of experimental protein biomarker MWs in the mass spectra with those of sequence-derived theoretical MWs of proteins in databases.

The bioinformatics approach to microorganism identification through ICM has been widely applied in the field of microbiology as discussed in this chapter. The rapidity and minimal sample preparation advantage of ICM, as well as the usefulness of internet-accessible protein databases demonstrates aspects of this approach that could be useful in rapid biomarker profiling for mammalian cell culture in bioprocessing as well as the rapid sorting of important protein biomarkers. A number of studies have already been reported in terms of mammalian cell lines. Feng's research group (Feng *et al.*, 2010; Feng *et al.*, 2011) in particular has successfully demonstrated the discrimination of monoclonal antibody producing CHO cell lines using ICM and PLS-DA. However, they did not use internet-accessible databases to identify protein biomarkers, an aspect that is explored and applied in this thesis.

'Intact-cell' MS by MALDI-ToF MS is a top-down-based proteomic approach and produces high throughput proteomic mass spectra data with high dimensionality. Preprocessing plays an important part in reducing the dimensionality of such multivariate mass spectra data sets as discussed in chapter 3. Even after performing preprocessing that reduces the dimensionality (number of *m/z* ratio values), such reductions are usually insufficient, hence chemometric techniques have to be applied to the data to further reduce the dimensionality of the data, mine the data, help classify the samples and identify protein biomarkers. The next chapter introduces multivariate data analysis techniques, including principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA), the algorithm primarily used in this thesis (PLS-DA). It also presents some results on the application of PLS-DA to data for a greater understanding of how this technique performs with respect to biomarker discovery/identification and classification.

Chapter 5

5. Partial Least Squares –Discriminant Analysis (PLS-DA) Scores and Loadings Plots

5.1. Overview

The objective of this chapter is to give a general introduction to and to discuss the multivariate data analytical techniques considered in this thesis, principal component analysis (PCA), partial least squares (PLS) and PLS –discriminant analysis (PLS-DA). These methods are appropriate for analysing high dimensional data sets such as mass spectra where the objectives are to obtain an overview of the data (PCA), data modelling (PLS) or classification (PLS-DA). PCA is a multivariate projection method that is designed to extract the systematic variation in a multivariate data set $\boldsymbol{X}$. A brief introduction to PCA including its mathematical basis is provided. PLS is an extension to PCA and is used to develop a model between two blocks of variables, $\boldsymbol{X}$ and $\boldsymbol{Y}$. A brief description of the PLS technique will be given followed by the associated algorithm. Particular attention will be given to PLS-DA as it forms a core aspect in terms of the mass spectra data analysis since it is used for classification and the eventual identification of biomarkers. A detailed explanation of the PLS-DA algorithm is given followed by a review of the application of PLS-DA method to other mass spectra studies. The chapter concludes with the results of where PCA and PLS-DA were used to analyse MALDI-ToF mass spectra data generated from cell lysate samples of *E. coli* K-12 cells during different growth phases. This is to demonstrate specifically how the scores and loadings plot for the PLS-DA model can be easily interpreted and utilised to subsequently identity biomarkers.

5.2. Introduction

The term chemometrics was first introduced in 1971 to describe the application of mathematical, multivariate statistical and other logic-based techniques in the field of chemistry, in particular analytical chemistry. The application of chemometrics has found considerable success in three areas, (a) calibration and validation of biological measurements (multivariate calibration); (b) optimisation of chemical measurements and experimental procedures; and (c) extraction of chemical information from analytical data (classification, pattern recognition, clustering) (Haswell, 1992).

Mass spectra-based proteomic experiments for biomarker discovery typically comprise a data generation stage, a data preprocessing step (described in section 3.4) and a data analysis phase that may include data mining, pattern extraction and peptide or protein identification. The raw MS data has two basic characteristics that serve to guide the mining technology to be applied; firstly, the quality of the mass spectra data and secondly the issue of high-dimensionality (Hilario and Kalousis, 2008). Preprocessing addresses the problem of quality and transforms the MS data into a representation that is then ready to be mined.

A mass spectrum typically comprises thousands of *m/z* ratios and since the sample size (e.g. the number of patients) is relatively small, this results in a so-called 'high dimensionality small sample problem'. This data structure is the feature of microarray and MS data and suffers from the 'curse of dimensionality', i.e. the number of samples needed to describe a (discrimination) problem increases exponentially with the number of dimensions (variables). (Smit *et al.*, 2007). To solve the high-dimensionality problem, it is important to use techniques that are capable of selecting a small number of discriminative variables from the thousands of variables in the spectrum.

5.3. Multivariate Projection Methods and Dimensionality Reduction

There are two approaches to overcome the high-dimensionality problem: variable selection or variable transformation. Variable selection is the extraction of a small subset of variables (*m/z* ratio peak selection) whilst variable transformation is the creation of new latent variables which express relationships between the original variables by applying a mathematical transformation (Fig. 5.1). Fig. 5.1(a) indicates how variable selection deletes some of the variables (X_1 and X_4) from the model, and Fig. 5.1(b) shows how in variable transformation all x–variables and transformed into linear combinations t_1 and t_2 which are related to y in a regression equation.

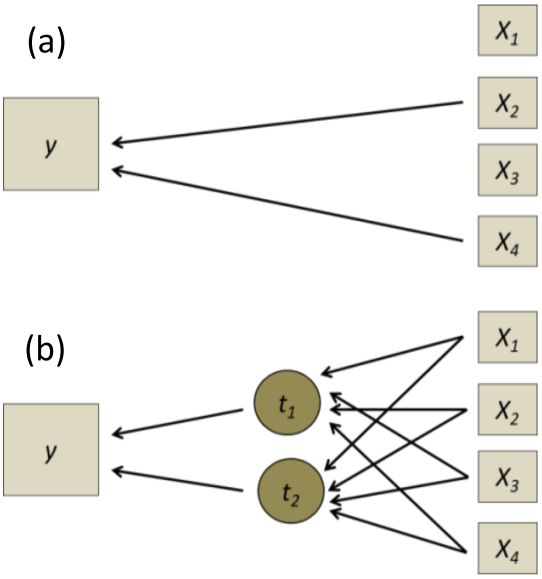


Figure 5.1: Conceptual illustration of the differences between variable selection and variable transformation (adapted from Naes et al., 2002)

A number of variable selection methods have been reported in the literature, including the *t*-test (Wu *et al*., 2003), the X^2-performance through neural-network analyses (Rogers *et al*., 2003) and the Wilcoxon test (Kozak *et al*., 2003). The benefits of variable selection are that it is simple and fast to apply, and the results are interpretable. Variable selection procedures reduce the size of the original variables by removing potentially uninformative variables. The output is a list of variables that contain potentially useful information (Hilario and Kalousis, 2008).

However, information contained in the data may be removed when variable selection is used to identify uninformative variables if interactions and correlations between variables are ignored. Variable transformation methods have the ability to handle the large amounts of data contained within the spectra data sets and overcome the issue of high-dimensionality. Unlike variable selection, such methods use all the variables included in the original data set. The data are projected onto a lower dimensional sub-space, and new components are attained that provide information underlying the structure of the data. Two of these are principal component analysis (PCA) (Pearson, 1901; Hotelling, 1933) and partial least squares analysis (PLS) (Wold *et al.*, 1984).

Supervised methods such as PLS use the class information to construct new components whilst for unsupervised methods, PCA, no class information is utilised (Eriksson *et al*., 1999). During multivariate data analysis of spectra data, an overview of the information contained in the data is usually carried out using PCA, with classification techniques later applied to identify those proteins considered as significant, and which can then be categorised and used as potential biomarkers. PLS is the method that is used in regression modeling between two data matrices (***X*** and ***Y***), with the aim of predicting ***Y*** from ***X*** for new observations.

5.4. Principal Component Analysis

5.4.1. Theory of Principal Component Analysis

Principal component analysis (PCA) is the method used by chemometricians for data compression, information extraction and preliminary visualisation of observations or samples (Hilario *et al*., 2004; Hilario and Kalousis, 2008). PCA's main function is the reduction of the high-dimensionality of the multivariate data to a few dimensions that capture the main source of variability in the data. The new space is defined in terms of principal components (PCs) that are a linear combination of the original variables.

The weights of the individual variables in the principal components are termed loadings. They are useful for identifying the important variables in individual PCs, and also contain information on how the variables relate to each other. Scores are the coordinates of the original data in the new space and contain information on how samples relate to each other with groups of samples indicating similar behaviour (Lee *et al*., 2003).

5.4.2. The PCA Algorithm

There are a number of PCA algorithms, including non-iterative partial least squares (NIPALS), the power method (POWER), singular value decomposition (SVD) and eigenvalue decomposition (EVD). The EVD algorithm is briefly described in this thesis. According to this algorithm, PCA is based on the eigenvalue decomposition of the covariance or correlation matrix of the original data (Wise *et al.,* 2005). For a given data matrix ***X*** with *m* rows and *n* columns, the covariance matrix of is defined as:

$$cov(\boldsymbol{X}) = \frac{\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}}{m-1} \qquad (5.1)$$

provided that the columns of $\boldsymbol{X}$ have been “mean centred”. If the columns of $\boldsymbol{X}$ have been “autoscaled”, equation 5.1 gives the correlation matrix of $\boldsymbol{X}$. PCA decomposes the data matrix $\boldsymbol{X}$ as the sum of the outer product of vectors $\boldsymbol{t_i}$ and $\boldsymbol{p_i}$:

$$\boldsymbol{X} = \boldsymbol{t_1 p_1^T} + \boldsymbol{t_2 p_2^T} + \ldots + \boldsymbol{t_m p_m^T} \qquad (5.2)$$

In equation 5.2, the $\boldsymbol{t_i}$ vectors are known as *score vectors* while the $\boldsymbol{p_i}$ vectors are known as the *loading vectors*. Equation 5.2 can be written in the following matrix form:

$$\boldsymbol{X} = \boldsymbol{TP^T} \qquad (5.3)$$

where $\boldsymbol{T} = [\boldsymbol{t_1}\ \boldsymbol{t_2}\ \ldots\ \boldsymbol{t_m}]^T$ is the score matrix and $\boldsymbol{P} = [\boldsymbol{p_1}\ \boldsymbol{p_2}\ \ldots\ \boldsymbol{p_m}]^T$ is the loading matrix. In the PCA decomposition, the $\boldsymbol{p_i}$ vectors are the *eigenvectors* of the covariance matrix;

$$cov(\boldsymbol{X})\boldsymbol{p_i} = \lambda_i \boldsymbol{p_i} \qquad (5.4)$$

where λ_i is the *eigenvalue* associated with the eigenvector $\boldsymbol{p_i}$.

The PCs are arranged in descending order based on the eigenvalues: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ i.e. the first PC explains the greatest amount of variability with the second PC explaining the next greater amount variability in $\boldsymbol{X}$. As many PCs as variables, if $m > n$, can be calculated but the majority of the variability will be captured in the first few PCs.

Therefore the PCA decomposition of $\boldsymbol{X}$, can be represented as:

$$\boldsymbol{X} = \boldsymbol{t_1 p_1^T} + \boldsymbol{t_2 p_2^T} + \ldots + \boldsymbol{t_k p_k^T} + \ldots + \boldsymbol{E} \qquad (5.5)$$

where $\boldsymbol{E}$ is the residual matrix. In practical applications k must be less than or equal to the smaller dimension of $\boldsymbol{X}$, i.e. $k \leq \min \{m, n\}$. Since $\boldsymbol{E}$ typically contains noise, it has the effect of noise filtering and will not cause any significant loss of useful information.

In this thesis, PCA was first applied to the mass spectra data sets during this project to help identify the major factors that may be useful in terms of differentiating between samples.

5.5. Partial Least Squares Analysis

Partial least squares (PLS) was first introduced by Herman Wold and co-workers, as a method for modeling data sets in terms of chains of matrices, known as path models. Wold then developed NIPALS (Non-linear Iterative Partial Least Squares) algorithm, an efficient way of estimating the parameters in the path models. The acronym PLS (Partial Least Squares) was thus used to refer to these models (Wold, 2001).

The goal of PLS is to relate two sets of variables, predictor variables, ***X***-block, and the response variables, ***y***-block (PLS1). In this thesis, the ***y***-block is a vector but the algorithm can be generalised to the case where ***Y*** is a matrix, PLS2. The basis of the algorithm is:

$$\boldsymbol{X} = \boldsymbol{T}.\boldsymbol{P} + \boldsymbol{E} \tag{5.6}$$

$$\boldsymbol{y} = \boldsymbol{T}.\boldsymbol{q} + \boldsymbol{f} \tag{5.7}$$

where ***q*** and ***f*** are the loadings and residual vectors of the response variables. ***T***, ***P*** and ***E*** are the scores, loadings and residual matrices of the response variables respectively. The dimensions of ***T***, ***P*** and ***q*** are $M \times A$, $A \times N$ and $N \times J$ (J = 1), where A is the number of PLS components (latent variables) retained in the model. The PLS scores are orthogonal (as in PCA) and each latent variable (LV) is obtained by maximising the covariance between ***y*** and the ***X***-variables. Fig. 5.2 is an illustration of the principles of PLS1. The PLS1 algorithm can be found in Fig. C.1, Appendix C.

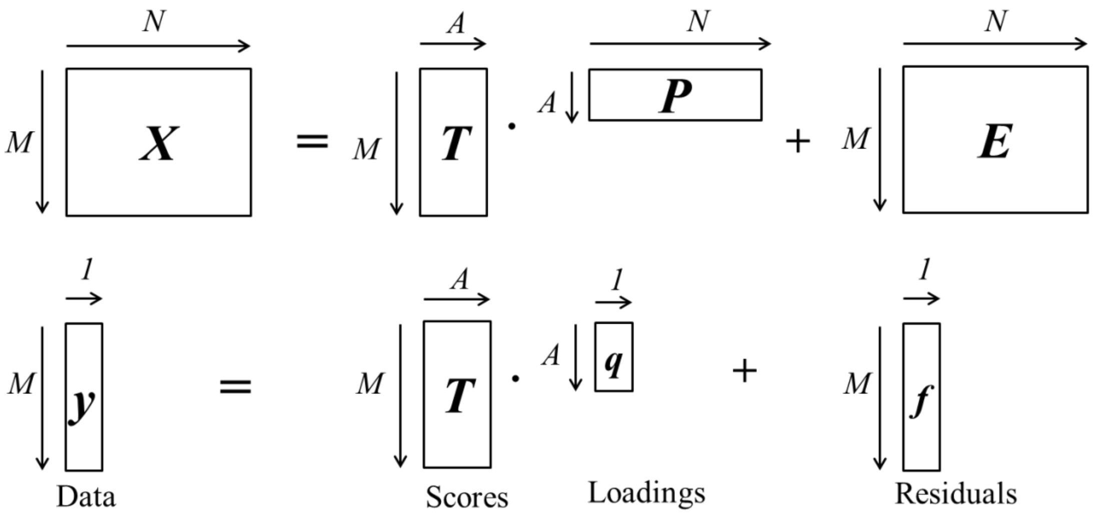


Figure 5.2: An illustration of the principles of PLS1 (adapted from Brereton, 2000)

5.6. Partial Least Squares - Discriminant Analysis (PLS-DA)

PLS-DA is a variant of PLS regression that is used for classification. PLS-DA is used in this thesis as it has a number of advantages over other commonly used classification tools such as soft independent modelling of class analogies (SIMCA), linear discriminant analysis (LDA), canonical correlation analysis (CCA), support vector machines (SVMs) and quadratic discriminant analysis (QDA).

In contrast to PLS-DA, SIMCA computes PCA submodels that captures variation within a class but it does not identify directions in the data space that directly discriminate between classes. The major classification advantage of PLS-DA is that it can handle the situation where the number of variables exceeds the number of samples (high-dimensionality). Overfitting may occur where such models may be describing noise rather than the underlying variability in the data set (Barker and Rayens, 2003). Techniques such as SVMs are more suited for discrimination analysis as opposed to determining the influence of variables (Brereton, 2009). With PLS-DA however, there is the potential to determine the important variables which are responsible for discrimination, and hence enable the identification of biomarkers.

In PLS-DA the prediction matrix, $\boldsymbol{Y}$ comprises the entries denoting the class with as many columns as classes and is termed a ‘dummy’ matrix. If there are two classes to be modeled, PLS-DA is based on the PLS1 algorithm, otherwise PLS2 algorithm is used where the number of class is greater than 2 i.e. $C > 2$ (where C = number of classes). Typically, ‘1’ denotes belonging to a class and ‘0’ not (Fig. 5.3). As shown in Fig. 5.3, the ‘dummy’ matrix contains ‘1’ and ‘0’ which describes class membership of samples in a calibration data set. The matrix has 3 columns (for 3 classes) such that 1st column is ‘1’ and the others are ‘0’ for samples belonging to class one. In practice, the model does not predict either a ‘1’ or ‘0’ precisely, so a threshold value is set, say 0.7, above which a sample is assumed to belong to a class otherwise not (Wise *et al.*, 2005).

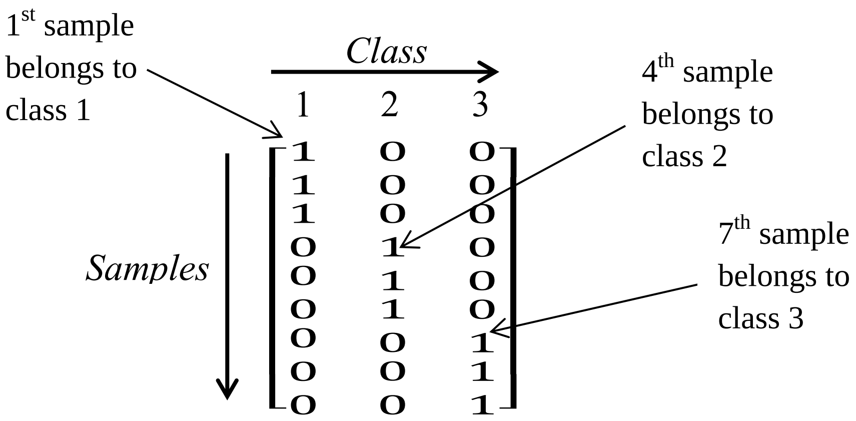

1st sample belongs to class 1
Class
1 2 3
4th sample belongs to class 2
7th sample belongs to class 3
Samples
1 0 0
1 0 0
1 0 0
0 1 0
0 1 0
0 1 0
0 0 1
0 0 1
0 0 1

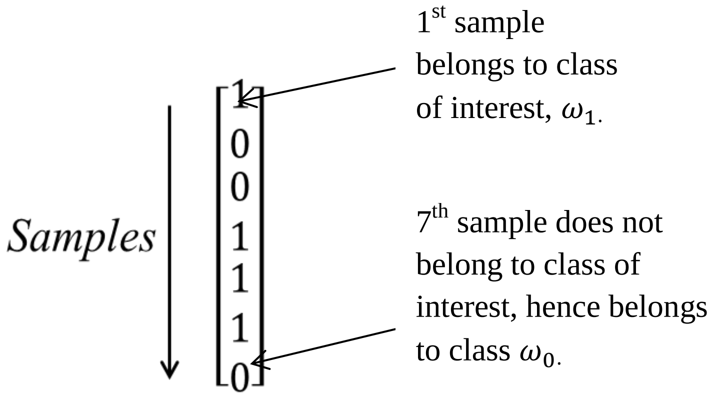
Figure 5.3: A 'dummy' matrix structure showing '1' and '0' denoting class of samples

In order to better define the threshold value for classes, a probabilistic version of PLS-DA has been developed (Pérez *et al.*, 2009). In this version (used in this thesis), the distribution of calibration sample predictions ($\hat{\boldsymbol{y}}$) obtained from a PLS model built for two or more classes to determine a threshold value which will best split the classes with the least probability for false classification for future predictions. If the calibration data contains more than two classes, the thresholds to distinguish each class are determined (Botella *et al.*, 2009).

The algorithm for the probabilistic PLS-DA version calculates the classification threshold based on a probability density function (PDF) with the mean and standard deviation of all the sample predicted responses, $\hat{\boldsymbol{y}}$, for each class. It assumes that the predicted responses for samples for each class in the calibration data set are approximately normally distributed. An empirical PDF is then used to derive posterior probabilities based on Bayes Theorem and a threshold is defined, the value of $\hat{\boldsymbol{y}}$ at which the posterior probabilities of both classes are equal. It is based on this threshold that a sample is assigned to a class. The probabilistic PLS-DA algorithm used in this thesis is described in detail in the following sections.

5.6.1. Summary of PLS-DA Algorithm

The algorithm started with the calculation of a PLS model (see PLS1 algorithm in Fig. C.1, Appendix C) for *A* latent variables (LVs) with spectra data sets (***X***-block) for calibration samples and ***y*** vector (Fig. 5.4). The PLS-DA algorithm applied performed classification on two classes at a time (*C* = 2), so the ***y*** vector contained '1' and '0' which describes class membership of calibration samples as belonging to class, ω_1 (the class of interest) otherwise $\omega_{0.}$ For a sample i (i = 1,2,3,…,m), the value predicted by the PLS model was, $\hat{y}_i$, calculated from equation 5.8, where, $b's$, were the regression coefficients of the PLS model for A LVs. With the $\boldsymbol{y}$ vector coding class membership, calibration or training samples with predicted values close to 1 were assigned to class, ω_1, whilst those with predicted values close to 0 were assigned to class $\omega_{0.}$



1st sample belongs to class of interest, $\omega_{1.}$

7th sample does not belong to class of interest, hence belongs to class $\omega_{0.}$

Figure 5.4: A $\boldsymbol{y}$ vector structure showing ‘1’ and ‘0’ denoting class of samples

The PLS-DA algorithm was as follows:

Step 1: Predicted Responses of Samples

For each training sample i (i = 1,2,3,…,m) the predicted response value, $\hat{y}_i$ was calculated:

$$\hat{y}_i \;=\; \boldsymbol{x}_i^T\hat{\boldsymbol{b}} \tag{5.8}$$

where $\boldsymbol{x}_i$ is the column vector of n m/z ratios measured for that sample and, $\hat{\boldsymbol{b}}$, is the vector of regression coefficients attained from the PLS model.

Step 2: **Gaussian Functions**

For each of the m training samples a Gaussian function centred at the predicted value, $\hat{y}_i$, was calculated:

$$f_i(\hat{y}) = \left(\frac{1}{SEP_i\sqrt{2\pi}}\right) exp^{-\left(\frac{(\bar{y}-\hat{y}_i)}{SEP_i}\right)^2} \tag{5.9}$$

where *SEP*$_i$ is the standard error of prediction of the samples (*i* = 1,2,3,…,*m*), $\bar{y}$ is the known mean of predicted values (for class ω_1 or ω_0), and $\hat{y}_i$ is the predicted value (Fig. 5.5). The figure shows a plot of the potential functions, $p(\hat{y}/\omega_c)$ (*c* = 1 or 0), as a function of the predicted value, which were the Gaussian functions centred at each $\hat{y}_i$, of samples of each class.

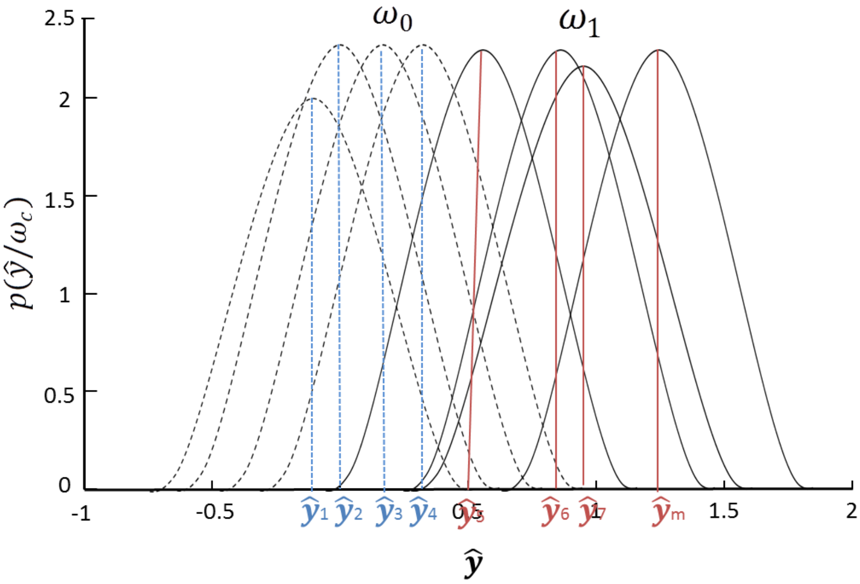


Figure 5.5: Calculation of the Gaussian functions for class $\boldsymbol{\omega_1}$ and class $\boldsymbol{\omega_0}$, centred at each $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$

Step 3: **Probability Density Functions (PDFs)**

The Gaussian functions of the training samples, m_1, belonging to class ω_1, the class of interest, were averaged to obtain the probability density function (PDF) of the class.

$$p(\hat{y}/\omega_1) = \frac{1}{m_1}\sum_{i=1}^{m_c} f_i(\hat{y}) \tag{5.10}$$

The procedure was repeated for the m_0 training samples belonging to the class ω_0.

$$p(\hat{y}/\omega_0) = \frac{1}{m_0}\sum_{i=1}^{m_0} f_i(\hat{y}) \tag{5.11}$$

where m_1 , $p(\hat{y}/\omega_1)$ and m_0 , $p(\hat{y}/\omega_0)$ are the number of samples and PDFs of class ω_1 and class ω_0, respectively (Fig. 5.6). Fig. 5.6 shows a plot of the PDFs, $p(\hat{y}/\omega_c)$ (*c* = 1 or 0), as a function of the predicted value. The PDFs were the averages of the Gaussian functions centred at each $\hat{y}_i$ of samples for the two classes, ω_1 and ω_0.

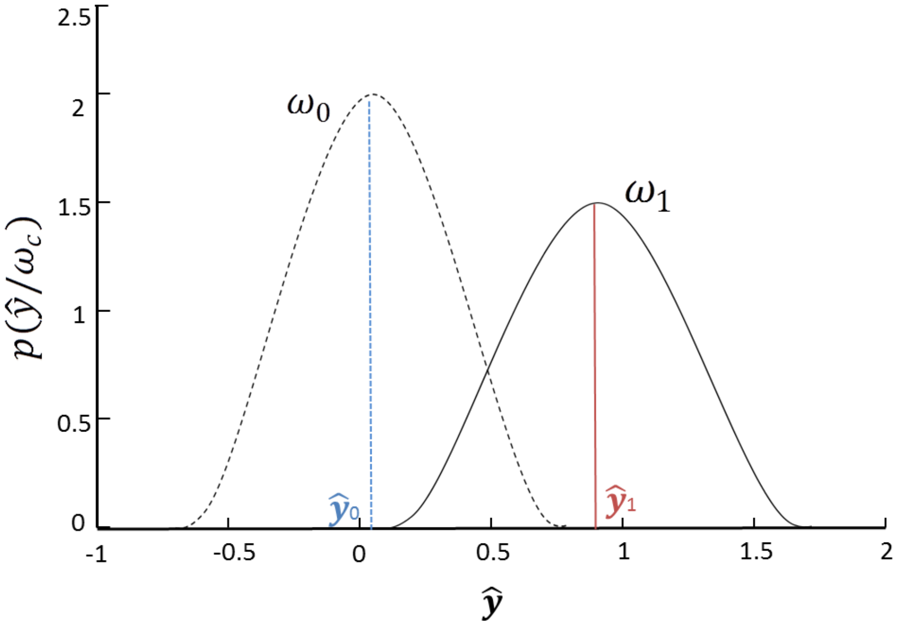


Figure 5.6: PDFs for classes $\boldsymbol{\omega_1}$ *and class* $\boldsymbol{\omega_0}$ *was calculated as the average the individual Gaussian functions for each class in Fig. 5.5*

Step 4: **Posterior PDFs**

Using Bayes theorem, the posterior PDFs were calculated:

$$P(\omega_1 / \hat{y}_i) = \frac{p(\hat{y}_i / \omega_1).P(\omega_1)}{p(\hat{y}_i)} \tag{5.12}$$

$$P(\omega_0 / \hat{y}_i) = \frac{p(\hat{y}_i / \omega_0).P(\omega_0)}{p(\hat{y}_i)} \tag{5.13}$$

where $p(\hat{y}/\omega_1)$ and $p(\hat{y}/\omega_0)$ are the conditional probabilities while $P(\omega_1)$ and $P(\omega_0)$ are the prior probabilities. Both the priors were estimated as the proportion of samples of each class in the training set, as the set was a representative of the total number of samples, *m*, that is:

$P(\omega_1) = m_1/m$ and $P(\omega_0) = m_0/m$, where $m = m_1 + m_0$.

The denominator of equation (5.12) and (5.13) was:

$$p(\hat{y}_i) = p(\hat{y}_i /\omega_0).\, P(\omega_0) + p(\hat{y}_i /\omega_1).\, P(\omega_1) \tag{5.14}$$

Step 5: **Classification Threshold**

The value of $\hat{y}$ (predicted value) corresponding to the point where the two posterior probability functions (for class ω_c and ω_0) are equal is identified as the threshold value, λ_c, for the class of interest. The threshold value, λ_c, for each class can then be used to classify the test samples (Fig. 5.7).

Fig. 5.7(a) shows a plot of the PDFs, $p(\hat{y}/\omega_c)$ multiplied by the prior probability $P(\omega_C)$ (c = 1 or 0), as a function of the predicted value, whilst Fig. 5.7(b) shows a plot of the posterior probability, $P(\omega_c/\hat{y})$, versus the predicted value. Fig. 5.7(b) also shows the point where posterior probabilities were equal, indicating the classification threshold.

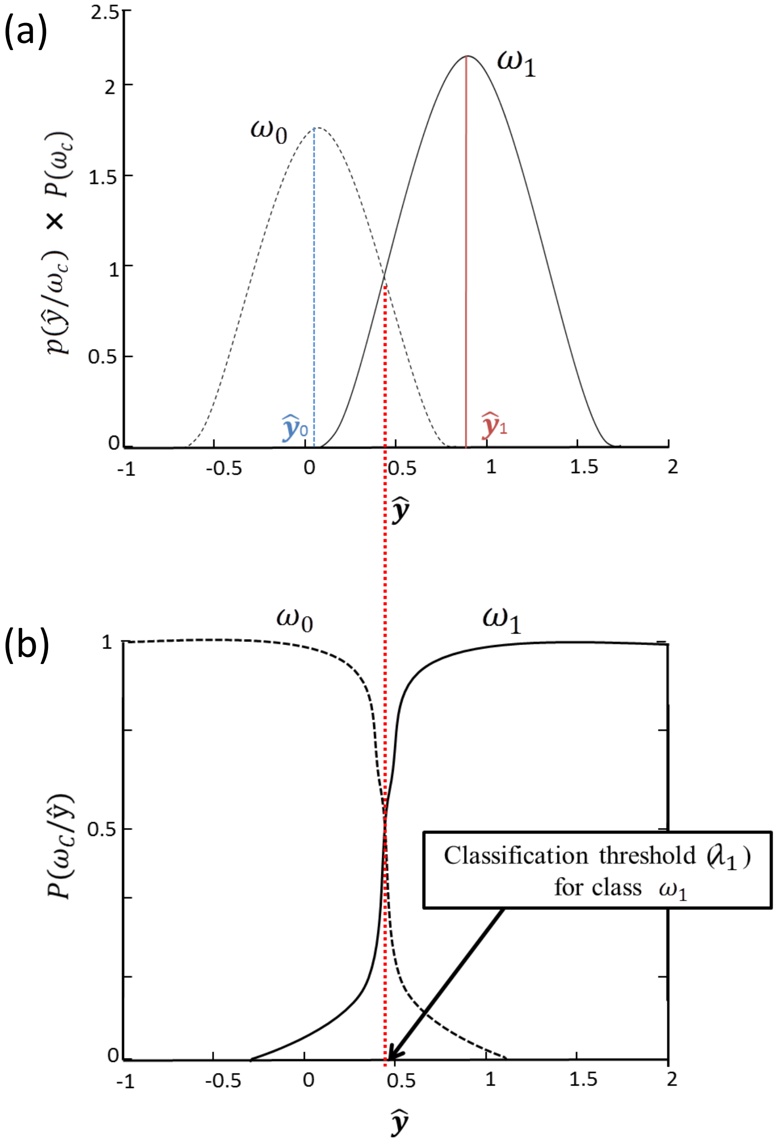


Figure 5.7: PDFs for classes $\boldsymbol{\omega_1}$ *and class* $\boldsymbol{\omega_0}$ *and classification threshold for class* $\boldsymbol{\omega_1}$

Step 7: **Classification Rule**

The predicted, $\hat{y}_{test}$, for a test sample was first calculated from equation 5.8, and the samples were then classified according to the following rules:

Assign the sample to: class ω_1 if $\hat{y}_{test} > \lambda_1$; otherwise assign to class ω_0.

A flow chart of the PLS-DA algorithm described above for the modelling of MALDI-ToF mass spectra data sets is shown in Fig. 5.8.

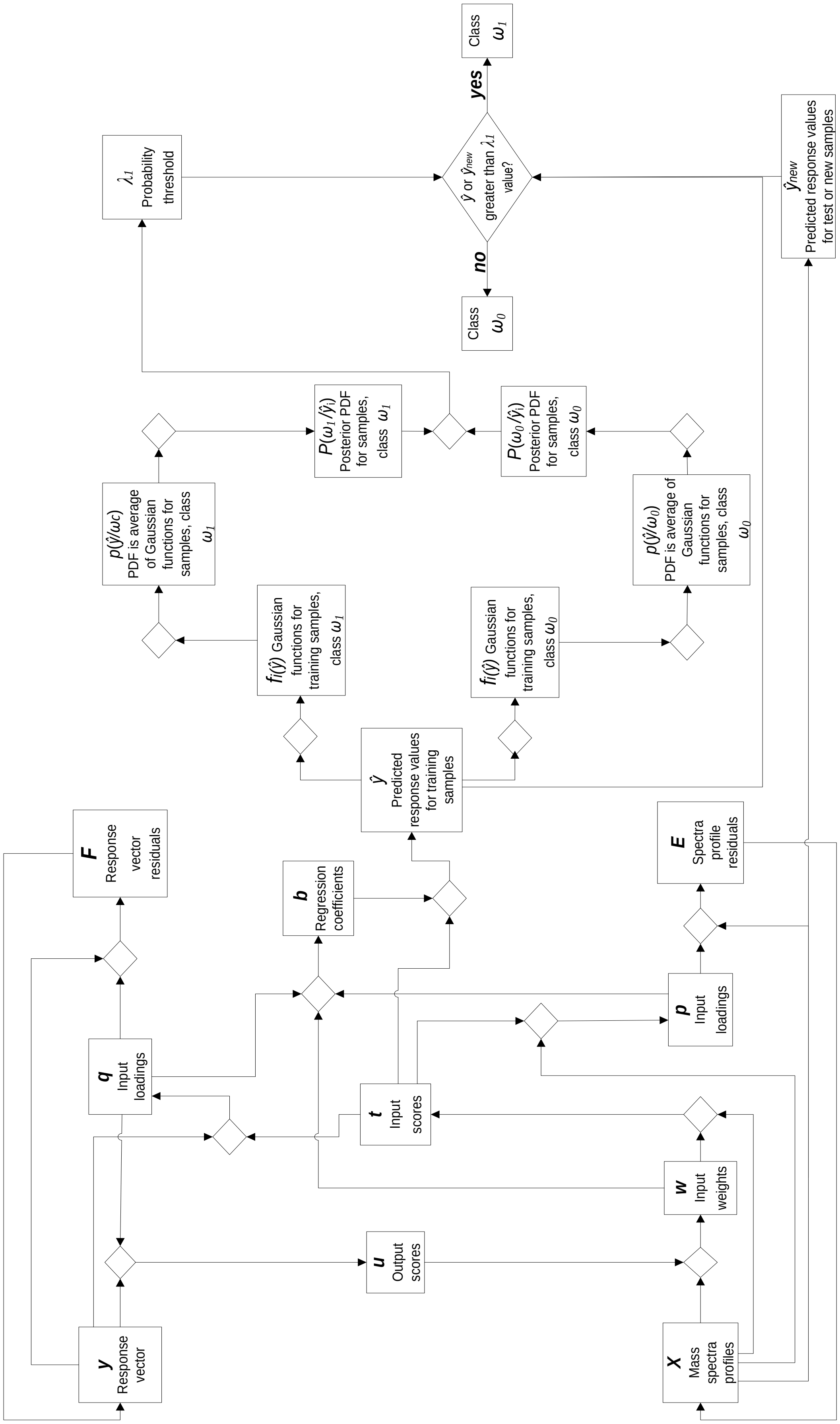

Class ω_1
yes
λ_1 Probability threshold
$\hat{y}$ or $\hat{y}_{new}$ greater than λ_1 value?
no
Class ω_0
$\hat{y}_{new}$ Predicted response values for test or new samples
$p(\hat{y}/\omega c)$ PDF is average of Gaussian functions for samples, class ω_1
$P(\omega_1/\hat{y}_i)$ Posterior PDF for samples, class ω_1
$P(\omega_0/\hat{y}_i)$ Posterior PDF for samples, class ω_0
$p(\hat{y}/\omega_0)$ PDF is average of Gaussian functions for samples, class ω_0
$f_i(\hat{y})$ Gaussian functions for training samples, class ω_1
$f_i(\hat{y})$ Gaussian functions for training samples, class ω_0
$\hat{y}$ Predicted response values for training samples
F Response vector residuals
b Regression coefficients
E Spectra profile residuals
q Input loadings
t Input scores
p Input loadings
w Input weights
u Output scores
y Response vector
X Mass spectra profiles

Figure 5.8: Flow chart of the PLS-DA algorithm for modeling of MALDI-ToF mass spectra data sets

5.6.2. Applications of the PLS-DA Algorithm

5.6.2.1. MALDI-ToF Mass Spectra Data of *E. coli* K-12 Cells at Different Growth Phases

The PLS-DA algorithm (described in section 5.6.1) was initially applied to MALDI-ToF MS data obtained from *E. coli* cell samples to determine if such an approach could be used to distinguish between and characterise different growth phases. For the application of PLS-DA, the training data set consisted of a matrix, $\boldsymbol{X}$, of *E. coli* culture MALDI spectra, comprising 7000 *m/z* ratio values measured for 300 samples, and a dependent variable vector, $\boldsymbol{y}$ (300 × 1) in which the class of each sample was coded as '0' or '1' depending on whether it belongs to the class of interest or not. The PLS-DA algorithm performed classification on two classes at a time and the classes were defined as follows: ω_1, class 1 represented the decline phase; $\omega_{2,}$ class 2 represented the exponential phase; and ω_3, class 3 represented the stationary phase. One of the classes was defined to be the one of interest, i.e. coded as '1' and the other two were coded as '0' i.e. class $\omega_{0.}$

5.6.2.2. MALDI-ToF Mass Spectra for IgG Monoclonal Antibody producing CHO Cell Lines

MALDI-ToF mass spectra were generated from IgG monoclonal antibody producing Chinese hamster ovaries (CHO) cell lines. The PLS-DA algorithm was also applied to model the data to discriminate between high and low producer cell lines. For this data set, the training sample consisted of a matrix, $\boldsymbol{X}$, the CHO cell line MALDI spectra, comprising 18092 *m/z* ratio values measured in 44 samples, and a dependent variable vector, *y* (44 ×1), in which the class of each sample was coded as '0' or '1' depending on whether it belongs to the class of interest or not. The classes were defined as follows: ω_H, class 1 represented the high producer cell lines (Hs) whilst $\omega_{L,}$ class 2 represented the low producer cell lines (Ls).

5.6.3. PLS-DA Model Quality and Performance Evaluation

5.6.3.1. Calibration

After the application of PLS-DA the quality and performance of the models is assessed mainly through their ability to predict unknown response (*y*) values. This is even more important when selecting the number of latent variables (LVs), *A*, to include in the model. It is important to select an appropriate number of LVs which guarantees good model quality and performance without overfitting i.e. when the model describes noise rather than the underlying variability in the data set. One measure used is the root mean square error of calibration (RMSEC) (Naes *et al*., 2002). It gives indication about the fit of the model to the calibration data set. It can be used to access the model quality and it is a measure of how well the model fits the data. It is the error of calibration and the smaller it is, the greater is the model quality.

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{I} \left(\hat{y}_i - y_i\right)^2}{M}} \qquad (5.15)$$

where y_i and $\hat{y}_i$ are the actual and predicted value of the response for the *i*th calibration sample respectively, and *M* is the number of calibration samples.

5.6.3.2. Cross-validation

Cross validation involves the removal of a subset of samples from the calibration set, and the construction of a model using the remaining samples, and the subsequent application of the resulting model to the samples withheld from the calibration set. This way the model is tested with samples that were not used to build the model (Naes *et al*., 2002). The process is repeated for a number of subsets of the calibration data set. Estimation of the root mean square error (RMSE) based on this technique is the RMSE of cross-validation (RMSECV). The RMSECV is a measure of the model's ability to predict samples that were not used to build the model. A good model compared to others is one that has the lowest RMSECV.

5.6.3.3. Prediction Testing

Prediction testing is an external validation technique based on splitting the samples into two, one set called the training or calibration set and the other called the test set. The prediction testing estimate is called the root mean square error of prediction (RMSEP). The latter is a RMSE assessment involving the use of a test set of samples that have known *y*-values. RMSEP is obtained when the model is applied to the test data.

5.6.3.4. Coefficient of Determination (R^2)

The coefficient of determination (R^2) of a model is a measure of the goodness of fit of the model and can be used to assess the quality of the model. It lies between 0 to 1 and the quality of the model improves as R^2 gets closer to 1. It is calculated as follows:

$$R^2 = \sqrt{\frac{\sum_{i=1}^{M}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{M}(y_i - \bar{y})^2}} \qquad (5.16)$$

where y_i and $\hat{y}_i$ are the actual and predicted value of the response for the *i*th sample respectively; is the average response for calibration (R^2), for validation (R^2_{CV}), or for prediction (R^2_p); and M is the number of samples.

5.7. Review of the Applications of PLS-DA on Mass Spectra Data

PLS-DA modelling has been applied in a number of areas of proteomic research and technology (Lee *et al*., 2003; Norden *et al.,* 2005; Liley and Cupree, 2006; Pierce *et al*., 2006; Feng *et al*., 2011). Lee *et al*. (2003) directly applied PLS-DA to address the high-dimensionality problem involving a proteomic mass spectra data set comprising 60,000 *m/z* ratio variables. The PLS-DA model performed well, reducing the data set to 545 *m/z* ratio variables, and clearly identifying biomarkers that potentially contributed to the discrimination between normal and diseased specimens of cancer patients.

In another biomarker identification study, Norden *et al.,* (2005) applied PCA and PLS-DA to mass spectra data from clinical urine samples. A number of peptide-biomarker fingerprints related to the diagnosis and progression of chronic obstructive pulmonary disease were identified. A similar approach was adopted by Liley and Dupree, (2006), to study plant organelles. Quantitative proteomic analysis using differential isotope tagging strategies coupled to non-gel-based LC-MS allowed proteins in different organelles to be discriminated between based on their differential fractionation in density gradients of the LC. PCA and PLS-DA scores plots showed clustering of proteins according to their subcellular localisation.

PLS-DA was applied to linear mode MALDI-ToF mass spectra data to confirm the identification and presence of the microorganism *Coxiella burnetti* as a category B bioterrorism agent in the U.S. (Pierce *et al*., 2006). The mass spectra data were preprocessed (normalisation, baseline-correction, filtering, and binarisation) prior to being modeled by PLS-DA using leave-one-out cross-validation. The model was validated by the prediction of unknown *C. burnetii* test samples resulting into a 100% sensitivity (proportion of actual *C. burnetii* samples correctly classified) and specificity (the proportion of non *C. burnetii* samples correctly classified), and successfully identifying five out of six strains of the microorganism.

A recent study conducted by Feng *et al*., (2011) demonstrated how biomarker profiling by 'intact cell' MALDI-ToF mass spectrometry in combination with chemometrics could be exploited to screen cultured mammalian cell lines in bioprocessing. They succeeded in distinguishing viabilities of IgG-producing CHO cell lines through different 'fingerprints' of mass spectra for the CHO cells. PLS-DA was used to discriminate between the cell lines with different productivities. In this study, *m/z* ratio peaks from the spectra generated were identified, and hypothesised as being associated with potential protein biomarkers which could be correlated to productivity of the cell lines.

5.8. Multivariate Data Analysis of MALDI-ToF Mass Spectra Data from *E. coli* K-12 Cell Lysate at Different Growth Phases Using PCA and PLS-DA

The multivariate projection methods (PCA and PLS-DA) were applied to MALDI-ToF mass spectra data generated from cell lysate samples of *E. coli* K-12 cells at different growth phases. The application of PCA and PLS-DA to the spectral data was carried out using the MATLAB® software v.7.6.0.324 (R2008a, The MathWorks, Inc.) and the MATLAB® PLS Toolbox v.3.5 (Eigenvector Research, Inc.).

5.8.1. Data Sampling

The cell lysate sample preparations as well as MALDI-ToF analysis of the samples to generate the spectra data sets has been described in chapter 3.3.7.1. The data were first divided into a training and a test data set, for each group of spectra samples (exponential, stationary and decline phases) using simple random sampling. There is no generally accepted rule as to the proportion of samples to assign to the test and training set (Brereton *et al*., 2009). In this application, 4/5 of the samples were placed in the training set whilst the rest were included in the test set.

From a total of 120 exponential phase cell lysate spectra samples, 100 were selected as training samples and 20 as test samples. For the stationary phase samples, the split was 100/16 training/test, and for the decline phase the split was 100/19 training/test. This gave a total of 300 spectra for training and 66 for testing. The data set was represented by three classes: class 1, decline phase; class 2, exponential phase; and class 3, stationary phase. The data sets were first preprocessed (section 3.6) and PCA was then applied to the preprocessed mass spectra.

5.8.2. Results

The results in this section are focused primarily on the interpretations of PCA and PLS-DA scores and loadings plot to emphasise their importance in this work, and to specifically demonstrate how the PLS-DA scores and loadings are graphical representations that can easily be interpreted and subsequently used in biomarker identification.

5.8.2.1. Interpretation of Principal Component Analysis Scores Plot

Prior to sampling the data into training and test sets, PCA was applied to get an overview of the 355 preprocessed cell lysate mass spectra to identify any groupings and to determine if differences occur due to the growth phase. The first three principal components (PCs) were retained and these account for 93.36% of the variability (Table B.1 in Appendix B). The PC scores plots are shown in Fig. 5.9(a) and (b). The Samples are colour-coded according to their growth phase or class and the ellipse (blue dashed line) represents the 95% confidence region based on Hotelling's T^2. The three-component model shows clear evidence that class 2 (exponential phase samples) are separate from the other two classes along the second principal component (Fig. 5.9(a)). Separation between class 1 and 3 is not evident from the three PCs. These preliminary observations suggest that a major part of the spectral variation is related to the growth phase.

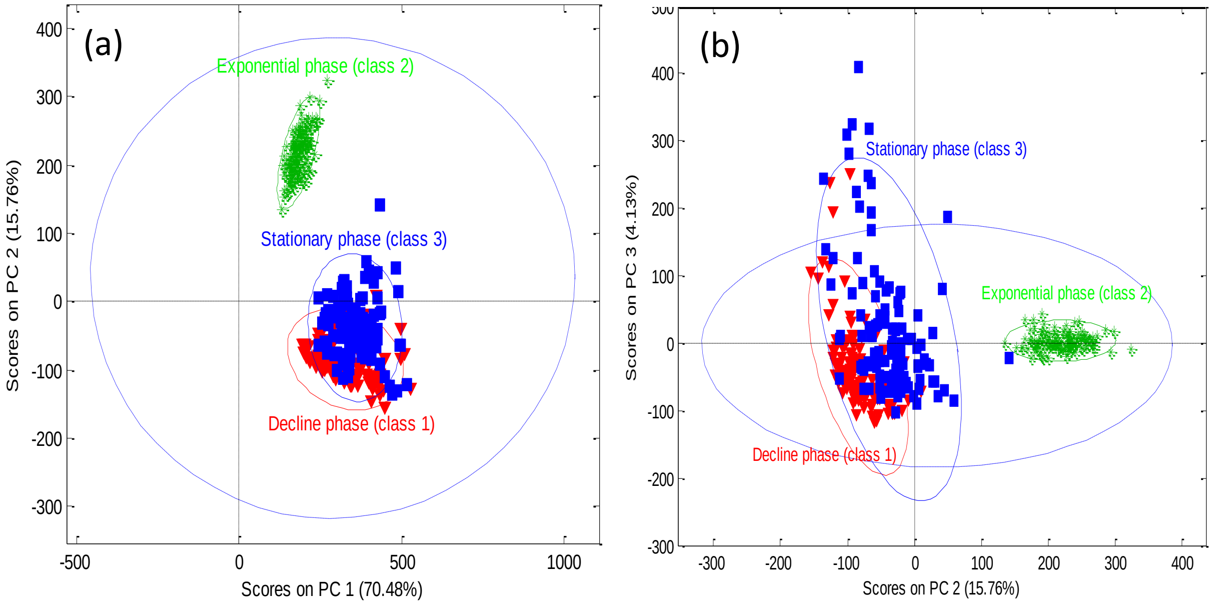


Figure 5.9: PCA overview of 355 mass spectra data sets for E. coli cell lysate samples

5.8.2.2. PLS-DA Model Quality and Performance

The next step was to apply PLS-DA to the full preprocessed 300 calibration mass spectral profiles from the *E. coli* cell lysate. During PLS-DA model calibration, leave-one-out-cross-validation (LOOCV). In practice, the cross-validation (CV) method used here can be termed a leave-class-out-cross-validation (LCOCV), a modified version of LOOCV, where all samples belonging to the same class were removed from the calibration data set and a sub-model based on the remaining samples was used to build the PLS-DA model and predict the left out samples.

The process was repeated with all the 300 calibration data set until each of the three classes had been left out once. Since samples of the same class were replicates, LCOCV helped to avoid the ‘replicate sample trap’; where the presence replicates of the same physical sample in both the calibration and test data sets may lead to overly optimistic cross-validation results giving a biased estimate of the error rate (Hansen *et al.*, 2009). A cross-validation was customised by creating a vector (Fig. B.1, Appendix B) and specifying how the cross-validation was to be performed. The validation residual variance, the RMSECV was then computed to help identify the number of LVs to retain. Fig. 5.10 shows a plot of RMSECV as a function of number of LVs. For the class 2 samples (RMSECV 2), the error stabilises after 4 LVs whilst for the other two classes

(1 and 3) the RMSECV is a minimum for LV1. From Fig. 5.10, as the results were not conclusive 3 LVs were retained.

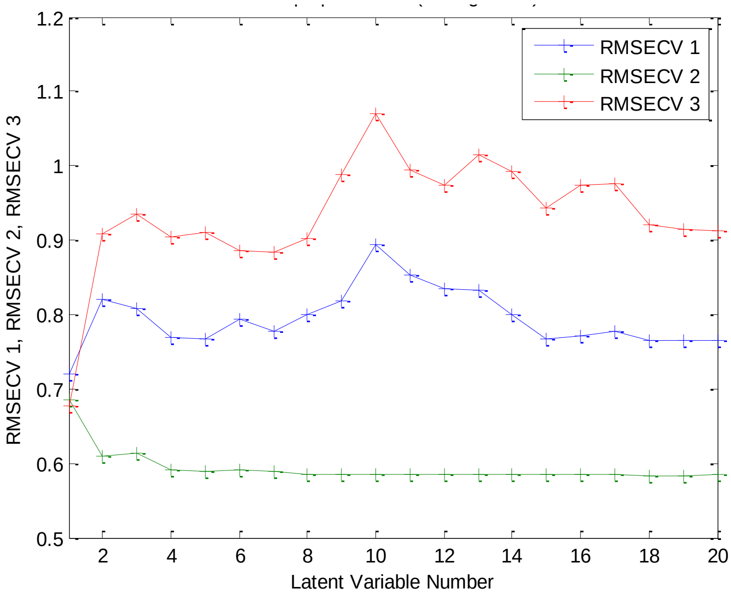


Figure 5.10: PLS-DA RMSECV for choosing latent variables

The percentage of variance explained for the individual LVs and the individual and cumulative explained is reported in Table B.2 (Appendix B), for the 3 LVs. The 3-LV PLS-DA model captured 90.50% of the ***X***-block variance and explained 85.25% of the ***Y***-block variance of the training data set. From the results reported in Table B.3 (Appendix B), the quality of this model was good, according to the values of R^2 calculated both fitting (73.1% for class 1, 93% for class 2 and 67.2% for class 3) and in cross-validation (86.3% for class 3). Table B.3 shows that the quality of the model was good both in calibration and CV with a sensitivity (proportion of samples correctly classified that belong to the class being modeled), and specificity (proportion of samples correctly classified that do not belong to the class being modeled) of at least 92%. This suggests that the model had a good practical value.

The same conclusions can be drawn from Fig. B.3 (Appendix B) representing the calculated *versus* the measured responses both in fitting and in prediction (using test sample). This suggests that all the useful information was taken into account by the model. The model was successfully validated using the test data sets, and results of the model performance are summarised in Fig. B.3 and Table B.4 (Appendix B). The successfully validation of the model means that the model parameters such as scores and loadings are valid.

5.8.2.3. Interpretations of PLS-DA Scores Plot

The practical value of the model was demonstrated by its ability to discriminate between the three classes as shown in the scores plot (Fig. 5.11). A plot of the first two components (LV1 vs LV2) shows that the model concentrates most of the discriminating information into the second LV (Fig. 5.11(a)). Exponential phase (class 2) samples appear quite distinct from the other two classes and have high positive scores along LV2 well separated from the other two classes with negative scores along the same LV. A few stationary phase samples however are borderline and have low positive scores. This observation is supported by a plot of scores on LV2 as a function of samples (Fig. 5.11(c)). A plot of LV2 vs LV3 (Fig. 5.11 (b) and (d)). shows that there is an indication of separation between stationary phase (class 3) and decline phase (class 1) samples along LV3, with overlapping. Majority of class 3 samples have a positive score whilst their class 1 counterparts have a negative score (Fig. 5.11(b)). This observation is supported by a plot of scores on LV3 as a function of samples (Fig. 5.11(d)).

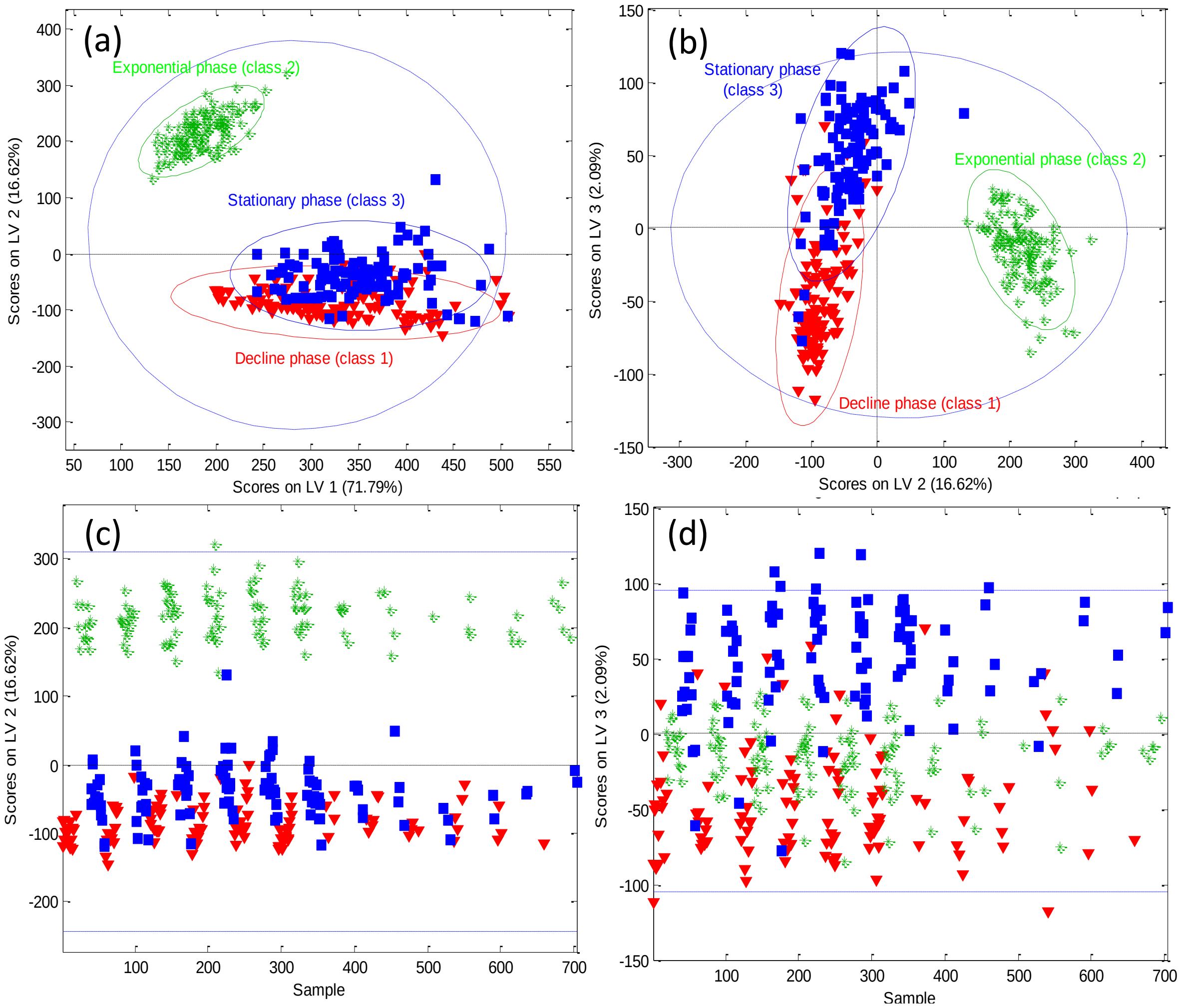


Figure 5.11: PLS-DA scores plots for the 300 preprocessed calibration spectra data sets for cell lysate

5.8.2.4. Interpretations of PLS-DA Loadings Plot

The PLS-DA loadings plot for a specific LV potentially contains valuable information regarding the regions of the mass spectra (*m/z* ratio peaks or signal ions) that contribute to the ability of the PLS-DA model to distinguish between classes. There is a direct geometric link between the scores and the loadings plot of PLS-DA, so interpreting scores and loadings plot together could provide the possibility of identifying the mass spectra regions (from the loadings plot) that influence the behavior of samples in the scores plot.

The loadings plot for LV2 and LV3 (Fig. 5.12) indicates which variable loadings (associated with *m/z* ratio signals in the MALDI data sets) that contain information that is causing the separation of the samples in the scores plot. The loadings describe the weighting coefficients for each *m/z* ratio signal ions (experimental MWs of proteins) with the magnitude of the variable loadings being indicative of how the expression of proteins varies with growth phase. In Fig. 5.12(a), it can be seen that variables 3578, 3462 and 2679 have absolute loadings of large positive magnitude. Since the exponential phase cell samples have positive scores along LV2, this implies that these samples can be differentiated from the other two classes by these *m/z* ratio signal ions associated with variables of large positive loadings in the LV2 loadings plot. These signal ions potentially identify those proteins which are differentially expressed in cells during the exponential phase and are most likely to be biomarkers of the exponential phase.

For the loadings associated with LV3 (Fig. 5.12(b)), the variables 3110, 2635 and 2293 have loadings with large positive magnitude. Most stationary phase cell samples can thus be differentiated from decline phase samples by the presence of *m/z* ratio signals associated with variables with positive loadings in the LV3 loadings plot. These *m/z* ratio signal ions are indicative of those proteins which are differentially expressed in cells during the stationary phase. Variables 3639, 3095 and 2684 have loadings with large negative magnitude on the LV3 loadings plot (Fig. 5.12(b)), and since the decline phase cell samples have negative scores along LV3, they can be distinguished from the stationary phase cell samples by the differential expression of proteins associated with signals with positive loadings in the LV3 loadings plot.

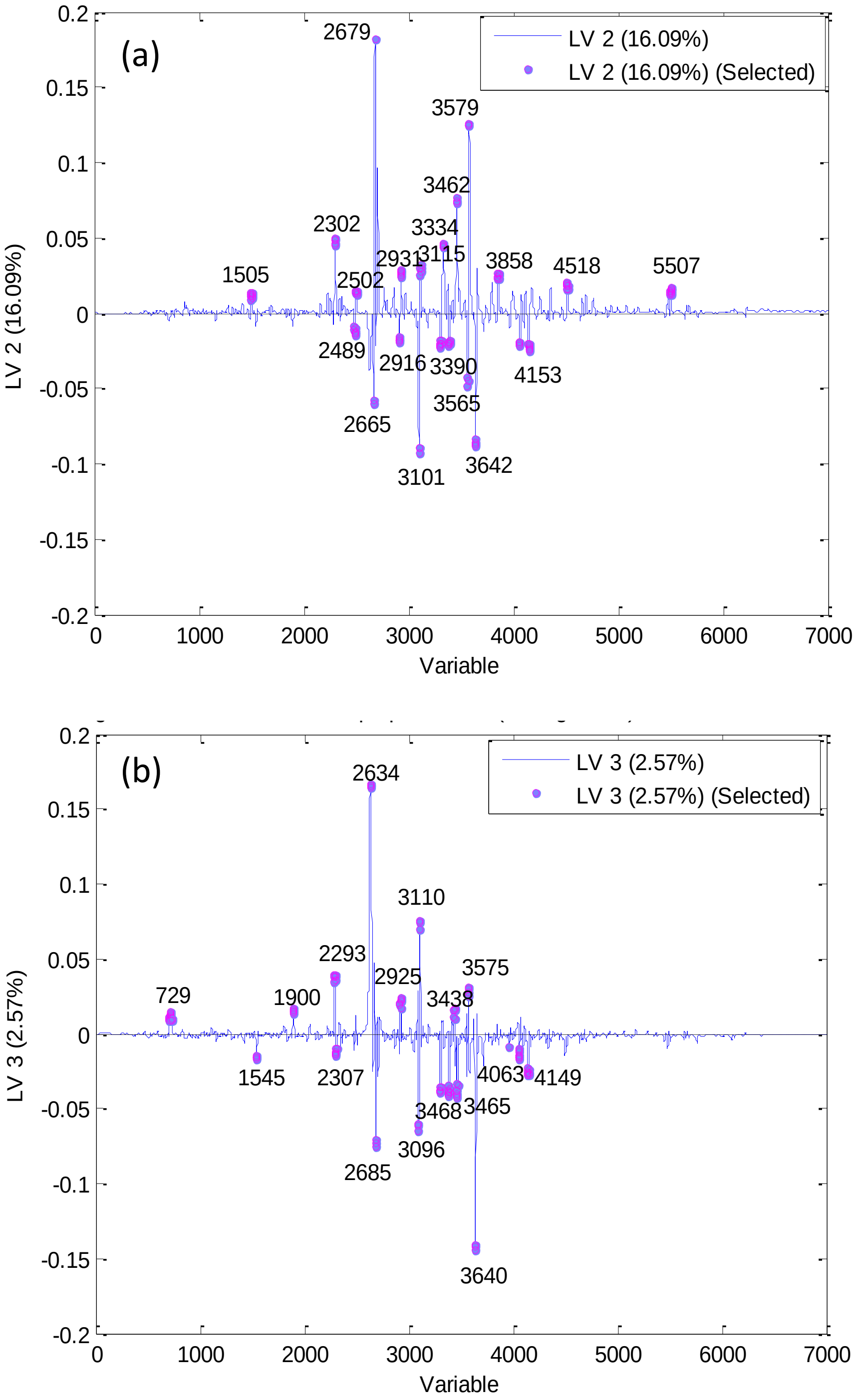


Figure 5.12: PLS-DA loadings plots on LV2 and LV3

To identify protein biomarkers through data base searches, PLS-DA loadings plots for LV2 and LV3 can be considered for further analysis. The variables from the loadings plot are associated to *m/z* ratio peaks or signal ions, and since most of the *m/z* ratio signal ions are singly charged protonated protein (MH^+) molecules, they represent the approximate MALDI experimental molecular weights (MWs) of the ionised proteins expressed by the cells at the different growth phases. The MWs of protein ions can be used as search parameters and matched with sequence derived theoretical MWs of proteins in the UniProtKB/Swiss-Prot database (http://expasy.org/proteomics). Details of how the database searches are carried out as well as the identified protein biomarkers are explained in section 6.5 (chapter 6). The biological implications of such identified biomarkers are also given subsequently.

5.9. Summary

In this chapter an introduction to multivariate data analysis was given focusing on the dimensionality reduction techniques of PCA and PLS-DA. It has been proposed in the literature that these methods are more realistic for multivariate data analysis of data sets such as mass spectra because of their ability to handle large amounts of data by reducing the high-dimensionality, handling correlated variables, ability to handle missing data, and providing graphical representations that are informative and easily interpretable. PCA, PLS and PLS-DA were introduced including the underpinning theory.

The probabilistic PLS-DA approach was explored and the PLS-DA algorithm used in this thesis was outlined. Applications of PLS-DA algorithm were demonstrated using MALDI-ToF mass spectra data obtained from *E. coli* cell samples to distinguish between and characterise different growth phases. A second application for the algorithm considered spectra from IgG monoclonal antibody-producing CHO cell lines.

The chapter is concluded with results of an example where PCA and PLS-DA were applied to MALDI-ToF mass spectra data generated from cell lysate samples of *E. coli* K-12 cells at different growth phases. Results of the models were shown and interpreted. Scores plot (shows relationships among the samples as in PCA) were also shown and interpreted, in order to demonstrate how results from these multivariate projection methods are easily interpretable. PLS-DA loadings (parameters that also supply information related to the variables) were also explained and interpreted. The importance of using loadings plots for database searches for biomarker identification as well as how this is carried out is explained in section 6.4.3 (chapter 6).

Chapter 6

6. CASE STUDY I: Biomarker Profiling Of *E. coli* Utilising the ICM, PLS-DA, and Protein Database Search Approach

6.1. Overview

In previous sections, the potential and applications of the methods of intact-cell MALDI-ToF MS (ICM-MS) (section 4.3.4.1), projection to latent structure – discriminant analysis (PLS-DA) (section 5.6), and protein database search (section 4.3.4.3) in various areas of proteomics have been discussed. In this chapter, the approach of combining ICM-MS, PLS-DA, and protein databases search is applied for growth phase-associated protein biomarker profiling of *E. coli* K-12 cultures. This case study was going serve as a proof-of-concept study for applying this approach for biomarker profiling of IgG-producing CHO cell lines during bioprocessing discussed in chapter 7.

6.2. Specific Aims of Chapter

The purpose of this study is to serve as a proof-of-concept study for applying the approach (ICM-MS and PLS-DA along with a database search) for biomarker profiling of IgG-producing CHO cell lines during bioprocessing. In this study a ICM-MS and PLS-DA along with a database search, is used to identify potential protein biomarkers associated with the different growth phases (exponential, stationary and decline) of unprocessed whole (or intact) *E. coli* K-12 culture.

Biologically, it is expected that *E. coli* cultures at the three different phases of growth exhibit different metabolic profiles and hence different protein expression patterns so that unique proteins can be induced and differentially expressed by these cultures. It is anticipated that the latter protein expression patterns correlate with the growth phase of the culture, and hence between-class (exponential, stationary and decline phase classes) variability will be evident. The multivariate classification method (PLS-DA) is used to capture these differences and hence classify the *E. coli* cells. A database search (based on the bioinformatics approach of microorganism identification), by matching the MALDI spectra experimental molecular weights (MWs) to sequence-derived theoretical MWs of *E. coli* cells using internet accessible protein databases, was then carried out to identify potential growth-phase-associated protein biomarkers.

Figure 6.1 shows an overview of the various steps involved in biomarker profiling using ICM-MS, PLS-DA and protein database search used in this *E. coli* growth phase-associated protein biomarker profiling model. It begins with preparation of the *E. coli* cell samples at different growth phases (section 6.1), the analysis of the samples by MALDI mass spectrometry, and preprocessing of the spectra data generated to remove to remove unwanted variation whilst preserving biological information. Sampling is carried out on the preprocessed data sets to separate the samples into training and test sets. The training set is overviewed using PCA to study initial trends and later analysed using PLS-DA, whilst the test sets are retained for external validation of the PCA and PLS-DA models built. PLS-DA scores and loadings plot are interpreted. The information from the scores and loadings plot is used for database searches to identify protein biomarkers.

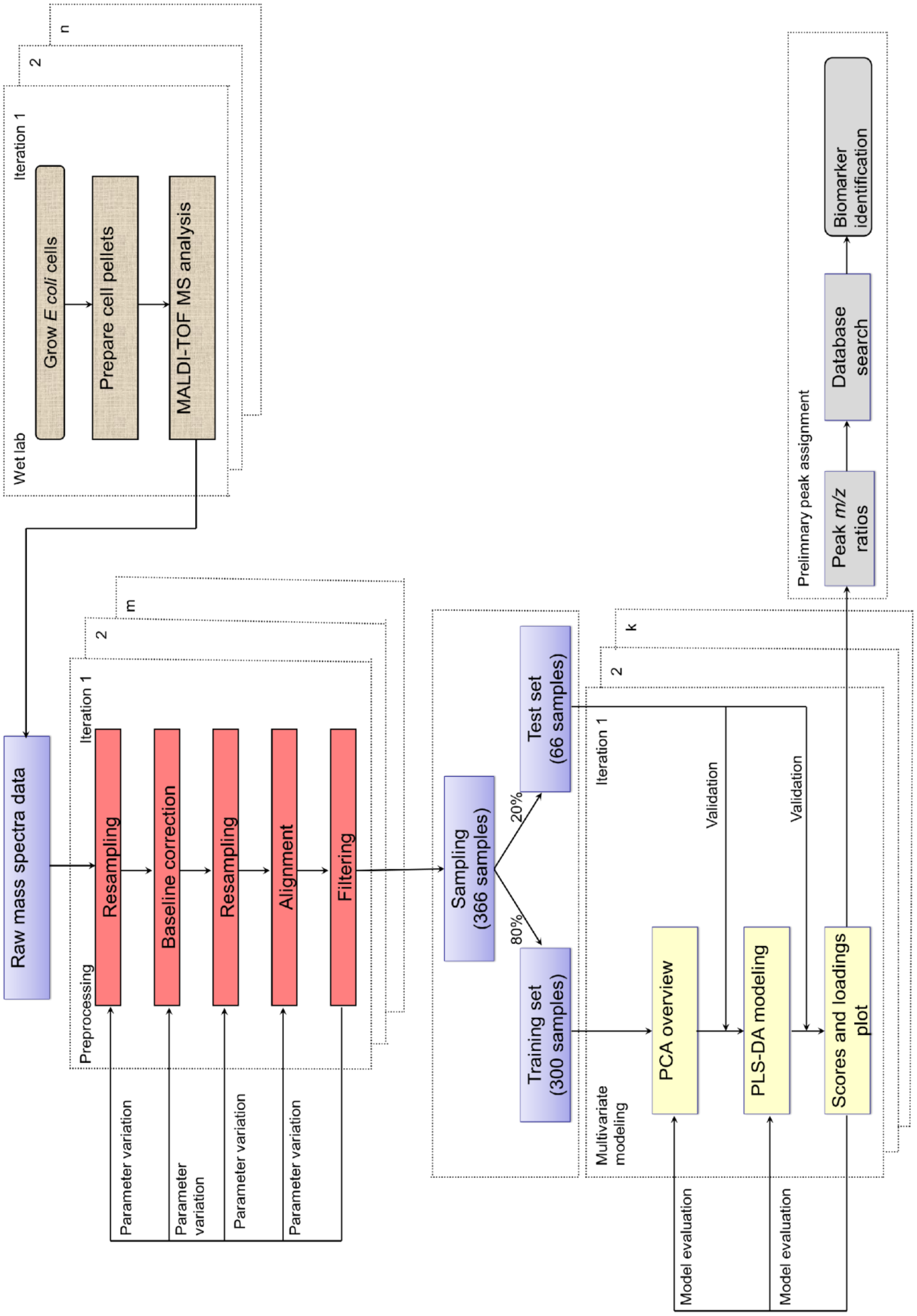


Figure 6.1: An overview of the various steps involved in the growth phase-associated protein biomarker profiling of E. coli K-12 cultures using ICM-MS, PLS-DA and protein database search

6.3. The Bacterial Growth Curve and Growth Measurements

The growth of bacteria in culture comprises a number of phases, characterised by a variation in the growth rate (Monod, 1949). The following definitions are growth phases usually distinguished in a bacterial culture, assuming a genetically homogeneous bacterial population. In Fig. 6.2, **A** is the lag phase of null growth rate (initial acclimatisation of cells to their new environment); **B**, the acceleration phase: growth rate increases; **C**, the exponential (logarithmic) phase: growth rate is constant (rapid growth as cell biomass increase linearly with time); **D**, the retardation phase: growth rate decreases; **E** is the stationary phase: null growth rate (nutrient becomes exhausted, rapid growth is halted); **F**. accelerated death phase; **G**, the logarithmic decline phase of negative growth rate negative (waste accumulates, cells die) (Stainier *et al.*, 1987). It is not uncommon for one or several of these growth phases to be absent. For instance under suitable conditions of abundant nutrients, the lag and acceleration phases may often be suppressed. It is also not uncommon to have very short retardation or stationary phases so that they are indiscernible (Monod, 1949). Thus some phases are usually ignored to give the main growth phases as lag, log, stationary and death phases.

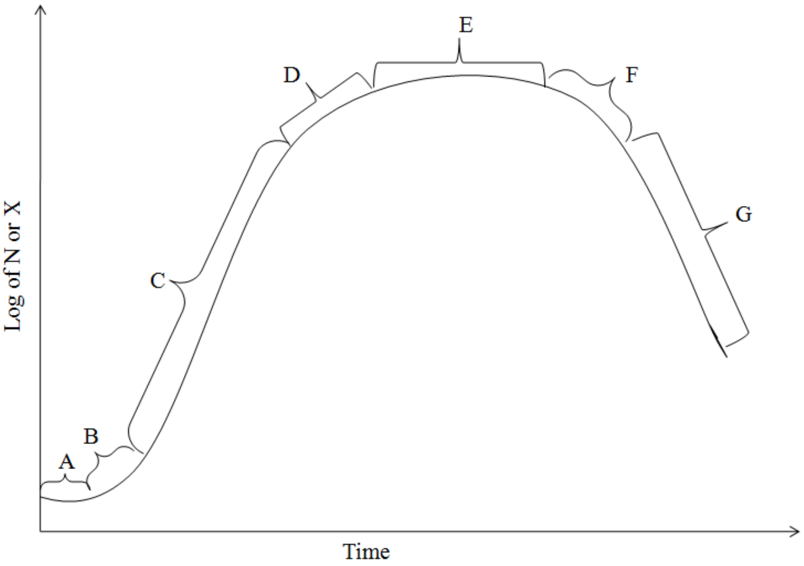


Figure 6.2: Schematic diagram of a bacterial culture growth curve showing the various growth phases

In this study, cells were grown in culture, and samples were prepared and MALDI-ToF mass spectra were recorded in a manner comparable with the previously reported studies (Saenz *et al*., 1999; Reilly *et al*., 1999; Holland *et al*., 2000; Harrington *et al*., 2008). Samples were collected and analysed from the cultures at points in the three specific growth phases (exponential, stationary and decline phases). To determine the sample collection times (in hours), the bacteria was first grown in culture and the growth curve was determined. The predictable timing of growth along the growth curve ensured that samples (bacteria cell pellets) were collected at specific time points during the three growth phases (Fig. 6.3). This increased the likelihood that differences between identified proteins will be related to the progression from one growth phase to another. The specific points (indicated by the thick arrows) of the growth curve at which samples were collected for MALDI analysis are shown on Fig. 6.3. To quantify the amount of bacteria to be analysed in the MALDI, the wet/dry cell weights (mg/mL) and bacterial number through the viable cell counts (CFU/mL) were also determined. The next sections describe the materials and methods used in this study, followed by the results, discussions and conclusions drawn.

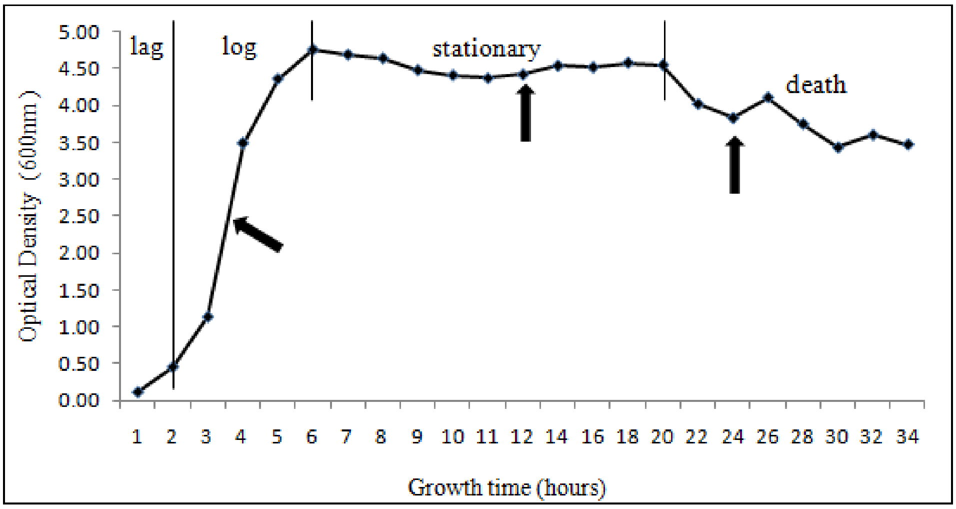


Figure 6.3: Growth curve (optical density vs growth time) of E. coli K-12 grown in 200mL LB growth media

6.4. Experimental Section

6.4.1. Materials

This section outlines the materials used to carry out the laboratory experiments. Freeze-dried stock cultures of *E. coli* K-12 strains ATCC 15223 were purchased from DSMZ GmbH, Germany. Acetonitrile/trifluoroacetic acid (ACN/TFA; 0.1%, v/v), absolute ethanol (Analytical grade reagent), glycerol (100% analytical grade reagent), hydrochloric acid (~36%), formic acid, and acetonitrile were obtained from Fisher Scientific, UK. Calcium chloride, disodium hydrogen phosphate, potassium dihydrogen phosphate, sodium chloride (Sigma Ultra, min 99.5%), were purchased from Sigma-Aldrich Co., USA. Distilled water was obtained from a Milli-Q Plus purification system (Millipore Corporation, Bedford, MA, USA). Nutrient broth was purchased from Oxoid (Basingstoke, Hampshire, United Kingdom) and nutrient agar was obtained from Merck, Germany. Sinapinic acid and the MALDI-ToF calibrant (a protein mixture containing insulin, ubiquitin I, cytochrome C, and myoglobin) was purchased from Bruker Daltonics, GmbH, Germany.

6.4.2. Culture Growth

This section briefly describes how the bacteria were cultured. Glycerol stocks were prepared from the freeze-dried stock cultures of *E. coli*. A growing culture was used to inoculate a 100 mL Luria Bertani (LB) medium in an Erlenmeyer flask and grown overnight at 37^{o}C in a shaker incubator set at 200 rpm. The volume required to produce an optical density at 600 nm (OD_{600}) of 0.05 OD units was then used to inoculate three flasks of 100 mL LB medium. The OD_{600} of the cultures were measured with a 6705 UV/Vis Spectrophotometer (JENWAY, Bibby Scientific Ltd, UK) at hourly intervals of incubation.

6.4.3. Determination of Bacteria Numbers, Dry and Wet Cell Weights

The bacteria was quantified using both standard plate counting on nutrient agar and by measuring the wet cell weight (WCW) and dry cell weight (DCW). Serial dilutions of bacterial culture were plated and incubated after which colonies were counted. For weight determination, ten 1 ml samples of culture were centrifuged at 17949 × *g* for 10 mins in a bench-top centrifuge in pre-weighed tubes. The supernatants were carefully removed by first pouring and then using stretched cotton wool swabs. The tubes containing cell pellets were then re-weighed both whilst wet and following drying, and the weight of the cell pellet was determined. Calibration curves determined from serial dilutions of the 12-hour culture using the bacteria numbers and weights were used to determine the amount of bacteria to be analysed by the MALDI.

6.4.4. MALDI-ToF MS Analysis and Data Preprocessing

'In-tact cell' sample preparation for MALDI analysis was carried out (on samples that had been stored at -70^{o}C) as described in section 3.3.7.1. Previously, the calibration curves were used to determine the amount of 'intact'bacteria cell pellets (approximately 3.3 × 10^{9}cells, 10.2 mgmL^{-1} (WCW), and 0.44 mgmL^{-1} (DCW)) to be analysed. All mass spectra were acquired with an Ultra Flex MALDI-ToF mass spectrometer (Bruker Daltonics, GmbH, Germany), located in the School of Biosciences, University of Kent, Canterbury, Kent, UK. The instrument parameters used as well as the analysis procedure has been described in section 3.3.7.1. Data preprocessing was carried out as described in section 3.6 using in-house scripts developed from MATLAB® v.7.6.0.324 (R2008a the MathWorks, Inc.) and functions from the Bioinformatics Toolbox of MATLAB® (v 3.1, R2008a, Eigenvector Research, Inc.). Preprocessing studies was carried out as described in section 3.6, to find the appropriate combination of preprocessing techniques.

6.4.5. Multivariate Data Analysis

The application of multivariate data analysis (PCA and PLS-DA) to the spectral data was carried out as described in section 5.8 using the PLS Toolbox v.3.5 and MATLAB® software v.7.6.0.324 (Eigenvector Research, Inc.). The first step was to divide the mass spectra data into training and a test data set for each group of spectra samples (exponential, stationary and decline phases). The PLS-DA algorithm used in modeling this data set has been described in detail in section 5.6.1.

Random sampling was undertaken and 80% of the samples were placed in the training set whilst the rest were included in the test set. For the exponential phase, 100 were denoted as training samples and 22 as test samples, with 100 training and 20 test samples for the stationary phase; and finally 100 selected as training set and 24 as test set samples for the decline phase. This resulted in a total of 300 training samples and 66 test set samples (from a total of 366 *E. coli* cell samples). PCA was then applied to the preprocessed mass spectral data sets to identify any groupings.

6.5. Database Search

The *m/z* ratio peak values (as experimental MWs) were submitted to a protein database search with the aim of assigning protein identities to mass spectra ion signals obtained from the PLS-DA modeling results. The searches were conducted using the MALDI mass spectra protein experimental MWs and the organism's theoretical MWs found in the UniProtKB/Swiss-Prot database, which is the Expert Protein Analysis System (ExPASy) of the Swiss Bioinformatics Institute (http://expasy.org/proteomics). With the query specified as *E. coli* K-12, the experimental MWs were matched to biomarker peaks contained in the Protein Knowledgebase (UniProt) and TrEMBL query form given in the database for the identification of the bacterial main biomarkers. Potential assignments from *m/z* ratio signal ions were possible for a number of *m/z* ratio values within the range 4 to 20kDa.

6.6. Results and Discussions

6.6.1. PCA Modeling of *E. coli* MALDI-TOF MS Data at different Growth Phases

PCA was applied to the 366 preprocessed mass spectral for preliminary data visualisation. The objective of applying PCA was to determine if the growth phase is the trend that differentiates samples into classes. The scores of the first three principal components (PC1 vs PC2 and PC2 vs PC3), which account for 93.73% of the variability in the original data set are shown in Fig. 6.4 and the variance explained is summarised in Table 6.1.

Principal component (PC) number	Eigenvalue of covariance (*X*)	% Variance captured for this PC	Total % variance captured
1	1.36×10^5	81.37	81.37
2	1.71×10^4	10.25	91.62
3	3.53×10^3	2.12	93.73

Table 6.1: Results of PCA model for all 366 mass spectra data sets for intact cells

In Fig. 6.4, the ellipse (blue dashed line) represents the 95% confidence region for the two PCs based on Hotelling's T^2. The samples are colour-coded according to their growth phase or class. The three-component model shows clear evidence of class 2 (exponential phase culture samples) being separate from the other two classes along the second principal component (Fig. 6.4a). The representation also showed an indication of a difference between class 1 and 3 from the PC2 vs PC3 plot (Fig. 6.4b). These preliminary observations suggest that a major part of the spectral variation is related to the growth phase.

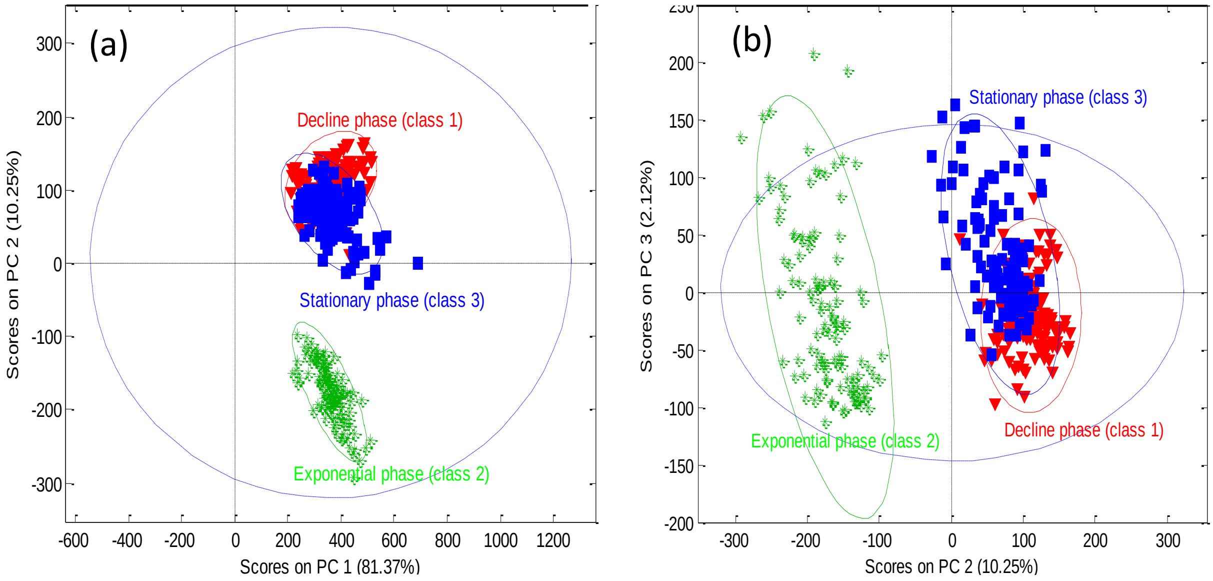


Figure 6.4: Principal component scores plot for 366 intact cell E. coli mass spectra data

6.6.2. PLS-DA Modeling of *E. coli* MALDI-TOF MS Data at different Growth Phases

6.6.2.1. Latent Variable (LV) Selection

The next step was to apply PLS-DA to the full preprocessed mass spectral profiles from the 'intact' *E. coli* cells. During PLS-DA model calibration, leave-class-out-cross-validation (LCOCV), a modified version of leave-one-out-cross-validation was carried out. In LCOCV, all samples belonging to the same class were removed from the training data set and a sub-model based on the remaining samples were used to build the PLS-DA model and predict the left out samples. Cross-validation was customised by creating a vector (Fig. A.3, Appendix A) and specifying how the cross-validation was to be performed.

Fig. 6.5 shows a plot of the root mean square error of cross-validation (RMSECV) for LCOCV as a function of number of the number of LVs and illustrates the impact of increasing the number of LVs under cross-validation conditions. The number of LVs was chosen to simultaneously maximise the percentage of explained systematic variation whilst achieving correlation with ***Y***. Fig. 6.5 indicates that exponential phase (class 2) samples behaved differently from the other two classes. For the exponential phase samples, the lowest RMSECV of 0.6 was attained with two LVs, with the RMSECV leveling out after 3 LVs. The other two classes achieved their lowest RMSECV for LV1. As a rule of thumb, the appropriate LV should have a minimum RMSECV (Li *et al.*, 2002). Based on these observations three LVs were retained for the subsequent model. The 3-LV model captured 86.62% of the ***Y***-block of the training data set.

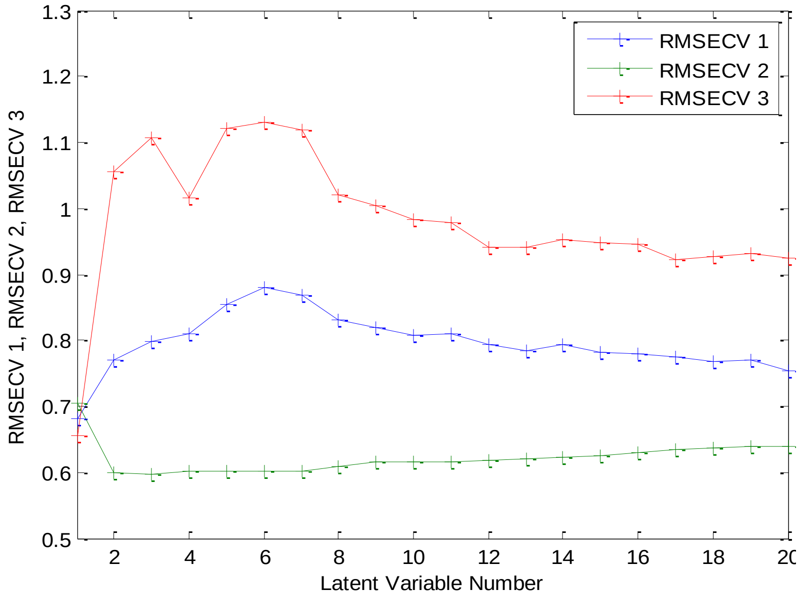


Figure 6.5: Root mean square error of cross validation (RMSECV) as a function of the number of LVs added to the PLS-DA model

6.6.2.2. Model Quality

The model parameters are summarised in Tables 6.2 and 6.3, and Fig. 6.6. The percentage of explained and cumulative explained variance of the ***X*** and ***Y***- variables are reported in Table 6.2 for the first six LVs. The results in the table suggest that three LVs was the appropriate choice or number of LVs. This is because starting from LV 6, the percentage variance captured on the ***Y***-block gets gradually larger until LV 4 is reached, then there is a sudden jump up to LV 3 (Table 6.2). The 3-LV PLS-DA model captured

86.7% of the ***Y***-block variance and explained 93.2% of the ***X***-block variance of the training data set under cross-validation.

Latent variable (LV) number	***X*-Block**		***Y*-Block**	
	% Variance captured for this LV	**Total % variance captured**	**% Variance captured for this LV**	**Total % variance captured**
1	81.18	81.18	32.36	32.36
2	10.49	91.66	31.68	64.04
3	1.56	93.22	22.60	86.65
4	1.57	94.79	4.17	90.81
5	0.60	95.39	3.39	94.17
6	0.71	96.10	1.01	95.18

*Table 6.2: PLS-DA percentage variance explained by the LVs calculated for **X** and **Y** variables*

Modeled class	**Class 1 (Decline phase)**	**Class 2 (Exponential phase)**	**Class 3 (Stationary phase)**
Sensitivity (Cal)	0.930	1.000	0.980
Specificity (Cal)	0.995	0.990	0.936
Sensitivity (CV)	1.000	1.000	1.000
Specificity (CV)	0.505	0.955	0.005
Classification error (Cal)	0.037	0.005	0.042
Classification error (CV)	0.248	0.023	0.498
RMSEC	0.238	0.148	0.254
RMSECV	0.773	0.599	1.038
R^2 Cal	0.745	0.904	0.708
R^2 CV	0.057	0.028	0.885

Table 6.3: *PLS-DA modeling results showing the quality of the model is good after calibration (Cal) and cross-validation (CV)*

Table 6.3 shows the parameters that describe the quality of the model after calibration (Cal) and cross-validation (CV). The Table shows that the quality of the model was good both in calibration and CV with a sensitivity (proportion of samples correctly classified that belong to the class being modeled), and specificity (proportion of samples correctly classified that do not belong to the class being modeled) of at least 93%. Furthermore, the model had a low classification error over the three classes, with the highest value being 3.7% (class 1). R^2 values calculated were high, both fitting (74% for class1, 90% for class 2 and 71% for class 3) and in cross-validation (81% for class 3).

The high values of R^2 in fitting (and the consequent small values of RMSEC (24% for class1, 15% for class 2, and 25% for class 3) show that the model is characterised by a large fitting ability.

Fig. 6.6 shows the calculated ***Y*** versus measured ***Y*** in fitting and prediction for the PLS-DA model after cross-validation. This figure also supports the assertion that the model is good. It represents the calculated versus the measured responses both in fitting and in prediction: no deviations can be identified along the y-axis in all three classes. This suggests that all the useful information is taken into account by the model.

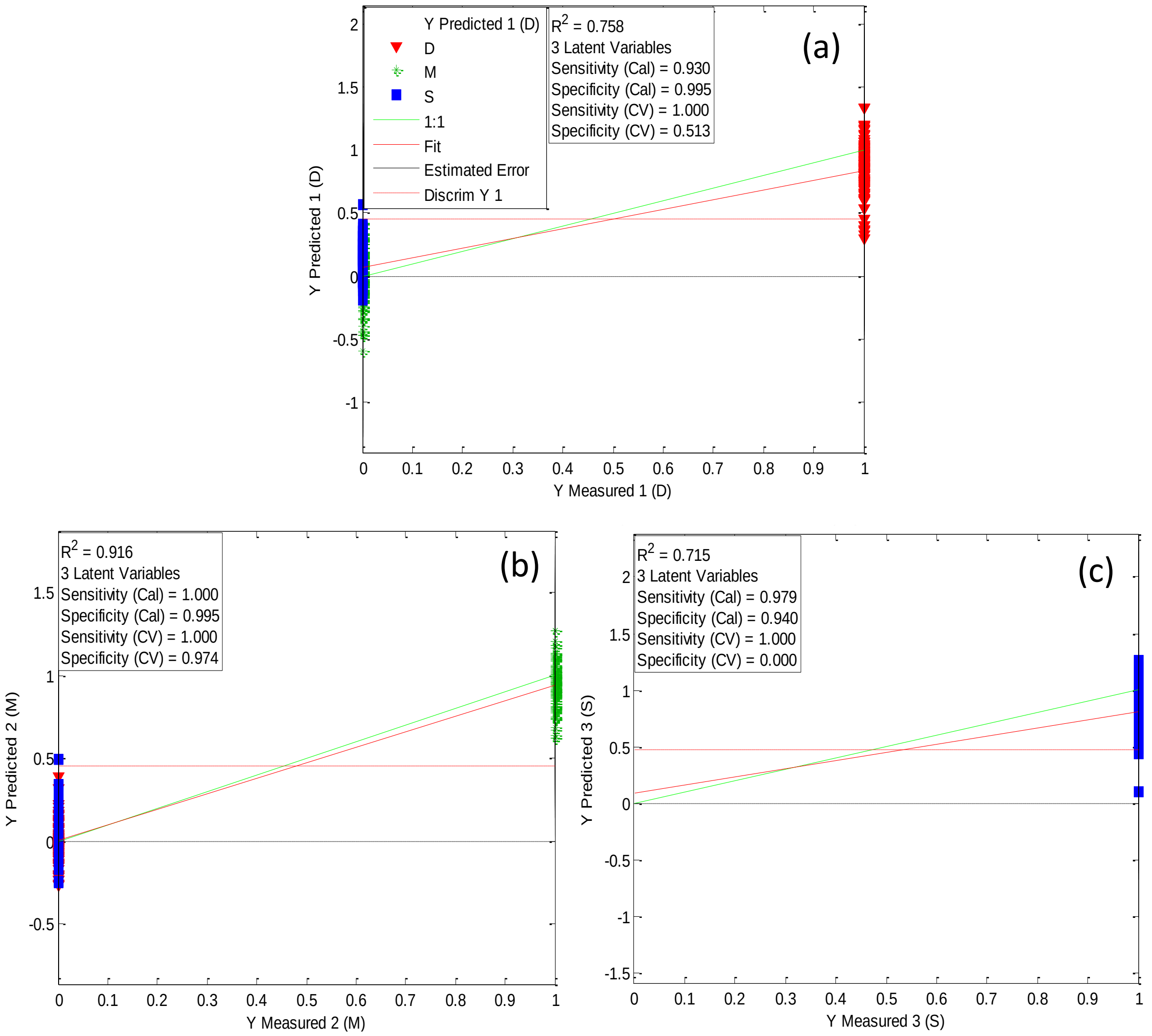


*Figure 6.6: PLS-DA calculated **Y** versus measured **Y** in fitting and prediction after cross-validation*

6.6.2.3. External Validation and Model Performance

The model was successfully validated with the 66 test data sets (Table 6.4; Fig. 6.7) indicating that the PLS-DA model is informative in terms of class separation. Table 6.4 shows that the model had a very high performance (95-100% sensitivity and specificity). Furthermore, the model had a low classification error over all the three classes, with the highest being 3.6% (class 3).

Modeled class	**Class 1 (Decline phase)**	**Class 2 (Exponential phase)**	**Class 3 (Stationary phase)**
Sensitivity (prediction)	0.958	1.000	0.950
Specificity (prediction)	1.000	1.000	0.978
Classification error (prediction)	0.021	0.000	0.036
RMSEP	0.228	0.156	0.228
Prediction Bias	-0.026	0.036	0.012
R^2 prediction	0.786	0.902	0.771

Table 6.4: PLS-DA modeling results showing the performance of the model after external validation

During the PLS-DA model building stage (calibration), the algorithm utilises Bayesian statistics to make a decision as to whether a future unknown sample will belong to a given class or not. During prediction of a test set, samples associated with a ***Y***-value (predicted value) above the decision line have a statistically significant probability of belonging to the target class; the samples associated with ***Y***-values below the decision line have a statistically significant probability of not belonging to the target class. As indicated in Fig. 6.7, the decision threshold shown in each case (middle dashed red line) is calculated using the distribution of predicted ***Y*** values obtained during model building. Misclassified samples have been labelled and coloured in pink.

As indicated in Fig. 6.7, just one sample from class 1 (1di01) was misclassified falling below the decision threshold. This gave an excellent prediction sensitivity and specificity of 95.8% and 100% respectively. No class 2 samples were misclassified giving 100% prediction sensitivity and specificity. Only one class 3 sample (4si13) was misclassified giving an excellent prediction sensitivity and specificity of 95% and 97.8% respectively. These results suggest that the model had an excellent performance and is valid. The successfully validation of the model means that the model parameters such as scores and loadings are valid. The model performance also validates the choices of the preprocessing methods/parameters used for the spectra data described in section 3.6.

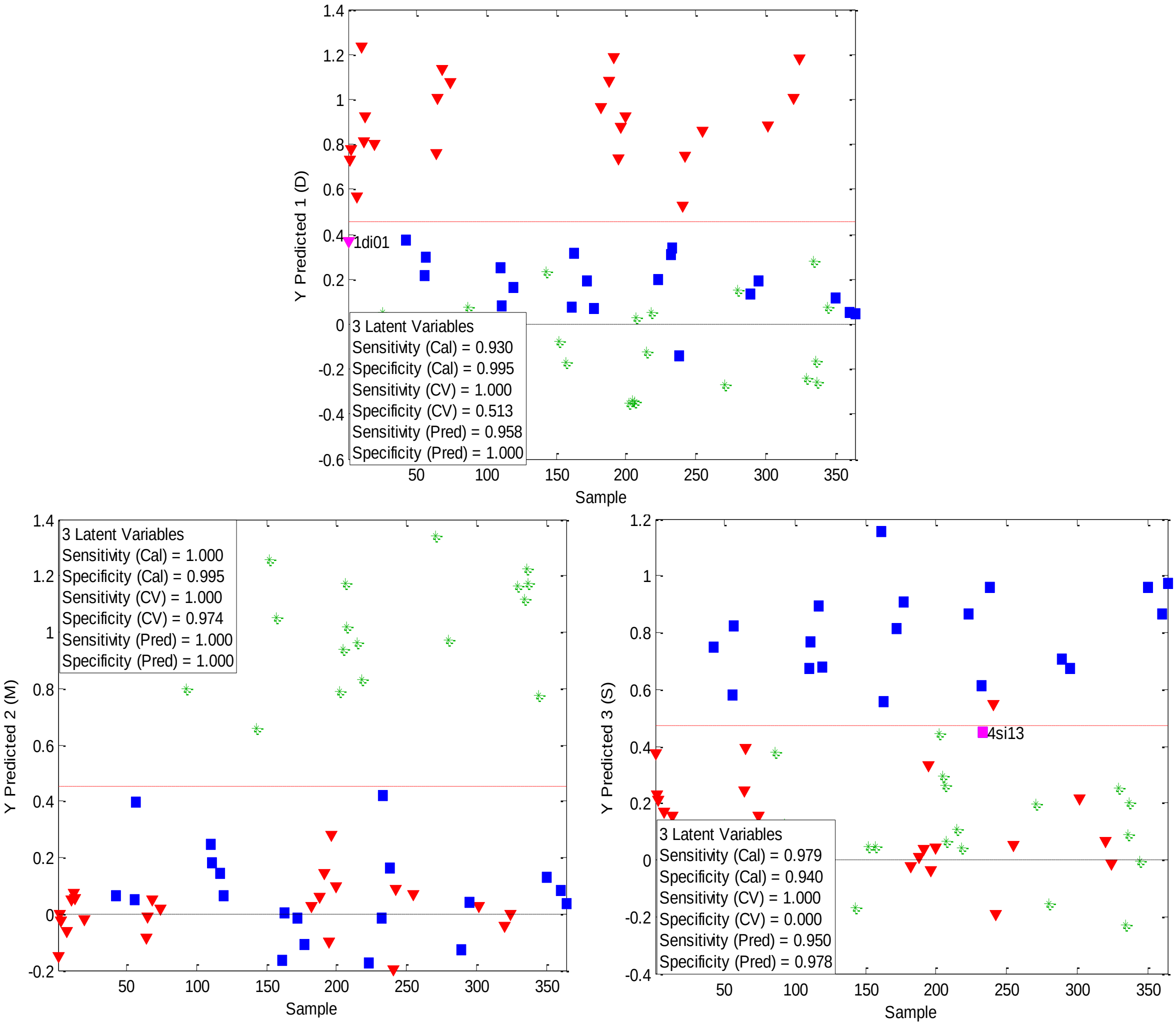


*Figure 6.7: PLS-DA calculated **Y** versus measured **Y** in fitting and prediction after external validation*

6.6.2.4. Scores Plot

The practical value of the model was demonstrated in its ability to separate the three classes as shown in the scores plot. Fig. 6.8 shows the resulting PLS-DA scores plot. A plot of the first two components (LV1 vs LV2) shows that the 3-LV model concentrates most of the discriminatory information into the second LV (Fig. 6.8(a)), with the exponential phase cell samples (class 2) exhibiting differences from the other two classes. This observation is supported by a plot of scores on LV2 as a function of samples (Fig. 6.8(c)). A plot of LV2 vs LV3 shows separation between the stationary phase (class 3) and decline phase (class 1) cell samples (Fig. 6.8(b)), an observation supported by Fig. 6.8(d). Stationary phase samples have positive scores (found in the left quadrant) whilst decline phase samples have positive scores (found in the right quadrant) in LV3 (Fig. 6.8(b)).

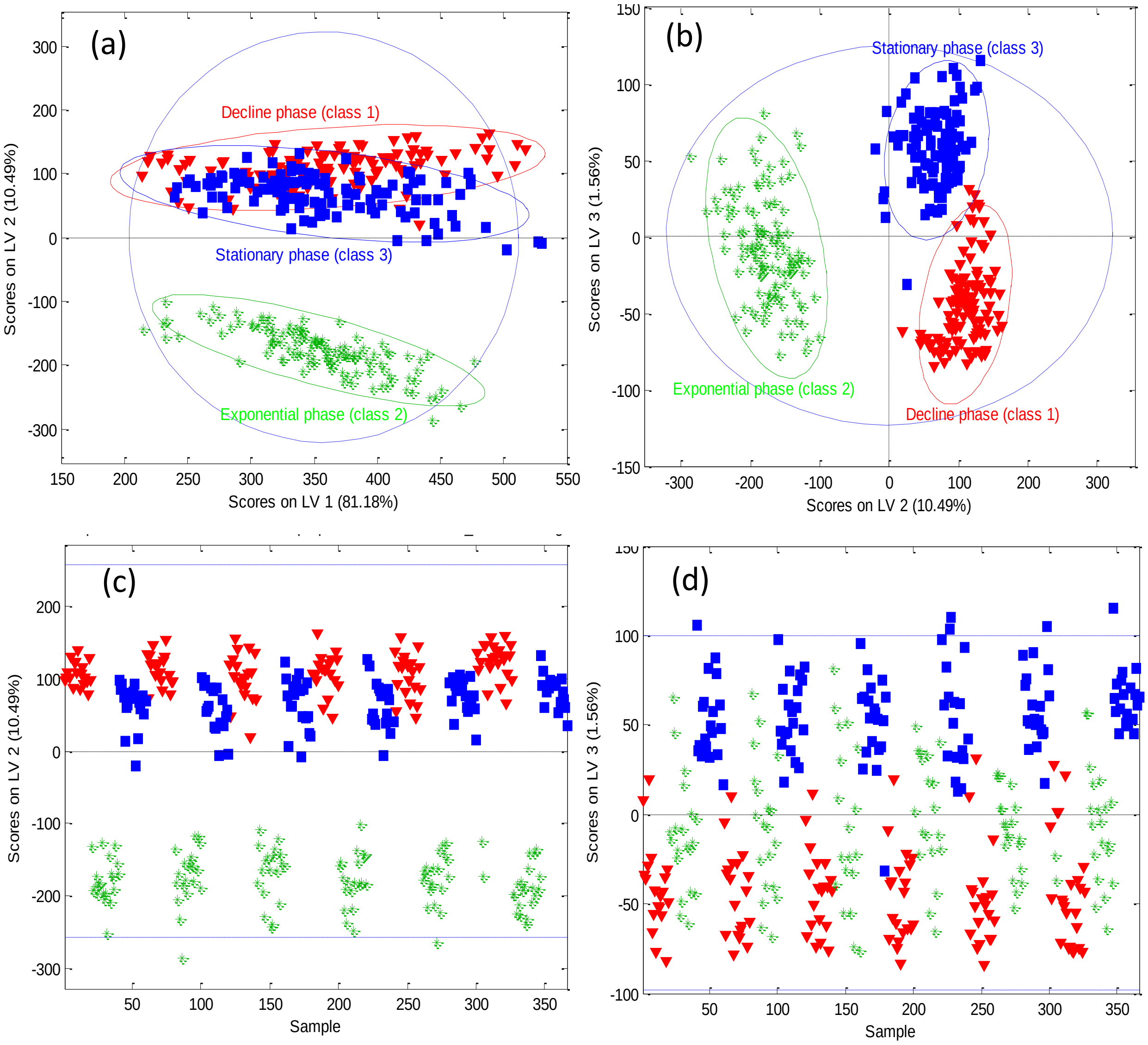


Figure 6.8: PLS-DA scores plot for the E. coli calibration mass spectra data

6.6.2.5. Loadings Plot

The PLS-DA model uses whole *m*/*z* ratio regions (without variable selection) to differentiate between classes so it is possible to investigate which *m*/*z* ratio regions contributing towards the discriminatory ability of the model for classifying a given sample as class 1, 2 or 3, that is, decline, exponential and stationary phases respectively**.** The PLS-DA loadings plot for a specific LV potentially contains valuable information regarding the regions of the mass spectra (*m/z* ratio peaks or ion signals) that contribute to the ability of the PLS-DA representation to distinguish between classes. There is a direct geometric link between the scores and the loadings plot of PLS-DA.

The loadings plot for LV2 and LV3 (Fig. 6.9 and 6.10) indicates which variable loadings (associated with *m/z* ratio ion signals in the MALDI data sets) that contain information which are causing the separation of the samples in the scores plot. The loadings describe the weighting coefficients for each *m/z* ratio ion signal (experimental MWs of proteins) with the magnitude of the variable loadings being indicative of how the expression of proteins varies with growth phase. In Fig. 6.9, it can be seen that variables 1702, 2271, 2344, 2729 and 3642 have loadings of large negative magnitude. Since the exponential phase cell samples have negative scores along LV2, this implies that these samples can be differentiated from the other two classes by these *m/z* ratio signals associated with variables of large negative loadings in the LV2 loadings plot. These ion signals potentially identify those proteins which are differentially expressed in cells during the exponential phase and are most likely to be biomarkers of the exponential phase.

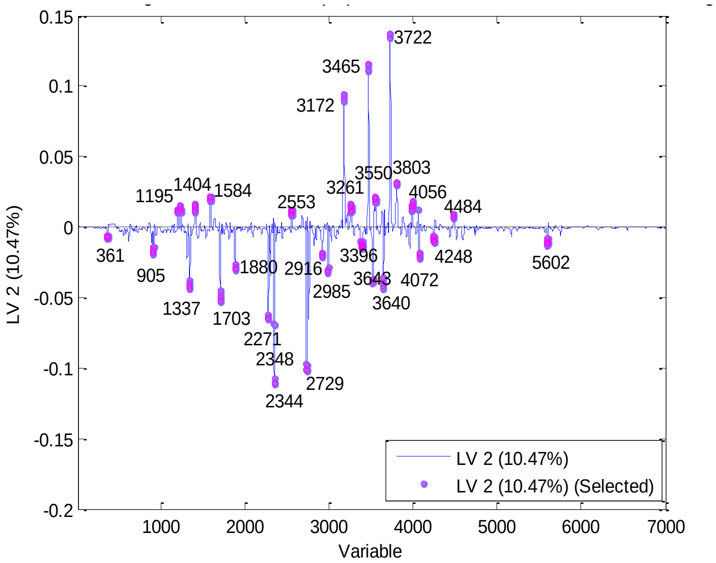


Figure 6.9: PLS-DA loadings plot on LV2

For the loadings associated with LV3 (Fig. 6.10), the variables 1701, 2271, 2344, 2915 and 3719 have loadings with large positive magnitude. Stationary phase cell samples can thus be differentiated from decline phase samples by the presence of *m/z* ratio signals associated with variables with positive loadings in the LV3 loadings plot. These *m/z* ratio signals are indicative of those proteins which are differentially expressed in cells during the stationary phase. Variables 2550, 2735, 3473 and 3528 have loadings with large negative magnitude on the LV3 loadings plot (Fig. 6.10), and since the decline phase cell samples have negative scores along LV3, they can be distinguished from the stationary phase cell samples by the differential expression of proteins associated with signals with positive loadings in the LV3 loadings plot.

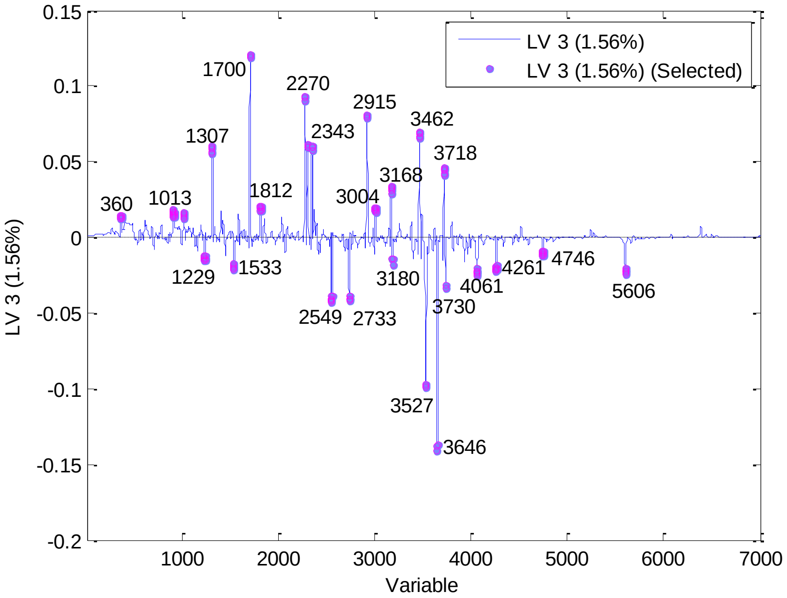


Figure 6.10: PLS-DA loadings plot on LV3

6.6.3. Protein Database Search for Database Search for the *E. coli* Cell Samples

The PLS-DA loadings plots for LV2 and LV3 are now considered for further analysis. The variables from the loadings plot are associated to *m/z* ratio peak or signal ions (Table A.6 to A.8, Appendix A). Since most of the *m/z* ratio signal ions are singly charged protonated protein (MH^+) molecules, they represent the approximate MALDI experimental MWs of the ionised proteins expressed by the cells at the different growth phases. Thus the *m/z* ratio signal ions (experimental MW of protein ions) were submitted to a protein database search to assign protein identities to them. The experimental MWs of protein ions were used as search parameters and were matched with sequence derived theoretical MW values in the UniProtKB/Swiss-Prot database (http://expasy.org/proteomics).

A number of matches resulted in protein biomarkers which have been predicted in the organism's genome, and there is experimental evidence that these proteins are expressed *in vivo* at the different phases of growth (Yoon *et al*., 2003; Nystrom, 2004; Han and Lee, 2006). Table 6.5 summarises the results of the database search and shows the compiled results of the protein matches obtained for the three growth phases. If no match was found after the initial search, the search was repeated assuming N-terminal Methionine (Met) is lost (experimental MW with less than 131 Da). In prokaryotes, it is estimated that more than 50% of *E. coli* proteins undergo Met loss as posttranslational modification (PTM) (Hirel *et al.*, 1989). *E. coli* have been well studied and N-terminal Met excision PTM is reflected in the UniProt/SwissPROT databases which contains proteins with or without N-terminal Met (Gibson *et al*., 1997; Demirev *et al*., 1999).

Most of the experimental MWs were matched with proteins in the mass range 4 – 20 kDa. Matches were not found for the majority of experimental MWs below 4 kDa. These MWs may represent the bacterial cell lipooligosaccharide and peptidoglycan molecules (Guo *et al*., 2002). Information relating to the synthesis of these non-protein biomolecules is not derived from the genetic code and hence would not be found in the protein database. MWs less than 4 kDa may also partly be a combination of abundant matrix-related ions since most of the currently used matrices have MWs less than 3 kDa. They may also act as their own matrices, producing a variety of matrix-related ions during laser ionisation (Ochoa *et al*., 2005). Thus any assignment to *m/z* ratio ion signals below the 4 kDa region is tentative, and necessitates further conclusive evidence. Such assignments were therefore excluded.

6.6.4. SwissPROT/TrEMBL Protein Molecular Weight Distribution for *E. coli* K-12, MALDI Mass Accuracy and Biomarker Identification

Biomarker assignment and identification was based on several factors. Biomarker identification was based on *m/z* ratio values or experimental MW ranges can correspond to variable count ranges in the PLS-DA loadings plot. An average experimental MW is assigned to a protein biomarker if it falls within this range and matches the theoretical

MW of the protein biomarker based on the mass accuracy of linear mode MALDI-ToF MS instrument. The assigned proteins to take into account the mass accuracy in the linear mode MALDI-ToF which is 100 ppm, i.e., for all the assigned proteins, experimental MW masses was within 0.01% of their theoretical MWs (+/- 1 mass unit for 10,000 MW protein). Tables A.9 to A.11 (Appendix A) shows variable count ranges and their experimental MW ranges as well as the average experimental MWs (within the latter range) matching theoretical MWs for the all the protein biomarkers shown in Table 6.5.

Secondly, assignments took into consideration the molecular mass distribution of 20653 protein sequences (providing theoretical MWs) of *E. coli* K-12 proteins derived from genomic open reading frame as well as nongenomic entries found in the SwissPROT/TrEMBL database. The molecular mass distribution of known *E. coli* K-12 proteins (Fig. 6.11) has a peak centered around 12 kDa. Fig. 6.11 shows the molecular mass distribution (in bins of 1 kDa) of *E. coli* K-12 proteins deposited in the SwissPROT/TrEMBL sequence database. Fig. 6.11 suggest that many *E. coli* K-12 proteins have masses in the range of mass range 4 – 20 kDa. Therefore, it may be also expected that unique combinations of protein masses in the mass range from 4 to 20 kDa can serve as protein biomarkers for *E. coli* K-12 (Arnold and Reilly, 1999; Ryzhov and Fenselau, 2001). For example most experimental MWs were assigned to ribosomal proteins because the latter are highly abundant in growing cells (almost half of the cell mass), relative to other cytosolic proteins and most have MWs of less than 20kDa. Moreover, ribosomal proteins are very basic, and basic proteins are more amenable to protonation during MALDI analysis. Experimental MWs that did not match ribosomal proteins were assigned to other abundantly produced non-ribosomal cytosolic proteins which are basic, and expressed at the different growth phases based on evidence from the literature (Arnold and Reilly, 1999; Ryzhov and Fenselau, 2001).

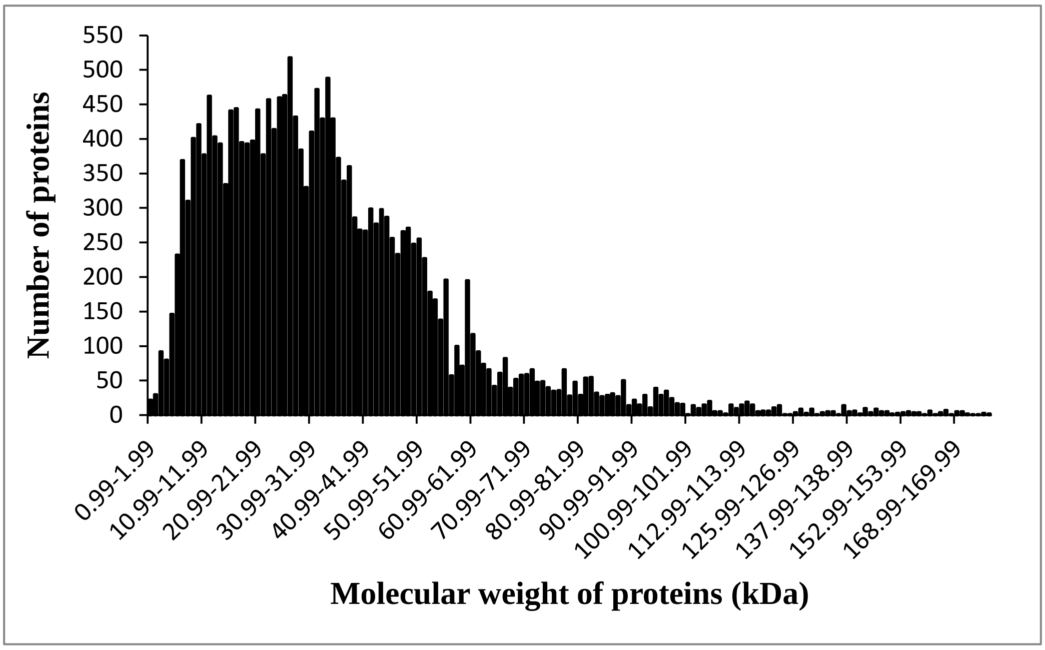


Figure 6.11: Molecular weight (MW) distribution (in bins of 1 kDa) of 20653 E. coli K-12 proteins deposited in the SwissPROT/TrEMBL sequence database

The full explanation of superscript letters assigned to column titles in Table 6.5 are as follows:

[a]Experimental MWs were the *m/z* ratio ion (MH^+) signals associated to the variables (associated to *m/z* ratio ion signals in the MALDI data set) from the PLS-DA loadings plot.

[b]The intensity of the variables (*m/z* ion signals) is the magnitude of their loadings in the loadings plot. High: intensity ≥ ±0.1 units; medium: intensity ≥ ±0.05 and ≤ ±0.1 units; low: intensity between 0 and ±0.05 units.

[c]Theoretical sequence MWs were calculated using the Compute pI/MW tool (http://web.expasy.org/compute_pi/). Assigned proteins took into account the mass accuracy in the linear mode MALDI-ToF which is 100 ppm, i.e., for the proteins, experimental MW masses should be within 0.01% of their theoretical MWs (+/- 1 mass unit for 10,000 MW protein).

[d]MWs of the proteins described in previous studies using MALDI-TOF analysis of *E. coli* K-12.

[ef]MW from studies by Arnold and Reilly (1999); and Ryzhov and Fenselau (2001) respectively. Proteins in bold with ‘none’ as paper match are proteins identified that were not reported in previous *E. coli* culture MALDI studies.

[g]Protein names, description, PI, accession numbers and remarks were from ExPASy Proteomics Server (http://www.uniprot.org/uniprot/).

Table 6.5: SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the E. coli K-12 Cell Samples

Experimental MW[a]	Intensity[b]	Match[c]	Paper match[d]	Protein name[g]	Protein description[g]	PI[g]	Accession Number[g]	Remarks[g]
Exponential phase								
5095.33	Medium	5095.82	5095.9[e,f]	Sra	Protein S22	11.04	P68191	
6240.06	Medium	6240.39	6254.1[ef]	RpmG	50S ribosomal protein L33	10.25	P0A7N9	Met loss
6314.92	Low	6315.19	6315.1[e]	RpmF	50S ribosomal protein L32	11.03	P0A7N4	Met loss
6385.91	Low	6385.04	8325.0[f]	Lpp	Major outer membrane lipoprotein	8.12	P69776	No signal peptide
6409.66	High	6410.60	6410.3[e]	RpmD	50S ribosomal protein L30	10.96	P0AG51	Met loss
6855.88	Low	6855.89	6856.0[f]	CsrA	Carbon storage regulator	8.16	P69913	
7273.32	High	7273.45	7273.0[ef]	RpmC	50S ribosomal protein L29	9.98	P0A7M6	
7271.02	High	7271.17	7271.0[f]	CspC	Cold shock-like protein CspC	6.82	P0A9Y6	Met loss
7333.30	Medium	7332.26	7333.8[ef]	CspE	Cold shock-like protein CspE	8.06	P0A972	
7535.79	Low	7536.55	None	**MarB**	Multiple antibiotic resistance protein	5.24	P31121	
7717.31	Medium	7716.72	None	**CspB**	Cold shock-like protein cspB	6.53	P36995	
7872.0	Medium	7871.06	7871.0[e]	RpmE	50S ribosomal protein L31	9.46	P0A7M9	
7968.32	Low	7968.97	None	**CspD**	Cold shock-like protein cspD	5.81	P0A968	
8118.45	Low	8118.38	None	**InfA**	Translation initiation factor IF-1	9.23	P69222	Met loss
8854.33	Low	8855.24	8897.0[e]	RpsR	30S ribosomal protein S18	10.60	P0A7T7	Met loss
9190.21	Medium	9190.56	9190.5[ef]	RpsP	30S ribosomal protein S16	10.54	P0A7T3	
9534.98	Medium	9534.98	9536.7[e]	HupA	DNA-binding protein HU-alpha	9.57	P0ACF0	
10298.28	Low	10299.09	10299.6[ef]	RpsS	30S ribosomal protein S19	10.52	P0A7U3	Met loss
10693.74	Low	10693.44	10694.0[ef]	RplY	50S ribosomal protein L25	9.60	P68919	
11199.26	Low	11199.12	11198.0[ef]	RplW	50S ribosomal protein L23	9.94	P0ADZ0	
11957.28	Low	11958.37	None	**EmrE**	Multidrug transporter emrE	7.72	P23895	
12227.40	Low	12226.29	12225.3[ef]	RplV	50S ribosomal protein L22	10.23	P61175	
12398.14	Low	12398.28	None	**MalE**	MalE	6.10	Q14F09	
13481.87	Low	13480.70	None	**RidA**	Enamine/imine deaminase	5.36	P0AF93	Met loss
13714.81	Low	13713.73	13727.7[e]	RpsK	30S ribosomal protein S11	11.33	P0A7R9	Met loss
9553.44	Medium	9553.20	9553.6[ef]	RpsT	30S ribosomal protein S20	11.18	P0A7U7	Met loss
15280.89	Low	15281.20	15326.0[e]	RplP	50S ribosomal protein L16	11.22	P0ADY7	
15596.17	Low	15596.78	None	**PsiE**	Protein psiE	8.85	P0A7C8	

Table 6.5: SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the E. coli K-12 Cell Samples

Table 6.5–(Continued): SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the E. coli K-12 Cell Samples

Experimental MW^{a}	Intensity^{b}	Match^{c}	Paper match^{d}	Protein name^{g}	Protein description^{g}	PI^{g}	Accession Number^{g}	Remarks^{g}
Stationary phase								
4309.30	Medium	4309.22	4364.2^{e,f}	RpmJ	50S ribosomal protein L36	11.06	Q9RSK0	
4581.04	Low	4580.12	None	**OsmB**	Osmotically inducible lipoprotein B	9.31	P0ADA7	
5095.33	High	5095.82	5095.9^{e,f}	Sra	Protein S22	11.04	P68191	
5380.55	Low	5380.39	5380.5^{e}	RpmH	50S ribosomal protein L34	13.00	P0A7P5	
6240.06	Medium	6240.39	6254.1^{ef}	RpmG	50S ribosomal protein L33	10.25	P0A7N9	Met loss
6385.91	Low	6385.04	8325.0^{f}	Lpp	Major outer membrane lipoprotein	8.12	P69776	No signal peptide
6507.28	Medium	6507.48	None	**Rmf**	Ribosome modulation factor	10.86	P0AFW2	
6855.88	Medium	6855.89	6856.0^{f}	CsrA	Carbon storage regulator	8.16	P69913	
7717.31	Medium	7716.72	None	**CspB**	Cold shock-like protein cspB	6.53	P36995	
7891.40	Medium	7891.88	None	**GlgS**	Glycogen synthesis protein glgS	5.38	P26649	
8854.33	Low	8855.24	8897.0^{e}	RpsR	30S ribosomal protein S18	10.60	P0A7T7	Met loss
9066.06	Medium	9065.24	9060.0^{f}	HdeB	Protein hdeB	4.94	P0AET2	No signal peptide
9190.21	Low	9191.00	9190.5^{ef}	RpsP	30S ribosomal protein S16	10.54	P0A7T3	
9553.44	Low	9553.20	9553.6^{ef}	RpsT	30S ribosomal protein S20	11.18	P0A7U7	Met loss
9741.69	Medium	9740.91	9742.0^{f}	HdeA	hns deletion induced protein A	4.68	P0AES9	Active protein
11354.05	Low	11353.93	None	**IhfA**	Integration host factor subunit α	9.34	P0A6X7	
10651.91	Low	10651.14	10650.0^{f}	IhfB	Integration host factor subunit β	9.34	P0A6Y1	
11857.12	Low	11858.00	None	**HdeA**	Chaperone-like protein hdeA	4.68	P0AES9	
11992.72	Low	11993.68	None	**BolA**	Protein bolA	6.19	P0ABE2	
12028.22	Low	12029.45	None	**CyoD**	Cytochrome o ubiquinol oxidase	6.56	P0ABJ6	
15738.09	Low	15738.58	None	**SodC**	Superoxide dismutase (Cu-Zn)	5.58	P0AGD1	
18154.23	Low	18153.47	None	**PpiB**	Peptidyl-prolyl *cis-trans* isomerise B	5.51	P23869	
18563.88	Low	18564.11	None	**Dps**	DNA protection during starvation	5.72	P0ABT2	Met loss
18161.50	Low	18161.15	None	**OsmY**	Osmotically inducible protein Y	5.42	P0AFH8	
Decline phase								
6507.28	Low	6507.48	None	**Rmf**	Ribosome modulation factor	10.86	P0AFW2	
6385.91	Medium	6385.04	8325.0^{f}	Lpp	Major outer membrane lipoprotein	8.12	P69776	No signal peptide
4309.30	Medium	4309.22	4364.2^{e,f}	RpmJ	50S ribosomal protein L36	11.06	Q9RSK0	

Table 6.5 – (Continued): SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the E. coli K-12 Cell Samples

Experimental MW[a]	**Intensity**[b]	**Match**[c]	**Paper match**[d]	**Protein name**[g]	**Protein description**[g]	**PI**[g]	**Accession Number**[g]	**Remarks**[g]
Decline phase – *(Continued)*								
6855.88	Medium	6855.89	6856.0[f]	CsrA	Carbon storage regulator	8.16	P69913	
7717.31	Low	7716.72	None	**CspB**	Cold shock-like protein cspB	6.53	P36995	
7891.40	Medium	7891.88	None	**GlgS**	Glycogen synthesis protein glgS	5.38	P26649	
8499.80	Low	8500.00	8368.8[e]	RpsU	30S ribosomal protein S21	11.15	P68679	
8544.65	Low	8543.83	8544.0[f]	DaaF	F1845 fimbrial adhesin operon regulatory protein daaF	9.34	Q47132	
8639.73	Low	8640.00	None	**AcP**	Acyl carrier protein	3.98	P0A6A8	
8854.33	Low	8855.24	8897.0[e]	RpsR	30S ribosomal protein S18	10.60	P0A7T7	Met loss
9066.06	Medium	9065.24	9060.0[f]	HdeB	Protein hdeB	4.94	P0AET2	No signal peptide
9071.57	Medium	9071.48	None	**RelB**	Antitoxin relB	4.81	P0C079	
9226.48	High	9226.00	9225.0[f]	HupB	DNA-binding protein HU-beta	9.70	P0ACF4	
9271.57	Low	9271.58	None	**ChpS**	Antitoxin of the ChpB-ChpS system	4.69	B1XDX3	
9307.03	Low	9307.58	None	**YefM**	Antitoxin yefM	5.07	P69346	
11857.12	Medium	11858.00	9742.0[f]	HdeA	hns deletion induced protein A	4.68	P0AES9	
9977.57	Low	9977.98	None	**YnfB**	Hypothetical protein	8.07	P76170	
10430.19	Low	10430.00	10299.6[ef]	RpsS	30S ribosomal protein S19	10.52	P0A7U3	
10386.13	Low	10386.95	None	**GroS**	10-kDa chaperonin	5.15	P0A6F9	
10651.91	Low	10651.14	10650.0[f]	IhfB	Integration host factor subunit β	9.34	P0A6Y1	
11239.29	Low	11239.93	None	**Fis**	DNA-binding protein fis	9.34	P0A6R3	
11224.98	Low	11225.22	None	**RelE**	mRNA interferase relE	9.67	P0C077	
11579.52	Low	11580.00	11449.3[e]	RpsN	30S ribosomal protein S14	11.16	P0AG59	
12227.40	Low	12226.29	12225.3[ef]	RplV	50S ribosomal protein L22	10.23	P61175	
12670.13	Low	12669.85	None	**FliO**	Flagellar protein fliO	10.59	P22586	
15407.96	Low	15408.44	None	**Hns**	DNA-binding protein H-NS	5.44	P0ACF8	

6.6.5. Identification of Growth Phase-associated Biomarkers

As seen in Table 6.5, most of the database matches identified corresponded to intracellular proteins of the bacteria suggesting that the *E. coli* cells from the culture were lysed releasing intracellular contents. Cell lysis may have occurred as a result of the sample preparation procedure as a consequence of the organic solvents used or possibly during the storage and freeze-thawing of the cell pellets at -70^{o}C prior to the MALDI analysis. The majority of protein matches corresponded to ribosomal proteins in the 50S and 30S subunits. This outcome was expected since ribosomal proteins have a high abundance of up to 45% of the total mass of *E. coli* cells and these proteins make up to 21% of the cell's protein content (Ryzhov and Fenselau, 2001; Ochoa *et al*., 2005). Previous MALDI studies on whole *E. coli* cultures have also matched several *m/z* ratio peaks to ribosomal proteins (Ryzhov and Fenselau, 2001; Jones *et al*., 2003; Ochoa and Harrington, 2005). Ribosomal proteins are basic (that is, pI greater than 9), and basic proteins are more amenable to MALDI analysis because they can easily be protonated as they comprise many basic functional groups (Ochoa and Harrington, 2005). It is interesting to observe that exponential phase cultures had more ribosomal protein matches than the other phases with the number of ribosomal protein matches dropping from the exponential to the decline phase. The predominance of ribosomal proteins in actively growing cells has previously been reported (Saenz *et al*., 1999; Harrington *et al*., 2008). This trend is in keeping with the biology of these cultures since during the exponential phase, cultures are metabolically at their maximum, actively growing and protein synthesis is higher than in the other phases. At later growth phases, cultures are less metabolically active and grow less so it is not necessary to retain a high level of ribosomes (Reilly *et al*., 1999). This suggests that ribosomal proteins may serve as biomarkers for exponential phase *E. coli* cultures.

Several proteins reported in other research studies as stationary phase proteins were also identified in the stationary phase samples (Table 6.13). Amongst these are the nucleoid-associated proteins (NAPs), IhfA and Dps; the DNA-binding morphogene (BolA); lipoproteins OsmY and OsmB; the glycogen synthesis protein (GlgS); and the bacteriocin MccB17 (Connell *et al*., 1987; Aldea *et al*., 1989; Jung *et al*., 1989; Yim and Villarejo, 1992; Hengge-Aronis and Fischer, 1992; Azam *et al*., 1999). Hence these proteins could be considered as potential stationary phase biomarker candidates. Identified proteins also indicated that a high number of toxin-antitoxin (TA) lociassociated proteins (TALAPs) were present in decline phase samples, ChpS antitoxins; antitoxin YefM, DinJ, RelB, RelE, and mRNA interferase. The presence of TALAPs is in keeping with the concept of either an increased culture die-off or increased culture stasis from the exponential to the decline phase (Aizenman *et al*., 1996; Pedersen *et al*., 2002). TALAPs are most likely to be biomarkers of the decline growth phase.

Since MALDI-MS produces mostly singly charged ion protein or peptide fragment, the MW of the molecular ion and hence protein can be directly determined, hence identification of the protein. However, since information from MW of the proteins is not sufficient to identify them (because of matrix effect, post-translational modification, and mass errors) protein identifications through database search in this project were tentative pending further analysis. Proteins identified in this project have been reported in other research studies (Reilly *et al*., 1999; Guo *et al*., 2002; Jones *et al*., 2003; Ochoa *et al*., 2005; Harrington *et al*., 2008).

6.7. Summary

Within this chapter, it is shown that PLS-DA is a potentially powerful tool for extracting systematic variation pertaining to biological effects of spectra data generated using intact-cell MALDI mass spectrometry (ICM-MS) method. The proposed approach has enabled the investigation and potential identification of the biological factors involved in a bacterial culture and the biomolecules contributing to the main trends of such biological factors. The PLS-DA has indicated that the spectra of bacterial cell samples from the exponential, stationary and decline growth phases exhibit differences due to differentially expressed proteins which can be identified from examining the PLS-DA loadings plot (that is, *m/z* ratio values) in conjunction with protein database search. Proteins such as ribosomal proteins are predominantly expressed during the exponential phase cultures since they are actively growing. This and other studies have indicated that these are potential biomarkers for cultures in the exponential phase. Other proteins previously identified in the literature were found to be differentially expressed in both the stationary and decline phases. The minimal sample pretreatment requirements for ICM which reduces sample processing time, the straight forward interpretability of the PLS-DA results, and the availability of internet accessible proteomic databases provide the potential to easily and rapidly identify biomarkers. This rapid identification of biomarkers demonstrates one aspect of this approach that could be useful for biomarker profiling in mammalian cell culture. This approach could also serve as a valuable tool in process development in the bioprocessing industry to enhance cell growth by facilitating the selection of high producing cell lines based on identified biomarkers, as well as identify potential targets for cell engineering. This is described in the subsequent chapter (chapter 7).

Chapter 7

7. CASE STUDY II: Biomarker Profiling Of Cultured Intact Mammalian Cell Lines Utilising the ICM, PLS-DA, and Protein Database Search Approach

7.1. Overview

In chapter 6, the application of the approach utilising intact-cell MALDI-ToF MS (ICM-MS), projection to latent structure – discriminant analysis (PLS-DA), and database search was demonstrated in the study of protein biomarker profiling, using an *E. coli* growth-phase-associated protein biomarker model. This case study acted as a proof-of-principle study for the ICM, PLS-DA and protein database search approach. The case study demonstrated that if between-class (exponential, stationary and decline phase classes in this case) variabilities exist in spectra data, PLS-DA is used to capture these differences and hence classify the samples based on these variabilities. The results obtained from the proof-of-concept study suggest that this approach was successful in identifying protein biomarkers.

In this chapter, the approach is applied in biomarker profiling of IgG-producing CHO cell lines during bioprocessing. It is hoped that a between-class variability (based on the cell line productivity) will exist within the CHO cell line spectra data so that PLS-DA can be used to classify the samples and identify potential protein biomarkers which are associated to the CHO cell line productivities.

7.2. Introduction

'Intact-cell' MALDI-ToF MS (ICM-MS) raises many possibilities for the analysis of mammalian cells (Chaurand *et al.,* 2006; Chaurand *et al.,* 2007; Crossman *et al.,* 2006; Khatib-Shahidi *et al.,* 2006; Reyzer *et al.,* 2007). These applications have been discussed in section 4.3.4.5 (chapter 4). Despite the application of ICM-MS to bacteria (section 4.3.4.4) and mammalian cells, there have been relatively few studies that have applied this approach to mammalian cell cultures (MCCs) in bioprocessing (Buchanan *et al.,* 2007; Feng *et al.,* 2010; Dong *et al.,* 2011; Feng *et al.,* 2011). The few applications that exist in the literature (to the best of the author's knowledge) have also been discussed in section 4.3.4.5 (chapter 4).

7.3. Specific Aims and Contributions of the Chapter

In a recent application Feng *et al.*, (2010) applied ICM-MS to a batch of CHO cell lines producing the monoclonal antibody, IgG. The mass spectra data sets obtained were modeled using PCA and PLS, discriminating between the cell lines (high/low producers) based on different expressed recombinant proteins (different productivities). As a follow-up to this work, the group more recently used PLS-DA to model spectra data sets obtained from another batch of monoclonal IgG-producing CHO cell lines (Feng *et al.,* 2011). In both studies mass spectra peaks of potential protein biomarkers associated with productivity were identified using the multivariate data analysis approaches of PCA, PLS and PLS-DA. Whilst these two studies represent an important step forward towards pattern-based biomarker profiling of mammalian cell culture (MCC) in bioprocessing, no attempts were made by the experimenters to provide the mass spectra peak assignments that were obtained to potential protein biomarkers.

In this chapter, ICM-MS combined with PLS-DA was used to distinguish between cell lines in terms of productivities (high/low producers; Hs/Ls). Protein database searches were then used to identify potential biomarkers associated with productivity, and whose differential presence may be useful in classifying between the two cell line classes. Furthermore, biological interpretations as to the presence of such protein biomarkers will be provided. The presence of such biomarkers may then be used as a basis for predicting cell line productivities; provide insight into biological mammalian cell lines used during bioprocessing; and may give indications to potential genetic engineering targets that may be exploited to engineer batter cell lines.

Figure 7.1 shows an overview of the various steps involved in using ICM-MS, PLS-DA and protein database searches for the biomarker profiling of IgG monoclonal antibody-producing CHO cell lines. It begins with preparation of the CHO cell samples, the analysis of the samples by MALDI mass spectrometry, and preprocessing of the spectra data generated to remove unwanted variation whilst preserving biological information. Sampling is then carried out on the preprocessed data sets to separate the samples into training and test sets. The training set is overviewed using PCA to study initial trends and later analysed using PLS-DA, whilst the test sets are retained for external validation of the PCA and PLS-DA models built. PLS-DA scores and loadings plot are then interpreted. The information from the scores and loadings plot is used for database searches to identify protein biomarkers.

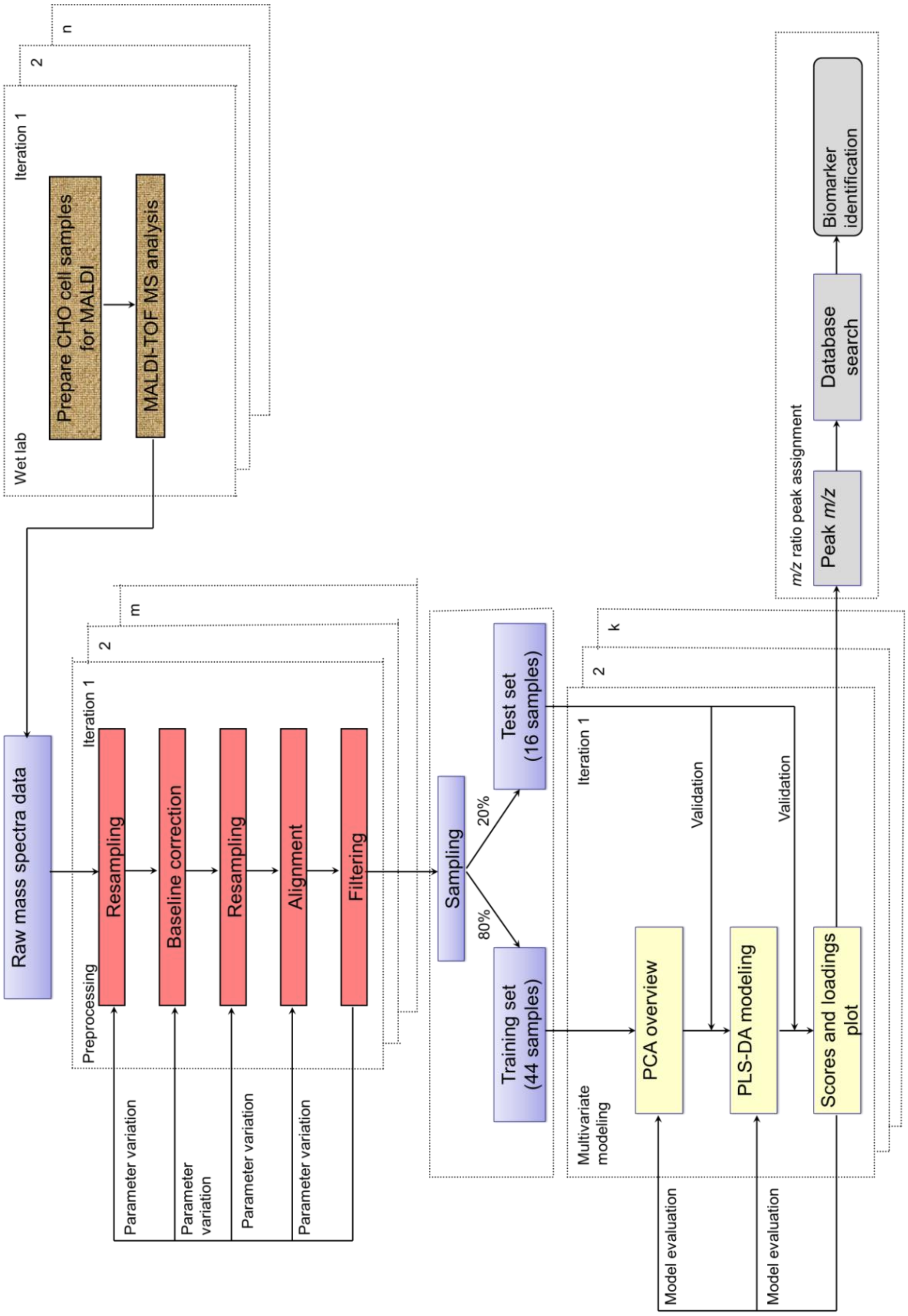


Figure 7.1: An overview of the various steps involved in the growth phase-associated protein biomarker profiling of IgG monoclonal antibody-producing CHO cell lines using ICM-MS, PLS-DA and protein database search

7.4. Experimental Section

7.4.1. Chemicals

This section outlines the chemicals used in this project, as provided by collaborators at the School of Bioscienes, University of Kent, to carry out the laboratory experiments. Acetonitrile/Trifluoroacetic Acid (ACN/TFA; 0.1%, v/v), Sinapinic acid (SA); Sucrose (>99.5%, Sigma Aldrich UK); Phosphate buffer saline (PBS) - Oxoid tablets (made up as directed, filter-sterilised and stored at 4^{o}C); and the MALDI-TOF calibrant (a protein mixture containing insulin, ubiquitin I, cytochrome C, and myoglobin) were purchased from Bruker Daltonics, GmbH, Germany

7.4.2. Cell Culture

This section presents information about the cell line used in this project, as provided by collaborators at the School of Bioscienes, University of Kent. Immunoglobulin G (IgG1 and IgG4) monoclonal antibody expressing mammalian cell lines (concealed identity) were generated in-house (Lonza Biologics, Slough, UK) through methothrexate amplification. Cell line suspensions of varying productivities were cultivated in a 96 deep well plate (96DWP), 24 well plate, and bioreactor. The cell line culture time was 5 days after which exponential phase cells were harvested for analysis. At the 96DWP scale of production, cell lines were attributed to a high or low class based on their antibody titres in the 96 deep well plate scale during production.

Samples of cell lines which had above 1000mgmL^{-1} antibody titres were classed as high producers (Hs). This was the threshold set by Lonza Biologics with respect to cell line productivity. Samples which had specific productivity (*q*P) titres less than 1000mgmL^{-1} were classed as low producers (Ls). Samples collected were retained for MALDI analysis. Viable cell concentration and percentage viable cells were measured using a Cedex automated cell counter (Innovatis, Bielefeld, Germany).

7.4.3. MALDI-ToF MS Analysis and Data Preprocessing

CHO cell line sample preparation for MALDI analysis was carried out as described in section 3.3.6.2. All mass spectra were acquired with an Ultra Flex MALDI-ToF mass spectrometer (Bruker Daltonics, GmbH, Germany), located in the School of Biosciences, University of Kent, Canterbury, Kent, UK. The instrument parameters used as well as the analysis procedure has been described in section 3.3.6.2. Data preprocessing was carried out using in-house scripts developed from MATLAB® v.7.6.0.324 (R2008a the MathWorks, Inc.) and functions from the Bioinformatics Toolbox of MATLAB® (v 3.1, R2008a, Eigenvector Research, Inc.). Preprocessing studies was carried out as described in section 3.6, to find the appropriate combination of preprocessing techniques.

7.4.4. Multivariate Data Analysis

The application of multivariate data analysis (PCA and PLS-DA) to the spectral data was carried out using the MATLAB® software v. 7.12.0.635 (R2011a The MathWorks, Inc.) and the PLS Toolbox v. 6.5.1 (Eigenvector Research, Inc.). The PLS-DA algorithm used in modeling the CHO cell line spectra data set has been described in detail in section 5.6.1. PCA was first applied to get an overview of the preprocessed mass spectral data sets to identify groupings in the data sets. Random sampling was used to separate the training set from the test set samples. In this study, 1/5 of the spectra data sets were used as the test set and 4/5 as training set. After random sampling, 44 spectra data sets (18 Hs and 26 Ls) were used as the training set whilst 16 samples (8 Hs and 8 Ls) comprised the test set.

7.4.5. Protein Database Search

Protein database searches were carried out to assign protein identities to mass spectra ion signals obtained from information in the PLS-DA loadings plots. The searches were conducted using the MALDI mass spectra protein experimental molecular weights (MWs) (*m/z* ratio peaks or signal ions) and the organism's theoretical MWs found in the UniProtKB/Swiss-Prot database. With the query specified as Chinese Hamster Ovary (CHO), the experimental MWs were matched to biomarker peaks contained in the Protein Knowledgebase (UniProt) and TrEMBL query form found in the database for the identification of the mammalian cell line's main protein biomarkers.

7.5. Results and Discussions

7.5.1. Principal Component Analysis Modeling of the Spectra Data Sets from CHO cell lines

Principal component analysis (PCA) was applied to get an overview of the 64 preprocessed CHO cell mass spectral data sets to help identify trends in the data and to sample the data into training and test sets. The first four principal components account for 46.60% of the original variation (Table 7.1). A plot of PC2 vs PC3 is shown in Fig. 7.2(a). The four-component model shows evidence of separation between the two classes along PC2, with most of the low producer cell line samples lying in the upper left quadrant and the high producer cell lines being found in the lower left quadrant. Four samples, 2 low producers (Ls) and 2 high producer cell lines (Hs) were outliers as they were located outside the 95% confidence region. After removal of the outliers and subsequent splitting of the data into training and test set samples, a 2-component PCA model was built with the 44 training data sets (Table 7.2). The two component model accounted for 93.73% of the variability in the original data set. A plot of PC1 vs PC2 of the 2-component model showed a strong evidence of separation between the two classes along PC2 (Fig. 7.2(b)). Overall, these suggest that the PCA representation has some practical value and that a major part of the spectral variation is related to class differences between the high and low producer CHO cell lines.

Principal component (PC) number	Eigenvalue of covariance (*X*)	% Variance captured for this PC	Total % variance captured
1	5.86×10^3	32.41	32.41
2	1.24×10^3	6.87	39.28
3	7.00×10^2	3.87	43.15
4	6.23×10^2	3.45	46.60
5	4.85×10^2	2.68	49.28
6	4.25×10^2	2.35	51.63
7	3.64×10^2	2.01	53.64
8	3.46×10^2	1.91	55.55

Table 7.1: Results of PCA model for all 64 CHO mass spectra data sets

Principal component (PC) number	Eigenvalue of covariance (*X*)	% Variance captured for this PC	Total % variance captured
1	3.28×10^5	90.13	90.13
2	1.36×10^4	3.73	93.86
3	4.91×10^3	1.35	95.21
4	3.89×10^3	1.07	96.28
5	2.43×10^3	0.57	96.94
6	1.53×10^3	0.42	97.36
7	9.49×10^2	0.26	97.62
8	8.79×10^2	0.24	97.89

Table 7.2: Results of PCA model for the 44 calibration CHO mass spectra data sets after outlier removal

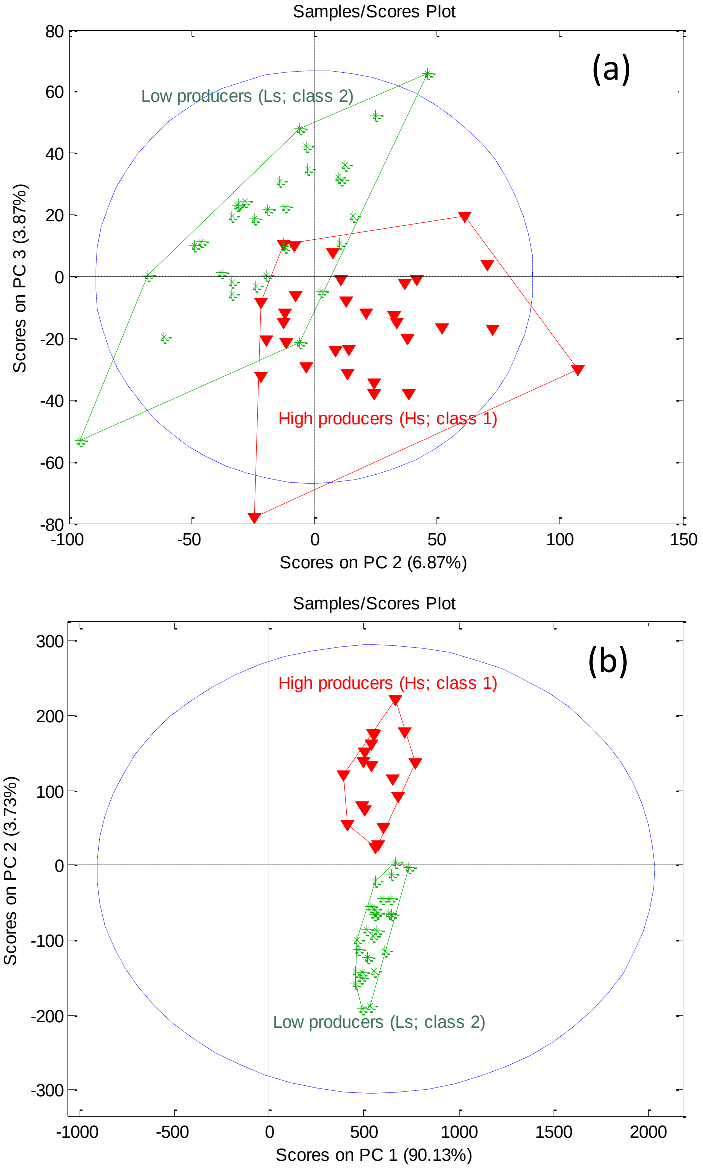


Figure 7.2: PCA scores plot for the Chinese hamster ovary mass spectra data before and after sampling

7.5.2. Partial Least Squares – Discriminant Analysis (PLS-DA) Modeling of the Spectra Data Sets from CHO cell lines

7.5.2.1. Latent Variable (LV) Selection

The 44 training mass spectra profiles from the CHO cell lines were then modeled by PLS-DA. The PLS-DA model use the whole *m/z* ratio region to differentiate between classes thus it is possible to investigate which regions were highly weighted in the model when classifying a given sample as a low or high producer cell line. Fig. 7.3 shows a plot of RMSECV as a function of the number of LVs after cross-validation. It captures the effect of increasing the number of LVs in the PLS-DA model. The optimum number of LVs was selected to simultaneously maximise the percentage of explained systematic variation while achieving correlation with ***Y***. Fig. 7.3 indicates that four LVs should be retained.

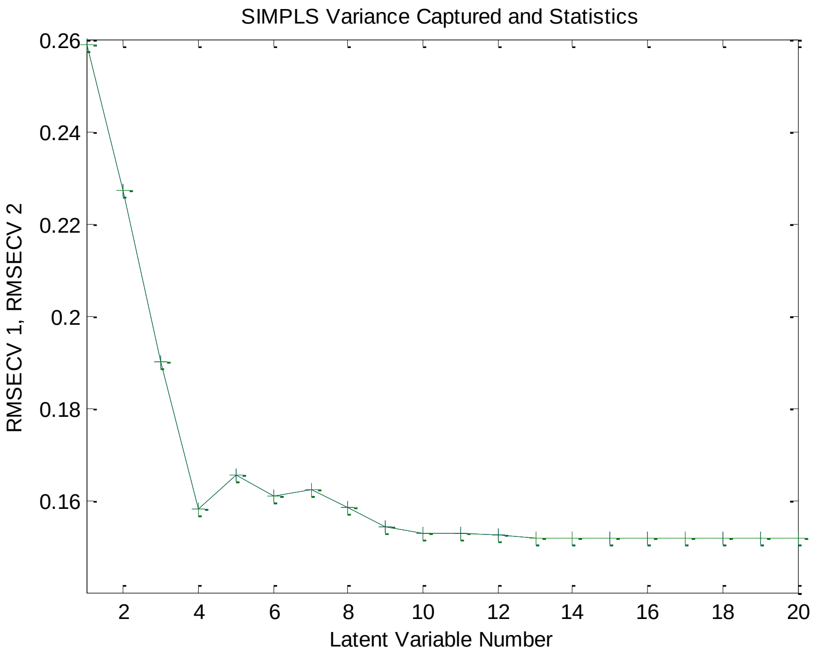


Figure 7.3: RMSECV as a function of the number of LVs added to the PLS-DA model for the CHO cell line data set

7.5.2.2. Model Quality

Results of the PLS-DA analysis are summarised in Table 7.3, 7.4 and Fig. 7.4. The 4-LV model captured 95.50% of the ***Y***-block of the training data set (Table 7.3), and the 100% specificity and sensitivity suggest that the model was excellent (Table 7.4). This was supported by the calculated ***Y*** versus measured ***Y*** in fitting and prediction for the PLS-DA model after cross-validation (Fig. 7.4). There was no deviation identified along the y-axis in both the high and the low producer classes, (Fig. 7.4(a)) and (Fig. 7.4(b)) respectively, suggesting that that all the useful information is taken into account by the model.

Latent variable (LV) number	***X*-Block**		***Y*-Block**	
	% Variance captured for this LV	**Total % variance captured**	**% Variance captured for this LV**	**Total % variance captured**
1	7.78	7.78	78.39	78.39
2	86.06	93.84	3.68	82.07
3	0.87	94.71	9.05	91.12
4	0.58	95.28	4.38	95.50
5	0.36	95.64	2.51	98.01
6	1.11	96.75	0.56	98.57
7	0.39	97.14	0.78	99.35
8	0.49	97.64	0.28	99.63

Table 7.3: PLS-DA results showing the quality of the model after calibration (Cal) and cross-validation

Modeled class	**Class 1 (High producers)**	**Class 2 (Low producers)**
Sensitivity (Cal)	1.000	1.000
Specificity (Cal)	1.000	1.000
Sensitivity (CV)	1.000	1.000
Specificity (CV)	1.000	1.000
Classification error (Cal)	0	0
Classification error (CV)	0	0
RMSEC	0.104	0.104
RMSECV	0.158	0.158
R^2 Cal	0.955	0.955
R^2 CV	0.897	0.897

Table 7.4: Results of PLS-DA model for the 44 calibration CHO mass spectra data sets

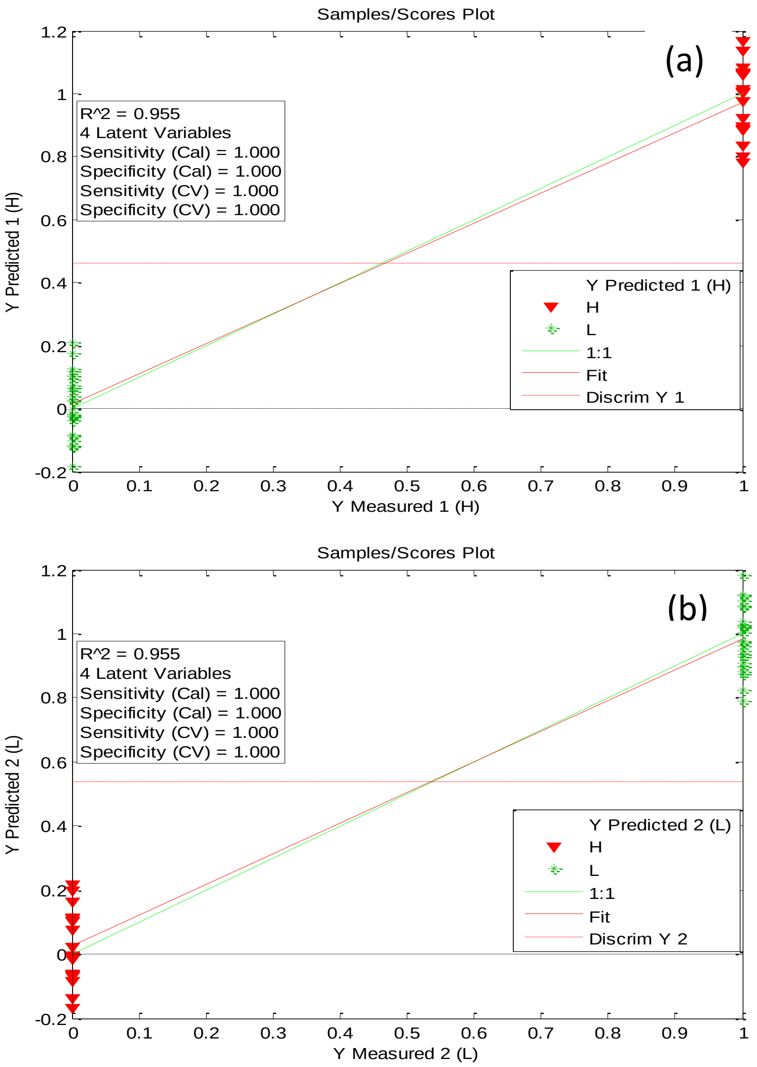


*Figure 7.4: PLS-DA calculated **Y** versus measured **Y** in fitting and prediction after cross-validation for the 44 CHO cell line calibration spectra; (a) High producer class, and (b) Low producer class*

7.5.2.3. External Validation and Model Performance

The model was successfully validated with 16 (8 Hs and 8 Ls) test data sets (Table 7.5; Fig. 7.5); indicating that the PLS-DA model is informative in terms of class separation. Table 7.5 shows that the model had a good performance (75-100% sensitivity and specificity). Furthermore, the model had a low classification error of 3.3% for both classes. As indicated in Fig. 7.5(a) two high producer samples (class 1) were misclassified (coloured in pink) falling below the decision threshold. This resulted into a good prediction sensitivity and specificity of 75% and 100% for class 1 respectively. No class 2 samples were misclassified giving 100% prediction sensitivity (Fig. 7.5(b)). The successfully validation of the model means that the model parameters such as scores and loadings are valid. The model performance also validates the choices of the preprocessing methods/parameters used for the spectra data described in section 3.6 (chapter 3).

Modeled class	**Class 1** **(High producers; Hs)**	**Class 2** **(Low producers; Ls)**
Sensitivity (prediction)	0.750	1.000
Specificity (prediction)	1.000	0.750
Classification error (prediction)	0.125	0.125
RMSEP	0.328	0.328
Prediction Bias	-0.067	0.067
R^2 prediction	0.596	0.596

Table 7.5: PLS-DA results showing the performance of the model after external validation with 16 test set samples

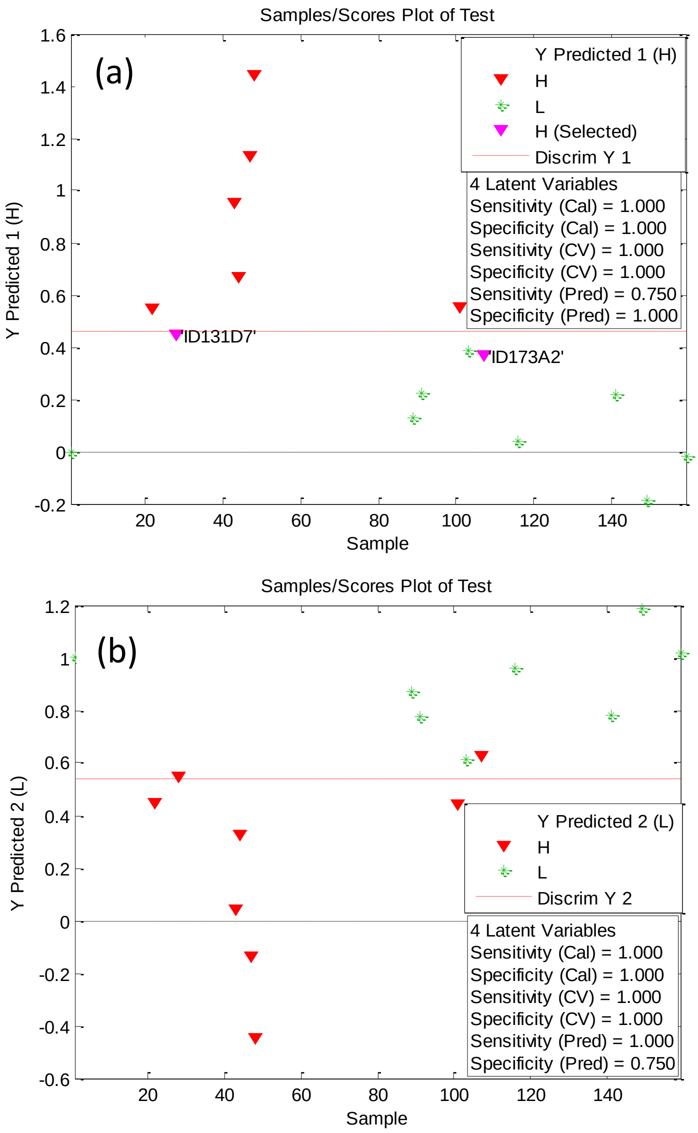


*Figure 7.5: PLS-DA calculated **Y** versus measured **Y** in fitting and prediction after external validation with 16 CHO cell line mass spectra test set samples; (a) High producer class, and (b) Low producer class*

7.5.2.4. Scores Plot

The model demonstrates the ability to separate the two classes (Fig. 7.6). As can be seen in Fig. 7.6, a plot of LV1 vs LV2 (a), and LVs 3 (b) indicates that the model concentrates most of the discriminatory information in the first LV. The two classes are approximately separated along LV1, with the high producer class having negative whilst the low producer class has positive scores.

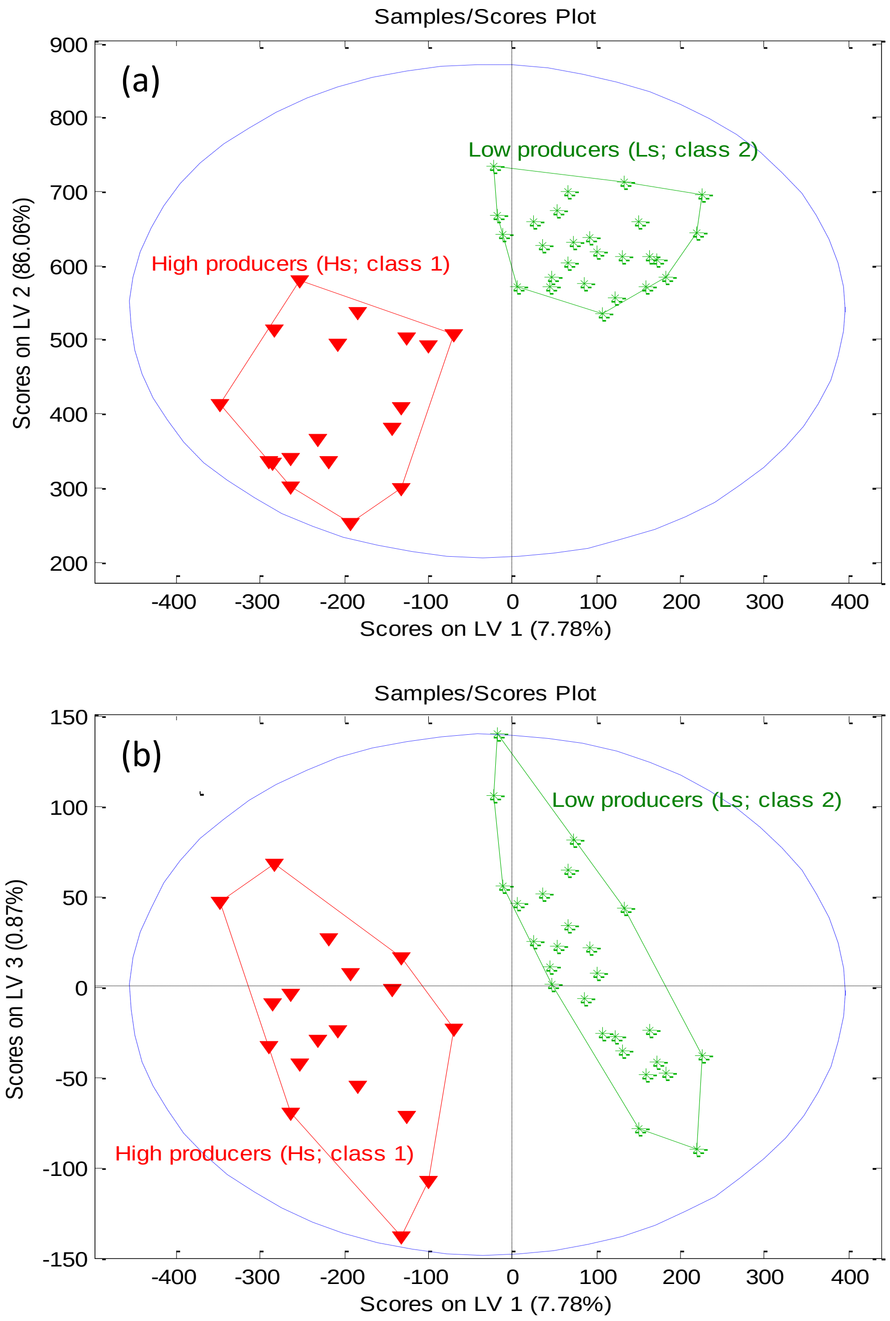


Figure 7.6: PLS-DA scores plot for the 44 CHO cell line calibration mass spectra data sets;(a) Scores on LV2, and (b) Scores on LV3

7.5.2.5. Loadings Plot

There is a relationship between the scores and the loadings plot for PLS-DA. The loadings plot for LV1 (Fig. 7.7) indicates which *m/z* ratio ion signals in the MALDI data set contain information that is driving the separation of the samples in the scores plot. More specifically, it describes the weighting coefficients for each *m/z* ratio ion signal (experimental MWs of proteins) which are associated to variables in the loadings plot.

In Fig. 7.7, it can be seen that variables at approximately 1217, 3056, 3708, 4781, and 4874 have high loadings with negative contribution. Since high producer CHO cell line samples have negative scores in LV1 in the scores plot, it implies that these samples can be distinguished from the low producer samples by the presence of *m/z* ratio signal ions associated with variables with negative loadings in the LV1 loadings plot. These signal *m/z* ratio signal ions potentially identify proteins which are differentially expressed in high producer cell lines.

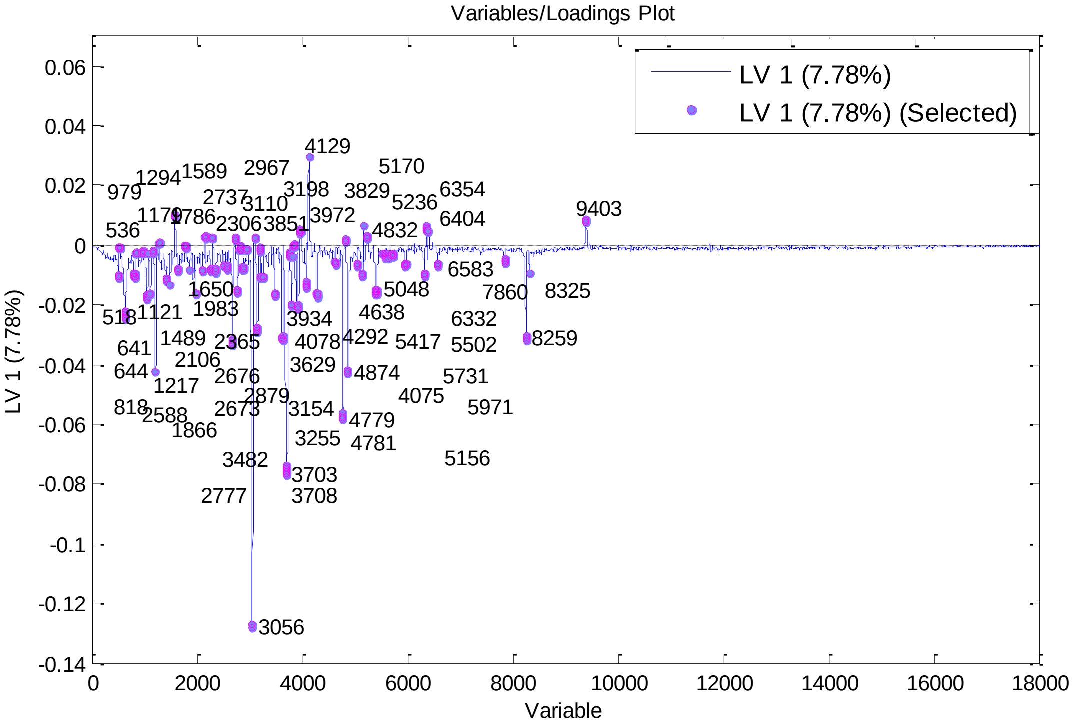


Figure 7.7: PLS-DA loadings plots on LV1 for the 44 CHO cell line calibration mass spectra data sets

7.5.3. Protein Database Search for CHO Cell Line Samples

The PLS-DA loadings plot on LV1 was considered for further analysis. The variables from the loadings plot are associated to *m/z* ratio signal ions (Table C.1 and C.2, Appendix C). Since the *m/z* ratio signal ions are singly charged protonated protein (MH^+) molecules, the MH^+ represents the MALDI experimental molecular weights (MWs) of the ionised proteins expressed by the cell lines. Thus the *m/z* ratio ion signals (experimental MW of protein ions) were submitted to a protein database search to assign protein identities to them. The experimental MWs of protein ions were used as search parameters and were matched with sequence derived theoretical MW values in the SwissPROT/TrEMBL database (http://expasy.org/proteomics). Tables 7.6 and 7.7 summarise the results of the database searches. A number of matches resulted in protein biomarkers which have been predicted in the CHO genome (*Cricetulus griseus*). Experimental evidence using 2D-PAGE and tandem-mass spectrometry, that some of the matched proteins are expressed in CHO cell line culture have been well described in the literature (Kaufmann *et al.,* 1999; Lee *et al.,* 2003; Van Dyk *et al.,* 2003; Krawitz *et al.,* 2006; Pascoe *et al.,* 2007; Meleady, 2007; Kim *et al.,* 2009; Carlage *et al.,* 2009).

When no protein was found to match an experimental MW, a further search was carried out for proteins of related mammalian species such as humans (*Homo sapiens*) or rodents (*Mus musculus*, mouse; *Rattus norvegicus*, Rat). For matched CHO proteins whose functions are not yet available in the database, the function was inferred from related proteins of the above close species and this was mentioned in the 'remarks' column of Table 7.6 and 7.7. Matches were not found for most experimental MWs below 4 kDa. These MWs may be a number of abundant matrix-related ions since most of the currently used matrices have MWs less than 4 kDa. They may also act as their own matrices, producing a variety of matrix-related ions during laser ionisation (Guo *et al.,* 2002).

In Table 7.6 and 7.7, the column 'protein existence' includes the value E = ' Evidence at transcript level level', indicating the existence of a protein that has not been proven but whose expression data (such as existence of cDNA(s), RT-PCR or Northern blots) indicates the existence of a transcript; I = 'Inferred by homology', indicates the existence of a protein is probable since clear orthologs exist in closely related species; P = 'Predicted', is used for entries without evidence at protein, transcript, or homology levels; and U = ‘Uncertain’, indicates the existence of the protein is unsure. Most of the proteins that were identified had a value of ‘P’, implying that they are only predicted and no MALDI experimental evidence exist that indicates that such proteins are expressed in vivo in CHO cells.

7.5.4. Swiss-Prot/TrEMBL Protein Molecular Weight Distribution for CHO cells, MALDI Mass Accuracy and Biomarker Identification

Biomarker assignment and identification for CHO cells was based on several factors. Similar to the *E. coli* K-12 cells (section 6.6.4), biomarker identification was based on *m/z* ratio values or experimental MW ranges can correspond to variable count ranges in the PLS-DA loadings plot. An average experimental MW is assigned to a protein biomarker if it falls within this range and matches the theoretical MW of the protein biomarker based on the mass accuracy of linear mode MALDI-ToF MS instrument as described in section 6.6.4. Tables C.3 to C.4 shows variable count ranges and their experimental MW ranges as well as the average experimental MWs matching theoretical MWs for the all the protein biomarkers shown in Tables 7.6 and 7.7.

In addition to the mass accuracy, assignments took into consideration the molecular mass distribution of 24049 protein sequences (providing theoretical MWs) of CHO cell proteins derived from genomic open reading frame as well as nongenomic entries found in the SwissPROT/TrEMBL database. The molecular mass distribution of known CHO cell proteins (Fig. 7.8) shows a positive skewness i.e. most proteins tend to cluster toward the lower end of the MW scale with increasingly fewer proteins at the upper end of the MW scale. Fig. 7.8 shows the molecular mass distribution (in bins of 1 kDa) of CHO cell proteins deposited in the SwissPROT/TrEMBL sequence database. The positive skewness in in Fig. 7.8 suggests that most of the CHO proteins found in the SwissPROT/TrEMBL database are lower MW proteins in the range 2 to 24kDa. Therefore, it may be expected that unique combinations of lower MW CHO protein in the range 2 to 24 kDa can serve as protein biomarkers. Since basic proteins are more amenable to protonation during MALDI analysis assignments also considered the basicity of the proteins.

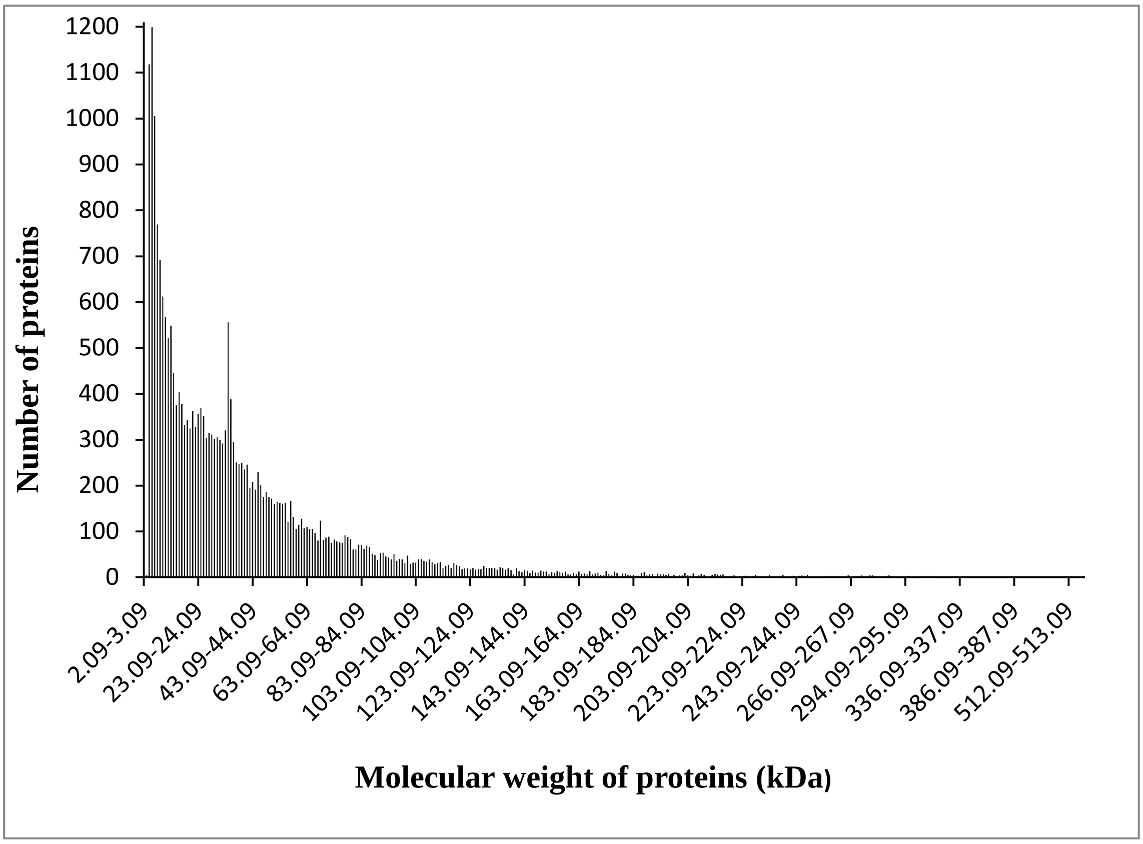


Figure 7.8: Molecular weight (MW) distribution (in bins of 1 kDa) of 24049 CHO cell proteins deposited in the SwissPROT/TrEMBL sequence database

The full explanation of superscript letters assigned to column titles in Table 6.5 are as follows: Tables 7.6 and 7.7:

[a]Variable counts were derived from PLS-DA loadings plot on LV1 (along which the two classes of cell lines were separated in the cores plot).

[b]The intensity of the variables (*m/z* ion signals) is the magnitude of their loadings in the loadings plot. High: intensity ≥ ±0.04 units; medium: intensity ≥ ±0.02 and ≤ ±0.04 units; low: intensity between 0 and ±0.02 units.

[c]Experimetal MWs were the *m/z* ratio ion (MH^+) signals associated to the variable counts of the from the PLS-DA loadings plot.

[d]Theoretical or sequence MWs were calculated using the Compute pI/MW tool (http://web.expasy.org/compute_pi/) of the SwissProt/TrEMBL Protein database.

[e]Protein names, existence, PI, accession numbers and remarks were from ExPASy Proteomics Server (http://www.uniprot.org/uniprot/) of the SwissProt/TrEMBL Protein database. Protein existence with value E = Evidence at transcript level; I = Inferred by homology; P = Predicted; and U = Uncertain. Remarks indicate protein functions from related mammalian species humans (*Homo sapiens*) or rodents (*Mus musculus*, mouse; *Rattus norvegicus*, Rat) for matched CHO proteins whose functions are not yet available in the database.

Table 7.6: SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the High Producer Chinese Hamster Ovaries Cell Line Samples

Variable count of loadings plot[a]	Intensity[b]	Exp'tal MW (Da)[c]	Sequence MW (Da)[d]	PI[e]	Database Accesion number[e]	Name[e]	Protein Existence[e]	Remarks[e]
Antibody and Protein Folding Proteins								
1233	High	6565.04	6565.38	5.09	G3HQE8_CRIGR	Stress-70 protein, mitochondrial	P	Function inferred from GRP75_HUMAN
2365	Medium	8575.57	8571.93	9.10	G3I0H0_CRIGR	Peptidyl-prolyl cis-trans isomerase NIMA-interacting 4	P	Function inferred from PIN4_HUMAN
3168	Medium	10166.96	10167.32	8.92	IGHG4_HUMAN	Ig gamma-4 chain C region (CH1)	E	Domain
3168	Medium	10166.96	10167.32	7.22	LAC3_HUMAN	Ig lambda-3 chain C regions	E	Domain
3480	Medium	10821.18	10820.97	5.58	IGKC_HUMAN	Ig kappa chain C region	E	Domain
3480	Medium	10821.18	10820.97	5.58	IGKC_HUMAN	Ig kappa chain C region	E	Domain
3674	High	11238.26	11237.50	6.91	LAC3_HUMAN	Ig lambda-3 chain C regions	E	
3703	High	11301.29	11301.05	9.27	G3H1E7_CRIGR	60 kDa heat shock protein, mitochondrial	I	
3720	High	11338.32	11337.91	4.48	G3GT32_CRIGR	Heat shock cognate protein HSP 90-beta	P	Function inferred from HS90B_HUMAN
3724	High	11347.04	11347.66	8.29	LAC1_HUMAN	Ig lambda-1 chain C regions	E	
3748	High	11399.43	11399.96	8.86	G3HH47_CRIGR	Peptidyl-prolyl cis-trans isomerase A	I	
3822	Medium	11561.75	11560.85	8.66	KV123_HUMAN	Ig kappa chain V-I region Walker	P	Chain
3936	Medium	11814.05	11815.14	8.72	KV311_HUMAN	Ig kappa chain V-III region IARC/BL41	E	Chain
4076	Low	12127.62	12126.51	5.52	IGHG4_HUMAN	Ig gamma-4 chain C region (CH3)	E	Domain
1353	Low	12658.31	12655.48	6.17	G3H731_CRIGR	Peptidyl-prolyl cis-trans isomerase A	I	
5429	Medium	15369.71	15371.00	9.37	G3I881_CRIGR	FK506-binding protein 2	P	
5431	Medium	15374.79	15375.35	4.97	Q7M080_CRIGR	DnaK-type molecular chaperone	I	Function inferred from HSP71_HUMAN
6340	Low	17768.76	17768.08	8.46	PPIA_CRIGR	Peptidyl-prolyl cis-trans isomerase A	E	Meth loss
6410	Medium	17960.29	17959.38	8.44	G3HIQ1_CRIGR	Peptidyl-prolyl cis-trans isomerase	I	
6418	Medium	17982.25	17982.30	6.28	G3HUK9_CRIGR	Peptidyl-prolyl cis-trans isomerase	I	
8214	Low	23250.62	23250.00	6.10	/	Humanised IgG light chain	E	
8268	Medium	23419.49	23419.30	6.23	HSPB1_CRILO	Heat shock protein beta-1 (Heat shock 27 kDa protein)	E	

Table 7.6: SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the High Producer Chinese Hamster Ovaries Cell Line Samples

Table 7.6 – (Continued): SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the High Producer Chinese Hamster Ovaries Cell Line Samples

Variable count of loadings plot[a]	Intensity[b]	Exp'tal MW (Da)[c]	Sequence MW (Da)[d]	PI[e]	Database Accesion number[e]	Name[e]	Protein Existence[e]	Remarks[e]
Protein Biosynthesis Proteins								
2371	Low	8588.35	8588.16	8.82	G3HJE6_CRIGR	40S ribosomal protein S15	I	
2659	Low	9143.43	9143.40	8.55	G3HGW7_CRIGR	40S ribosomal protein S21	P	
3038	High	9900.38	9900.14	7.09	G3IMY3_CRIGR	40S ribosomal protein S2	P	
3146	Medium	10121.59	10122.12	9.85	G3IED0_CRIGR	40S ribosomal protein S15	I	
3660	High	11207.90	11206.82	4.31	G3ILE6_CRIGR	60S acidic ribosomal protein P1	P	Function inferred from RLA1_HUMAN
3677	Low	11244.77	11243.66	4.95	G3GZV3_CRIGR	60S acidic ribosomal protein P2	P	
4873	High	13978.88	13978.00	/	G3GWA7_CRIGR	40S ribosomal protein S6	P	
6580	Low	18429.73	18430.73	10.31	G3I004_CRIGR	40S ribosomal protein S11	I	
642	Low	5621.33	5620.59	9.92	G3ICB6_CRIGR	Elongation factor 1-alpha 1	P	
658	Low	5645.91	5646.66	9.12	G3H513_CRIGR	Ribosomal protein S27	I	
1668	Low	7306.41	7306.42	10.11	G3HX33_CRIGR	60S ribosomal protein L7	P	
1979	Low	7860.75	7859.99	10.54	G3GX65_CRIGR	60S ribosomal protein L37	P	Function inferred from RL37_MOUSE
2769	Low	9360.02	9359.33	9.90	DPM2_CRIGR	Dolichol phosphate-mannose biosynthesis regulatory protein	I	Meth loss
2868	Low	9557.13	9557.88	8.89	G3ILV7_CRIGR	Eukaryotic translation initiation factor 3 subunit E	P	Function inferred from EIF3E_HUMAN
3630	Medium	11142.98	11143.87	8.91	G3IN64_CRIGR	60S ribosomal protein L11	I	
3865	Medium	11656.60	11657.32	4.53	G3IPA3_CRIGR	Tryptophanyl-tRNA synthetase, cytoplasmic	I	
3876	Medium	11680.92	11680.94	4.38	G3I3H2_CRIGR	60S acidic ribosomal protein P2	P	
3961	Medium	11869.75	11868.82	8.95	G3IF50_CRIGR	Polymerase delta-interacting protein 3	P	
4628	Low	13404.01	13403.20	8.47	G3IIH3_CRIGR	Eukaryotic translation initiation factor 4E	I	
4888	High	14027.33	14027.07	6.29	G3I8A7_CRIGR	Protein S100-A9	P	
5151	Low	14697.06	14696.99	9.59	G3HV18_CRIGR	60S ribosomal protein L12	I	
5169	Low	14716.92	14716.18	9.88	G3I078_CRIGR	60S ribosomal protein L30	P	
5654	Low	15946.06	15945.34	9.05	C1D_CRIGR	Nuclear nucleic acid-binding protein C1D	E	
5978	Low	16794.64	16793.20	5.38	G3I948_CRIGR	Eukaryotic translation initiation factor 5A-2	P	

Table 7.6 – (Continued): SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the High Producer Chinese Hamster Ovaries Cell Line Samples

Variable count of loadings plot[a]	**Intensity**[b]	**Exp'tal MW (Da)**[c]	**Sequence MW (Da)**[d]	**PI**[e]	**Database Accesion number**[e]	**Name**[e]	**Protein Existence**[e]	**Remarks**[e]
Cytoskeleton/Structural Related Proteins								
2666	Medium	9157.17	9156.60	10.08	G3H687_CRIGR	Calponin-1	P	
2880	Medium	9581.16	9580.84	5.40	G3IC49_CRIGR	Annexin A7	I	
3622	Low	11125.69	11124.79	9.52	G3I126_CRIGR	Catenin alpha-2	P	
5659	Low	15958.99	15957.94	5.23	G3GRX4_CRIGR	Coactosin-like protein	P	
6584	Low	18440.84	18439.69	4.71	G3H511_CRIGR	Tropomyosin alpha-1 chain	I	
DNA and RNA Metabolism Proteins								
814	Low	5888.43	5888.10	9.60	G3HCT7_CRIGR	DNA polymerase subunit gamma-2, mitochondrial	P	Function inferred from DPOG2_HUMAN
1229	High	6558.40	6557.43	4.71	G3GS73_CRIGR	Heterogeneousnuclear ribonucleoprotein A3-like 1	P	Function inferred from ROA3_HUMAN
1647	Low	7269.70	7268.97	9.69	G3HFW8_CRIGR	Non-histone chromosomal protein HMG-14	P	
1654	Low	7281.93	7282.41	10.62	G3IPZ7_CRIGR	ATP-dependent RNA helicase DHX8	P	
1868	Low	7660.57	7659.96	8.71	G3HSM6_CRIGR	Guanine nucleotide-binding protein subunit gamma	I	
1973	Low	7849.86	7850.14	7.78	G3IAE2_CRIGR	Guanine nucleotide-binding protein subunit gamma	I	
2784	Low	9387.77	9388.76	7.56	G3IJP9_CRIGR	Transcription elongation factor 1-like	P	
2863	Low	9547.12	9547.56	6.82	G3H7G8_CRIGR	Splicing factor 3B subunit 5	P	
3066	High	9957.50	9957.40	11.99	G3GUE8_CRIGR	Putative uncharacterized protein	P	
3190	Low	10212.42	10212.93	8.96	G3HV32_CRIGR	Histone H3.3 type 1	I	DNA binding
3484	Low	10829.70	10829.73	11.51	G3HDU7_CRIGR	Histone H4	I	
3621	Low	11123.54	11123.00	10.28	G3ILX7_CRIGR	Histone H2A	I	
3624	Low	11128.35	11128.02	6.26	G3GS83_CRIGR	Transcription elongation factor B polypeptide 2	P	Function inferred from ELOB_HUMAN
3722	Medium	11342.68	11342.16	10.29	G3GYE5_CRIGR	Histone H2A	I	
3723	Medium	11344.86	11345.29	11.20	G3HDT9_CRIGR	Histone H4	I	
3803	Medium	11519.97	11519.17	7.69	XPA_CRIGR	DNA repair protein complementing XP-A	I	
3803	Medium	11519.97	11519.17	7.69	XPA_CRIGR	DNA repair protein complementing XP-A cells homolog	I	
4783	High	13773.90	13774.96	10.48	G3H3H8_CRIGR	Histone H2A	I	

Table 7.6 – (Continued): SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the High Producer Chinese Hamster Ovaries Cell Line Samples

Variable count of loadings plot[a]	Intensity[b]	Exp'tal MW (Da)[c]	Sequence MW (Da)[d]	PI[e]	Database Accesion number[e]	Name[e]	Protein Existence[e]	Remarks[e]
General Metabolism/Glycolysis Proteins								
1064	Low	6287.70	6287.07	4.75	G3I1B8_CRIGR	Phosphoglycerate kinase	I	
1352	Low	6763.91	6763.95	9.80	G3H767_CRIGR	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 1	P	Function inferred from NDUA2_HUMAN
1490	Low	6998.25	6998.21	9.65	G3I1H8_CRIGR	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 1	P	Function inferred from NDUBA_HUMAN
1496	Low	7008.53	7009.15	4.88	G3IMT0_CRIGR	Thioredoxin, mitochondrial	P	
1867	Low	7658.78	7658.87	7.72	G3HCJ4_CRIGR	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 1	P	Function inferred from NDUA2_HUMAN
2669	Low	9163.01	9163.25	6.69	G3IJW5_CRIGR	Nucleoporin SEH1	P	
2911	Low	9643.38	9644.23	9.50	G3H0I8_CRIGR	Alpha-enolase	P	Function inferred from ENOA_MOUSE
3160	Medium	10150.45	10150.96	8.86	G3HY36_CRIGR	Glyceraldehyde-3-phosphate dehydrogenase	E	
3623	Low	11127.85	11128.93	6.02	G3HHJ9_CRIGR	ATP synthase lipid-binding protein, mitochondrial	I	Function inferred from AT5G3_HUMAN
3725	Medium	11349.22	11348.85	5.55	G3HN28_CRIGR	Glyceraldehyde-3-phosphate dehydrogenase	P	
3809	Medium	11533.15	11533.17	8.98	G3GUF3_CRIGR	GMP reductase 1	P	
3949	Low	11845.00	11842.51	8.79	G3IF45_CRIGR	Carbonyl reductase [NADPH] 1	P	
4079	Medium	12134.39	12134.91	4.73	G3IM81_CRIGR	Choline/ethanolamine kinase	P	
5051	Low	14425.31	14425.43	6.17	Q6PW16_CRIGR	ATPase 3	E	
5140	Low	14644.98	14645.55	5.59	G3HMX1_CRIGR	Glyceraldehyde-3-phosphate dehydrogenase	I	
5163	Low	14702.02	14703.23	10.04	G3H1V3_CRIGR	ATP synthase lipid-binding protein, mitochondrial	I	Function inferred from AT5G2_HUMAN
5175	Low	14731.82	14730.96	5.89	G3IKJ4_CRIGR	GTP-binding nuclear protein Ran	P	
5175	Low	14731.82	14730.96	5.89	G3IKJ4_CRIGR	GTP-binding nuclear protein Ran	P	Function inferred from RAN_MOUSE
5977	Low	16791.98	16792.23	5.11	G3IJY1_CRIGR	Aldose reductase	I	
6605	Low	18599.27	18498.06	10.04	C560_CRIGR	Succinate dehydrogenase cytochrome b560 subunit, mitochondrial	E	

Table 7.6 – (Continued): SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the High Producer Chinese Hamster Ovaries Cell Line Samples

Variable count of loadings plot[a]	Intensity[b]	Exp'tal MW (Da)[c]	Sequence MW (Da)[d]	PI[e]	Database Accesion number[e]	Name[e]	Protein Existence[e]	Remarks[e]
Cell Growth/Death Proteins								
2915	Low	9651.42	9652.09	5.50	G3HVF3_CRIGR	Prohibitin	I	Function inferred from PHB_HUMAN
3206	Low	10245.55	10244.89	8.11	G3I308_CRIGR	26S proteasome non-ATPase regulatory subunit 2	P	Function inferred from PSD10_MOUSE
3277	Low	10393.20	10392.95	9.33	G3IKC4_CRIGR	Voltage-dependent anion-selective channel protein 1	P	Function inferred from VDAC1_MOUSE
3675	Low	11240.43	11240.92	6.56	G3HUU6_CRIGR	Protein S100-A11	P	Function inferred from S10A6_MOUSE
5137	Low	14637.55	14636.72	6.18	G3HUD3_CRIGR	Proteasome activator complex subunit 1	P	Function inferred from PSME1_HUMAN
5151	Low	14672.25	14671.61	5.49	LEG1_CRIGR	Galectin-1	E	Meth loss
5982	Low	16805.25	16804.80	8.26	G3GUG8_CRIGR	E3 ubiquitin-protein ligase RNF144B	P	Function inferred from TPM3_HUMAN
5996	Low	16842.43	16842.07	5.69	G3ILI7_CRIGR	26S proteasome non-ATPase regulatory subunit 3	P	Function inferred from PSMG4_HUMAN
6604	Low	18496.49	18497.33	4.62	G3GVS0_CRIGR	Nucleosome assembly protein 1-like 1	I	
7912	Low	22317.46	22318.79	8.36	G3HVP2_CRIGR	Stathmin	I	Function inferred from STMN1_HUMAN
Other proteins								
4292	Low	12619.48	12619.54	8.48	G3HVI1_CRIGR	Ubiquitin-conjugating enzyme E2 D2	I	Function inferred from UB2D2_MOUSE
5433	Low	15379.86	15379.68	4.94	G3HJ48_CRIGR	Rap guanine nucleotide exchange factor 5	P	Function inferred from RPGF5_MOUSE
5648	Low	15930.55	15931.27	4.62	G3I482_CRIGR	ADP-ribosylation factor-related protein1	P	Function inferred from ARFRP_MOUSE
5730	Low	16143.13	16142.85	6.45	Q8VHL2_CRIGR	Transient receptor potential-like protein	E	
7870	Low	22189.20	22189.44	5.07	TMED2_CRIGR	Transmembrane emp24 domain-containing protein 2	E	

Table 7.7: SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the Low Producer Chinese Hamster Ovaries Cell Line Samples

Variable count of loadings plot[a]	Intensity[b]	Exp'tal MW (Da)[c]	Sequence MW (Da)[d]	PI[e]	Database Accesion number[e]	Name[e]	Protein Existence[e]	Remarks[e]
Antibody and Protein Folding Proteins								
3123	Low	10074.28	10074.39	5.42	G3HIB2_CRIGR	Prefoldin subunit 2	P	Aids protein folding
Protein Biosynthesis Proteins								
3038	Low	9900.38	9900.14	7.09	G3IMY3_CRIGR	40S ribosomal protein S2	P	Translational elongation
1283	Low	6648.23	6647.86	12.15	RS30_CRIGR	40S ribosomal protein S30	I	Translational elongation
902	Low	10250.44	10250.77	6.90	G3HHT8_CRIGR	40S ribosomal protein S5	P	Translation/ required for rRNA maturation
902	Low	10250.44	10250.77	6.90	G3HHT8_CRIGR	40S ribosomal protein S5	P	Translational elongation
1092	Low	11234.29	11243.66	4.95	G3GZV3_CRIGR	60S acidic ribosomal protein P2	P	Translational elongation
5165	High	14704.98	14696.99	9.59	G3HV18_CRIGR	60S ribosomal protein L12	I	Translational elongation
5230	Low	14866.78	14865.44	10.51	G3H5W4_CRIGR	60S ribosomal protein L23	I	Translational elongation
5170	Low	14726.08	14717.39	9.88	G3I078_CRIGR	60S ribosomal protein L30	P	Translation
396	Low	7851.03	7859.99	10.54	G3GX65_CRIGR	60S ribosomal protein L37	P	Translational elongation
3222	Medium	10277.15	10275.25	10.44	G3H3Z5_CRIGR	60S ribosomal protein L37a	P	Translation
3110	Low	10046.03	10036.92	10.00	G3IFR0_CRIGR	60S ribosomal protein L7	P	Translation
1092	Low	11250.18	11243.09	8.45	G3I8P2_CRIGR	Eukaryotic translation initiation factor 1	P	Translational initiation
1781	Low	7504.19	7506.30	4.13	G3GVF4_CRIGR	Eukaryotic translation initiation factor 3 subunit B	P	Translational elongation
Cytoskeleton/Structural Related Proteins								
3442	Medium	10738.79	10732.77	10.89	G3H9H8_CRIGR	Catenin alpha-3	P	
1092	Low	11234.29	11240.92	6.56	G3HUU6_CRIGR	Protein S100-A11	P	
3110	Low	10046.03	10050.62	5.30	G3HC31_CRIGR	Protein S100-A6	P	
3851	Low	11623.96	11626.57	5.39	G3H328_CRIGR	Protein S100-Z	P	
1589	Low	7167.57	7160.51	9.51	G3H3L9_CRIGR	Tubulin alpha chain	P	

Table 7.7: SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the High Producer Chinese Hamster Ovaries Cell Line Samples

Table 7.7 – (Continued): SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the Low Producer Chinese Hamster Ovaries Cell Line Samples

Variable count of loadings plot[a]	Intensity[b]	Exp'tal MW (Da)[c]	Sequence MW (Da)[d]	PI[e]	Database Accesion number[e]	Name[e]	Protein Existence[e]	Remarks[e]
DNA and RNA Metabolism Proteins								
893	Low	10205.27	10212.93	8.96	G3HV32_CRIGR	Histone H3.3 type 1	I	
905	Low	10265.92	10263.02	6.27	G3HKX5_CRIGR	Centromere protein T	P	
1589	Medium	7167.57	7161.27	9.43	G3I3E3_CRIGR	Ubiquitin-conjugating enzyme E2 variant 2	P	
3113	Low	10053.74	10058.58	5.81	G3I2F5_CRIGR	Barrier-to-autointegration factor	P	
3222	Medium	10277.15	10272.74	6.27	G3H2J4_CRIGR	Transcription elongation factor B polypeptide 2	P	
5170	Low	14717.39	14719.03	8.72	G3IMU7_CRIGR	Mediator of RNA polymerase II transcription subunit 20	P	
5230	Low	14866.78	14870.59	10.02	G3HDS1_CRIGR	Histone H4	I	
6354	Low	17804.70	17800.67	9.33	G3IFA1_CRIGR	Heterogeneous nuclear ribonucleoprotein A1	P	
General Metabolism/Glycolysis Proteins								
187	Low	6953.20	6954.64	5.49	OFUT1_CRIGR	GDP-fucose protein O-fucosyltransferase 1	E	
902	Low	10250.44	10256.55	5.26	G3GUT3_CRIGR	Ras GTPase-activating protein-binding protein 1	P	
1466	Low	6955.98	6954.94	11.77	G3H586_CRIGR	Ras GTPase-activating protein-binding protein 1	P	
1697	Medium	7355.97	7363.49	7.71	G3I0W0_CRIGR	Alpha-enolase	P	
2967	Low	9754.76	9758.35	6.07	G3H5R7_CRIGR	S-methyl-5'-thioadenosine phosphorylase	P	
2967	Low	9754.76	9758.35	6.07	G3H5R7_CRIGR	S-methyl-5'-thioadenosine phosphorylase	P	
11749	Low	35590.87	35512.52	8.61	Q9Z2J2_CRIGR	Apurinic/apyrimidinic endonuclease	P	
11749	Low	35590.87	35587.23	8.17	G3I0I7_CRIGR	S-methyl-5'-thioadenosine phosphorylase	P	
9403	Low	27107.20	27103.70	7.12	G3ID79_CRIGR	Superoxide dismutase [Cu-Zn]	I	

Table 7.7 – (Continued): SwissProt/TrEMBL Database Proteins Matching Experimental Biomarker Masses in MALDI-ToF MS for the Low Producer Chinese Hamster Ovaries Cell Line Samples

Variable count of loadings plot[a]	Intensity[b]	Exp'tal MW (Da)[c]	Sequence MW (Da)[d]	PI[e]	Database Accesion number[e]	Name[e]	Protein Existence[e]	Remarks[e]
Cell Growth/Death Proteins								
278	Medium	7337.50	7346.29	9.35	EGR2_CRIGR	Early growth response protein 2	P	
711	Low	9307.20	9302.32	3.94	DPH3_CRIGR	DPH3 homolog	I	
3205	Low	10243.47	10244.89	8.11	G3I308_CRIGR	26S proteasome non-ATPase regulatory subunit 2	P	
3851	Low	11623.96	10618.86	10.31	HMGA1_CRIGR	High mobility group protein HMG-I/HMG-Y	I	
5230	Low	14866.78	14862.37	6.31	G3IB15_CRIGR	Nucleolar phosphoprotein p130	P	
5236	Low	14881.76	14871.25	8.65	G3H3B3_CRIGR	Annexin A10	I	
5993	Low	16832.26	16842.07	5.69	G3ILI7_CRIGR	26S proteasome non-ATPase regulatory subunit 3	P	
10007	Low	29181.17	29177.26	5.31	G3H303_CRIGR	Proteasome subunit beta type	I	
Other Proteins								
275	Low	7324.59	7318.48	9.90	G3H7D1_CRIGR	Guanine nucleotide-binding protein subunit gamma	I	
4823	Low	13868.24	13862.86	8.83	G3HM25_CRIGR	V-type proton ATPase subunit G 3	P	
9403	Low	27107.20	27109.33	7.48	G3I5Q8_CRIGR	E3 ubiquitin-protein ligase	P	

7.5.5. Differential Protein Expression on High and Low Producer CHO Cell Lines

Differentially expressed proteins were identified and were characterised with respect to upregulation as high, medium or low, based on the magnitude of the PLS-DA loadings which are related to experimental MWs (*m/z* ratio ion signals; MH^+). A total of 123 differentially expressed proteins were matched for high producer cell lines while 62 proteins were matched for low producer cell lines as indicated in Tables 7.6 and 7.7. The major functionalities of these proteins include protein folding, protein biosynthesis, cytoskeletal structure, DNA and RNA metabolism, glycolysis, and cell growth as discussed in the following sections.

7.5.6. Protein Folding/Processing Proteins

Ten proteins which function in protein folding or processing were identified in the high producer cell lines whilst just one was identified in the low producers. As can be seen from Table 7.6, half of the protein-processing proteins in high producers are peptidyl-prolyl cis-trans isomerases (PPIases). PPIases accelerate the folding of proteins. These proteins have been shown to catalyse the cis-trans isomerisation of proline imidic peptide bonds in oligopeptides. They are thought to be implicated in the folding, transport, and assembly of proteins during cellular protein synthesis (Göthel and Marahiel, 1999). Other proteins of interest include the humanised IgG recombinant antibody light chain, protein processing related proteins such as 60 kDa heat shock protein (HSP60), DnaK-type molecular chaperone (DnaK), FK506-binding protein 2 (FKBP2), Heat shock cognate protein HSP 90-β (HSCP90), and Heat shock protein β-1 (HSPB1). These protein processing related proteins were identified only in the high producer cell lines. The presence of these proteins only in high producers may suggest that they are upregulated in high producers.

HSP60 is part of a class of proteins known as molecular chaperones that are required to promote folding and assembly of both misfolded and newly synthesised proteins preventing their aggregation during translation, and under conditions of cellular stress (Itoh *et al.*, 2002). DnaK belongs to the chaperones of the Hsp70 family. The latter proteins have also been shown to be involved in a variety of cellular activities such as protein transport across membranes and anti-apoptosis (Mosser *et al.*, 2000; Garrido *et al.*, 2006). HSPB1 belongs to another class of molecular chaperones known as Small

Heat Shock Proteins (sHSP). Apart from their chaperonin function, they also have a protective effect on cell viability at elevated temperatures (Jakob *et al.*, 1993). For high producer cell lines (Table 7.6) *m/z* ratio signal ions representing these chaperon proteins all had medium intensity based on the magnitude of the loadings (except for FKBP2 which had a low intensity) indicating perhaps an upregulation for the high producer cell lines. An upregulation of chaperons in high producers suggests that the cells are overburdened due to high levels of translation thus more foldases are needed to clear up the accumulating proteins by accelerating their folding.

IgG recombinant antibody light chain (LC) showed differential expression with a medium intensity in high producers (Table 7.6) based on magnitude of the loadings whilst the low producers (Table 7.7) showed a low intensity. Moreover, a higher number of human IgG heavy and light chain domain proteins were matched in high producers than low producers (19 as opposed to 5) suggesting more production and secretion of IgG in high producers than in low producers . This is an important finding, as it corroborates proposed theories from previous SDS-PAGE studies which indicated a direct link between the amount of secreted antibody on cell surface and cellular productivity (Alete *et al.,* 2003; van Dyk *et al.,* 2003; Pascoe *et al.,* 2007). It may be that the high producers may be involved in higher antibody production, transport and secretion than their low producer counterparts, hence have high intracellular concentration, with eventually high amounts being secreted, of the recombinant protein. It can thus be concluded that this observation supports the high productivity trends associated with high producing cell lines during cell culture.

7.5.7. Protein Biosynthesis Proteins

Proteins involved with translation or protein biosynthesis were amongst the most abundant class of proteins matched. Several proteins, mainly 40S and 60S ribosomal subunits, and translation initiation/elongation factors, were identified both in high producers (25 proteins) (Table 7.6) and low producers (13 proteins) (Table 7.7). The abundance of protein biosynthetic proteins in high producers, notably with proteins such as 40S ribosomal protein S15, 40S ribosomal protein S2, 60S acidic ribosomal protein P1, 60S ribosomal protein S7, and Protein S100-A9 showing medium intensity in the loadings plot implies they were upregulated in high producers. This suggests that high producers perhaps have higher rate of protein synthesis than low producers during culture. As expected, the important roles played by these proteins during protein synthesis supports the high productivity trend observed in high producers.

7.5.8. Cytoskeleton Proteins

A number of structural and growth-associated proteins were among the proteins matched in the database. Five proteins each were identified in both the high (Table 7.6) and low producer (Table 7.7) cell lines. All the cytoskeleton proteins included Annexin A7, β-actin, Calmodulin, Calponin-1, Catenin α-2, Coactosin-like protein, Tropomyosin α-1 chain (for high producers only), and Catenin α-3, Calmodulin, Protein S100-A11, Protein S100-A6, Protein S100-Z, Tubulin α-chain (for low producers only). These proteins function in stabilising the cell cytoskeletal. For example β-Actin is an actin filament component at the golgi complex in mammalian cells. It is part of the actin cytoskeleton plays an essential role both in endocytic and secretory pathways. Tropomyosin is an actin-binding protein that associates with actin filament to regulate its stability (Egea *et al.,* 2006).

Proteomic studies with mouse myeloma (NS0) cell lines showed a direct link between increased productivity a general increase in cellular cytoskeletal framework (Smales *et al.*, 2004; Dinnis *et al.*, 2006). Experimental evidence has suggested a functional interaction between cellular cytoskeletal apparatus, where a disruption of actin filaments led to a profound negative effect in translation (Stapulionis *et al.*, 1997). The identification of cytoskeleton proteins in both high and low producers may suggest a normal cellular physiological phenomenon where the global protein synthesis network is supported and stabilised by the actin cytoskeletal framework. Hence it is possible that the upregulation of cytoskeleton proteins in high producers may be advantageous in providing these cell lines with a more stable and efficient protein synthesis framework, hence a guarantee of high product yield.

7.5.9. Metabolism Proteins

A large number of metabolism proteins were also matched, amongst which are a number involved in glycolysis while others are associated with nucleic acid metabolism. The high producer cell lines matched 20 glycolysis proteins (Table 7.6) whilst the low producer cell lines matched just 9 (Table 7.7). Glycolysis proteins were mainly enzymes with the matched ones including aldose reductase, α-enolase, adenosine triphosphate (ATP) synthase, glyceraldehyde-3-phospate dehydrogenase (GAPDH), carbonyl reductase, NADH dehydrogenase, phosphoglycerate kinase (PK), succinate dehydrogenase (SDH), and superoxide dismutase (SOD). The α-enolase protein, matched in both high and low producers, is a key enzyme in the glycolytic pathway and hence it is ubiquitously present in abundance in the biological world. It is multifunctional, serving as a cell surface plasminogen receptor for a variety of hematopoetic, epithelial and endothelial cells; heat-shock protein property; cytoskeletal and chromatin structure binding properties suggests that α-enolase may play a crucial role in transcription and a variety of physiological processes in the cell (Pancholi 2001). With its multifunctional role in cellular physiological processes, the presence of α-enolase in both high and low producers was expected.

Another important metabolic protein matched in high and low producer cell lines is the enzyme SOD. In aerobic respiration, many oxidative metabolic processes (e.g cellular respiration, xanthine oxidase, NADPH oxidase, lipoxygenase) produce reactive oxygen species (ROS) (superoxide radicals, hydrogen peroxide, and singlet oxygen) as by products which are deleterious to the cell. The phenomenon of oxidative stress arises as a result of hydroxyl radicals (produced from ROS in the presence of metal ions) interacting with cellular macromolecules to cause lipid peroxidation, protein denaturation, DNA mutation, and eventual cell death. SOD plays a crucial role in the defence mechanism against oxidative stress in the cell by reacting with superoxide radicals to produce harmless hydrogen peroxide (Bowler *et al.*, 1992). The presence of α-enolase in both high and low producers was expected as the cell lines are all involved in aerobic respiration.

Important metabolic proteins matched only in high producers include ATP synthase (synthesizes ATP from adenosine diphosphate (ADP) and inorganic phosphate) (Nakamoto *et al.,* 2008); carbonyl reductase (cellular protective role by reduction of xenobiotic carbonyls and quinones) (Oppermann, 2007); GAPDH (breakdown of glucose into energy during glycolysis) (Sirover, 1997); NADH dehydrogenase (catalyses electron transfer from NADH to coenzyme Q in the electron transport chain) (Weiss *et al.*, 1991); PK (catalyses the high-energy phosphoryl transfer of the acyl phosphate of 1,3-bisphosphoglycerate to ADP to produce ATP) (Blake and Rice, 1981);

thioredoxin (antioxidant enzyme - major cellular protein disulfide reductases - with growth factor properties responsible for maintaining intracellular proteins in their reduced state); glutaredoxin (catalyse glutathione-disulfide oxidoreductions overlapping the functions of thioredoxins and using electrons from NADPH via glutathione reductase) (Arnér and Holmgren, 2000); SDH (couples the Krebs cycle to the electron transport chain by the oxidation of succinate and the reduction of ubiquinone respectively) (Oyedotun and Lemire, 2004). This trend of differential expression may be related to metabolic differences in the cell lines which may in turn influence cellular productivity. The exclusive identification of these proteins for high producer cell lines suggests that these cell lines were metabolically more active than low producers.

7.5.10. Nucleic Acid Metabolism Proteins

Several proteins involved in DNA and RNA metabolism were also matched. Approximately half the number of proteins matched for high producers (18 proteins) was matched for the low producers (8 proteins). Proteins matched only in high producers included transcription elongation factor 1-like, ATP-dependent RNA helicase DHX8, DNA polymerase subunit gamma-2, DNA repair protein complementing XP-A, DNA repair protein complementing XP-A cells homolog, guanine nucleotide-binding protein subunit gamma, non-histone chromosomal protein HMG-14, and splicing factor 3B subunit 5. In addition, high producers matched more histone proteins than low producers.

The matching of more histones in high producers may represent upregulation of these proteins. This result differs from previous studies that found a downregulation of histones in high-producing CHO cultures (Nissom *et al.,* 2006; Carlage *et al.,* 2009). Histones condense DNA into chromatin structures reducing the accessibility of DNA for transcription, hence downregulation of histones makes biological sense as transcription is enhanced from chromatin templates of high producers. However, though results here unexpectedly suggest an upregulation of histones in high producers, the matching of transcription, splicing factors and enzymes (DNA polymerase subunit gamma-2, splicing factor 3B subunit 5, and transcription elongation factor 1-like) exclusively in high producers was comforting and makes biological sense as this suggests more active transcription and more efficient mRNA processing in high producers.

7.5.11. Cell Cycle Proteins

Several proteins that are involved in cell cycle regulation were differentially expressed in this study. Fourteen cell cycle proteins were matched for high producers (Table 7.6) whilst just 9 were matched for low producers (Table 7.7). Important cell cycle proteins that were matched only in high producers include Galectin-1, Protein S100-A11 (Calgizzarin), Stathmin, Bcl-2-like protein 10, E3 ubiquitin-protein ligase RNF144B, and Nucleosome assembly protein 1-like 1.

Calgizzarin is a member of a family of calcium-modulated proteins that is involved in a variety of cellular processes such as proliferation and differentiation. Studies have implicated this protein as a cellular growth inhibitor on the basis of its down-regulation in immortalized compared to fibroblast cells (Donato, 2001). Another protein of interest is Galectin-1. Galectins are a group of sugar-binding proteins specific for their carbohydrate moieties. They modulate a wide range of cellular activities such as tumor progression, cell differentiation, cell growth, and apoptosis (Yang and Liu, 2003). Galectin-1 has specifically been shown to have growth modulation properties - having both negative and positive effects on cell proliferation.

The identification of Galectin-1 and Calgizzarin only in high producers may suggest an upregulation of these proteins in high producers, hence a negative growth modulation. These results agree well with previous shortgun proteomic and quantitative proteomic profiling of high-producing CHO cell lines (Nissom *et al.*, 2006; Carlage *et al.*, 2009). This is not surprising with respect to the growth kinetics of the high producers compared to the low producers as low producers will tend to grow faster using up much needed energy needed for translation. From a productivity perspective, it is hypothesised that the presence of growth inhibitory proteins such as Calgizzarin and Galectin-1 in high producer cell lines will enhance product yield as the high producers will grow slower and commit all their energy into protein production.

7.6. Summary

In conclusion, MALDI-ToF MS was used for intact cell profiling of IgG-producing CHO cell lines during biopharmaceutical bioprocessing. The spectral data generated were preprocessed to reduce experimental variabilities that might otherwise have masked biological trends in the data. PLS-DA was used to model the spectra data and distinguish between cell lines with respect to productivity (high/low producers; Hs/Ls).

The theory was that variability exists within cell lines, based on their productivity (titre of IgG produced) and that PLS-DA can be used to help understand this behaviour. Specific *m/z* ratio regions were identified (with large absolute loadings) and their ability to act as discriminatory molecules between Hs and Ls were investigated, with the aim of identifying differentially expressed protein biomarkers associated with cell line productivity after a protein database searches.

A total of 185 (123 Hs and 62 Ls) differentially expressed proteins were matched and identified after SwissProt/TrEMBL protein database search. The identified proteins revealed that more proteins involved in biological processes such as protein biosynthesis, protein folding, glycolysis and cytoskeleton architecture were upregulated in Hs. These findings are consistent with finding in the literature. A subset of these protein biomarkers such as molecular chaperons (heat shock protein families), α-enolase, and superoxide dismutase, and translation initiation/elongation factors have already been identified from mammalian cell lines in a series of publications; and found to be correlated with specific antibody productivity. These results provide important insights into the overall cellular protein biology, and the gene of these protein biomarkers may represent valuable genetic engineering targets aimed at improving cell line productivity. It is clear that there is a relationship between the upregulation of some proteins in high producers and increases in protein productivity. The genes of such proteins can be targeted and genetically engineered to produce enhanced cell lines.

This study demonstrates that PLS-DA if combined with linear mode MALDI-ToF MS can be a valuable tool for biomarker discovery in the biopharmaceutical bioprocessing industry. More specifically to CHO cell lines in culture, this study provides a foundation for rapid biomarker profiling of CHO cell lines with the completion of the sequencing of the CHO genome. Although the aforementioned points clearly demonstrate that this approach has a good potential in the area of mammalian cell culture during bioprocessing, further confirmatory studies are needed before the full potential of this approach can be realised. This is discussed in the future work section (section 8.2).

Chapter 8

8. Conclusions and Future Work

The findings of this thesis are summarised in this chapter. A summary of the chapters is presented and additional approaches beyond the scope of this work, that could provide avenues of investigation for future work, are suggested. The key contributions of this thesis are as follows:

- This thesis has investigated the potential to utilise the approach of ‘intact-cell’ MALDI-ToF MS (ICM-MS) combined with PLS-DA to distinguish between IgG monoclonal antibody-producing CHO mammalian cell lines based on their productivities and identify protein biomarkers through protein database searches that are differentially expressed in these cell lines. Although a number of studies have been carried out recently to demonstrate how biomarker profiling by ICM-MS combined with PLS and PLS-DA could be exploited to screen cultured mammalian cell lines in bioprocessing, no attempt was made to assign the mass spectra ion signals to potential protein biomarkers (Feng *et al.,* 2010; Feng *et al*., 2011). Together with the appropriate use of internet accessible protein data base searches, ICM-MS combined with PLS-DA has been shown to be effective for classifying CHO cell lines based on their productivities and identify protein biomarkers associated with the cell line productivities.

- A proof-of-concept study applied to *E. coli* K-12 cells at different growth phases utilising the same methodology identified potential protein biomarkers associated with to the different growth phases of the cells.

- Preprocessing is data dependent so preprocessing studies have been carried out for each data set used in this work. A number of combinations of data preprocessing techniques/parameters have been considered and empirically investigated to enable the most appropriate selection of the combination of methods/parameters. The parameters of the preprocessing methods were modified systematically and applied to the spectra data which was subsequently used to calibrate PLS-DA models. The combination of preprocessing techniques/parameters that gave an improved and optimum model performance was selected as the appropriate preprocessing method. The successful identification of protein biomarkers from the spectra data suggested that the preprocessing methods used were valid, as it eliminated differences between spectra profiles as a consequence of experimental and instrumental procedures, whilst preserving the inherent biological information within the spectra profiles.

8.1. Thesis Summary

In chapter 1 an overview of the project was presented, including a summary of the main aims and objectives of the thesis. Chapter 2 proceeded to provide an overview of biopharmaceutical therapeutic proteins, the science and technology of *in vitro* mammalian cell culture which are used in their production as well as a short review on the methods that are currently being used for improving large-scale production of heterologous proteins from mammalian cell lines.

MALDI-ToF MS was presented in chapter 3 along with a description of how the instrument was used to generate data from *E. coli* cells at different growth phases as well as from IgG monoclonal antibody producing CHO cell lines during culturing. Aspects relating to mass spectrometry data preprocessing was also discussed and the techniques were applied to the two mass spectra data sets. Finally, applications of mass spectra data preprocessing reported in the literature were discussed.

In chapter 4, an introduction to proteomics and biomarker discovery relevant to this thesis was given. A summary of the approaches used in biomarker discovery highlighting the top-down proteomics based approach of ICM-MS, and its applications to microorganisms and mammalian cell line analysis for biomarker discovery in the biopharmaceutical bioprocessing industry was presented. The chapter also explored the importance of bioinformatics and internet accessible protein databases for biomarker identification in top-down proteomics. A literature review on applications of proteomic biomarker discovery was also presented.

Chapter 5 focused on the multivariate data analysis techniques of principal component analysis (PCA), partial least squares (PLS) and PLS – discriminant analysis (PLS-DA). It explored the PLS-DA algorithm involved in modeling MALDI-ToF mass spectra data collected from *E. coli* cells at different growth phases as well as IgG monoclonal antibody producing CHO cell lines with the aim of identifying protein biomarkers. The chapter also presented results of an example where PCA and PLS-DA were used to analyse the mass spectra data generated from cell lysate samples of *E. coli* K-12 cells at different growth phases.

After discussing the advantages of utilising ICM-MS, PLS-DA, and protein databases in various areas of proteomics in chapters 4 and 5, chapter 6 served to provide a proof-of-principle study of ICM-MS, PLS-DA and a database search to identify protein biomarkers associated with the growth phases of the *E. coli.* Firstly, PLS-DA was applied to the MALDI-ToF MS data to determine if such an approach could be used to distinguish between the cells at different growth phases. The application of PLS-DA resulted in the successful classification of the samples according to the growth phase of the cells. A further outcome of the analysis was that it was possible to identify the mass-to-charge (*m/z*) ratio ion signals that contributed to the classification of the samples. The Swiss-Prot/TrEMBL database and primary literature was then used to assign a number of these *m/z* ion signals to proteins and these assignments revealed that the major contributors from the exponential phase were ribosomal proteins. Additional assignments were possible for the stationary phase and the decline phase cells where the proteins identified were consistent with observed biological interpretation. In summary, the results suggested that MALDI-ToF and PLS-DA can be used in combination to discriminate between *E. coli* cells in different growth phases and thus could potentially be used as a tool in process development in the bioprocessing industry to enhance cell growth and cell engineering strategies.

After the proof-of-concept study successfully demonstrated in chapter 6, in chapter 7 the approach was applied to the mass spectra data of IgG monoclonal antibody-producing CHO cell lines. The cell lines were classified according to their productivities into high and low producer cell lines. The *m/z* ratio ion signals that contributed to the classification of the cell lines were subjected to Swiss-Prot/TrEMBL database search and primary literature. These searches revealed a number of *m/z* ratio ion signals that could be assigned to proteins. The identified proteins classified revealed that more proteins were in the high than in the low producer cell lines. These proteins are involved in biological processes such as protein biosynthesis, protein folding, glycolysis and cytoskeleton architecture. The upregulation of these proteins in high producer cell lines were findings that are consistent with those reported in the literature. The ability to identify proteins that correlate to cell line productivity may be important in predicting the likelihood of a cell line being a high or low producer, provide insight into the biology of mammalian cell lines during biopharmaceutical bioprocessing, and may give indications of potential genetic engineering targets that may be exploited to engineer better cell lines. Chapter 7 provided the main contribution since no previous studies have adapted this approach to the best of the author's knowledge.

Appendix A and B presents additional information relating to the generation and analysis of the *E. coli* K-12 'intact' cell and cell lysate mass spectra data sets respectively. Appendix A provides a summary of the chemicals, reagents and laboratory instruments used; the modeled population growth curve of *E. coli* K-12, ATCC 15223 to show important growth parameters; the results of average wet and dry cell weight measurements for the *E. coli* cells as well as the calibration curves to determine these parameters; the results of standard plate count *E. coli* cells; approximate amount of bacterial cell pellets analysed using MALDI MS; the design of experiment results of preprocessing techniques applied to all *E. coli* MALDI mass spectra data sets; the preprocessing algorithms; the variables from the loadings plot that are associated with *m/z* ratio ion signals; and the MALDI-ToF instrumental parameters used in this work. Appendix B presents additional information relating to the multivariate data analysis of the *E. coli* K-12 cell lysate spectra data sets. Appendix C provides a summary of the PLS algorithm used in this thesis as well as information of the variables from the loadings plot that are associated with *m/z* ratio ion signals for the high and low producer CHO cell lines.

8.2. Recommendations for Future Work

The approach using ICM-MS, PLS-DA, and protein database searches has been shown to offer much promise as an effective tool in protein biomarker profiling and potentially useful as a tool in process development in the bioprocessing industry to enhance cell growth and cell engineering strategies. There are however issues that require further investigation before the approach can be considered ready to be used in real world situations.

The sequencing of the CHO genome and compilation of sequence-derived theoretical MWs of CHO proteins in internet accessible protein databases has created the possibility of identifying potential protein biomarkers of CHO cells by matching the sequence-derived theoretical MWs of the database proteins with experimental MWs derived from MALDI-TOF mass spectrometry. Assigning *m/z* ratio peaks or ion signals to protein biomarkers in internet accessible databases is facilitated by the fact that most of the *m/z* ratio peaks from MALDI generated mass spectra data represent singly charged protein ions, whose molecular weight can be directly inferred as that of the protein molecule.

8.2.1. Further Studies for Protein Identification

Nevertheless, the provision of molecular weight of the proteins by MALDI-MS is not enough information to identify a protein. Therefore all protein assignments in this project were tentative for a number of reasons including the fact that there may be more than one potential match in the protein database corresponding to a molecular weight. For example in the *E. coli* spectra data sets, the *m/z* ratio peak of 7332 could be assigned to the ribosomal protein RL29 and the cold shock protein-E (CPSE) as both have MWs of 7273 Da. Therefore, it is not possible to provide a definitive conclusion about the assignment of the peaks. Moreover, adduct ions, mass errors and post-translational modifications involved with some proteins may impede the assignment process. Matrix effects during the MALDI experiment may lead to ion suppression where protein ions are masked by matrix ones leading to improper assignments.

For a proper assignment and thus a positive identification of an individual protein more studies would be required. These usually involve tryptic digestion of the samples which are then subjected to tandem mass spectrometry analysis (such as two-dimensional polyacrylamide gel electrophoresis with matrix assisted laser desorption ionisation time of flight MS; 2D-PAGE/MALDI-TOF-MS or liquid chromatography with electrospray ionisation MS (nano-LC-ESI-MS/MS)), providing sequence-specific fragments of the individual proteins.

8.2.2. Integration of ‘Omics’ Data for Biomarker Discovery

This research involves a proteomics technology (MALDI-ToF) that has the potential of identifying or discovering biomarkers. However, to enhance the contextualisation of the proteomics results, it can be integrated with of data from other platforms such as metabolomics or transcriptomics. Thus having the possibility of working with transcriptomics data from the same system (*E. coli* or CHO cell lines) would be more interesting as there would be the linkage of known proteins with their identified genes or metabolites to relevant biochemical pathways. Correlating the changes in protein abundance from proteomics and gene expression from transcriptomics with changes in the cell function will permit the exploration of a broad range of biological processes. This will help provide insight into the biology of mammalian cell lines during biopharmaceutical bioprocessing.

8.2.3. Future Studies using an alternative PLS-DA Algorithm

Whilst the PLS-DA algorithm improves the arbitrary selection of classification which now depends on the distribution of sample predictions than other methods which sets threshold at an arbitrary value, that is, 0.5, the algorithm assumes that each sample will have a predicted ***y***-value greater/less than the threshold making it to fall to one of the class or the other. However this may not be the case as we may have a ***y***-value equal to the threshold hence no class and misclassification will be inevitable where sample will be included in one of the classes. Consequently, the algorithm needs a ‘Rejection Rule’ where such a sample is rejected altogether (Such an algorithm has been proposed by Botella *et al*. (2009). It is recommended that future study should involve using an algorithm which provides ‘Rejection Rule’ option in order to avoid the risk of misclassification and improve the quality of the models.

8.2.4. Predictive Modelling for Process Development

There is another dimension to this approach not considered in this project that could have a direct contribution to process development in the bioprocessing industry. The MALDI-ToF mass spectra data generated from IgG monoclonal antibody-producing CHO cell lines can be used to train predictive models of PLS-DA. These models can then be used to predict the likelihood of other CHO cell lines being high or low producers. Predicting productivity earlier on during biotechnological process development may lead to early screening of high producing cell lines which will have the desired high productivity during manufacturing (bioreactor stage).

Appendices

Appendix A Additional information for the ‘intact’ cell *E. coli* K-12 project

Appendix A provides a summary of the chemicals, reagents and laboratory instruments used; the modeled population growth curve of *E. coli* K-12, ATCC 15223 to show important growth parameters; the results of average wet and dry cell weight measurements for the *E. coli* cells as well as the calibration curves to determine these parameters; the results of standard plate count *E. coli* cells; approximate amount of bacterial cell pellets analysed using MALDI MS; the design of experiment results of preprocessing techniques applied to all *E. coli* MALDI mass spectra data sets; the preprocessing algorithms; the variables from the loadings plot that are associated with *m/z* ratio ion signals; and the MALDI-ToF instrumental parameters used in this work.

Figures

A.1 Preprocessing algorithm used for the *E. coli* mass spectra profiles

For the optimal combination of preprocessing techniques and associated parameters, the baseline correction quantile value was at 0.1, baseline correction window size was set to 200, normalisation quantile value, 0.1 or 0.2, and smoothing span value was at 25

```
files=dir('*dat');
for i = 1:366;
files(i).data = load(files(i).name);
MZ = files(i).data(:,1);
Y(:,i)=[files(i).data(:,2)];
Ynew=Y(4300:end,:);
MZnew=MZ(4300:end,:);
[MZR, YR(:,i)]= msresample(MZnew,Ynew(:,i),7000,'range',[2350 max(MZnew)]);
YB(:,i)=msbackadj(MZR,YR(:,i),'WINDOWSIZE',200,'QUANTILE',0.1,'PRESERVEHEIGHTS',true);
P=[6411 6855 7273 7333 7869 9061 9218 9532 9736];
YA(:,i)=msalign(MZR,YB(:,i),P);
YN(:,i) = msnorm(MZR,YA(:,i),'Quantile',0.1,'MAX',100);
YS(:,i) = mssgolay(MZR,YN(:,i),'SPAN',25);
YStcal=[YS']; End
P=mspeaks(MZR,YS,'Denoising',false,'HeightFilter',10,'SHOWPLOT',false);
```

Figure A.1: Preprocessing algorithm used for the E. coli mass spectra

A.2 Preprocessing algorithm used for the CHO cell line mass spectra profiles

For the optimal combination of preprocessing techniques and associated parameters, the baseline correction quantile value was at 0.2, baseline correction window size was set to 100, normalisation quantile value, 0.1, and smoothing span value was at 25

```
files=dir('*txt');
for i = 1:80;
files(i).data = load(files(i).name);
MZ = files(i).data(:,1);
Y(:,i)=[files(i).data(:,2)];
Ynew=Y(6200:end,:);
MZnew=MZ(6200:end,:);
[MZR, YR(:,i)]= msresample(MZnew,Ynew(:,i),7000,'range',[6200
max(MZnew)]);
YB(:,i)=msbackadj(MZR,YR(:,i),'WINDOWSIZE',200,'QUANTILE',0.2,'PRESE
RVEHEIGHTS',true);
P=[6411 6855 7273 7333 7869 9061 9218 9532 9736];
YA(:,i)=msalign(MZR,YB(:,i),P);
YN(:,i) = msnorm(MZR,YA(:,i),'Quantile',0.1,'MAX',100);
YS(:,i) = mssgolay(MZR,YN(:,i),'SPAN',25);
YStcal=[YS']; End
P=mspeaks(MZR,YS,'Denoising',false,'HeightFilter',10,'SHOWPLOT',false);
```

Figure A.2: Preprocessing algorithm used for the CHO cell line mass spectra profiles

A.3 Cross-validation vector used for the intact cell *E. coli* data sets

The cross-validation vector was used for cross-validation of 300 ‘intact’ cells *E. coli* MALDI mass spectra calibration data sets. Leave-class-out-cross-validation was performed. The vector was coded as ‘1’ to indicate that the sample involved belonged to class 1; ‘2’ for class 2; and ‘3’ for class 3. All samples coded with ‘1’, belonging to the same class 1 were removed from the calibration data set and a sub-model based on the remaining samples was used to build the PLS-DA model and predict the left out samples. The process was repeated with all the 300 calibration data set of classes 2 and 3 had been left out once.

```
[1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  1
1  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  1  1  1
1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  1  1  1  1
1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  31  1  1  1  1  1  1
1  1  1  1  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  1  1  1  1  1  1  1  1
1  1  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3]
```

Figure A.3: Cross-validation vector used for the intact cell E. coli data sets

Tables

A.1 Results of MALDI-ToF instrument parameter optimisation

Table A.1 summarises the results of MALDI-ToF instrument parameter optimisation when the parameters were altered based on the design matrix. As can be seen from the table, the parameters that were changed were ion source voltage 1 (accelerating voltage), ion source voltage 2 (grid voltage), pulse ion extraction (PIE), and the laser beam focus. All other parameters were maintained at a fixed value. For good signals, the grid voltage was always set at approximately 88% of the accelerating voltage. The two values for the accelerating and grid voltage used were 24 and 21, and 24.24 and 21.23 respectively. The PIE was altered at intervals of about 50 units ranging from 150-500ns. Only two values (35 or 39%) of the laser beam were used. Results were based on the signal-to-noise (s/n) ratio of the peaks observed. This was rated from bad to excellent as follows. As can be seen from the table, results suggest that higher PIE values in the range 350-500ns gave very good signals observed only in taller spectra peaks. PIE value of 300ns gave very good signals observed only in smaller peaks while PIE values of 300 and 350ns gave excellent signals across all spectra peaks. Consequently, trial 10 was adopted as condition for MALDI analysis.

Trial number	**Instrument parameters**										**Result (s/n)**	**Remarks**
	Laser				**Spectrometer**							
	Laser beam attenuation	Laser beam focus (%)	Laser repetition rate (Hz)	Number of shots	Suppress up to (Da)	Positive voltage polarity	Pulse ion extraction (PIE) (ns)	Ion source voltage 1 (kV)	Ion source voltage 2 (kV)	Lens voltage (kV)		
1	20	35	20	20	2000	Positive	150	24	21	5.5	++	Poor signals
2	20	35	20	20	2000	Positive	300	24	21	5.5	++	
3	20	39	20	20	2000	Positive	250	24	21	5.5	++	
4	20	39	20	20	2000	Positive	350	24	21	5.5	+++	Good signals
5	20	39	20	20	2000	Positive	370	24	21	5.5	+++	
6	20	39	20	20	2000	Positive	400	24	24	5.5	+++	
7	20	35	20	20	2000	Positive	150	24.24	21.23	5.5	++++	Very good signals; observed for smaller peaks
8	20	35	20	20	2000	Positive	200	24.24	21.23	5.5	++++	
9	20	39	20	20	2000	Positive	250	24.24	21.23	5.5	++++	
10	**20**	**39**	**20**	**20**	**2000**	**Positive**	**300**	**24.24**	**21.23**	**5.5**	**+++++**	Excellent signals observed across all peaks
11	20	35	20	20	2000	Positive	350	24.24	21.23	5.5	+++++	
12	20	35	20	20	2000	Positive	370	24.24	21.23	5.5	++++	Very good signals; observed across taller peaks
13	20	35	20	20	2000	Positive	400	24.24	21.23	5.5	++++	
14	20	35	20	20	2000	Positive	450	24.24	21.23	5.5	++++	
15	20	35	20	20	2000	Positive	500	24.24	21.23	5.5	++++	

Table A.1: Results of MALDI-ToF instrument parameter optimisation

A.2 MALDI ground steel target plate layout showing the *E. coli* cell pellet sample arrangement

Table A.2 shows the MALDI ground steel target plate layout of the *E. coli* cell pellet sample arrangement for analysis. STD (standard) represented the calibrant. About 1µl of calibrant was spotted onto of the MALDI target plate onto six specific sample spots surrounded by 20 spots containing samples to be analysed. Each calibrant within each group of sample was used to calibrate the instrument before analysing all samples in the group. The samples were named as di (intact cell pellets for cultures at decline phase), mi (intact cell pellets for cultures at mid-log or exponential phase), si (intact cell pellets for cultures at stationary phase), dl (cell lysate for cultures at decline phase), ml (cell lysate for cultures at exponential phase), and sl (cell lysate for cultures at stationary phase). To explain the labeling, for example the 20 different MALDI spots of intact cells at the decline growth phase were labeled di01, di02, di03,...., di018, di19, di20; intact cells at the exponential growth phase mi01, mi02, mi03,...., mi018, mi19, mi20. All the other sample spots were labeled following this format.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	test	test	test	test								mi01	mi02	mi03	mi04									
B	test	test	test	test								mi05	mi06	mi07	mi08									
C	test	test	STD	test								mi09	mi10	STD	mi11									
D	test	test	test	test								mi12	mi13	mi14	mi15									
E	test	test	test	test								mi16	mi17	mi18	mi19									
F	test	test	test	test								mi20												
G																								
H	si01	si02	si03	si04		di01	di02	di03	di04		ml01	ml02	ml03	ml04		sl01	sl02	sl03	sl04		dl01	dl02	dl03	dl04
I	si05	si06	si07	si08		di05	di06	di07	di08		ml05	ml06	ml07	ml08		sl05	sl06	sl07	sl08		dl05	dl06	dl07	dl08
J	si09	si10	STD	si11		di09	di10	STD	di11		ml09	ml10	STD	ml11		sl09	sl10	STD	sl11		dl09	dl10	STD	dl11
K	si12	si13	si14	si15		di12	di13	di14	di15		ml12	ml13	ml14	ml15		sl12	sl13	sl14	sl15		dl12	dl13	dl14	dl15
L	si16	si17	si18	si19		di16	di17	di18	di19		ml16	ml17	ml18	ml19		sl16	sl17	sl18	sl19		dl16	dl17	dl18	dl19
M	si20					di20					ml20	tmi02	tmi02	tmi03		sl20	tsi01	tsi02	tsi03		dl20	tdi01	tdi02	tdi03
N																								
O																								
P																								
Q																								

Table A.2: MALDI ground steel target plate layout showing the E. coli cell pellet sample arrangement

A.3 Data used for graph of Figure 3.19 showing the effect of cropping and resampling on quality of PLS-DA models across LV2

Table A.3 presents the data used in plotting graph in Figure 3.19 which shows the effect of cropping and resampling on quality of PLS-DA models across latent variables 2 to 11. Signal resampling was carried out after cropping the data. The data sets were resampled by down sampling to 7000 data points from 2350 to 20400 *m/z* ratio values.

Number of latent variables	Average R^2 (%)			Average RMSECV (%)		
	Raw data	Data after cropping	Data after cropping and resampling	Raw data	Data after cropping	Data after cropping and resampling
2	2.9	32.5	32.6	39.3	39.4	39.3
3	34.8	48.2	48.2	34.8	34.7	34.7
4	55	64.3	64.6	29.2	28.7	28.8
5	62.3	66.9	67.1	28.5	28.4	27.9
6	60.3	68.5	68.6	27.7	27.6	27.5
7	64	70	70	27.2	27.4	27.4
8	65.7	70	70.1	27.2	27.2	27.2
9	67.4	69.9	69.9	26.9	27.7	27.7
10	68.8	70.6	70.5	26.9	27.6	27.7
11	70.5	70.1	70.2	27.6	27.9	27.9

Table A.3: Data used for graph of Figure 3.19

A.4 Data used for graph of Figure 3.21 showing the effect of alignment across a number of latent variables

Table A.4 shows the effect of alignment across a number of latent variables. Alignment of the spectra data was carried out along five *m/z* ratio peaks 6411, 6855, 7273, 7333, 7869, 9061, 9218, 9532, 9736 which were found to be common among most of the spectra profiles. The results suggest that applying alignment on the data slightly improved the qualities of the PLS-DA models built using the mass spectra data, on average.

Number of latent variables	Average R^2 (%)		Average RMSECV (%)	
	Raw data	Data after alignment	Raw data	Data after alignment
2	2.9	35.7	39.3	38.2
3	34.8	54.2	34.8	32.6
4	55.0	65.6	29.2	28.5
5	62.3	67.5	28.5	28.0
6	60.3	69.3	27.7	27.4
7	64.0	71.3	27.2	26.2
8	65.7	73.0	27.2	25.9
9	67.4	72.7	26.9	26.3
10	68.8	72.8	26.9	26.2
11	70.5	73.2	27.6	26.2

Table A.4: Data used for graph of Figure 3.21

A.5 Design of experiment (DOE) results of preprocessing techniques applied to all *E. coli* spectra data

Table A.5 summarises the results for the main effects and interaction plots for the preprocessing techniques with output being the RMSEP. RMEP was used because prediction error is an absolute measure.

A 2^4 full factorial design matrix (with 2 center points per block) set up with four factors, i.e. baseline correction quantile value, baseline correction window size, normalisation quantile value and smoothing span value. All these factors were held at 2 levels (upper and lower levels). Eighteen runs were performed with baseline correction quantile value at 0.1-0.2, baseline correction window size at 300-500, normalisation quantile value at 0.1-1 and smoothing span value at 20-25. These intervals were those that gave optimal results (high R^2 and low RMSECV) when the corresponding preprocessing techniques were used in isolation. A random design was subsequently set up with baseline correction quantile value at 0.1-0.2, baseline correction window size at 100-200, and normalisation quantile value at 0.1-0.2, evaluated against a fixed smoothing span value of 25. The qualities of the models were evaluated through the R^2 and the RMSECV whilst the performance was evaluated against the root mean square error of prediction (RMSEP) and the R^2 of prediction.

A seen in Fig. 3.22 (chapter 3), the main effects plot suggests that all the preprocessing techniques are significant, with baseline correction quantile value and smoothing span value being the most significant preprocessing techniques affecting the model performance. This goes to support the earlier view that all these techniques were essential for preprocessing the spectra data. The main effects for average RMEP are maximised when baseline correction quantile value was set at 0.2 and smoothing span value to 20.

From the interaction plots, the following interactions could be observed;

- baseline correction window size and normalisation quantile value;
- smoothing span value and normalisation quantile value;
- baseline correction window size and smoothing span value; and
- baseline correction window size and quantile value.

Estimated Effects and Coefficients for Average RMSEP (coded units)

```
Term                              Effect    Coef     SE Coef     T       P
Constant                                   17.8744   0.1111   160.87   0.000
Baseline correction window size  0.0800    0.0400    0.1179    0.34     0.767
Baseline correction quantile va   1.8200   0.9100    0.1179   7.72     0.016
Normalisation quantile value    -0.1650   -0.0825    0.1179   -0.70    0.556
Smoothing span value            -1.4750   -0.7375    0.1179   -6.26    0.025
Baseline correction window size*  0.0100   0.0050    0.1179    0.04     0.970
  Baseline correction quantile va
Baseline correction window size* -0.1650  -0.0825    0.1179   -0.70    0.556
  Normalisation quantile value
Baseline correction window size*  -0.0950  -0.0475   0.1179   -0.40    0.726
  Smoothing span value
Baseline correction quantile va*   -0.1650  -0.0825   0.1179  -0.70   0.556
  Normalisation quantile value
Baseline correction quantile va*    0.0450   0.0225   0.1179   0.19    0.866
  Smoothing span value
Normalisation quantile value*       0.1650   0.0825   0.1179   0.70    0.556
  Smoothing span value
Baseline correction window size*  -0.1650   -0.0825   0.1179  -0.70    0.556
```

Baseline correction quantile va* Normalisation quantile value					
Baseline correction window size* Baseline correction quantile va* Smoothing span value	-0.2250	-0.1125	0.1179	-0.95	0.441
Baseline correction window size* Normalisation quantile value* Smoothing span value	0.1650	0.0825	0.1179	0.70	0.556
Baseline correction quantile va* Normalisation quantile value* Smoothing span value	0.1650	0.0825	0.1179	0.70	0.556
Baseline correction window size* Baseline correction quantile va* Normalisation quantile value* Smoothing span value	0.1650	0.0825	0.1179	0.70	0.556

S = 0.471405   PRESS = 1024.44
R-Sq = 98.11%   R-Sq(pred) = 0.00%   R-Sq(adj) = 83.95%

Analysis of Variance for Average RMSEP (coded units)

Source	**DF**	**Seq SS**	**Adj SS**	**Adj MS**	**F**	**P**
Main Effects	4	22.0866	22.0866	5.52165	24.85	0.039
2-Way Interactions	6	0.3713	0.3713	0.06188	0.28	0.906
3-Way Interactions	4	0.5292	0.5292	0.13230	0.60	0.705
4-Way Interactions	1	0.1089	0.1089	0.10890	0.49	0.556
Residual Error	2	0.4444	0.4444	0.22222		
Lack of Fit	1	0.4444	0.4444	0.44444		
Pure Error	1	0.0000	0.0000	0.00000		
Total	17	23.5404				

Unusual Observations for Average RMSEP

		Average				
Obs	**StdOrder**	**RMSEP**	**Fit**	**SE Fit**	**Residual**	**St Resid**
1	1	16.1600	16.1044	0.4698	0.0556	1.41 X
2	2	20.4200	20.3644	0.4698	0.0556	1.41 X
3	3	16.1600	16.1044	0.4698	0.0556	1.41 X
5	5	19.3500	19.2944	0.4698	0.0556	1.41 X
6	6	17.7500	17.6944	0.4698	0.0556	1.41 X
7	7	18.0100	17.9544	0.4698	0.0556	1.41 X
8	8	18.2400	18.1844	0.4698	0.0556	1.41 X
9	9	17.7500	17.6944	0.4698	0.0556	1.41 X
10	10	17.8100	17.7544	0.4698	0.0556	1.41 X
11	11	16.3600	16.3044	0.4698	0.0556	1.41 X
12	12	16.3600	16.3044	0.4698	0.0556	1.41 X
13	13	18.0100	17.9544	0.4698	0.0556	1.41 X
14	14	19.3500	19.2944	0.4698	0.0556	1.41 X
15	15	17.8100	17.7544	0.4698	0.0556	1.41 X
16	16	18.2400	18.1844	0.4698	0.0556	1.41 X
18	18	19.1000	19.0444	0.4698	0.0556	1.41 X

X denotes an observation whose X value gives it large leverage.

Estimated Coefficients for Average RMSEP using data in uncoded units

Term	**Coef**
Constant	38.7794
Baseline correction window size	-0.0460167
Baseline correction quantile va	-127.750
Normalisation quantile value	-11.0000
Smoothing span value	-1.02800
Baseline correction window size* Baseline correction quantile va	0.405167
Baseline correction window size* Normalisation quantile value	0.0366667
Baseline correction window size* Smoothing span value	0.00196667
Baseline correction quantile va* Normalisation quantile value	110.000
Baseline correction quantile va* Smoothing span value	6.20000
Normalisation quantile value* Smoothing span value	0.44000
Baseline correction window size* Baseline correction quantile va* Normalisation quantile value	-0.366667
Baseline correction window size* Baseline correction quantile va* Smoothing span value	-0.0170667
Baseline correction window size* Normalisation quantile value* Smoothing span value	-0.00146667
Baseline correction quantile va* Normalisation quantile value* Smoothing span value	-4.40000
Baseline correction window size* Baseline correction quantile va* Normalisation quantile value* Smoothing span value	0.0146667

Table A.5: Design of experiment (DOE) results of preprocessing techniques applied to all E. coli spectra data

A.6 Variables from the loadings with their corresponding *m/z* ratio signal ions for exponential phase samples

Table A.6 shows the variables from the loadings plots which are associated to *m/z* ratio signal ions for exponential phase samples. Since most of the *m/z* ratio ion signals are singly charged protonated protein (MH^+) molecules, they represent the approximate MALDI experimental MWs of the ionised proteins expressed by the cultures at the exponential phase samples. The magnitude of the loadings may reveal how expression of proteins varies with growth phase. Loadings were considered to have high intensity if the magnitude was greater than 0.2; medium intensity if the magnitude was greater than 0.05, but less than 0.2; and low intensity if the magnitude was between 0 and 0.05.

Variables from loadings plot	Intensity of variable from loadings plot	Associated *m/z* (MH^+) or experimental MW	Intensity of variable from loadings plot	Intensity	Associated *m/z* (MH^+) or experimental MW
362	Low	2846.1	2911	Low	7700.7
615	Low	3222.2	2912	Low	7703.1
903	Low	3678.6	2913	Low	7705.5
904	Low	3680.2	2914	Low	7707.8
905	Low	3681.9	2915	Low	7710.2
906	Low	3683.5	2983	Low	7872.0
1326	Low	4403.7	2984	Low	7874.4
1327	Low	4405.5	2985	Low	7876.8
1335	Low	4419.9	3302	Low	8654.8
1337	Medium	4423.5	3396	Low	8882.3
1443	Medium	4615.8	3520	Medium	9211.0
1533	Low	4782.4	3521	Medium	9213.6
1701	Medium	5100.7	3522	Medium	9216.2
1703	Medium	5104.5	3641	Medium	9527.2
1810	Medium	5313.0	3642	Medium	9529.9
1878	Medium	5447.7	3643	Medium	9532.5
1879	Medium	5449.6	3961	Low	10389.2
1880	Medium	5451.6	4073	Low	10699.8
1881	Medium	5453.6	4248	Low	11197.0
2270	Medium	6256.7	4520	Low	11983.9
2297	Medium	6314.5	4596	Low	12210.4
2301	Medium	6323.1	4741	Low	12645.8
2344	High	6415.8	5175	Low	13997.6
2538	Medium	6842.2	5599	Low	12219.3
2539	Medium	6844.4	5604	Low	15403.2
2540	Medium	6846.6	2756	Medium	7337.7
2541	Medium	6848.9	2818	Low	7482.0
2728	High	7273.1	2845	Low	7545.5

Table A.6: Variables from the loadings with their corresponding m/z ratio ion signals for exponential phase samples

A.7 Variables from the loadings with their corresponding *m/z* ratio signal ions for stationary phase samples

Table A.7 shows the variables from the loadings plots which are associated to *m/z* ratio signal ions for stationary phase samples. The magnitude of the loadings may reveal how expression of proteins varies with growth phase. Loadings were considered to have high intensity if the magnitude was greater than 0.2; medium intensity if the magnitude was greater than 0.05, but less than 0.2; and low intensity if the magnitude was between 0 and 0.05.

Variables from loadings plot	**Intensity of variable from loadings plot**	**Associated *m/z* (MH$^+$) or experimental MW**	**Intensity of variable from loadings plot**	**Intensity**	**Associated *m/z* (MH$^+$) or experimental MW**
411	Low	2917.15	2915	Medium	7710.20
628	Low	3242.11	2991	Medium	7891.22
902	Low	3676.94	3005	Low	7925.00
909	Low	3688.07	3180	High	8351.03
1014	Low	3862.56	3168	Low	8321.55
1070	Low	3956.71	3264	Low	8559.58
1236	Low	4243.58	3267	Low	8567.07
1305	Medium	4366.20	3381	Low	8854.33
1308	Medium	4371.55	3387	Low	8869.58
1407	Low	4549.61	3463	Medium	9063.84
1422	Low	4576.96	3553	Low	9296.72
1538	Low	4791.26	3554	Low	9299.32
1583	Low	4876.15	3555	Low	9301.92
1591	Medium	4890.81	3654	High	9561.51
1701	High	5101.12	3719	Medium	9733.70
1816	Low	5324.82	3730	High	9763.22
2271	Medium	6259.30	3731	High	9765.89
2304	Medium	6329.57	3800	Low	9950.62
2305	Medium	6331.71	3810	Low	9977.82
2306	Medium	6333.86	3970	Low	10414.02
2307	Medium	6336.01	4097	Low	10766.89
2344	Medium	6416.10	4488	Low	11890.24
2553	Medium	6875.70	4522	Low	11989.77
2738	Medium	7296.09	5250	Low	14238.17
2739	Medium	7298.40	5256	Low	14259.00
2920	High	7721.85	5615	Low	15440.04
2921	High	7724.23	6393	Low	18157.87

Table A.7: Variables from the loadings with their corresponding m/z ratio ion signals for stationary phase samples

A.8 Variables from the loadings with their corresponding *m/z* ratio signal ions for decline phase samples

Table A.8 shows the variables from the loadings plots which are associated to *m/z* ratio signal ions for decline phase samples. The magnitude of the loadings may reveal how expression of proteins varies with growth phase. Loadings were considered to have high intensity if the magnitude was greater than 0.2; medium intensity if the magnitude was greater than 0.05, but less than 0.2; and low intensity if the magnitude was between 0 and 0.05.

Variables from loadings plot	**Intensity of variable from loadings plot**	**Associated *m/z* (MH^+) or experimental MW**	**Intensity of variable from loadings plot**	**Intensity**	**Associated *m/z* (MH^+) or experimental MW**
1222	Low	4219.00	4053	Low	10643.97
1223	Low	4220.76	4054	Low	10646.76
1224	Low	4222.51	4055	Low	10649.54
1225	Low	4224.26	4056	Low	10652.33
1441	Low	4611.73	4259	Low	11225.56
1517	Low	4752.10	4474	Low	11849.06
1574	Low	4858.77	4742	Low	12649.85
2069	Low	5834.88	5598	Low	15383.06
2335	Low	6396.33	5599	Low	15386.41
2540	High	6846.62	5600	Low	15389.76
2541	High	6848.86	5601	Low	15393.11
2542	High	6851.09	5602	Low	15396.46
2676	Low	7153.81	6400	Low	18186.28
2723	Medium	7261.54	3524	High	9221.37
2906	Low	7688.67	3641	High	9527.22
2976	Medium	7855.29	3715	High	9723.24
3165	High	8069.70	3795	Low	9937.40
3373	Low	8834.02	3954	Low	10369.98
3457	High	9048.47	3986	Low	10458.15
3458	High	9051.04	4052	Low	10641.19

Table A.8: Variables from the loadings with their corresponding m/z ratio ion signals for decline phase samples

A.9 Ranges of variables from the loadings with their corresponding *m/z* ratio signal ion ranges for exponential phase sample

The full explanation of superscript letters assigned to column titles in Tables A.9 to A.11 are as follows:

[a] Variable count range were the variables (associated to *m/z* ratio values or ion signals in the MALDI data set) from the PLS-DA loadings plot. The loadings plot help provide the *m/z* ratio values (which are experimental MWs) of potential protein biomarkers.

[b]Exp'tal MW (*m/z* ratio value) range mapped to variable count range. Located within a range is an average exp'tal MW which can be matched to the theoretical MW of a protein in the database.

[c]Average variable count was a variable located within a variable count range from the PLS-DA loadings plot that was mapped to a *m/z* ratio value (exp'tal MW). The latter can be matched to the theoretical MW of a protein in the database.

[d]Average exp'tal MW was the *m/z* ratio value (within an exp'tal MW range) that was matched to the theoretical MW of a protein in the database taking into account the mass accuracy in the linear mode MALDI-ToF which is 100 ppm, i.e., exp'tal MW masses should be within 0.01% of their theoretical MWs (+/- 1 mass unit for 10,000 MW protein)

[e]The intensity of the variables (*m/z* ion signals) is the magnitude of their loadings in the loadings plot. High: intensity ≥ ±0.1 units; medium: intensity ≥ ±0.05 and ≤ ±0.1 units; low: intensity between 0 and ±0.05 units.

[f]Match or theoretical sequence MW was calculated using the Compute pI/MW tool (http://web.expasy.org/compute_pi/).

[g]Error was the difference between the theoretical sequence MW of the protein in the database and the exp'tal MW of the MALDI experiment obtained from the PLS-DA loadings plot

[h]Proteins names were those of proteins whose theoretical sequence MW in database tentatively matches exp'tal MW of the MALDI experiment obtained from the PLS-DA loadings plot.

Table A.9: Variable Counts from Loadings Plot and their corresponding Experimental MWs (m/z ratios) for selected SwissProt/TrEMBL Database Proteins for E. coli K-12 Cell Samples that were Tentatively Matched

Variable count range from loadings plot[a]	**Exp'tal MW (m/z ratio) range mapped to variable count range**[b]	**Average variable count within range**[c]	**Average Exp'tal MW (*m/z* ratio) within range**[d]	**Intensity**[e]	**Match (Theoretical MW in database)**[f]	**Error (Da) (Theoreical MW – Exp'tal MW)**[g]	**Protein name**[h]
			Exponential phase				
1695 - 1703	5089.56 – 5104.97	1698	5095.33	Medium	5095.82	0.49	Protein S22
2255 - 2270	6225.14 – 6257.13	2262	6240.06	Medium	6240.39	0.33	50S ribosomal protein L33
2297 - 2301	6314.92 – 6323.50	2297	6314.92	Low	6315.19	0.27	50S ribosomal protein L32
2327 - 2331	6379.44 – 6388.06	2330	6385.91	Low	6385.04	-0.87	Major outer membrane lipoprotein
2341 - 2343	6409.66 - 6413.98	2341	6409.66	High	6410.60	0.94	50S ribosomal protein L30
2543 - 2553	6853.65 – 6876.02	2544	6855.88	Low	6855.89	0.01	Carbon storage regulator
2726 - 2728	7268.72 – 7273.32	2727	7271.02	High	7271.17	0.15	Cold shock-like protein CspC
2726 - 2728	7268.72 – 7273.32	2728	7273.32	High	7273.45	0.13	50S ribosomal protein L29
2754 - 2756	7333.30 – 7337.96	2754	7332.30	Medium	7332.26	-0.04	Cold shock-like protein CspE
2840 - 2845	7533.45 – 7545.17	2841	7535.79	Low	7536.55	0.76	Multiple antibiotic resistance protein
2909 - 2919	7695.98 – 7719.68	2918	7717.31	Medium	7716.72	-0.59	Cold shock-like protein cspB
2980 - 2983	7872.22 - 7965.04	2983	7872.22	Medium	7871.06	-1.16	50S ribosomal protein L31
3021 - 3024	7963.50 – 7970.73	3023	7968.32	Low	7968.97	0.65	Cold shock-like protein cspD
3074 - 3086	8091.69 – 8120.85	3085	8118.42	Low	8118.38	-0.04	Translation initiation factor IF-1
3381 - 3383	8854.33 – 8859.41	3381	8854.33	Low	8855.24	0.91	30S ribosomal protein S18
3510 - 3516	9185.04 – 9200.57	3512	9190.21	Medium	9190.56	0.35	30S ribosomal protein S16
3634 - 3645	9508.64 – 9537.62	3644	9534.98	Medium	9534.98	0	DNA-binding protein HU-alpha
3632 - 3651	9503.38 – 9553.44	3651	9553.44	Medium	9553.20	-0.24	30S ribosomal protein S20
3905 - 3928	10235.37 – 10298.28	3928	10298.28	Low	10299.09	0.81	30S ribosomal protein S19
4065 - 4073	10677.00 – 10699.32	4071	10693.74	Low	10693.44	-0.3	50S ribosomal protein L25
4231 - 4251	11173.56 – 11202.12	4250	11199.26	Low	11199.12	-0.14	50S ribosomal protein L23
4509 - 4516	11951.37 – 11972.04	4511	11957.28	Low	11958.37	1.09	Multidrug transporter emrE
4569 - 4603	12129.09 – 12230.39	4602	12227.40	Low	12226.29	-1.11	50S ribosomal protein L22
4658 - 4661	12395.13 – 12404.15	4659	12398.14	Low	12398.28	0.14	MalE
4996 - 5015	13431.76 – 13491.27	5012	13481.87	Low	13480.70	-1.17	Enamine/imine deaminase
5080 - 5089	13695.85 – 13724.29	5086	13714.81	Low	13713.73	-1.08	30S ribosomal protein S11
5564 - 5574	15267.55 – 15300.92	5568	15280.89	Low	15281.20	0.31	50S ribosomal protein L16
5657 - 5666	15579.32 – 15609.66	5662	15596.17	Low	15596.78	0.61	Protein psiE

Table A.9: Variable count ranges from loadings plot and their corresponding experimental MW ranges for exponential phase samples

A.10 Ranges of variables from the loadings with their corresponding *m/z* ratio signal ion ranges for stationary phase sample

Table A.10: Variable Counts from Loadings Plot and their corresponding Experimental MWs (m/z ratios) for selected SwissProt/TrEMBL Database Proteins for E. coli K-12 Cell Samples that were Tentatively Matched

Variable count range from loadings plot[a]	Exp'tal MW (*m/z* ratio) range mapped to variable count range[b]	Average variable count within range[c]	Average Exp'tal MW (*m/z* ratio) within range[d]	Intensity[e]	Match (Theoretical MW in database)[f]	Error (Da) (Theoreical MW – Exp'tal MW)[g]	Protein name[h]
Stationary phase							
1271 - 1274	4305.76 – 4311.08	1273	4309.30	Medium	4309.22	-0.08	50S ribosomal protein L36
1409 - 1429	4553.67 – 4590.18	1424	4580.04	Low	4580.12	-0.08	Osmotically inducible lipoprotein B
1696 - 1703	5091.48 – 5104.97	1698	5095.33	High	5095.82	0.49	Protein S22
1842 - 1849	5374.59 – 5390.45	1844	5380.55	Low	5380.39	-0.16	50S ribosomal protein L34
2255 - 2270	6225.14 – 6257.13	2262	6240.06	Medium	6240.39	0.33	50S ribosomal protein L33
2327 - 2331	6379.44 – 6388.06	2330	6385.91	Low	6385.04	-0.87	Major outer membrane lipoprotein
2382 - 2398	6498.57 – 6533.44	2386	6507.28	Medium	6507.48	0.2	Ribosome modulation factor
2543 - 2553	6853.65 – 6876.02	2544	6855.88	Medium	6855.89	0.01	Carbon storage regulator
2907 - 2919	7691.25 – 7719.68	2918	7716.61	Medium	7716.72	-0.11	Cold shock-like protein cspB
2982 - 2992	7869.83 – 7893.79	2991	7891.40	Medium	7891.88	0.48	Glycogen synthesis protein glgS
3369 - 3384	8823.88 – 8861.95	3381	8854.33	Low	8855.24	0.91	30S ribosomal protein S18
3456 - 3464	9040.86 – 9066.42	3464	9065.42	Medium	9065.24	-0.18	Protein hdeB
3507 - 3513	9177.28 – 9192.80	3512	9190.21	Low	9191.00	0.79	30S ribosomal protein S16
3636 - 3653	9513.91 – 9558.72	3651	9553.44	Low	9553.20	-0.24	30S ribosomal protein S20
3710 - 3723	9709.75 – 9744.36	3722	9741.69	Medium	9740.91	-0.78	hns deletion induced protein A
4054 - 4058	10646.34 – 10657.48	4056	10651.91	Low	10651.14	-0.77	Integration host factor subunit β
4301 - 4307	11345.43 – 11362.68	4304	11354.05	Low	11353.93	-0.12	Integration host factor subunit α
4466 - 4477	11824.81 – 11857.12	4477	11857.12	Low	11858.00	0.88	Chaperone-like protein hdeA
4508 - 4526	11948.42 –12001.59	4523	11992.72	Low	11993.68	0.96	Protein bolA
4535- 4555	12028.22 – 12087.51	4535	12028.22	Low	12029.45	1.23	Cytochrome o ubiquinol oxidase
5703 - 2714	15734.70 – 15771.97	5704	15738.09	Low	15738.58	0.49	Superoxide dismutase (Cu-Zn)
6382 - 6395	18117.87 – 18165.14	6392	18154.23	Low	18153.47	-0.76	Peptidyl-prolyl *cis-trans* isomerise B
6382 - 6395	18117.87 – 18165.14	6394	18161.50	Low	18161.15	-0.35	Osmotically inducible protein Y
6488 - 6508	18505.07 – 18578.59	6504	18563.88	Low	18564.11	0.23	DNA protection during starvation

Table A.10: Variable count ranges from loadings plot and their corresponding experimental MW ranges for exponential phase samples

A.11 Ranges of variables from the loadings with their corresponding *m/z* ratio signal ion ranges for decline phase sample

Table A.11: Variable Counts from Loadings Plot and their corresponding Experimental MWs (m/z ratios) for selected SwissProt/TrEMBL Database Proteins for E. coli K-12 Cell Samples that were Tentatively Matched

Variable count range from loadings plot[a]	**Exp'tal MW (*m/z* ratio) range mapped to variable count range**[b]	**Average variable count within range**[c]	**Average Exp'tal MW (*m/z* ratio) within range**[d]	**Intensity**[e]	**Match (Theoretical MW in database)**[f]	**Error (Da) (Theoreical MW – Exp'tal MW)**[g]	**Protein name**[h]
Decline phase							
2382 - 2398	6498.57 – 6533.44	2386	6507.28	Low	6507.48	0.2	Ribosome modulation factor
2327 - 2331	6379.44 – 6388.06	2330	6385.21	Medium	6385.04	-0.17	Major outer membrane lipoprotein
2543 - 2553	6853.65 – 6876.02	2544	6855.88	Medium	6855.89	0.01	Carbon storage regulator
2907 - 2919	7691.25 – 7719.68	2918	7717.31	Low	7716.72	-0.59	Cold shock-like protein cspB
2985 - 299	7877.02 – 7891.40	2991	7891.40	Medium	7891.88	0.48	Glycogen synthesis protein glgS
3236 - 3240	8489.85 – 8499.80	3240	8499.80	Low	8500.00	0.2	30S ribosomal protein S21
3241 - 3261	8502.29 – 8552.14	3258	8544.65	Low	8543.83	-0.82	F1845 fimbrial adhesin operon regulatory protein daaF
3292 - 3305	8629.70 – 8662.33	3296	8639.73	Low	8640.00	0.27	Acyl carrier protein
3380 - 3400	8851.79 – 8902.66	3381	8854.33	Low	8855.24	0.91	30S ribosomal protein S18
3456 - 3464	9040.86 – 9066.42	3464	9066.42	Low	9065.24	-1.18	Protein hdeB
3466 - 3474	9071.57 – 9092.14	3466	9071.57	Medium	9071.48	-0.09	Antitoxin relB
3525 - 3527	9223.88 – 9229.07	3526	9226.48	Medium	9226.00	-0.48	DNA-binding protein HU-beta
3540 - 3552	9262.81 – 9294.02	3543	9270.91	Low	9271.58	0.67	Antitoxin of the ChpB-ChpS system
3556 - 3559	9304.43 – 9312.24	3557	9307.03	Low	9307.58	0.55	Antitoxin yefM
3806 - 3815	9966.78 – 991.05	3810	9977.57	Low	9977.98	0.41	Hypothetical protein
3958 - 3965	10380.63 – 10399.89	3960	10386.13	Low	10386.95	0.82	10-kDa chaperonin
3951 - 3956	10361.38 – 10375.13	3976	10430.19	Low	10430.00	-0.19	30S ribosomal protein S19
4054 - 4058	10646.34 – 10657.48	4056	10651.91	Low	10651.14	-0.77	Integration host factor subunit β
4254 - 4261	11210.69 – 11230.70	4259	11224.98	Low	11225.22	0.24	mRNA interferase relE
4249 - 4266	11196.40 – 11245.01	4264	11239.29	Low	11239.93	0.64	DNA-binding protein fis
4372 - 4382	11550.49 – 11579.52	4382	11579.52	Low	11580.00	0.48	30S ribosomal protein S14
4477 - 4490	11857.12 – 11895.37	4477	11857.12	Medium	11858.00	0.88	hns deletion induced protein A
4590 - 4606	12191.61 – 12239.34	4602	12227.40	Low	12226.29	-1.11	50S ribosomal protein L22
4731 - 4751	12615.49 – 12676.21	4749	12670.13	Low	12669.85	-0.28	Flagellar protein fliO
5594 - 5610	15367.78 – 15421.36	5606	15407.96	Low	15408.44	0.48	DNA-binding protein H-NS
5703 - 2714	15734.70 – 15771.97	5704	15738.09	Low	15738.58	0.49	Superoxide dismutase (Cu-Zn)

Table A.11: Variable count ranges from loadings plot and their corresponding experimental MW ranges for exponential phase samples

Appendix B Additional information for the *E. coli* K-12 cell lysate project

Appendix B presents additional information relating to the multivariate data analysis of the *E. coli* K-12 cell lysate spectra data sets. It provides a summary of the cross-validation vector used; figure of calculated ***Y*** versus measured ***Y*** in fitting and prediction after leave-class-out cross-validation for PLS-DA model built with the data sets; figure of calculated ***Y*** versus measured ***Y*** in fitting and prediction after external validation for PLS-DA model built with the data sets; results of PCA model for the mass spectra data sets; PLS-DA results for the data sets showing percentage variance explained by the LVs calculated for ***X*** and ***Y*** variables; PLS-DA results for the data sets showing the quality of the model after calibration and cross-validation; and PLS-DA results for the data sets showing the performance of the model after external validation.

Figures

B.1 Cross-validation vector used for the *E. coli* cell lysate data sets

The cross-validation vector was used for cross-validation of 300 *E. coli* cell lysate MALDI mass spectra calibration data sets. Leave-class-out-cross-validation was performed. The vector was coded as ‘1’ to indicate that the sample involved belonged to class 1; ‘2’ for class 2; and ‘3’ for class 3. All samples coded with ‘1’, belonging to the same class 1 were removed from the calibration data set and a sub-model based on the remaining samples was used to build the PLS-DA model and predict the left out samples. The process was repeated with all the 300 calibration data set of classes 2 and 3 had been left out once.

```
[1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 1
1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 1 1 1
1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 1 1 1 1
1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 31 1 1 1 1 1 1
1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1
1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3]
```

Figure B.1: Cross-validation vector used for the E. coli cell lysate data sets

B.2 Calculated *Y* versus measured *Y* in fitting and prediction after leave-class-out cross-validation

Figure B.2 shows the PLS-DA calculated ***Y*** versus measured ***Y*** in fitting and prediction after leave-class-out cross-validation 300 preprocessed calibration spectra data sets for cell lysate. No significant deviations can be identified along the *y*-axis in all three classes suggesting that that all the useful information is taken into account by the model.

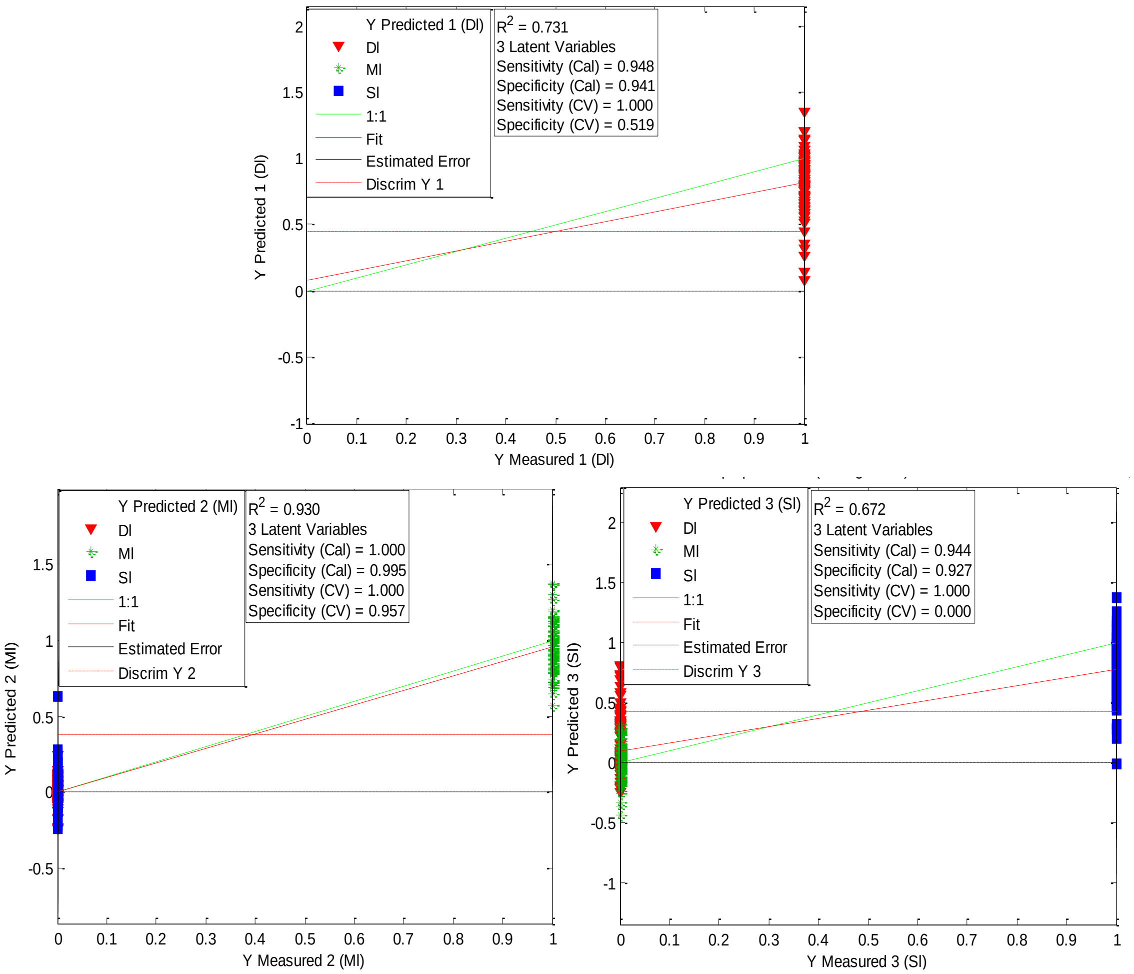


Figure B.2: Calculated Y versus measured Y in fitting and prediction after leave-class-out cross-validation

B.3 PLS-DA calculated *Y* versus measured *Y* in fitting and prediction after external validation

Figure B.3 shows the PLS-DA calculated ***Y*** versus measured ***Y*** in fitting and prediction after external validation with 55 test set samples of cell lysate. The decision threshold shown in each case (middle dashed red line) is calculated using the distribution of predicted ***Y*** values obtained during model building. As seen from the model, just one class 1 (decline phase) sample (1dl18) was misclassified falling below the decision threshold. This gave an excellent prediction sensitivity and specificity of 95.7% and 95.2% respectively. No class 2 (exponential phase) samples were misclassified giving 100% prediction sensitivity and specificity. Only two class 3 (stationary phase) samples (1sl18 and 3sl16) were misclassified giving very good prediction sensitivity and specificity of 88.9% and 85.1% respectively.

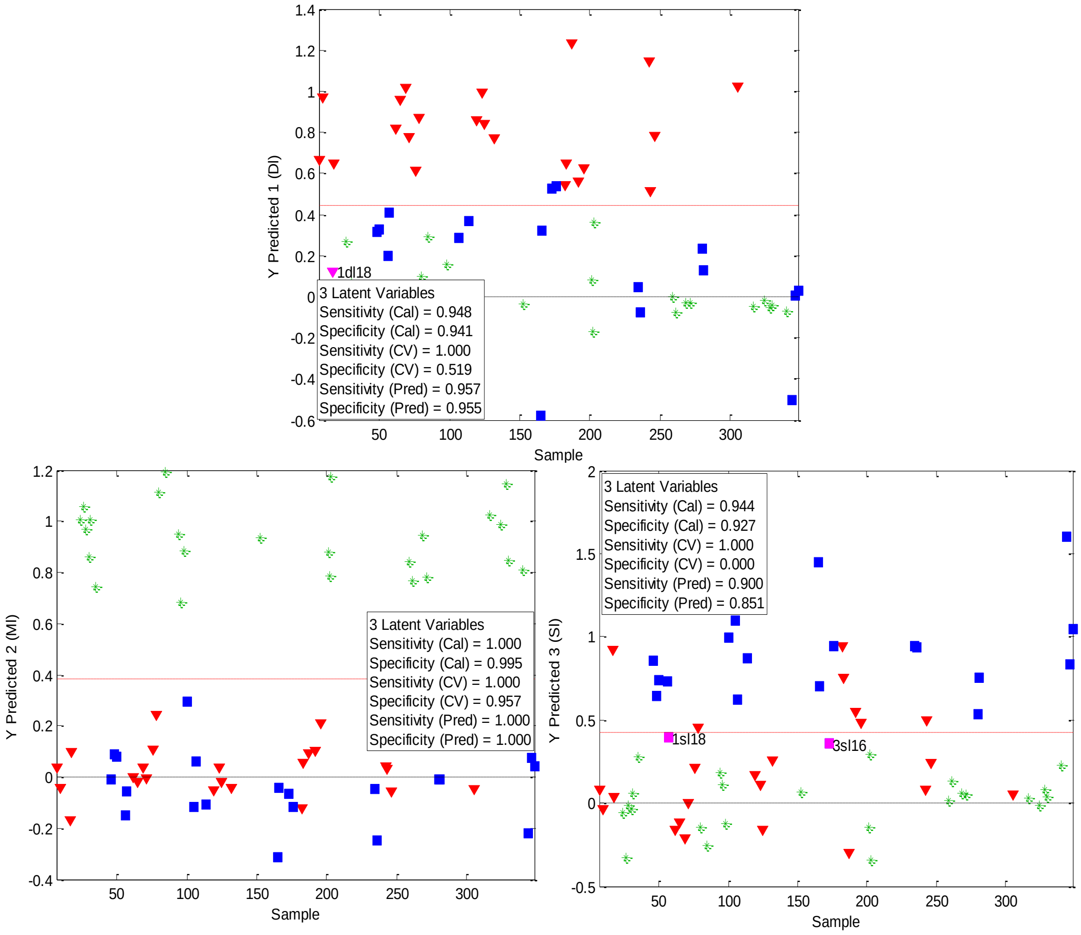


Figure B.3: PLS-DA calculated ***Y*** *versus measured* ***Y*** *in fitting and prediction after external validation*

Tables

B.1 PCA model of the cell lysate mass spectra data sets

Table B.1 summarises the results of PCA model for the 355 preprocessed *E. coli* cell lysate mass spectra data sets. PCA was applied to get an overview of the mass spectra to identify any groupings and to determine if differences occur due to the growth phase. The first three principal components were retained and these account for 93.36% of the variability.

Principal component (PC) number	**Eigenvalue of covariance (*X*)**	**% Variance captured for this PC**	**Total % variance captured**
1	9.14×10^4	70.48	70.48
2	2.04×10^4	15.76	86.24
3	5.35×10^3	4.13	93.36

Table B.1: PCA model of the cell lysate mass spectra data sets

B.2 PLS-DA percentage variance explained by the LVs calculated for *X* and *Y* variables

Table B.2 shows PLS-DA percentage variance explained by the LVs calculated for ***X*** and ***Y*** variables for the 300 preprocessed calibration spectra data sets for E. coli cell lysate. The 3-LV PLS-DA model captured 90.50% of the ***X***-block variance and explained 85.25% of the ***Y***-block variance of the training data

Latent variable (LV) number	***X*-Block**		***Y*-Block**	
	% Variance captured for this LV	**Total % variance captured**	**% Variance captured for this LV**	**Total % variance captured**
1	71.79	71.79	32.06	32.06
2	16.62	88.41	32.31	64.38
3	2.09	90.50	20.87	85.25

*Table B.2: PLS-DA percentage variance explained by the LVs calculated for **X** and **Y** variables*

B.3 PLS-DA results showing the quality of the model after calibration and cross-validation

Table B.3 summarises the PLS-DA results showing the quality of the model after calibration (Cal) and cross-validation (CV) for the 300 preprocessed calibration spectra data for *E. coli* cell lysate. The quality of this model is good, according to the values of R^2 calculated both fitting (73.1% for class 1, 93% for class 2 and 67.2% for class 3) and in CV (86.3% for class 3). The quality of the model was good both in calibration and CV with a sensitivity (proportion of samples correctly classified that belong to the class being modeled), and specificity (proportion of samples correctly classified that do not belong to the class being modeled) of at least 92%. The high values of R^2 in fitting (and the consequent small values of RMSEC, (0.245 for class1, 0.125 for class 2, and 0.267 for class 3) shows that the model is characterised by a large fitting ability.

Modeled class	**Class 1 (Decline phase)**	**Class 2 (Exponential phase)**	**Class 3 (Stationary phase)**
Sensitivity (Cal)	0.948	1.000	0.944
Specificity (Cal)	0.941	0.995	0.927
Sensitivity (CV)	1.000	1.000	1.000
Specificity (CV)	0.519	0.957	0.000
Classification error (Cal)	0.05545	0.00269	0.06405
Classification error (CV)	0.24064	0.02150	0.50000
RMSEC	0.24555	0.12586	0.26716
RMSECV	0.79172	0.61399	0.90605
Bias	-0.00566	-0.01060	-0.01018
CV Bias	-0.20691	-0.31593	0.25343
R^2 Cal	0.73125	0.93043	0.67168
R^2 CV	0.03328	0.01120	0.86356

Table B.3: PLS-DA results showing the quality of the model after calibration and cross-validation

B.4 PLS-DA results showing the performance of the model after external validation

Table B.4 summarises the PLS-DA results showing the performance of the model after external validation with 55 test set samples for cell lysate. This model has an excellent prediction sensitivity and specificity of 95.7% and 95.2% respectively. These results are supported by those in Fig. B.3.

Modeled class	Class 1 (Death phase)	Class 2 (Exponential phase)	Class 3 (Stationary phase)
Sensitivity (prediction)	0.957	1.000	0.889
Specificity (prediction)	0.952	1.000	0.851
Classification error (prediction)	0.04555	0.00000	0.01300
RMSEP	0.25455	0.12477	0.30432
Prediction Bias	-0.01623	-0.02150	0.01605
R^2 prediction	0.71893	0.93518	0.56058

Table B.4: PLS-DA results showing the performance of the model after external validation

Appendix C Additional information for the CHO cell line spectra data sets

Appendix A provides a summary of the PLS algorithm used in this work as well as information of the variables from the loadings plot that are associated with *m/z* ratio ion signals for the high and low producer CHO cell lines.

Figures

C.1 The non-iterative partial least squares (NIPALS) algorithm for PLS1 algorithm

The non-iterative partial least squares (NIPALS) algorithm for PLS1 algorithm was a prerequisite for the PLS-DA algorithm used in this work. The PLS-DA algorithm started with the calculation of a PLS model of *A* latent variables (LVs) with spectra data sets (***X***-block) for calibration samples and ***y*** vector (Fig. 5.4). From step 12 (equation 13), the calculated predicted response value, $\hat{y}_i$ from PLS1 is the starting point of the PLS-DA algorithm (section 5.6.1).

The non-iterative partial least squares (NIPALS) algorithm for PLS1 is presented as follows:

Step 1:

Using a temporary ***y*** factor as output scores, ***u*** (that summarises the remaining variability in ***y***), calculate the loading weights, ***w***, by regressing columns of the ***X*** block data of calibrated samples on ***u***. Give some initial values to ***u*** e.g. the column of ***y*** with the largest sum of squares.

$$\boldsymbol{w^T} = \frac{\boldsymbol{u^T X}}{\boldsymbol{u^T u}} \tag{1}$$

Step 2:

Normalise ***w*** to unit length

$$\boldsymbol{w} = \frac{\boldsymbol{w}}{||\boldsymbol{w}||} \tag{2}$$

Step 3: Estimate the input scores

$$\boldsymbol{t} = \frac{\mathbf{X\,w}}{\mathbf{w}^{\boldsymbol{T}}\mathbf{w}} \tag{3}$$

Step 4:

Estimate the output loadings

$$\boldsymbol{q^T} = \frac{\boldsymbol{u^T\,Y}}{\boldsymbol{u^T\,u}} \tag{4}$$

Step 5:

Normalise input loadings, $\boldsymbol{q}$**,** to unit length

$$\boldsymbol{q} = \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \tag{5}$$

Step 6:

Estimate new output scores, $\boldsymbol{u}$,

$$\boldsymbol{u} = \frac{\boldsymbol{q\,Y}}{\boldsymbol{q^Tq}} \tag{6}$$

Step 7:

Test the occurrence of convergence on $\boldsymbol{u}$, by checking that the $\boldsymbol{t}$ has no longer changed meaningfully since the last iteration, that is:

$$\frac{\|\boldsymbol{t_{old}} - \boldsymbol{t}\|}{\|\boldsymbol{t}\|} < \boldsymbol{\varepsilon} \tag{7}$$

where $\boldsymbol{\varepsilon}$ is “small” e.g. 10^{-6} or 10^{-8}. If convergence has occurred it means the PLS component has been adequately modelled, proceed to step 8. If not, put $\boldsymbol{t}$ as $\boldsymbol{t}_{old}$ and calculate a new $\boldsymbol{u}$ vector by repeating step 1.

Step 8:

Estimate the input loadings

$$\boldsymbol{p^T} = \frac{\boldsymbol{t^T\,X}}{\boldsymbol{t^T\,t}} \tag{8}$$

Step 9:

Compute the estimated regression coefficient of the relation, $\boldsymbol{b}$.

$$\boldsymbol{b} = \frac{\boldsymbol{t^T\,u}}{\boldsymbol{t^T\,t}} \tag{9}$$

where $\boldsymbol{W}$ is the input weights matrix and $\boldsymbol{P}$ is input loading matrix

Step 10:

Estimate the input and output residual matrices

$$\boldsymbol{E = X - tp^T} \tag{10}$$

$$\boldsymbol{F = y - btq^T} \tag{11}$$

If additional PLS components are necessary, replace $\boldsymbol{X}$ and $\boldsymbol{Y}$ by $\boldsymbol{E}$ and $\boldsymbol{F}$, respectively and repeat steps 1-9.

Step 11:

Compute the regression coefficient vector used for the linear PLS predictor

$$\widehat{\boldsymbol{b}} = \boldsymbol{W}(\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{W})^{-\boldsymbol{1}}\boldsymbol{q} \qquad (12)$$

Step 12:

Calculate the $\boldsymbol{y}$ predicted values, $\widehat{\boldsymbol{y}}$**,** for calibration data by fitting a linear relationship between the independent and dependent variable scores.

$$\widehat{\boldsymbol{y}} = \boldsymbol{t}\,\boldsymbol{b} \qquad (13)$$

Figure C.1: The non-iterative partial least squares (NIPALS) algorithm for PLS1 algorithm

Tables

C.1 The variables from the loadings plot and their corresponding to *m/z* ratio signal ions for high producer cell line samples

Since most of the *m/z* ratio signals ions are singly charged protonated protein (MH^{+}) molecules, they represent the approximate MALDI experimental MWs of the ionised proteins expressed by the high producer cell lines. The magnitude of the loadings may reveal how expression of proteins varies with growth phase. Loadings were considered to have high intensity (H) if the magnitude was greater than 0.04; medium intensity (M) if the magnitude was greater than 0.02, but less than 0.04; and low intensity (L) if the magnitude was between 0 and 0.02.

Variables from loadings plot	**Intensity of variable from loadings plot**	**Associated *m/z* (MH$^+$) or experimental MW**	**Variables from loadings plot**	**Intensity of variable from loadings plot**	**Associated *m/z* (MH$^+$) or experimental MW**
518	M	5431.57	3165	M	10160.76
641	L	5618.72	3817	L	11550.75
644	L	5623.33	3936	M	11814.05
818	L	5893.61	3724	M	11347.04
1121	L	6379.42	6608	L	18507.62
1217	H	6537.35	6406	L	17949.32
1489	L	6995.31	6339	L	17766.03
1650	L	7273.68	2918	L	9657.46
1866	L	7655.69	6409	L	17957.55
1983	L	7866.69	4888	M	14027.33
2106	L	8091.60	5434	L	15382.40
2365	M	8575.57	660	L	5648.99
2673	L	9169.39	2782	L	9385.79
2676	L	9175.27	2775	L	9385.78
2777	L	9374.39	6329	L	17738.76
2879	L	9577.65	3205	L	10243.47
3048	H	9900.38	3113	L	10053.74
3154	M	10136.51	4638	L	13425.84
3255	L	10345.75	4779	L	13762.37
3482	L	10825.94	4781	L	13767.18
3629	L	11139.15	4873	H	13978.88
3620	L	11121.38	5048	L	14415.97
3703	H	11299.61	5159	L	14692.09
3708	H	11310.49	5165	L	14706.99
3934	M	11807.87	5172	L	14724.37
3874	L	11676.50	5417	L	15337.20
4075	L	12123.61	5502	L	15553.46
4078	M	12130.37	5720	L	16130.13
4292	L	12617.67	5652	L	15940.89
3938	L	11818.50	5731	L	16143.60
4631	L	13411.12	5971	L	16773.89
5434	L	15382.40	6329	L	17738.76
3161	M	10152.51	6583	L	18435.72
4631	L	13411.12	2907	L	9635.34
4783	H	13773.90	8257	M	23385.04
1112	M	11340.44	8327	L	23601.91

Table C.1: The variables from the loadings plot and their corresponding to m/z ratio ion signals for high producer cell line samples

C.2 The variables from the loadings plot and their corresponding to *m/z* ratio signal ions for low producer cell line samples

Since most of the *m/z* ratio signal ions are singly charged protonated protein (MH$^+$) molecules, they represent the approximate MALDI experimental MWs of the ionised proteins expressed by the high producer cell lines. The magnitude of the loadings may reveal how expression of proteins varies with growth phase. Loadings were considered to have high intensity (H) if the magnitude was greater than 0.04; medium intensity (M) if the magnitude was greater than 0.02, but less than 0.04; and low intensity (L) if the magnitude was between 0 and 0.02.

Variables from loadings plot	**Intensity of variable from loadings plot**	**Associated *m/z* (MH$^+$) or experimental MW**	**Variables from loadings plot**	**Intensity of variable from loadings plot**	**Associated *m/z* (MH$^+$) or experimental MW**
536	L	5458.76	3817	L	11550.75
768	L	5815.29	3830	L	11579.37
979	L	6149.35	3835	L	11590.39
1179	L	6474.60	534	L	5455.73
1294	L	6665.42	1180	L	6476.25
1589	L	7167.57	1293	L	6663.75
2306	L	8464.08	1466	L	6955.98
2737	L	9295.27	1589	L	7167.57
2967	L	9754.76	1781	L	7504.19
3110	L	10046.03	2308	L	8467.85
3198	L	10227.40	4783	H	13773.90
3829	L	11575.46	2967	L	9754.76
3851	L	11623.96	3110	L	10046.03
3972	L	11892.56	3654	L	11193.23
4129	M	12245.63	3970	L	11888.09
5170	L	14717.39	4126	M	12238.84
5236	L	14881.76	4830	L	13885.12
6354	L	17804.70	5230	L	14866.78
6404	L	17941.54	6205	L	17400.02
9403	L	27107.20	1112	M	11340.44
47	L	6382.26	9403	L	27107.20
72	L	6482.42	711	L	9307.20
115	L	6656.52	902	L	10250.44
187	L	6953.20	1091	L	11228.61
275	L	7324.59	1135	L	11462.73
396	L	7851.03	1275	L	12223.72
516	L	8391.16	2202	L	17879.63
2829	L	22306.94			

Table C.2: The variables from the loadings plot and their corresponding to m/z ratio ion signals for low producer cell line samples

Bibliography

Bibliography

2011 Biopharmaceutical Approvals: Low Number And Economic Impact, But A Record Number Of Orphan Approvals And Many Innovations, Cited February 3, 2012. Available from: http://www.pharmaceuticalonline.com/doc.mvc/2011-FDA-Biopharmaceutical-Approvals-Low-0001

Aggarwal Saurabh (2011). What's fueling the biotech engine—2010 to 2011. *Nat. Biotechnol.* **29**, 1083–1089

Aizenman, E. Engelberg-Kulka, H. *et al.* (1996). An *Escherichia coli* chromosomal "addiction module" regulated by guanosine 3',5'-bispyrophosphate: a model for programmed bacterial cell death. *Proc. Natl. Acad. Sci.* **93**(12): 6059-6063.

Aldea, M.T., Garrido, C. *et al.* (1989). Induction of a growth-phase-dependent promoter triggers transcription of *bolA*, an *Escherichia coli* morphogene. *EMBO J.* **8**:3923-393

Alete, D.E., Racher, A.J. *et al.* (2005). Proteomic analysis of enriched microsomal fractions from GS-NS0 murine myeloma cells with varying secreted recombinant monoclonal antibody productivities. *Proteomics*, **5**, 4689-4704.

Andersen, D.C., & Krummen, L. (2002) Recombinant protein expression for therapeutic applications. *Curr. Opinion Biotechnol.*, **13**, 117-123.

Andrade, L. & Manolakos, E.S. (2003) Signal Background Estimation and Baseline Correction Algorithms for Accurate DNA Sequencing. *The J. VLSI Signal Process.*, **35**, 229-243.

Arnér, E.S.J., & Holmgren, A. (2000). "Physiological functions of thioredoxin and thioredoxin reductase." *European J. Biochem.* **267**(20): 6102-6109.

Arnold, R.J., & Reilly, J.P. (1999). Observation of *Escherichia coli* Ribosomal Proteins and Their Posttranslational Modifications by Mass Spectrometry. *Anal. Biochem.*, **269**(1): 105-112.

Azam, T.A., Iwata, A. *et al.* (1999). Growth Phase-Dependent Variation in Protein Composition of the *Escherichia coli* Nucleoid. *J. Bacteriol.* **181**(20): 6361-6370.

Baggerly, K.A., Morris J.S. *et al.* (2004). Reproducibility of SELDITOF protein patterns in serum: Comparing data sets from different experiments. *Bioinformatics*, **20**:777–785.

Bakry, R., Rainer, M. *et al.* (2011). "Protein profiling for cancer biomarker discovery using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and infrared imaging: A review." *Anal. Chim. Acta*. **690**(1): 26-34.

Barak P. (1995). Smoothing and differentiation by an adaptive-degree polynomial filter. *Anal. Chem*. 67, 2758–2762

Barker, M., & Rayens, W. (2003) Partial least squares for discrimination. *J. Chemometrics*, **17**, 166-173.

Bebbington, C. R., Renner, G. *et al*. (1992) High-Level Expression of a Recombinant Antibody from Myeloma Cells Using a Glutamine Synthetase Gene as an Amplifiable Selectable Marker. *Nat. Biotechnol.*, **10**, 169-175.

Bergquist, J., Håkansson, P. *et al.* (2007). Mass spectrometry of proteins--Uppsala perspectives on past and present. *Int. J. Mass Spectrom.* **268**, 73-82

Biopharmaceuticals: Current Market Dynamics & Future Outlook, AS Insights, Cited 1 May 2005, Report Code: ASI0505-5, 66 Pages. Available from: http://www.bioportfolio.com/cgi-bin/acatalog/Biopharmaceuticals_Current_Market_Dynamics_Futur.html

Biotechnology Report: Monoclonal Therapeutics and Companion Diagnostic Products. Report Code: BIO016G, Published: January 2008 http://www.bccresearch.com/report/monoclonal-therapeutics-products-bio016g.html

Birch, J.R. & Onakunle, Y., Biopharmaceutical Proteins. (2005); Vol. 308, pp 1-16.

Birch, J.R., & Racher, A.J. (2006). "Antibody production." *Adv. Drug Deliv. Rev.* **58**(5–6): 671-685.

Blake, C. C. F. and D. W. Rice (1981). "Phosphoglycerate Kinase." Philosophical Transactions of the Royal Society of London. B, *Biol. Sci.* **293**(1063): 93-104.

Blattner F.R., Plunkett G. III, *et al.*, (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**(5331): 1453–1474.

Botella, C., Ferré, J. *et al.* (2009). Classification from microarray data using probabilistic discriminant partial least squares with reject option. *Talanta*. **80**(1): 321–328.

Bowler, C., Montagu, M.V. *et al.* (1992). "Superoxide Dismutase and Stress Tolerance." *Annual Rev. Plant Physiol. Plant Mol. Biol.* **43**(1): 83-116.

Breen E.J., Femia, G.H. *et al.* (2000). Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis*, **21**, 2243-2251.

Brereton, R.G. (2000). "Introduction to multivariate calibration in analytical chemistry." *Analyst* **125**(11): 2125-2154.

Brereton, R.G. (2009) Chemometrics for Pattern Recognition. John Wiley & Sons, Ltd p 453-58

Browne, S.M., & Al-Rubeai, M. (2007) Selection methods for high-producing mammalian cell lines. *Trends in Biotechnol.,* **25**, 425-432.

Bubunenko, M., Baker, T. *et al.* (2007). Essentiality of ribosomal and transcription antitermination proteins analyzed by systematic gene replacement in *Escherichia coli*. *J. Bacteriol.* **189**(7):2844-53.

Buchanan, C.M., Malik, A.S. *et al.,* (2007). "Direct visualisation of peptide hormones in cultured pancreatic islet alpha- and beta-cells by intact-cell mass spectrometry." *Rapid Commun. Mass Spectrom.* **21**(21): 3452-3458.

Butler, M. (2005). Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. *Appl. Microbiol. Biotechnol*., **68**(3):283-29

Carlage, T., M. Hincapie, *et al.* (2009). "Proteomic Profiling of a High-Producing Chinese Hamster Ovary Cell Culture." *Anal. Chem.* **81**(17): 7357-7362.

Chang, E.J., Archambault, V. *et al.* (2004) Analysis of protein phosphorylation by hypothesis-driven multiple-stage mass spectrometry. *Anal. Chem.*, **76**(15), 4472–4483.

Chaurand, P., Norris, J.L. *et al.*, (2006) New Developments in Profiling and Imaging of Proteins from Tissue Sections by MALDI Mass Spectrometry. *J. Proteome Res.,* **5**, 2889-2900.

Chaurand, P., Schriver, K.E., *et al*. (2007) Instrument design and characterisation for high resolution MALDI-MS imaging of tissue sections. *J. Mass Spectrom.,* **42**, 476-489.

Chen, C.C., Buckland, B. *et al.* (1997) Fed-batch culture of recombinant NS0 myeloma cells with high monoclonal antibody production. *Biotechnol. Bioeng.*, **55**, 783-792.

Christner, M., Rohde, H., *et al*., (2010). Rapid Identification of Bacteria from Positive Blood Culture Bottles by Use of Matrix-Assisted Laser Desorption-Ionization Time of Flight Mass Spectrometry Fingerprinting. *J. Clin. Microbiol*., **48**, 1584-1591.

Chu, L., & Robinson, D.K. (2001). "Industrial choices for protein production by large-scale cell culture." *Curr. Opinion Biotechnol.,* **12**(2): 180-187.

Chu, L., Blumentals, I. *et al.* (2005). Production of Recombinant Therapeutic Proteins by Mammalian Cells in Suspension Culture. *Methods Mol. Biol.,* **308**: 107-121.

Chung, J.Y., Lim, S.W. *et al.,* (2004). "Effect of doxycycline-regulated calnexin and calreticulin expression on specific thrombopoietin productivity of recombinant chinese hamster ovary cells." *Biotechnol. Bioeng.,* **85**(5): 539-546.

Connell, N., Han, Z. *et al*. (1987). An *E. coli* promoter induced by the cessation of growth. *Mol. MicrobioI.* **1**:1 95-201

Coombes, K.R., Fritsche, Jr H.A. *et al.* (2003). Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionisation. *Clinical Chem.*, **49**, 1615– 23.

Cozzolino, D., Chree, A. *et al.* (2005). Usefulness of near-infrared reflectance (NIR) spectroscopy and chemometrics to discriminate fishmeal batches made with different fish species. *Agric. Food Chem.* **53**, 4459–4463.

Crossman, L., McHugh, N.A. *et al*. (2006) Investigation of the profiling depth in matrix-assisted laser desorption/ionisation imaging mass spectrometry. *Rapid Commun. in Mass Spectrom.,* **20**, 284-290.

Dalmasso, E., Caseñas, D. *et al.* (2009). Top-down, Bottom-up. The merging of two high performance technologies. Biorad Laboratories, Inc.

De Hoffmann, E., & Stroobant, V. (2008). Mass Spectrometry: Principles and Applications, John Wiley & Sons.

Demirev, P.A., Ho, Y.P. *et al.,* (1999). "Microorganism Identification by Mass Spectrometry and Protein Database Searches." *Anal. Chem.* **71**(14): 2732-2738.

Demirev, P.A., Lin, J.S. *et al.,* (2001). "Bioinformatics and Mass Spectrometry for Microorganism Identification: Proteome-Wide Post-Translational Modifications and Database Search Algorithms for Characterization of Intact H. pylori." *Anal. Chem.* **73**(19): 4566-4573.

Desperyrous, D. Phillpots, R. *et al*. (1996). Electrospray Mass Spectrometry for Detection and Characterization of Purified Cricket Paralysis Virus (CrPV). *Rapid Commun. Mass Spectrom.* **10**, 937

Dieckmann, R., Helmuth, R., *et al*., (2008). Rapid Classification and Identification of Salmonellae at the Species and Subspecies Levels by Whole-Cell Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. *Applied and Environ. Microbiol.*, **74**, 7767-7778.

Dietmair, S., Nielsen, L.K. *et al*. (2012). "Mammalian cells as biopharmaceutical production hosts in the age of omics." *J. Biotechnol.* **7**(1): 75-89

Dietmair, S., Nielsen, L.K. *et al.,* (2011). "Engineering a mammalian super producer." *J. Chem. Technol. Biotechnol.,* **86**(7): 905-914.

Dinnis, D.M., Stansfield, S.H. *et al.* (2006). "Functional proteomic analysis of GS-NS0 murine myeloma cell lines with varying recombinant monoclonal antibody production rate." *Biotechnol. Bioeng.,* **94**(5): 830-841.

Domin, M.A., Welham, K.J. *et al*. (1999). The effect of solvent and matrix combinations on the analysis of bacteria by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry *Rapid Commun. Mass Spectrom.* **13**, 222-226.

Donato, R. (2001). "S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles." *Int. J. Biochem. Cell Biol.,* **33**(7): 637-668.

Dong, H., Shen, W. *et al.,* (2011). "Rapid detection of apoptosis in mammalian cells by using intact cell MALDI mass spectrometry." *Analyst* **136**(24).

Du, P., Kibbe, W.A. *et al*. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, **22**, 2059-2065.

Egea, G., Lázaro-Diéguez, F. *et al*. (2006). "Actin dynamics at the Golgi complex in mammalian cells." *Current Opinion in Cell Biology* **18**(2): 168-178.

Eriksson, Å., Persson Waller, K. *et al*. (2005). Detection of mastitic milk using a gas-sensor array system (electronic nose). *Int. Dairy J.* **15**, 1193–1201.

Eriksson, L., Antti, H. *et al.,* (2004). "Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm)." *Anal. Bioanal. Chem.,* **380**(3): 419-429.

Eriksson, L., Johansson, E. *et al*. (1999) Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS). Umetrics AB, SE90719 Umea, Sweden.

Evason, D.J., Claydon, M.A. *et al*. (2000). Effects of ion mode and matrix additives in the identification of bacteria by intact cell mass spectrometry *Rapid Commun. Mass Spectrom.,* **14,** 669-672.

Farid, S. (2006). Established Bioprocesses for Producing Antibodies as a Basis for Future Planning Cell Culture Engineering. W.-S. Hu, Springer Berlin / Heidelberg. 101: 1-42.

Feng, H., Wong, N.S.C. *et al.*, (2010) Rapid characterisation of high/low producer CHO cells using matrix-assisted laser desorption/ionisation time-of-flight. *Rapid Commun. Mass Spectrom.,* **24**, 1226-1230.

Feng, H.-t., Sim, L.C. *et al*., (2011). "Rapid characterization of protein productivity and production stability of CHO cells by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry." *Rapid Commun. Mass Spectrom.,* **25**(10): 1407-1412.

Fenselau, C., & Demirev, P.A. (2001). "Characterization of intact microorganisms by MALDI mass spectrometry." *Mass Spectrom. Rev.,* **20**(4): 157-171.

Fenselau, C., & Ryzhov, V. (2001) Characterization of intact microorganisms by MALDI mass spectrometry. *Anal. Chem.*, 73, 746-750.

Frank, R. and Hargreaves, R. (2003). "Clinical biomarkers in drug discovery and development." *Nat. Rev. Drug Discov.,* **2**(7): 566-580.

Fussenegger, M., Schlatter, S. *et al.* (1998) Controlled proliferation by multigene metabolic engineering enhances the productivity of Chinese hamster ovary cells. *Nat. Biotechnol.,* **16**, 468-472.

Gantt, S.L., Valentine, N.B. *et al.* (1999) Use of an internal control for matrix-assisted laser desorption/ionisation Time-of-Flight mass spectrometry analysis of bacteria. *J. Am. Soc. Mass Spectrom.*, **10**, 1131-1137.

Garrido, C., Brunet, M. *et al.,* (2006). Heat shock proteins 27 and 70: Anti-apoptotic proteins with tumorigenic properties. *Cell Cycle*, **5**, 2592–2601

Geladi, P. (2003). "Chemometrics in spectroscopy. Part 1. Classical chemometrics." *Spectrochimica Acta Part B: Atomic Spectroscopy* **58**(5): 767-782.

Gibbs, R.A., Weinstock, G.M. *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**(6982):493–521.

Gibson, B.W., Engstrom, J. *et al.* (1997). "Characterization of bacterial lipooligosaccharides by delayed extraction matrix-assisted laser desorption ionization time-of-flight mass spectrometry." *J. Am. Soc. Mass Spectrom.,* **8**(6): 645-658.

Global Biopharmaceutical Market Report (2010-2015). Cited 28 June 2012. Available from: http://www.imarcgroup.com/global-biopharmaceutical-market-report-2010-2015/

Gombotz, W.R. & Shire, S.J. (2010). Current Trends in Monoclonal Antibody Development and Manufacturing. S. J. Shire, W. Gombotz, K. Bechtold-Peters and J. Andya, Springer New York. XI: 1-5.

Göthel, S.F., & Marahiel, M.A. (1999). "Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts." *Cell Mol. Life Sci.*, **55**(3):423-436

Guo, Z., Zhang, Q. *et al.* (2002). "A Method for the Analysis of Low-Mass Molecules by MALDI-TOF Mass Spectrometry." *Anal. Chem.*, **74**(7): 1637-1641.

Gygi, S.P., Rochon, Y. *et al.* (1999). "Correlation between protein and mRNA abundance in yeast." *Mol. Cell. Biol.*, **19**(3): 1720-1730.

Hammond, S., Swanberg, J. *et al.*, (2011). "Genomic sequencing and analysis of a Chinese hamster ovary cell line using Illumina sequencing technology." *BMC Genomics* **12**(1): 67.

Han, M.J., & Lee, S.Y. (2006). "The *Escherichia coli* Proteome: Past, Present, and Future Prospects." *Microbiol. Mol. Biol. Rev*. **70**(2): 362-439

Han, X., Aslanian, *et al.* (2008) Mass spectrometry for proteomics. *Curr. Opinion in Chemical Biol.*, **12**, 483-490.

Hansen, K., Rathke, F. *et al.* (2009). "Bias-Correction of Regression Models: A Case Study on hERG Inhibition." *J. Chem. Inf. Model.*, **49**(6): 1486-1496.

Harrington, P.B. Chen, P., *et al.* (2008). "Biomarker Profiling and Reproducibility Study of MALDI-MS Measurements of *Escherichia coli* by Analysis of Variance-Principal Component Analysis." *Anal. Chem.*, **80**(5): 1474-1481.

Harrington, P.D., Vieira, N.E. *et al*. (2005). Analysis of variance-principal component analysis: A soft tool for proteomic discovery, *Anal. Chimic. Acta.*, **544**, 118-127.

Haswell, S.J. (1992). A practical guide to Chemometrics. Marcel Dekker, New York, ISBN: 0-8247-8597-5.

Heller, D., Cotter, R. *et al.* (1987) Mass spectral analysis of complex lipids desorbed directly from lyophilized membranes and cells. *Biochem. Biophys. Res. Commun.* **142:**194-199.

Heller, D., Murphy, C. *et al.* (1988) Constant neutral loss scanning for the characterization of bacterial phospholipids desorbed by fast atom bombardment. *Anal. Chem.* **60**, 2787 (1988).

Hengge-Aronis, R., & Fischer, D. (1992). "Identification and molecular analysis of *glgS*, a novel growth-phase-regulated and *rpoS*-dependent gene involved in glycogen synthesis in *Escherichia coli*." *Mol. Microbiol.*, **6**(14): 1877-1886

Hilario, M., & Kalousis, A. (2008) Approaches to dimensionality reduction in proteomic biomarker studies. *Brief. Bioinform.*, **9**, 102-118.

Hilario, M., Kalousis, A. *et al.* (2006) Processing and classification of protein mass spectra. *Mass Spectrom. Rev.*, **25**, 409-449.

Hirel, P.H., Schmitter, M.J. *et al.* (1989). "Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid." *Proc. Natl. Acad. Sci.*, **86**(21): 8247-8251.

Horneffer, V., Glückmann, M. *et al.* (2006). "Matrix–analyte-interaction in MALDI-MS: Pellet and nano-electrospray preparations." *International J. Mass Spectrom.*, **249–250**(0): 426-432.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, Warwick & York.

Hotta, Y., Sato, J., *et al.*, (2011). Classification of the Genus Bacillus Based on MALDI-TOF MS Analysis of Ribosomal Proteins Coded in *S10* and *spc* Operons. *J. Agric. Food Chem.*, **59**, 5222-5230.

http://expasy.org/proteomics (2012, 08/03/2012). "UniProtKB/Swiss-Prot database." Retrieved 08/03/2012.

http://web.expasy.org/compute_pi/

http://www.uniprot.org/uniprot/

Hunt, I. (2005). "From gene to protein: a review of new and enabling technologies for multi-parallel protein expression." *PREP.* **40**(1): 1-22

Hwang, S.O., Chung, J.Y. *et al.*, (2003). "Effect of Doxycycline-Regulated ERp57 Expression on Specific Thrombopoietin Productivity of Recombinant CHO Cells." *Biotechnol. Progress* **19**(1): 179-184.

Ibarra, N., Watanabe, S. *et al.* (2003) Modulation of Cell Cycle for Enhancement of Antibody Productivity in Perfusion Culture of NS0 Cells. *Biotechnol. Progress*, **19**, 224-228.

Ilina, E. N., Borovskaya, A. D., *et al.*, (2009). Direct Bacterial Profiling by Matrix-Assisted Laser Desorption−Ionization Time-of-Flight Mass Spectrometry for Identification of Pathogenic Neisseria. *J. Mol. Diagn.*, **11**, 75-86.

Ishihama, A. (1999). "Modulation of the nucleoid, the transcription apparatus, and the translation machinery in bacteria for stationary phase survival." *Genes to Cells* **4**(3): 135-143.

Jakob, U., Gaestel, M. *et al.* (1993). "Small heat shock proteins are molecular chaperones*." J. Biol. Chem.,* **268**(3): 1517-1520.

Jayapal, K., Wlaschin, K. *et al.* (2007). "Recombinant Protein Therapeutics from CHO Cells - 20 Years and Counting." CHO Consortium: SBE Special Edition: 40-47.

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

Jones, J. J., Stump, M.J. *et al.* (2003). Investigation of MALDI-TOF and FT-MS Techniques for Analysis of *Escherichia coli* Whole Cells. *Anal. Chem.* **75**:1340-1347.

Jung, J.U., Gutierrez, C. *et al.* (1989). "Sequence of an osmotically inducible lipoprotein gene." *J. Bacteriol.* **171**(1): 511-520.

Kaufmann, H., Mazur, X. *et al.*, (1999). "Influence of low temperature on productivity, proteome and protein phosphorylation of CHO cells." *Biotechnol. Bioeng.,* **63**(5): 573-582.

Khatib-Shahidi, S., Andersson, M. *et al.,* (2006) Direct Molecular Analysis of Whole-Body Animal Tissue Sections by Imaging MALDI Mass Spectrometry. *Anal. Chem.,* **78**, 6448-6456.

Kim, J.Y., Kim, Y.G. *et al*. (2009) "A proteomic approach for identifying cellular proteins interacting with erythropoietin in recombinant Chinese hamster ovary cells." *Biotechnol. Progress* **26**(1): 246-251

Koller, A., Washburn, M.P. *et al*. (2002). "Proteomic survey of metabolic pathways in rice." *Proc. Natl. Acad. Sci.,* **99**(18): 11969-11974.

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

Krawitz, D.C., Forrest, W. *et al.,* (2006). "Proteomic studies support the use of multi-product immunoassays to monitor host cell protein impurities." *Proteomics* **6**(1): 94-110.

Lander, E.S., Linton L.M. *et al*., (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822):860–921.

Lay, J. (1996). Rapid Identification of Intact Whole Bacteria Based on Spectral Patterns using Matrix-assisted Laser Desorption/Ionisation with Time-of-flight Mass Spectrometry. *Rapid Commun. Mass Spectrom.,* **10**, 1227-1232.

Lay, J. O. (2001). "MALDI-TOF mass spectrometry of bacteria." *Mass Spectrom. Rev.,* **20**(4): 172-194.

Lee, K.R., Lin, X. *et al.,* (2003). "Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method." *Proteomics* **3**(9): 1680-1686.

Lee, M.S., Kim, K.W. *et al.,* (2003). "Proteome Analysis of Antibody-Expressing CHO Cells in Response to Hyperosmotic Pressure." *Biotechnol. Progress* **19**(6): 1734-1741.

Li, B., Morris, J. *et al.* (2002). "Model selection for partial least squares regression." *Chemometr. Intell. Lab.,* **64**(1): 79-89.

Li, L.J., Romanova, E.V. *et al.,* (2000). "Peptide Profiling of Cells with Multiple Gene Products: Combining Immunochemistry and MALDI Mass Spectrometry with On-Plate Microextraction." *Anal. Chem.* **72**(16): 3867-3874.

Liland, K.H., & Indahl, U.G. (2009). "Powered partial least squares discriminant analysis." *J. Chemometrics* **23**(1): 7-18.

Lilley, K.S., & Dupree, P. (2006) Methods of quantitative proteomics and their application to plant organelle characterisation. *J. Exp. Bot.,* **57**, 1493-1499.

Lim, Y., Wong, N.S.C., *et al.,* (2010) Engineering mammalian cells in bioprocessing – current achievements and future perspectives. *Biotechnol. Appl. Biochem*., **55**, 175–189.

Link, A.J., Eng, J. *et al.* (1999). "Direct analysis of protein complexes using mass spectrometry." *Nat Biotech* **17**(7): 676-682.

López, J.L, (2007). "Two-dimensional electrophoresis in proteome expression analysis." *J. Chromatogr. B* **849**(1–2): 190-202.

Martens, H. (2001). "Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression." *Chemometr. Intell. Lab.,* **58**(2): 85-95.

MATLAB, The Math Works, Inc., Natick, MA (2008).

Maurya, P., Meleady, P. *et al.* (2007). "Proteomic Approaches for Serum Biomarker Discovery in Cancer." *Anticancer Res.,* **27**(3A): 1247-1255.

Mazzeo, M.F., Sorrentino, A. *et al.,* (2006) Matrix-Assisted Laser Desorption Ionisation-Time of Flight Mass Spectrometry for the Discrimination of Food-Borne Microorganisms. *Applied Environ. Microbiol.,* **72**, 1180-1189.

Mcguire, J.N., Overgaard, J. *et al.* (2008) Mass spectrometry is only one piece of the puzzle in clinical proteomics. *Brief. Funct. Genomic Proteomic*. **7**, 74 -83.

Meleady, P. (2007). "Proteomic profiling of recombinant cells from large-scale mammalian cell culture processes." *Cytotechnology* **53**(1): 23-31.

Mohan, C., Kim, Y.G. *et al*. (2008). Assessment of cell engineering strategies for improved therapeutic protein production in CHO cells. *Biotechnol. J.*, **3**, 624–630.

Mohan, C., Park, S.H. *et al.,* (2007). "Effect of doxycycline-regulated protein disulfide isomerase expression on the specific productivity of recombinant CHO cells: Thrombopoietin and antibody." *Biotechnol. Bioeng.,* **98**(3): 611-615.

Monchamp, P., Andrade-Cetto, L. *et al.* (2007) Signal Processing Methods for Mass Spectrometry. In Systems Bioinformatics: An Engineering Case-Based Approach. G. Alterovitz and M.F. Ramoni, eds. (Artech House Publishers).

Monod, J. (1949). "The Growth of Bacterial Cultures." *Annual Rev. Microbiol.,* **3**(1): 371-394

Moskovets, E., & Karger, B.L. (2003) Mass calibration of a matrix-assisted laser desorption/ionization time-of-flight mass spectrometer including the rise time of the delayed extraction pulse. *Rapid Commun. Mass Spectrom.,* **17,** 229-237.

Mosser, D.D., Caron, A.W. *et al.,* The chaperone function of hsp70 is required for protection against stress-induced apoptosis. (2000). *Mol. Cell. Biol.*, **20**, 7146–7159.

Naes, T., Isaksson, T., *et al*. (2002). A user-friendly guide to Multivariate Calibration and Classification. NIR Publications, Charlton, Chichester, UK. pp. 22–23

Nakamoto, R.K., Baylis Scanlon, J.A. *et al.* (2008). "The rotary mechanism of the ATP synthase." *Arch. Biochem. Biophys.,* **476**(1): 43-50

Nissom, P., Sanny, A. *et al.* (2006). "Transcriptome and proteome profiling to understanding the biology of high productivity CHO cells." *Mol. Biotech.* **34**(2): 125-140.

Norden, B., Broberg, P., *et al.* (2005) Analysis and Understanding of High-Dimensionality Data by Means of Multivariate Data Analysis. *Chem. Biodivers.,* **2**, 1487-1494.

Nystrom, T. (2004). "Stationary-Phase Physiology." *Annual Rev. Microbiol.,* **58**(1): 161-181.

Ochoa, M.L., & Hrrington, P.B. (2005). Immmunomagnetic Isolation of Enterohemorrhagic *Escherichia coli* O157:H7 from Ground Beef and Identification by matrix-assisted laser desorption ionization time-of-flight mass spectrometry and database searches. *Anal. Chem.* **77**:5258-5267.

Oppermann, U. (2007). "Carbonyl Reductases: The Complex Relationships of Mammalian Carbonyl- and Quinone-Reducing Enzymes and Their Role in Physiology." *Annual Rev. Pharm. Toxicol.* **47**(1): 293-322.

Ouedraogo, R., Flaudrops, C. *et al.,* (2010). "Global Analysis of Circulating Immune Cells by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry." *PLoS ONE* 5(10): e13691.

Oyedotun, K.S., & Lemire B.D. (2004). "The Quaternary Structure of the Saccharomyces cerevisiae Succinate Dehydrogenase." *J. Biol. Chem.* **279**(10): 9424-9431.

Pancholi, V. (2001). "Multifunctional a-enolase: its role in diseases." *Cell. Mol. Life Sceinc.,* **58**(7): 902-920.

Paredes, C., Prats, E. *et al.,* (1999). "Modification of glucose and glutamine metabolism in hybridoma cells through metabolic engineering." *Cytotechnol.* **30**(1): 85-93.

Parisi, D., Magliulo, M., *et al.,* (2008) Analysis and classification of bacteria by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry and a chemometric approach. *Anal. Bioanal. Chem,* **391**, 2127-2134.

Pascoe, D.E., Arnott, D. *et al.,* (2007) Proteome analysis of antibody-producing CHO cell lines with different metabolic profiles. *Biotech. Bioeng.,* **98**, 391-410.

Paulovich, A.G., Whiteaker, J.R. *et al.* (2008). "The interface between biomarker discovery and clinical validation: The tar pit of the protein biomarker pipeline." *Proteomics – Clin. Appl.* **2**(10-11): 1386-1402

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2(6): 559-572.

Pedersen, K., Christensen, S.K. *et al.* (2002) Rapid induction and reversal of a bacteriostatic condition by controlled expression of toxins and antitoxins. *Mol. Microbiol.* **45**:501-510

Pereira, M., Andrade, L. *et al.* (2000) Statistical Learning Formulation of the DNABase-Calling Problem and its Solution Using a Bayesian EM Framework," Discrete Applied Mathematics, vol. 104, no. 1-3, pp. 229-258.

Perera, I.K., Perkins, J. *et al*. (1995). Spin-coated samples for high resolution matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry of large proteins. *Rapid Commun. Mass Spectrom.,* **9**, 180-187.

Pérez, N.F., Ferré, J. *et al*. (2009). Calculation of the reliability of classification in discriminant partial least-squares binary classification. *Chemometr. Intell. Lab. Syst*. **95**, 122–128

Phatak, A. & De Jong, S. (1997). "The geometry of partial least squares." *J. Chemometr.* **11**(4): 311-338

Pierce, C.Y., Barr, J.R. *et al.,* (2007) Strain and phase identification of the U.S. category B agent Coxiella burnetii by matrix assisted laser desorption/ionisation time-of-flight mass spectrometry and multivariate pattern recognition. *Anal. Chimic. Acta,* 583, 23-31.

Pothur, R., Srinivas, S.S. *et al.* (2001) Proteomics in Early Detection of Cancer. *Clin. Chem.*, 47, 1901-1911.

Rai, M., & Padh, H. (2001). Expression systems for production of heterologous proteins. *Current* **80**, 1121-1128

Reichert, J. M. (2012). "Marketed therapeutic antibodies compendium." mAbs 4(3): 413-415.

Reilly, J.P., Arnold, R.J., *et al.* (1999) Monitoring the Growth of a Bacteria Culture by MALDI-MS of Whole Cells. *Anal. Chem.,* **71,** 1990-1996.

Reyes-Ruiz, J.M., & Barrera-Saldana, H.A. (2006) Proteins in a DNA world: expression systems for their study. *Rev. Invest. Clin*. 58:47–55

Reyzer, M.L. & Caprioli, R.M. (2007) MALDI-MS-based imaging of small molecules and proteins in tissues. *Curr. Opinion Chem. Biol,* 11, 29-35.

Rifai, N., Gillette, M.A. *et al.* (2006). "Protein biomarker discovery and validation: the long and uncertain path to clinical utility." *Nat. Biotechnol.* 24(8): 971-983.

Riley, M.T., Abe T., *et al.* (2005). "*Escherichia coli* K-12: a cooperatively developed annotation snapshot*." Nucl. Acids Res.* **34**(1): 1-9.

Rogers, M.A., Clarke, P. *et al.* (2003). "Proteomic Profiling of Urinary Proteins in Renal Cancer by Surface Enhanced Laser Desorption Ionization and Neural-

Network Analysis: Identification of Key Issues Affecting Potential Clinical Utility." *Cancer Res.* **63**(20): 6971-6983.

Ryzhov, V. & Fenselau, C. (2001). Characterization of the protein subset desorbed by MALDI from whole bacterial cells. *Anal. Chem.* **73**:746±750.

Ryzhov, V., and Fenselau, C. (2000). Characterization of the protein subset desorbed by MALDI from whole bacterial cells. *Anal. Chem.* **73**:746±750.

Saenz, A.J., Petersen, C.E. *et al.,* (1999) Reproducibility of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for replicate bacterial culture analysis. *Rapid Commun. Mass Spectrom.,* **13,** 1580-1585.

Satten, G.A., Datta, S. *et al*. (2004) Standardisation and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinform.*, **20**, 3128-3136.

Schroeder, M. (2008). Engineering eukaryotic protein factories. *Biotechnol. Lett*. **30**, 187–196.

Sekhon, B.S. (2010). Biopharmaceuticals: an overview *Thai J. Pharm. Sci*. **34** 1-19

Shin, H. & Markey, M.K. (2006) A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *J. Biomed. Inform.,* **39,** 227-248.

Simonsen, C.C. & McGrogan, M. (1994). "The Molecular Biology of Production Cell Lines." *Biologicals* 22(2): 85-94.

Sirover, M.A. (1997). "Role of the glycolytic protein, glyceraldehyde-3-phosphate dehydrogenase, in normal cell function and in cell pathology." *J. Cell. Biochem*. 66(2): 133-140.

Smales, C.M. & James, D.C. Therapeutic Proteins : Methods and Protocols. 2005; Vol. 308 Vol. 308, pp 1-16

Smales, C.M., Dinnis, D.M. *et al.*, (2004). "Comparative proteomic analysis of GS-NS0 murine myeloma cell lines with varying recombinant monoclonal antibody production rate." *Biotechnol. Bioeng*. **88**(4): 474-488.

Smith, P., Snyder, A. *et al*. (1995) Characterization of bacterial phospholipids by electrospray ionization tandem mass spectrometry *Anal. Chem.* **67**, 1824

Sorace, J.M. & Zhan, M.A. (2003). Data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinform.*, **4**, 24

Srinivas, P.R., Verma, M. *et al.* (2002). "Proteomics for Cancer Biomarker Discovery." *Clin. Chem.* **48**(8): 1160-1169.

Stainier, R.Y., Ingraham, J.L. *et al.* (1987). General Microbiology, Fifth Edition, Englewood Cliffs, New Jersey, U.S.A

Stapulionis, R., Kolli, S. *et al.* (1997). "Efficient Mammalian Protein Synthesis Requires an Intact F-Actin System." *J. Biol. Chem.* **272**(40): 24980-24986

Stenberg F., Chovanec P. *et al.* (2005) "Protein complexes of the *Escherichia coli* cell envelope."*J. Biol. Chem.* **280**:34409-34419

Subramaniam, J.M., Whiteside, G. *et al.* (2012). "Mogamulizumab: First Global Approval." *Drugs* **72**(9): 1293-1298

Thieringer, H.A., Jones, P.G. *et al.* (1998). Cold shock and adaptation. *BioEssays.* **20**: 49-57.

Tibshirani, R., Hastie. T. *et al.* (2004). Sample classification from protein mass spectrometry by ‘‘peak probability contrasts’’. *Bioinform.,* **20**, 3034–3044.

Tillier, E.R.M., & Charlebois, R.L. (2009). The human protein coevolution network. *Genome Res.* **19**(10): 1861–1871.

Touchon, M., Hoede, C. *et al.* (2009). "Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths." *PLoS Genet.,* **5**(1): e1000344.

Van Dyk, D.D., Misztal, D.R. *et al.,* (2003). "Identification of cellular changes associated with increased production of human growth hormone in a recombinant Chinese hamster ovary cell line." *Proteomics* **3**(2): 147-156.

van Veelen P.A., Jimenez C.R. *et al.,* (1993). "Direct peptide profiling of single neurons by matrix-assisted laser desorption-ionisation mass spectrometry." *Organic Mass Spectrom.* **28**(12): 1542-1546.

Veltri, P. (2008) Algorithms and tools for analysis and management of mass spectrometry data. *Brief. Bioinform.*, 9, 144-155.

Venter, J. C., M. D. Adams, *et al.* (2001). "The Sequence of the Human Genome." *Science* **291**(5507): 1304-1351

Vlek, A. L. M., Bonten, M. J. M. *et al.* (2012). Direct Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry Improves Appropriateness of Antibiotic Treatment of Bacteremia. *PLoS ONE*, **7**, e32589.

Wagner, M., Naik, D., *et al.* (2003) Protocols for disease classification from mass spectrometry data. *Proteomics*, 3, 1692-1698.

Wang, P., Tang, H., *et al*. (2006). Normalisation Regarding Non-Random Missing Values in High-Throughput Mass Spectrometry Data. Pacific Symposium on Biocomputing **11**:315-326

Wang, J., Zhou, N., *et al.*, (2012). Identification and Cluster Analysis of *Streptococcus pyogenes* by MALDI-TOF Mass Spectrometry. *PLoS ONE*, **7**, e47152.

Wang, W., Zhou, H., *et al*. (2003). Quantification of proteins and metabolites by mass spectrometry without isotopic labelling or spiked standards. *Anal. Chem.*, **75**, 4818-4826.

Warscheid, B. & Fenselau, C. (2004) A targeted proteomics approach to the rapid identification of bacterial cell mixtures by matrix-assisted laser desorption/ionisation mass spectrometry. *Proteomics*, **4**, 2877-2892.

Warscheid, B. & Fenselau, C. (2004). A targeted proteomics approach to the rapid identification of bacterial cell mixtures by matrix-assisted laser desorption/ionisation mass spectrometry. *Proteomics*, **4**, 2877-2892.

Waterston R.H, Lindblad-Toh K, *et al.*, (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915):520–562.

Watson, J.T & Sparkman, O.D. (2007). Introduction to Mass Spectrometry: instrumentation, applications and strategies for data intepratation. 4TH ed. John Woley & Sons Ltd. The Atrium, Southern Gate Chichester, West Sussex PO19 8SQ, Englang

Wehr, T. (2006). Top-Down versus Bottom-Up Approaches in Proteomics, Bio-Rad Laboratories, Hercules, California. "Directions in Discovery," LCGC, Woodbridge Corporate Plaza, 485 Route 1 South, Building F, First Floor, Iselin, NJ 08830,

Weiss, H., Friedrich, T. *et al.* (1991). "The respiratory-chain NADH dehydrogenase (complex I) of mitochondria." *European J. Biochem*. **197**(3): 563-576.

Wery, J., (2012) http://www.basinc.com/

Wilkins, C., L. & Lay, J., O. Jr. (2006). Identification of Micrororganisms by Mass Spectrometry. Vol. 169. University of Fayetteville, AR. John Woley & Sons, Inc., publication.

Wilkins, C.L., & Lay, J.O. Jr. (2006). Identification of Micrororganisms by Mass Spectrometry. Vol. 169. University of Fayetteville, AR. John Woley & Sons, Inc., publication.

Wise, B.M., Gallagher, N.B. *et al*. (2005) PLS_Toolbox Version 3.5 for use with MATLAB™, Eigenvector Research, Inc., Manson, WA, USA, pp. 185–189.

Wu, B., Abbott, T. *et al.* (2003). "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data." *Bioinformatics* **19**(13): 1636-1643.

Wunschel, S.C., Jarman, K.H. *et al*. (2005) *J. Am. Soc. Mass Spectrom.*, **16**, 456-462.

Xu, X., Nagarajan, H. *et al.,* (2011). "The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line." *Nat. Biotech.* **29**(8): 735-741.

Yang, R.Y., and Liu, F.T. (2003). "Galectins in cell growth and apoptosis." *Cell Mol. Life Sci*. **60**(2): 267-276.

Yim, H.H., & Villarejo, M. (1992). "*osmY*, a new hyperosmotically inducible gene, encodes a periplasmic protein in *Escherichia coli*." *J. Bacteriol.* **174**(11): 3637-3644.

Yoon, S.H., Han, M.J. *et al*. **(**2003). Combined transcriptome and proteome analysis of *Escherichia coli* during high cell density culture. *Biotechnol. Bioeng.* **81:**753–767.

Zhang, X., Scalf, M. *et al*., (2006). "Identification of Mammalian Cell Lines Using MALDI-TOF and LC-ESI-MS/MS Mass Spectrometry." *J. Am. Soc. Mass Spectrom*. **17**(4):490-499.

Zhou, W., Chen, C.C. *et al*. (1997) Fed-batch culture of recombinant NS0 myeloma cells with high monoclonal antibody production. *Biotechnol. Bioeng.*, **55**, 783-792